Lecture notes for
Advanced Data Analysis 2 (ADA2)
Stat 428/528
University of New Mexico

Edward J. Bedrick Ronald M. Schrader Erik B. Erhardt

Spring 2012

Contents

Course details **vii**

1 SAS statistical software **1**
1.1 SAS Basics . . . 1
1.2 Post Tutorial Discussion . . . 2
1.3 A Short Discussion of the Unix Version of SAS . . . 2
1.4 A Little More SAS . . . 5
1.5 Debugging a SAS Program . . . 6
1.6 Variable Creation, Case Selection and Deletion . . . 7
1.7 Another SAS Example . . . 9
1.7.1 Creating better graphs . . . 18
1.8 Selected Details on Procs . . . 21
1.9 One-Way ANOVA in SAS . . . 24
1.9.1 Additional Multiple Comparison Methods for ANOVA . . . 27
1.10 Additional SAS Topics on Data Management . . . 28
1.11 Correlation in SAS . . . 31
1.12 Regression in SAS . . . 33
1.12.1 Regression Diagnostics in SAS: Rocket Propellant Data . . . 35
1.13 Features of SAS for Windows . . . 41
1.13.1 Using the Import and Export Wizards in Windows . . . 41
1.13.2 SAS/Graph . . . 43
1.13.3 High Resolution Graphs in SAS for Windows . . . 43
1.13.4 SAS/Assist . . . 43
1.13.5 SAS/Analyst and SAS/Insight . . . 43

2 Introduction to Multiple Linear Regression **45**
2.1 Indian systolic blood pressure example . . . 45
2.1.1 Taking Weight Into Consideration . . . 48
2.1.2 Important Points to Notice About the Regression Output . . . 49
2.1.3 Understanding the Model . . . 51
2.2 GCE exam score example . . . 53
2.2.1 Some Comments on GCE Analysis . . . 59

3 A Taste of Model Selection for Multiple Regression **63**
3.1 Model . . . 63
3.2 Backward Elimination . . . 63

3.3 Example: Peru Indian blood pressure . . . 64
3.3.1 Analysis for Selected Model . . . 69
3.3.2 Residual and Diagnostic Analysis for Selected Model . . . 72
3.4 Options with Backward Elimination . . . 73
3.5 Example: Dennis Cook's Rat Data . . . 73

4 One Factor Designs and Extensions 81

5 Paired Experiments and Randomized Block Experiments 85
5.1 Analysis of a Randomized Block Design . . . 86
5.2 Extending the One-Factor Design to Multiple Factors . . . 93
5.2.1 Example: Beetle insecticide two-factor design . . . 94
5.2.2 The Interaction Model for a Two-Factor Experiment . . . 95
5.2.3 Example: Survival Times of Beetles . . . 100
5.2.4 Example: Output voltage for batteries . . . 105
5.2.5 Checking assumptions in a two-factor experiment . . . 109
5.2.6 A Remedy for Non-Constant Variance . . . 112
5.2.7 glm: MEANS versus LSMEANS . . . 117
5.2.8 Unbalanced Two-Factor Designs and Analysis . . . 120

6 A Short Discussion of Observational Studies 127

7 Analysis of Covariance: Comparing Regression Lines 133
7.1 ANCOVA . . . 135
7.2 Generalizing the ANCOVA Model to Allow Unequal Slopes . . . 139
7.2.1 Relating Models to Two-Factor ANOVA . . . 148
7.2.2 Choosing Among Models . . . 149
7.2.3 Comments on Comparing Regression Lines . . . 150
7.3 Using GLM to Fit Models . . . 151
7.3.1 GLM Output for the Tool Lifetime ANCOVA Model . . . 157

8 Polynomial Regression 159
8.1 Polynomial Models with One Predictor . . . 159
8.1.1 Example: Cloud point and percent I-8 . . . 161
8.2 Polynomial Models with Two Predictors . . . 164
8.2.1 Example: Mooney viscosity . . . 164
8.2.2 Example: Mooney viscosity on log scale . . . 166
8.3 Using glm to Fit Polynomial Models . . . 168

9 Discussion of Response Models with Factors and Predictors 171
9.1 Some Comments on Building Models . . . 176
9.2 Example: The Effect of Sex and Rank on Faculty Salary . . . 177
9.2.1 A Three-Way ANOVA on Salary Data . . . 179
9.2.2 Using Year and Year Since Degree to Predict Salary . . . 183
9.2.3 Using Factors and Predictors to Model Salaries . . . 184
9.2.4 Discussion of the Salary Analysis . . . 188

10 Automated Model Selection for Multiple Regression **191**
10.1 Forward Selection . . . 192
10.2 Backward Elimination . . . 192
10.3 Stepwise Regression . . . 193
10.3.1 Example: Indian systolic blood pressure . . . 193
10.4 Other Model Selection Procedures . . . 200
10.4.1 R^2 Criterion . . . 200
10.4.2 Adjusted-R^2 Criterion, maximize . . . 201
10.4.3 Mallows' C_p Criterion, minimize . . . 201
10.5 Illustration with Peru Indian data . . . 202
10.5.1 R^2, $\overline{R}^2$, and C_p Summary for Peru Indian Data . . . 204
10.5.2 Peru Indian Data Summary . . . 205
10.6 Example: Oxygen Uptake . . . 206

11 Logistic Regression **215**
11.1 Example: Age of Menarche in Warsaw . . . 215
11.2 The Simple Logistic Regression Model . . . 217
11.2.1 Estimating Regression Parameters . . . 218
11.2.2 Maximum Likelihood Estimation for Logistic Regression Model . . . 219
11.2.3 Fitting the Logistic Model by Maximum Likelihood . . . 221
11.2.4 Example: Leukemia white blood cell types . . . 223
11.3 Example: The UNM Trauma Data . . . 228
11.3.1 Selecting Predictors in the Trauma Data . . . 229
11.3.2 Checking Predictions for the Trauma Model . . . 231
11.4 Historical Example: O-Ring Data . . . 232
11.5 Using genmod for generalized linear models . . . 237

12 An Introduction to Multivariate Methods **241**
12.1 Linear Combinations . . . 242
12.2 Vector and Matrix Notation . . . 244
12.3 Matrix Notation to Summarize Linear Combinations . . . 246

13 Principal Component Analysis: PCA **249**
13.1 Example: Temperature Data . . . 250
13.2 PCA on Correlation Matrix . . . 252
13.3 Interpreting Principal Components . . . 253
13.4 Example: Painted turtle shells . . . 253
13.4.1 PCA on covariance matrix . . . 254
13.4.2 PCA on correlation matrix . . . 255
13.5 Why is PCA a Sensible Variable Reduction Technique? . . . 256
13.5.1 A Warning on Using PCA as a Variable Reduction Technique . . . 257
13.5.2 PCA is Used for Multivariate Outlier Detection . . . 259
13.6 Example: Sparrows, for Class Discussion . . . 260
13.7 PCA for Variable Reduction in Regression . . . 263

14 Cluster Analysis **269**
14.1 Example: Mammal teeth 272
14.2 Identifying the Number of Clusters 275
14.3 Example: 1976 birth and death rates 277

15 Multivariate Analysis of Variance **285**

16 Discriminant Analysis **295**
16.1 Canonical Discriminant Analysis Using CANDISC 296
16.2 Example: Owners of riding mowers 296
16.3 Discriminant Analysis on Fisher's Iris Data 301
16.3.1 Variable Selection Using STEPDISC 305
16.3.2 Stepwise Selection of Features and MANOVA 307

17 Classification **309**
17.1 Classification in SAS 310
17.2 Evaluating the Accuracy of a Classification Rule 312
17.3 Example: Error Estimation on the Carapace Data 312
17.4 Example: Cross-Validation Using Fisher's Iris Data 317
17.5 Example: Analysis of Admissions Data 320
17.5.1 Further Analysis of the Admissions Data 321
17.5.2 Classification Using Unequal Prior Probabilities 324
17.5.3 Classification With Unequal Covariance Matrices 326

Course details

Course Stat 428/528.001, Spring 2012, TR 14:00–15:15, Anthro 163

Instructor Erik B. Erhardt, SMLC 312, (505)750-4424, `erike+stat.unm.edu`

Office Hours MSLC 312, Wed 2:00-2:50 PM and Fri 10:30-11:20 AM

Prerequisite ADA1 Stat 427/527

Course materials `statacumen.com/teaching/ADA2/` has information on the software, course notes, homework, and other materials.

1. (required) Instructor's notes on the website.

   I will try to have a final version of the current chapter ready a week ahead of time.

2. (optional) The Statistical Sleuth by Ramsey and Schafer.

   I will not make reference to this text. If you like to see things presented in multiple ways, then it may be conforting to have. Also, used copies are not expensive.

Syllabus A continuation of 427/527 that focuses on methods for analyzing multivariate data and categorical data. Topics include MANOVA, principal components, discriminant analysis, classification, factor analysis, analysis of contingency tables including log-linear models for multidimensional tables and logistic regression.

Grading

75% Homework (late assignments will not be accepted)

25% Final/Project (no make-ups)

Chapter 1

SAS statistical software

1.1 SAS Basics

SAS is a programming language for data management and statistical analysis. Version 9 of SAS for Windows is available in the SMLC basement computer lab. The **Unix** version of SAS is available on UNM's **linux.unm.edu** server. I will primarily discuss the PC (i.e., Windows) version.

```
> ssh -XYC linux.unm.edu -l NETID                          <--- your UNM NETID

[NETID@vulcan ~]$ sas &
Starting SAS ....
The http://Fastinfo.unm.edu  entry for SAS is:
    http://unm.custhelp.com/cgi-bin/unm.cfg/php/enduser/std_adp.php?p_faqid=6341
```

SAS is more difficult to master than most statistical packages (including **Minitab**) for the following reasons: (**1**) SAS does not, in general, provide a point-and-click environment for statistical analysis. Rather, SAS uses syntax-based programs (i.e., code) to define, transform, and read data, and to define the procedures for analyzing data. (**2**) SAS does not really have a spreadsheet environment for data management. Rather, data are entered directly within a SAS program, read from a file, or imported from a spreadsheet. All manipulation, transformation, and selection of data is coded in the SAS program. Well done, this means that all the steps of the analysis are available to be understood.

To illustrate, the following code is a SAS program that defines a data set containing three variables. The only analysis performed is that the data are printed.

```
* rabbit - first example;
options ls=79 nodate nocenter;
data rabbit;
   input dose nres ntotal;
   datalines;
   50   1  5
   200  4 10
   250  2  5
   400  6 10
  2000  4  5
;
run;

proc print;
run;
```

Nonetheless, SAS has its virtues — it provides a much richer environment for data management (especially of large databases) and data analysis than almost any other statistical package (R has only in the last couple years become more popular and in some ways more powerful).

I will begin our SAS discussion by focusing on two of the tutorials on the ADA2 webpage: the SAS learning Module at `support.sas.com/training/elearn/tutorials/tutsas9.0/menunew.htm` and the UCLA ATS site at `http://www.ats.ucla.edu/stat/sas/`.

Let's go to the support.sas.com site and look at the four tutorials under "Getting to know SAS".

1.2 Post Tutorial Discussion

Seven important points that we should have learned from the tutorials, with some embellishments:

1. Four primary windows in the SAS Display Manager are the **program editor** window, where you enter and edit SAS program lines, the **log** window, which provides notes and error messages associated with your SAS code, the **output** window which contains any output generated by your SAS code, and the **explorer** window, which provides access to files, libraries, and output. The contents of the program editor, log, and output windows can be saved to disk. I will discuss the **graph window** later.

2. You should make extensive use of the SAS online help — click **HELP** on the **SAS MENU BAR**.

3. SAS for Windows has an extensive online tutorial. Just follow the path: SAS MENU BAR > HELP > GETTING STARTED WITH SAS SOFTWARE.

4. SAS programs are entered from the console in the program editor window. Alternatively, you can retrieve a SAS program in a stored file by following the steps: click on (activate) the program editor window, then go to SAS MENU BAR > FILE > OPEN and specify the file.

5. A SAS program will typically consist of one or more **data steps** that define the data for analysis, and one or more **proc steps** that specify the analyses to conduct.

6. After submitting a program, you can retrieve the program by activating the program editor window. The program can be modified, debugged, and expanded.

7. Two or more program editor windows can be used in one SAS session. A new program window can be initialized by following the steps: SAS MENU BAR > VIEW > ENHANCED (or PROGRAM) EDITOR, or by importing a SAS program from your PC, as described above.

1.3 A Short Discussion of the Unix Version of SAS

In the (batch) **Unix** version of SAS you must use an editor, such as **pico**, to create a file that contains your SAS program. The file is given a **.sas extension** to keep track of the contents. The SAS program on the first page of the handout can be found in the file `ada2_01_firstprog.sas` in the main directory of my linux account. I can list the file (ls) and the contents of the file (cat) with the following commands at the linux prompt ($):

```
[erike@mizar ~/ada2]$ ls

ada2_01_firstprog.sas

[erike@mizar ~/ada2]$ cat ada2_01_firstprog.sas

* rabbit - first example;
options ls=79 nodate nocenter;
data rabbit;
   input dose nres ntotal;
   datalines;
   50   1  5
   200  4 10
   250  2  5
   400  6 10
  2000  4  5
;
run;

proc print;
run;
```

To **run** the SAS program `ada2_01_firstprog.sas`, type in the command
`sas ada2_01_firstprog.sas`
at the $ prompt. You will get a new $ prompt after the program is completed, and your directory will (usually) contain two new files: one with a `.log` extension, and the other with a `.lst` extension. For us, those files will be named `ada2_01_firstprog.log` and `ada2_01_firstprog.lst`. The .log file plays the role of the log window in the PC version. The .lst file plays the role of the PC output window. In general, you would import the output file into an editor, or (gulp) MS Word, to write a discussion of the analysis.
Run program:

```
[erike@mizar ~/ada2]$ sas ada2_01_firstprog.sas

Starting SAS ....

The http://Fastinfo.unm.edu  entry for SAS is:
    http://unm.custhelp.com/cgi-bin/unm.cfg/php/enduser/std_adp.php?p_faqid=6341
```

List files:

```
[erike@mizar ~/ada2]$ ls
ada2_01_firstprog.log  ada2_01_firstprog.lst  ada2_01_firstprog.sas
```

Type log file to screen (three commands that do this are cat, less, and more):

```
[erike@mizar ~/ada2]$ cat ada2_01_firstprog.log
1      The SAS System                              18:23 Sunday, January 8, 2012

NOTE: Copyright (c) 2002-2008 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) Proprietary Software 9.2 (TS2M3)
      Licensed to UNIVERSITY OF NEW MEXICO - T&R - ADMIN, Site 70092051.
NOTE: This session is executing on the Linux 2.6.18-274.12.1.el5 (LIN X64) platform


You are running SAS 9. Some SAS 8 files will be automatically converted
by the V9 engine; others are incompatible.  Please see
http://support.sas.com/rnd/migration/planning/platform/64bit.html

PROC MIGRATE will preserve current SAS file attributes and is
recommended for converting all your SAS libraries from any
SAS 8 release to SAS 9.  For details and examples, please see
http://support.sas.com/rnd/migration/index.html
```

```
This message is contained in the SAS news file, and is presented upon
initialization.  Edit the file "news" in the "misc/base" directory to
display site-specific news and information in the program log.
The command line option "-nonews" will prevent this display.



NOTE: SAS initialization used:
      real time           0.16 seconds
      cpu time            0.02 seconds

1          options ls=79 nodate nocenter;
2          data rabbit;
3             input dose nres ntotal;
4             datalines;

NOTE: The data set WORK.RABBIT has 5 observations and 3 variables.
NOTE: DATA statement used (Total process time):
      real time           0.01 seconds
      cpu time            0.00 seconds


10         ;
11         run;
12
13         proc print;
14         run;

NOTE: There were 5 observations read from the data set WORK.RABBIT.
NOTE: The PROCEDURE PRINT printed page 1.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.04 seconds
      cpu time            0.02 seconds


15

NOTE: SAS Institute Inc., SAS Campus Drive, Cary, NC USA 27513-2414
NOTE: The SAS System used:
      real time           0.23 seconds

2 The SAS System

      cpu time            0.04 seconds
```

Type the output lst file to screen:

```
[erike@mizar ~/ada2]$ cat ada2_01_firstprog.lst
The SAS System                                                          1

Obs    dose    nres    ntotal

 1       50       1         5
 2      200       4        10
 3      250       2         5
 4      400       6        10
 5     2000       4         5
```

Additional details on running SAS on UNM's linux system are available on the ADA2 webpage. At the end of the tutorial you will find information on running SAS interactively in linux. This uses a pseudo-windows type environment that you might find to be more palatable than batch SAS if you are on the linux system.

1.4 A Little More SAS

You have seen how to input data in a stream as part of a SAS program. The variables can be either numeric, or categorical (use a $ to indicate a categorical (or string) variable in the **input** statement). The following example illustrates mixed numeric/categorical input and several other items available in a **data step**:

1) You can enter multiple cases per line if you use the @@ option at end of **input**.

2) Variables can be labeled for later use with a **label** statement.

3) Missing values are allowed in a data set. The default missing value label is a period.

4) The **title** statement.

The contents of the SAS program, as seen in the program editor window, and the results of running the program, as contained in the output window, are given below.

Before examining the code, let us re-emphasize some SAS rules. In general, a SAS procedure (or **proc**) has the form

```
proc PROCEDURE_NAME options;
  more options;
run;
```

where certain options are global (i.e., work for all **proc**s) and some are specific to a given **proc**. For the **print** procedure, we specified the `data=DATASETNAME` (global) option, and the **label** option. The **label** option (which is only needed with the **print** procedure), produces output with the variables labeled as defined in the **data step**. The `data=DATASETNAME` option specifies which data set will be analyzed by the **proc**. As noted earlier, several data sets can be defined in a SAS program, and certain **proc**s generate their own output data sets, so this option is recommended to keep things straight. If you omit the `data=DATASETNAME` option, SAS will operate on the most recently created data set. If you are sure you have but one data set floating around then you can omit this option without worry. My practice is to always include the dataset name.

Depending on the procedure, one or more statements may follow the **proc** statement. Below, we used the **title** statement to provide a title that will appear above the output for the **proc**, and a **var** statement, which identifies the variables to be included in the analysis (i.e., printing). If you specify **var amt;**, then you will only get a printout for the **amt** variable. If the **var** statement is omitted, all variables in the data set will be analyzed. (Including the **var** statement is unnecessary here, right?)

Program editor contents:

```
* speeding tickets - using labels;

* text between an asterisk and a semicolon is a comment;
options ls=79 nodate nocenter;    * define sas options;
data tickets;
   input state $ amt @@;
   label state = 'STATE'
         amt   = 'AMOUNT OF SPEEDING TICKET';
   datalines;
    al 60 hi 35 de 31.5 il 20 ak 20 ct 60
    ar 47 ia 33 fl 44.5 nm 20 az 15 ca 50
    dc .
;
```

```
run;

title 'cost of speeding ticket for selected states';
proc print data=tickets label;  * use labels rather than variable names;
   var state amt;
run;
```

Output contents:

```
COST OF SPEEDING TICKET FOR SELECTED STATES                                    1

                 AMOUNT OF
                  SPEEDING
Obs    STATE      TICKET

  1     al          60.0
  2     hi          35.0
  3     de          31.5
  4     il          20.0
  5     ak          20.0
  6     ct          60.0
  7     ar          47.0
  8     ia          33.0
  9     fl          44.5
 10     nm          20.0
 11     az          15.0
 12     ca          50.0
 13     dc            .
```

1.5 Debugging a SAS Program

Debugging a SAS program can be difficult. The log window contains useful information but is sometimes difficult to unravel because of the variety of possible errors you can make. The error messages are often placed after the position of the error. Here is a version of an earlier program with errors. Let us look at the log window and see if we can locate the errors.

Program editor contents:

```
* debugging example;

options ls=79 nodate nocenter;
data d1;
  input pulse height sex;
  datalines;
   62 72 0
   66 70 1
   70 66 1
   49 63 1
   76 76 0
;
run;

proc prnt;
   var pulse sex;
run;

proc print
   var pulse sex;
run;
```

Log windown contents:

```
71   options ls=79 nodate nocenter;
72   data d1;
73     input pulse height sex;
74     datalines;
```

```
NOTE: The data set WORK.D1 has 5 observations and 3 variables.
NOTE: DATA statement used (Total process time):
      real time           0.01 seconds
      cpu time            0.01 seconds


80   ;
81   run;
82
83   proc prnt;
ERROR: Procedure PRNT not found.
84      var pulse sex;
85   run;

NOTE: The SAS System stopped processing this step because of errors.
NOTE: PROCEDURE PRNT used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds

86


87   proc print
88      var pulse sex;
        --- -----
        22  202
ERROR 22-322: Syntax error, expecting one of the following: ;, DATA, DOUBLE,
              HEADING, LABEL, N, NOOBS, OBS, ROUND, ROWS, SPLIT, STYLE,
              UNIFORM, WIDTH.
ERROR 202-322: The option or parameter is not recognized and will be ignored.
89   run;

NOTE: The SAS System stopped processing this step because of errors.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds
```

1.6 Variable Creation, Case Selection and Deletion

In many examples, you will need to transform variables, define new variables from old variables, or select certain cases for analysis. These features are implemented in the **data step**, by placing appropriate commands between the **input** and **datalines** statements.

The following simple program illustrates case selection and variable creation. Let us try to figure out what was done by looking at the program and the output. Note that once a variable is defined it can be referred to later in the **data step**.

Program editor contents:

```
* variable creation, case selection and deletion;

options ls=79 nodate nocenter;
data d1;
  input pulse height sex;
  if height le 69 then delete;   * delete short people;
  htft = height/12;              * height in feet;
  pulpht = pulse + height;       * nonsense variable;
  lacko = .25*(height + height**2)/(pulpht - 12);    * another;
datalines;
62 72 0
66 70 1
70 66 1
 . 78 0
86 70 0
49 63 1
```

```
76 76 0
;
run;

proc print;
run;
```

Output window contents:

```
Obs    pulse    height    sex      htft     pulpht     lacko

 1       62       72       0     6.00000      134     10.7705
 2       66       70       1     5.83333      136     10.0202
 3        .       78       0     6.50000        .        .
 4       86       70       0     5.83333      156      8.6285
 5       76       76       0     6.33333      152     10.4500
```

The rules on creating variables and selecting cases are fairly straightforward. Mathematical operations are performed on a case-by-case basis. We can use standard mathematical operators: $+$, $-$, $*$, $/$, $**$ (exponentiation), and numerous built-in mathematical functions. For example, we can create these new variables in **data step** (assume we have created a data set with variables `v1`, `v2`, and `v3`):

```
v4 = abs(v1);      * absolute value;
v4 = sqrt(v1);     * square root;
v4 = log(v1);      * natural logarithm;
v4 = log10(v1);    * logarithm base 10;
v4 = exp(v1);      * natural exponental;
v4 = sin(v1);      * or cos tan acos atan, etc (in radians);
```

We can create variables from `v4`, once `v4` is defined. For example:

```
v4 = abs(v1);      * create v4;
v5 = 2*v4 - 123;   * create v5 from v4;
```

Selecting and deleting cases is done on a case-by-case basis using comparison operators with variations of the **if** statement. You saw an example above. Valid comparison operators are:

```
lt  or  <
gt  or  >
eq  or  =
le  or  <=
ge  or  >=
ne  or  ^=    (not equal)
```

Here are examples of comparison operators used with an **if** statement:
To include in the (internal SAS) data set only the cases where $v3 = 7$:

```
if v3 = 7;    * keep cases where v3=7 ;
if v3 eq 7;   * this is equivalent;
```

To exclude cases where either $v1 > 20$ or $8 \leq v2 \leq 15$:

```
if ( v1 gt 20 ) or ( 8 <= v2 <= 15  ) then delete;
```

To exclude cases with observation numbers 20 or less use the internal SAS variable for case ID:

```
if _n_ <= 20 then delete;
```

To create a categorical variable $v4$ from $v1$ with levels "low" if $v1 < 20$ and "high" if $v1 \geq 20$:

```
if (v1 < 20)
  then v4 = 'low';
  else v4 = 'high';
```

This is just the tip of the iceberg on ways you can create variables and select cases in the **data step**. Some of these ideas are illustrated in the following program. Looking at the program and output, let us try to figure out what variables have been created and what cases have been deleted. Program editor contents:

```
* if statement for conditional variable assignment;

options ls=79 nodate nocenter;
data d1;
  input pulse height sex;
  if pulse ne .;       * delete missing values;
  htft = height/12;    * height in feet;
  loght = log(htft);   * log of height;
  if htft < 5.5
    then vertchal='yes'; * short;
    else vertchal='no';  * not short;
  if (vertchal = 'yes') or (pulse > 75)
    then agh = 6;      * short or fast pulse, agh=6;
    else agh = loght;  * otherwise, agh=log of height;
  datalines;
  62 72 0
  66 70 1
  70 66 1
   . 78 0
  86 70 0
  49 63 1
  76 76 0
;
run;

proc print;
run;
```

Output window contents:

Obs	pulse	height	sex	htft	loght	vertchal	agh
1	62	72	0	6.00000	1.79176	no	1.79176
2	66	70	1	5.83333	1.76359	no	1.76359
3	70	66	1	5.50000	1.70475	no	1.70475
4	86	70	0	5.83333	1.76359	no	6.00000
5	49	63	1	5.25000	1.65823	yes	6.00000
6	76	76	0	6.33333	1.84583	no	6.00000

1.7 Another SAS Example

This example illustrates several **procs** for data summarization and analysis. The data for this example are from 15 turkeys raised on farms in either Virginia or Wisconsin. The **procs** you will see here are:

1. PRINT - print out data set.
2. SORT - sort data.
3. UNIVARIATE - gives univariate summary information.
4. TTEST - two sample t-test.
5. PLOT - make 2-D plots.

6. FREQ - frequency distributions and contingency tables for categorical variables.

7. CHART - charts and bar graphs.

After we discuss the example, I will mention certain options associated with these **proc**s. You should use **help** to get more information on these (and other) **procs**. Click on **index** in the left panel of the **help** window and enter **PROCNAME proc** to get help on a specific procedure (i.e., type **sort proc** to get more help on the **sort** procedure). A list of topics will then appear. As a first step, I usually display information on the procedure **syntax**.

Many of the lines in the program have (simple) comments, which helps anyone reading the program to more easily follow the logic. SAS has two styles for comments:

```
* the star-asterisk is a good comment style for individual comment lines;

/*
  this slash-star/star-slash style is better
  for big blocks of lines of comments
  or for commenting out blocks or sections of code
*/
```

Comments will appear in the log window but not in the output window. I **strongly recommend** commenting any line of code that isn't absolutely obvious what is being done and why it is being done. I also recommend placing comments before blocks of code in order to describe what the code below is *meant* to do.

```
* turkey example - sort, by, univariate, ttest, plot, freq, chart, boxplot, gplot, gchar

options ls=79 nodate nocenter;
data turkey;
  input age weight orig $;              * Note orig is character $ var ;
  if age gt 25
    then yt = 1; * older than 25;
    else yt = 0; * not older than 25;
  label age    = 'age at death'
        weight = 'weight at death'
        orig   = 'site of farm'
        yt     = 'coded age';
  datalines;
  28 13.3 va
  20  8.9 va
  32 15.1 va
  25 13.8 wi
  23 13.1 wi
  22 10.4 va
  29 13.1 va
  27 12.4 va
  28 13.2 va
  26 11.8 va
  21 11.5 wi
  31 16.6 wi
  27 14.2 wi
  29 15.4 wi
  30 15.9 wi
;
run;

* Print entire data set with title;
title 'print input data set with created variables';
proc print data=turkey;
run;
```

Output window contents:

```
print input data set with created variables                                5

Obs    age    weight    orig    yt

  1     28     13.3      va      1
  2     20      8.9      va      0
  3     32     15.1      va      1
  4     25     13.8      wi      0
  5     23     13.1      wi      0
  6     22     10.4      va      0
  7     29     13.1      va      1
  8     27     12.4      va      1
  9     28     13.2      va      1
 10     26     11.8      va      1
 11     21     11.5      wi      0
 12     31     16.6      wi      1
 13     27     14.2      wi      1
 14     29     15.4      wi      1
 15     30     15.9      wi      1
```

```
* Sort data by orig in ascending order ;
*   (could use BY ORIG WT if needed)   ;
* proc sort "quietly" sorts the data, producing no output;
proc sort data=turkey;
  by orig;
run;

* Separate print for each origin. Data must be presorted by orig. ;
title 'data by state of origin';
proc print data=turkey;
  var age weight;
  by orig;                    * using "by" requires pre-sort;
run;
```

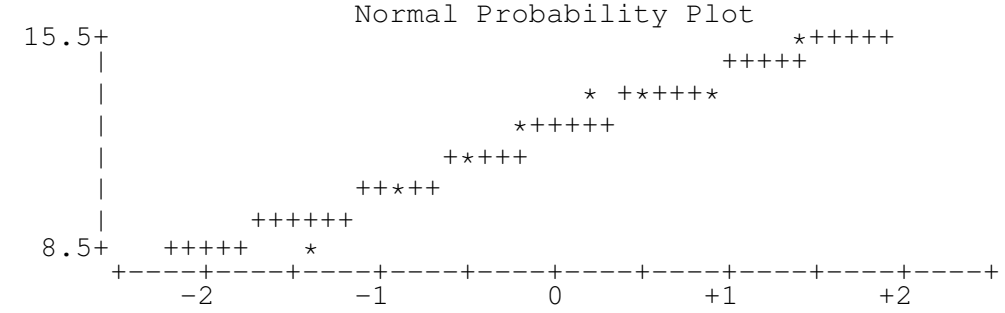
```
data by state of origin                                                    6

site of farm=va

Obs    age    weight

  1     28     13.3
  2     20      8.9
  3     32     15.1
  4     22     10.4
  5     29     13.1
  6     27     12.4
  7     28     13.2
  8     26     11.8


site of farm=wi

Obs    age    weight

  9     25     13.8
 10     23     13.1
 11     21     11.5
 12     31     16.6
 13     27     14.2
 14     29     15.4
 15     30     15.9
```

```
* Plot option gives pictures + default summaries. ;
* specify /normal option to get a formal test of normality. ;
title 'summary of weight data by state';
proc univariate data=turkey plot;
  var weight;
  by orig;            * this creates the side-by-side boxplots;
run;
```

Under the "Tests for Location" are a t-test for the mean=0, a sign test for the median=0, and a Wilcoxon signed rank test that the population median (or mean)=0. Quantiles are actually percentiles. The plots proc univariate produces are among the worst ever made.

```
summary of weight data by state                                          7

site of farm=va

The UNIVARIATE Procedure
Variable:  weight  (weight at death)

                          Moments

N                          8    Sum Weights               8
Mean                    12.275    Sum Observations        98.2
Std Deviation      1.91665631    Variance          3.67357143      s and s^2
Skewness           -0.5455211    Kurtosis          0.38817271
Uncorrected SS        1231.12    Corrected SS          25.715
Coeff Variation    15.6143081    Std Error Mean    0.67764034


             Basic Statistical Measures

    Location                      Variability

Mean     12.27500     Std Deviation            1.91666
Median   12.75000     Variance                 3.67357
Mode        .         Range                    6.20000
                      Interquartile Range      2.15000


          Tests for Location: Mu0=0

Test            -Statistic-    -----p Value------

Student's t     t  18.11433    Pr > |t|    <.0001
Sign            M         4    Pr >= |M|   0.0078
Signed Rank     S        18    Pr >= |S|   0.0078


Quantiles (Definition 5)

Quantile       Estimate

100% Max         15.10
99%              15.10
95%              15.10
90%              15.10
75% Q3           13.25
50% Median       12.75
25% Q1           11.10
10%               8.90
5%                8.90
1%                8.90
0% Min            8.90


        Extreme Observations

----Lowest----        ----Highest---

Value      Obs        Value      Obs

  8.9        2         12.4        6
 10.4        4         13.1        5
 11.8        8         13.2        7
 12.4        6         13.3        1
 13.1        5         15.1        3
```

```
   Stem Leaf                     #  Boxplot
     15 1                        1      |
     14                                 |
     13 123                      3  +-----+
     12 4                        1  *--+--*
     11 8                        1  +-----+
     10 4                        1      |
      9                                 |
      8 9                        1      |
        ----+----+----+----+
```



summary of weight data by state 9

site of farm=wi

The UNIVARIATE Procedure
Variable: weight (weight at death)

Moments

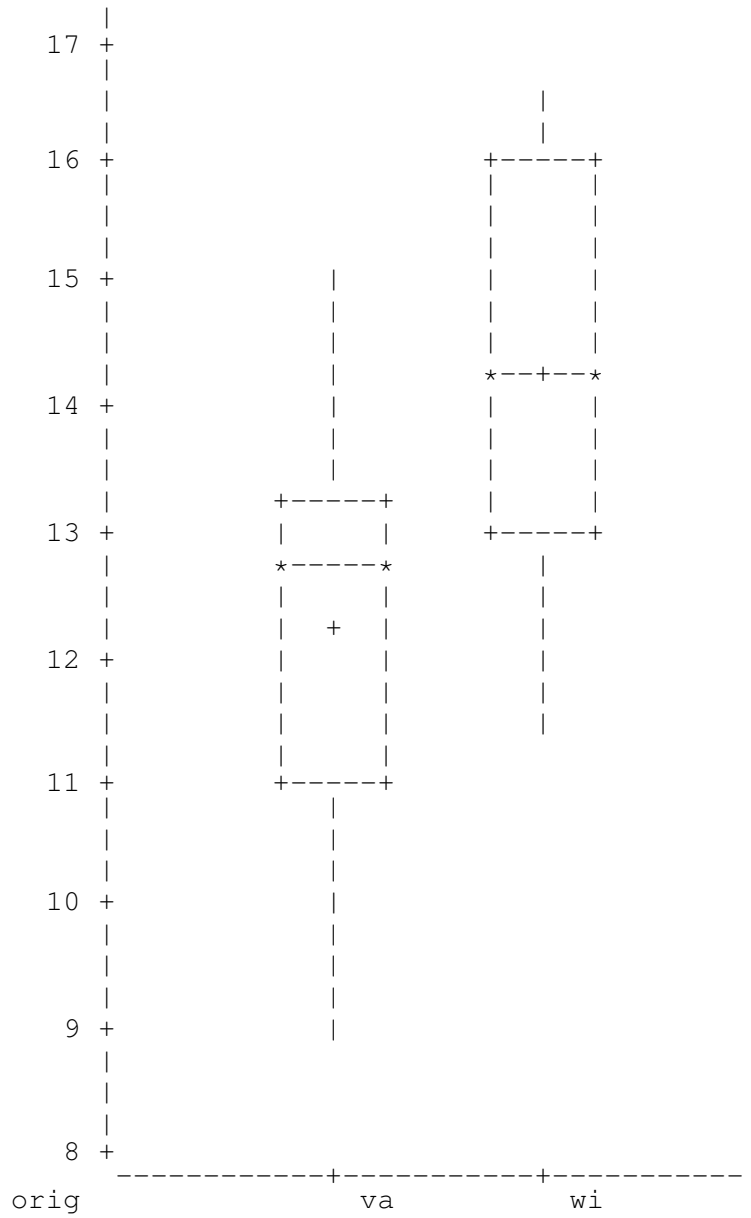
N	7	Sum Weights	7
Mean	14.3571429	Sum Observations	100.5
Std Deviation	1.75959952	Variance	3.09619048
Skewness	-0.4025027	Kurtosis	-0.4644383
Uncorrected SS	1461.47	Corrected SS	18.5771429
Coeff Variation	12.2559171	Std Error Mean	0.66506611

OMITTING THE REST OF THE WISCONSIN SUMMARY TO SAVE SPACE
...

summary of weight data by state 11

The UNIVARIATE Procedure
Variable: weight (weight at death)

Schematic Plots




```
* 2-sample t-tests on both age and wt.                          ;
* Class statement identifies indicator variable for 2 samples. ;
title '2 sample t-tests for age and weight';
proc ttest data=turkey;
  var age weight;
  class orig;
run;
```

2 sample t-tests for age and weight 12

The TTEST Procedure

Statistics

Variable	orig	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev

age	va	8	23.278	26.5	29.722	2.5485	3.8545
age	wi	7	23.117	26.571	30.026	2.407	3.7353
age	Diff (1-2)		-4.32	-0.071	4.1773	2.7548	3.7999
weight	va	8	10.673	12.275	13.877	1.2672	1.9167
weight	wi	7	12.73	14.357	15.985	1.1339	1.7596
weight	Diff (1-2)		-4.146	-2.082	-0.018	1.3381	1.8458

Statistics

Variable	orig	Upper CL Std Dev	Std Err	Minimum	Maximum
age	va	7.8449	1.3628	20	32
age	wi	8.2254	1.4118	21	31
age	Diff (1-2)	6.1219	1.9667		
weight	va	3.9009	0.6776	8.9	15.1
weight	wi	3.8748	0.6651	11.5	16.6
weight	Diff (1-2)	2.9737	0.9553		

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
age	Pooled	Equal	13	-0.04	0.9716
age	Satterthwaite	Unequal	12.8	-0.04	0.9715
weight	Pooled	Equal	13	-2.18	0.0483
weight	Satterthwaite	Unequal	13	-2.19	0.0472

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
age	Folded F	7	6	1.06	0.9552
weight	Folded F	7	6	1.19	0.8509

```
* First plot weight by age,                        ;
* then use orig (wv or va) as plotting symbol.    ;
*    Vpos and Hpos control plot size.             ;
title 'illustration of plot features';
proc plot data=turkey;
  plot weight*age      /vpos=16 hpos=32;
  plot weight*age=orig /vpos=16 hpos=32;
run;
```

```
illustration of plot features                                          13

            Plot of weight*age.  Legend: A = 1 obs, B = 2 obs, etc.

w       |
e       |
i 17.5 +
g       |                         A                      * I replaced the
h       |                       A                          axes manually
t 15.0 +                     A     A                       since the
        |           A   A                                  "graphical"
a       |      A          B A                              characters used
t 12.5 +                 A                                in the output
        |  A          A                                    window do not
d       |                                                  render correctly
e 10.0 +    A                                            as (not) seen
a       |A                                                 below.
t       |
h  7.5 +
        |
        -+---------+---------+---------+-
        20         25         30         35

                    age at death
```

```
illustration of plot features                                          14

                   Plot of weight*age.  Symbol is value of orig.

w
e
i 17.5
g                                 w
h                               w
t 15.0                        w     v
                    w   w
a             w           v v
t 12.5                    v
          w           v
d
e 10.0      v
a       v
t
h  7.5

        Ł
        20         25         30         35

                    age at death

NOTE: 1 obs hidden.
```

```
* Contingency table of origin by categorized age ;
title 'frequency distribution for orig by age';
proc freq data=turkey;
  tables orig*yt/chisq;
run;
```

Below is the cross-classification of **orig** by **yt**, generated by the **freq** procedure. The 4 entries in each cell are: the cell frequency; the % of the entire data set in cell; the % of the row in the cell; and the % of the column in the cell. The `/chisq` option gives us the Pearson and likelihood ratio tests for independence in the two-way table. Output for Fisher's exact test (which does not require large samples) is also provided.

```
frequency distribution for orig by age                                 15
```

```
The FREQ Procedure

Table of orig by yt

orig(site of farm)     yt(coded age)

Frequency|
Percent  |
Row Pct  |
Col Pct  |       0|       1|  Total
---------+--------+--------+
va       |      2 |      6 |      8
         |  13.33 |  40.00 |  53.33
         |  25.00 |  75.00 |
         |  40.00 |  60.00 |
---------+--------+--------+
wi       |      3 |      4 |      7
         |  20.00 |  26.67 |  46.67
         |  42.86 |  57.14 |
         |  60.00 |  40.00 |
---------+--------+--------+
Total          5       10       15
            33.33    66.67   100.00
```

* I manually put the lines back in this table and the two tables and plot below, too

```
Statistics for Table of orig by yt

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1      0.5357    0.4642
Likelihood Ratio Chi-Square    1      0.5373    0.4635
Continuity Adj. Chi-Square     1      0.0335    0.8548
Mantel-Haenszel Chi-Square     1      0.5000    0.4795
Phi Coefficient                      -0.1890
Contingency Coefficient               0.1857
Cramer's V                           -0.1890

 WARNING: 75% of the cells have expected counts less
          than 5. Chi-Square may not be a valid test.


       Fisher's Exact Test
---------------------------------
Cell (1,1) Frequency (F)          2
Left-sided Pr <= F           0.4266
Right-sided Pr >= F          0.8998

Table Probability (P)        0.3263
Two-sided Pr <= P            0.6084

Sample Size = 15
```

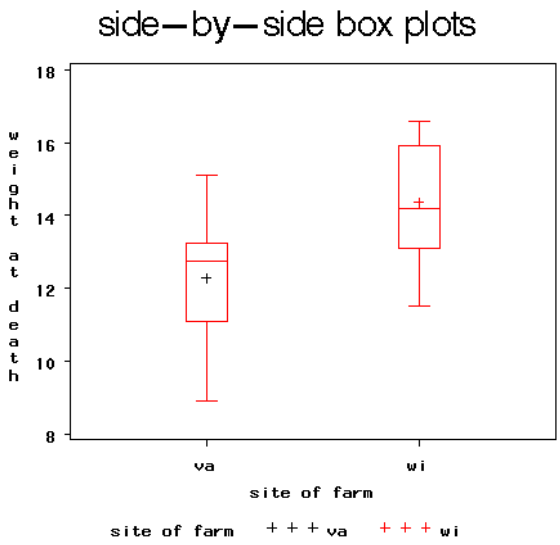
```
* a horizontal bar chart ;
proc chart data=turkey;
  hbar orig/type=percent;
run;
```

```
frequency distribution for orig by age                                  16

site of farm                                    Cum.                Cum.
                                        Freq    Freq    Percent     Percent
        |
va      |**************************       8       8     53.33       53.33
        |
wi      |***********************          7      15     46.67      100.00
        |
        +---------------------------
             10   20   30   40   50

                  Percentage
```

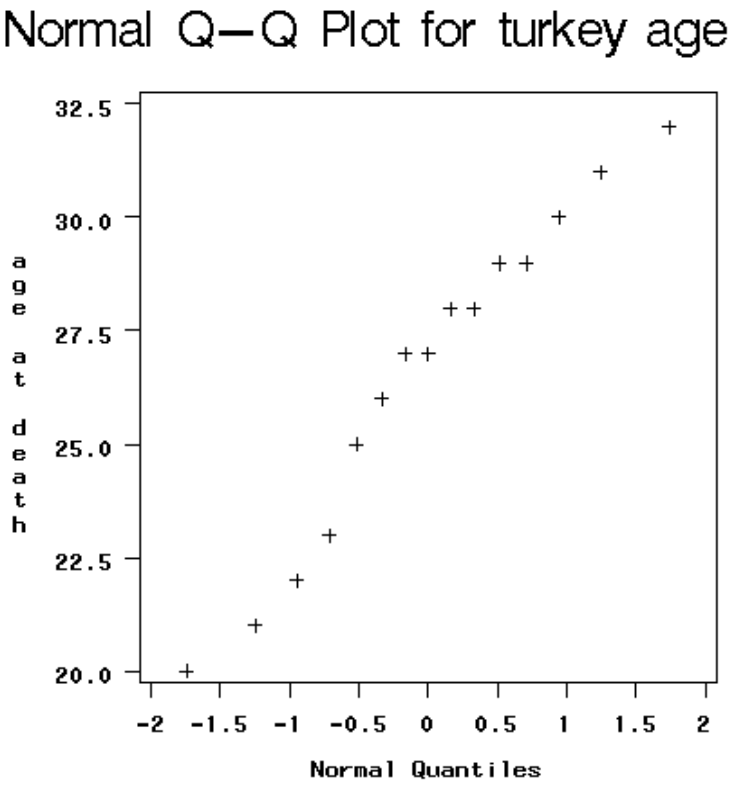
1.7.1 Creating better graphs

Here are a selection of plots redone as graphics to replace those seen above in text format.

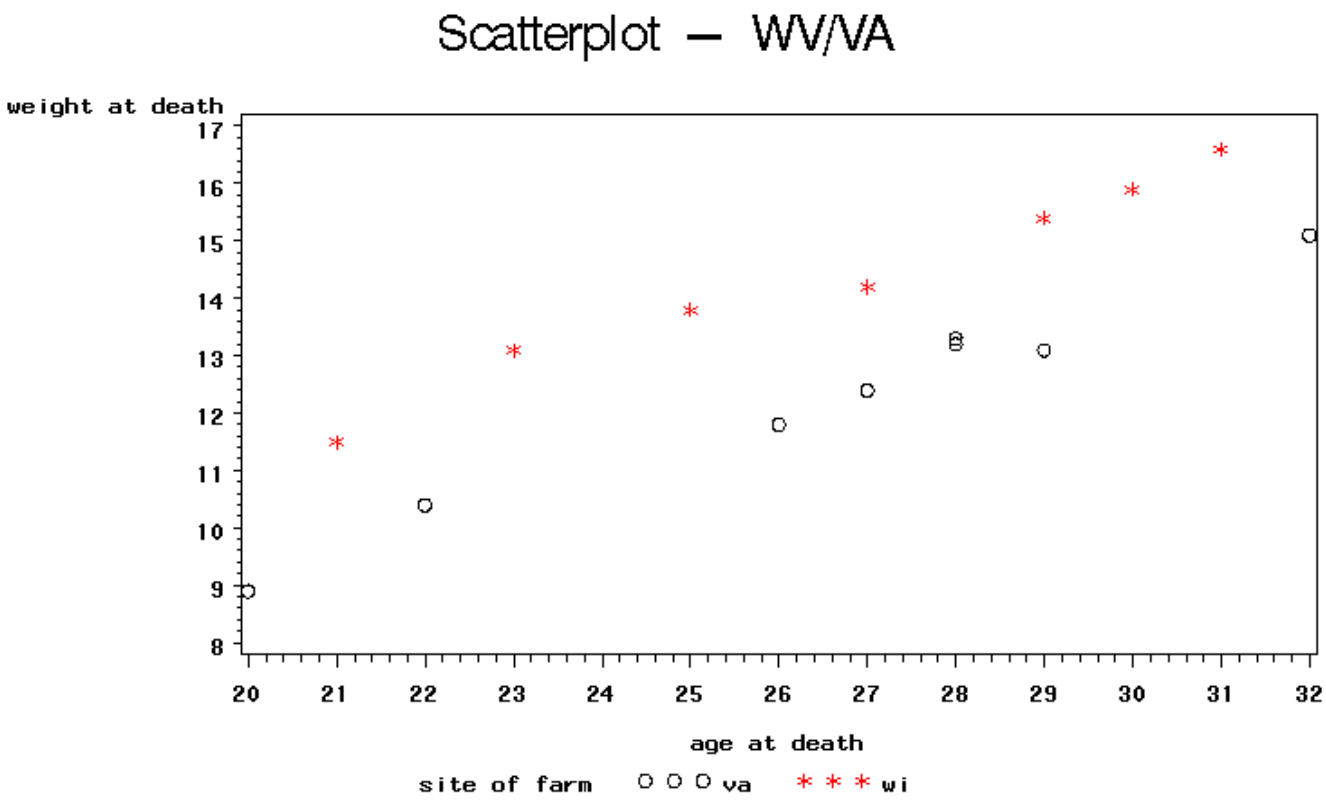
```
title 'side-by-side box plots';
proc boxplot data=turkey ;
  plot weight*orig=orig /boxstyle=schematic;
run;
```



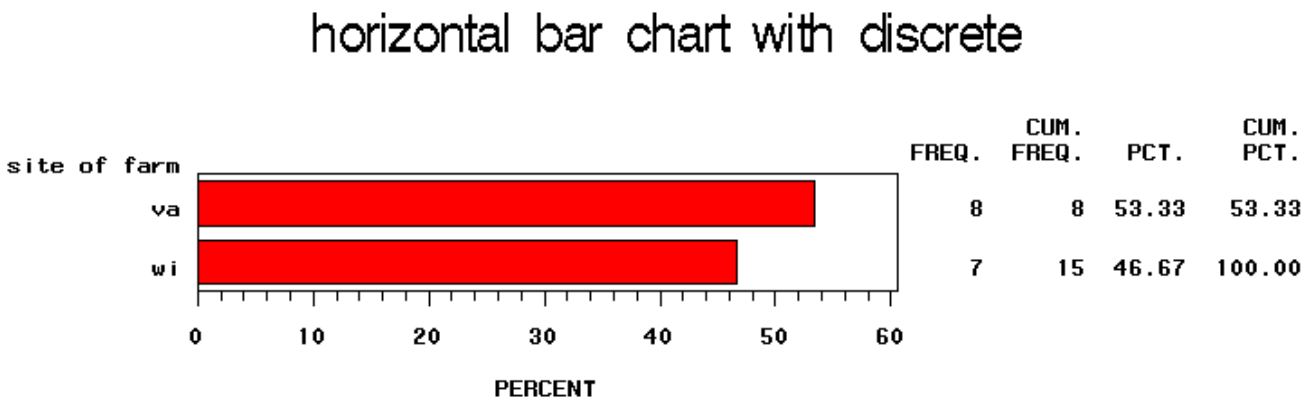

```
symbol v=plus;
title 'Normal Q-Q Plot for turkey age';
proc univariate data=turkey noprint;
   qqplot age / normal square vaxis=axis1;
   axis1 label=(a=90 r=0);
run;
```




```
title 'Scatterplot - WV/VA';
symbol1 v=circle c=black i=none;  * define v=symbol, c=color, ;
symbol2 v=star   c=red   i=none;  *   and i=interpolation     ;
  *symbol1 v=circle c=blue i=r;  * i=r makes a regression line;
proc gplot data=turkey;
  plot weight*age=orig;  * weight vert, age horiz, orig symbol;
run;
```




```
title 'horizontal bar chart with discrete';
proc gchart data=turkey;
  hbar orig /type=percent discrete;
run;
```



1.8 Selected Details on Procs

The form and some common options for five of the SAS **procs** you have learned are given below.

proc print *options* ;
Three common options for the **print** procedure:

> **data=DATASETNAME** – names the data set to be analyzed. **The most recently created data set is analyzed if this option is omitted!** This option is available for all **procs**.
>
> **label** – uses variable labels in printout. The variable name is used if this option is not specified.
>
> **double** – double spaces the output.

Statements that can follow the **print** statement:

> **var** *variables* ; – names the variables to be printed. If this is omitted then entire data set is printed.
>
> **by** *variables* ; – separate output for each level of **by** variables. The data MUST be sorted beforehand on the **by** variables.

proc sort *options* ;
Three common options for the **sort** procedure:

data=DATASETNAME – as above.

nodup – checks and eliminates duplicate records.

out=DATASETNAME – gives a name to the sorted data set. If omitted, SAS provides a name (see the log window for chosen name).

One statement **must** follow the **sort** statement:

by *option variable-name option variable-name* ; – sorts data sequentially using listed variables. Unless you specify the option of **descending** before the variable name, the data is sorted in increasing alphanumeric order.

proc ttest *option* ;
The only available option is **data=DATASETNAME**. Statements that can follow the **ttest** procedure statement:

class *variable* ; – name of the grouping variable, which must assume two and only two levels. This statement is required. The **class** variable can be numeric or character.

var *variables* ; – names the variables on which the groups are compared. If this is omitted then all numeric variables in the data set (except the class variable if it is numeric) are considered.

by *variables* ; – results in separate analyses for all combinations of **by** variables. You must sort the data beforehand on the **by** variables.

proc freq *options* ;
Three common options for the **freq** procedure:

data=DATASETNAME – as above.

noprint – suppresses all printed output.

order = – specifies ordering of levels for variables. Two choices are **order=data** (by order of occurrence in data set) and **order=formatted** (by ascending format order).

Statements that can follow the **freq** procedure statement:

tables *a a*b a*b*c* ; – examples of how to specify one, two, or three variables in frequency distribution. This statement has options for computing chi-square tests and measures of association for two-way tables.

weight *variable* ; – uses specified variable as a column of frequencies to generate counts for requested tables. A weight variable is needed for analyzing summary tables of counts — think back to how we entered contingency tables in **Minitab**).

by *variables* ; – results in separate analyses for all combinations of **by** variables. You must sort the data beforehand on the **by** variables.

proc univariate *options* ;
Five common options for the **univariate** procedure:

data=DATASETNAME – as above.

noprint – suppresses all printed output.

plot – gives graphical summaries.

normal – test statistic for testing normality.

cibasic – basic CIs for means, standard deviation, assuming normality.

Statements that can follow the **univariate** procedure statement:

var *variables* ; – names the variables to be analyzed.

histogram *variables* ; – creates **high resolution** histograms (for Windows version) for each variable listed (must be a subset of those listed on the **var** statement)

qqplot *variables* ; – creates **high resolution** normal quantile (rankit) plots (for Windows version) for each variable listed (must be a subset of those listed on the **var** statement). Other distributional plots are possible.

by *variables* ; – gives separate analyses for each combination of **by** variables. You must sort the data beforehand on the **by** variables.

output out=DATASETNAME *keyword=names* ; – saves summary output to the named data set. The keyword option identifies the summaries that you want in the new data set, using one variable name to the right of the equal sign to identify that feature for each variable in the **var** statement. Commonly used keywords are MEAN, SUM, STD (standard deviation), VAR, NOBS (sample size), NMISS (number of missing values), MEDIAN, MIN, MAX, P1, P5, P10, P90, P95, P99 (percentiles), Q1, Q3, and QRANGE (the IQR).

For example, the SAS program below saves the mean, median, and standard deviations of HT and WT for each level of SEX in a data set called SHOWIT. This data set could be used in a subsequent analysis, but is printed here for illustration. You should use the **noprint** option if you are primarily interested in saving selected summaries and do not want to slog through a forest of **univariate** output.
Program editor contents:

```
* creating an output dataset from proc univariate;

data d1;
  input ht wt sex;
  datalines;
  4 5 1
  3 7 2
  5 7 1
  3 2 2
  2 5 2
;
run;

proc sort data=d1;
```

```
  by sex;
run;

proc univariate data=d1 noprint;
  var ht wt;
  by sex;
  output out    = showit
         mean   = mht mwt
         median = medht medwt
         std    = stdht stdwt;
run;

proc print data=showit;
run;
```

Output window contents:

Obs	sex	mht	mwt	stdht	stdwt	medht	medwt
1	1	4.50000	6.00000	0.70711	1.41421	4.5	6
2	2	2.66667	4.66667	0.57735	2.51661	3.0	5

1.9 One-Way ANOVA in SAS

SAS has two procedures for general analysis of variance: **anova** and **glm**. The **anova** procedure is restricted to one-way ANOVA and balanced multi-factor ANOVA. The **glm** procedure is more general, allowing us to fit fairly complex unbalanced ANOVA and regression type models, so there is not much lost by focusing on **glm**.

I will use the fat absorption in donuts data from last semester to illustrate one-way ANOVA in SAS. For a one-way ANOVA you must define 2 variables in the data set. Here, **fat** is a grouping variable that identifies the type of fat used and **abs** gives the amount of fat absorbed for a given batch of 24 doughnuts. In **glm**, use the **class** statement to identify the variable that contains sample identifiers. The **model** statement has the form **response var name = grouping variable** for a one-way ANOVA. The **means** statement provides the mean response (abs) at each fat level. The **bon** and **lsd** options give Bonferroni and FSD (Fisher) multiple comparisons. The default level for the individual comparisons is $\alpha = 0.05$ for the FSD method and 0.05/#comparisons for the Bonferroni method. The default 0.05 level can be changed, say, to 0.01 via **means fat / bon lsd alpha=0.01**. Other multiple comparison methods are available in SAS. See below.

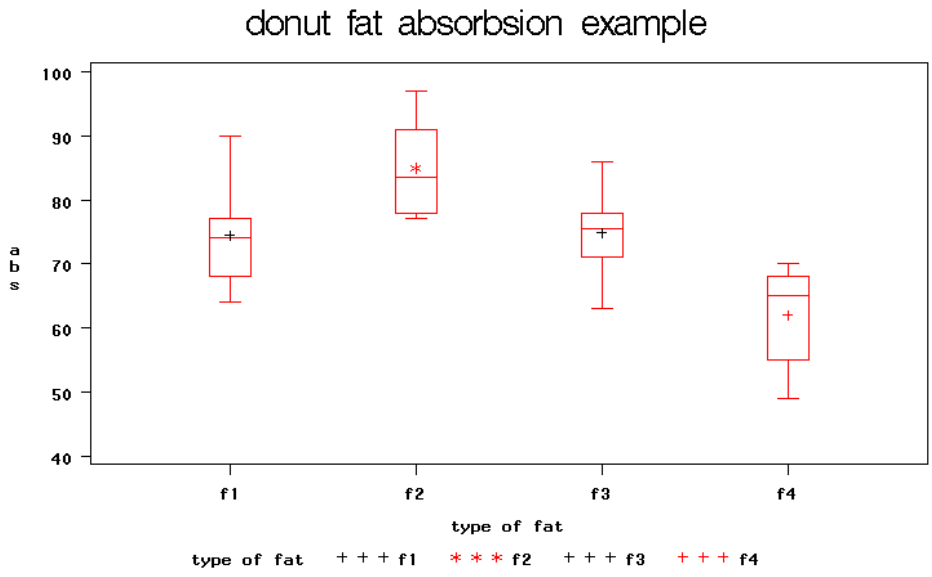
Program editor contents:

```
* using @@ for multiple observations on a line;
* donut - anova analysis;

options ls=79 nodate nocenter;
title 'donut fat absorbsion example';
data donut;
  input fat $ abs @@;  * fat type is a character string $ ;
  label fat = 'type of fat'
        abs = 'amt of fat absorbed in grams';
datalines;
f1 64 f1 72 f1 68 f1 77 f1 90 f1 76
f2 78 f2 91 f2 97 f2 82 f2 85 f2 77
f3 75 f3 86 f3 78 f3 71 f3 63 f3 76
f4 55 f4 66 f4 49 f4 64 f4 70 f4 68
;
run;

proc print data=donut;
run;
```

```
proc boxplot data=donut;
  plot abs*fat=fat /boxstyle=schematic;
run;

* ANOVA of absorbsions by fat type;
proc glm data=donut;
  class fat;   * Identifies samples. The class var could be numeric ;
  model abs = fat;
  means fat / bon lsd;
run;
```

Output window contents:

```
donut fat absorbsion example                                    30

Obs     fat     abs

  1     f1       64
  2     f1       72
  3     f1       68
  4     f1       77
  5     f1       90
  6     f1       76
  7     f2       78
  8     f2       91
  9     f2       97
 10     f2       82
 11     f2       85
 12     f2       77
 13     f3       75
 14     f3       86
 15     f3       78
 16     f3       71
 17     f3       63
 18     f3       76
 19     f4       55
 20     f4       66
 21     f4       49
 22     f4       64
 23     f4       70
 24     f4       68
```




```
The GLM Procedure
      Class Level Information
Class          Levels    Values
```

```
fat                 4    f1 f2 f3 f4                  <- # of levels and
                                                        labels for class var

Number of Observations Read          24
Number of Observations Used          24

Dependent Variable: abs   amt of fat absorbed in grams

                                    Sum of
Source                      DF      Squares    Mean Square   F Value   Pr > F
Model                        3   1595.500000     531.833333      7.95   0.0011
Error                       20   1338.333333      66.916667
Corrected Total             23   2933.833333

                                        Note: Mean Square error is the
                                              pooled variance estimate,
                                              whereas Root MSE is the
                                              pooled standard deviation.
                                              Can ignore bottom output.

R-Square     Coeff Var      Root MSE      abs Mean
0.543828      11.04197      8.180261      74.08333

Source                      DF      Type I SS    Mean Square   F Value   Pr > F
fat                          3   1595.500000     531.833333      7.95   0.0011

Source                      DF    Type III SS    Mean Square   F Value   Pr > F
fat                          3   1595.500000     531.833333      7.95   0.0011



t Tests (LSD) for abs

NOTE: This test controls the Type I comparisonwise error rate, not the
experimentwise error rate.

Alpha                              0.05
Error Degrees of Freedom             20
Error Mean Square             66.91667       <- pooled var
Critical Value of t            2.08596
Least Significant Difference    9.8517       <- same for each
                                              comparison

Means with the same letter are not significantly different.

t

G
r
o
u
p
i
n
g          Mean      N    fat

A        85.000      6    f2

B        74.833      6    f3
B
B        74.500      6    f1

C        62.000      6    f4
```

Orders fats by sample means. Groups defined by letters, i.e.

```
  f4   f1   f3   f2
  --   -------   --
```

```
Bonferroni (Dunn) t Tests for abs
```

```
NOTE: This test controls the Type I experimentwise error rate, but it
generally has a higher Type II error rate than REGWQ.

Alpha                              0.05
Error Degrees of Freedom             20
Error Mean Square              66.91667
Critical Value of t             2.92712
Minimum Significant Difference   13.824

Means with the same letter are not significantly different.

 Bon
Groupi
  ng             Mean      N    fat

     A         85.000      6    f2
     A
B    A         74.833      6    f3
B    A
B    A         74.500      6    f1
B
B              62.000      6    f4
```

Two groups, with overlap, that is,

```
  f4   f1   f3   f2
  ------------
       ------------
```

1.9.1 Additional Multiple Comparison Methods for ANOVA

If the **means** statement in the **glm proc** is

```
means fat / bon lsd cldiff;
```

then the FSD and Bonferroni output is summarized using all possible CI for differences in means, as in **Minitab**. This is the default output style when the group sample sizes are unequal.

If the **means** statement in the **glm proc** is

```
means fat / bon lsd snk;
```

then you also get multiple comparisons based on Student-Newman-Keuls multiple range test. This is a commonly used alternative to the Bonferroni and FSD methods. Other possibilities are **regwq** for Ryan-Einot-Gabriel-Welsh multiple range comparisons, **tukey** for Tukey's honest significant difference method, **scheffe**, **waller**, **sidak**, etc. The Bonferroni method is usually the most conservative multiple comparison method.

The Student-Newman-Keuls output looks like:

```
Student-Newman-Keuls Test for abs

NOTE: This test controls the Type I experimentwise error rate under the
complete null hypothesis but not under partial null hypotheses.


Alpha                          0.05
Error Degrees of Freedom         20
Error Mean Square          66.91667


Number of Means                2             3             4
Critical Range       9.8517518     11.948791     13.219032


Means with the same letter are not significantly different.
```

```
S
N
K

G
r
o
u
p
i
n
g          Mean      N    fat

A        85.000      6    f2
A
A        74.833      6    f3
A
A        74.500      6    f1

B        62.000      6    f4
```

This gives two groups, with no overlap. Contrast this with the other comparisons.

```
  f4   f1   f3   f2
  --  --------------
```

1.10 Additional SAS Topics on Data Management

Using an INFILE Statement It is often easier to store your data in a (**text**) file and keep the SAS program, with the **proc** steps, free from data. The data file can be read into SAS using an **infile** statement. In the fats problem, the raw data was put into a file called **fats.dat**, as shown below. Any valid file name can be used. The **infile** statement must preceed the **input** statement. The data file must be referenced using a valid path, as shown below. Friendly's handout shows an alternative way to use the **infile** statement, following a **filename** statement.

```
  f1 64 f1 72 f1 68 f1 77 f1 90 f1 76
  f2 78 f2 91 f2 97 f2 82 f2 85 f2 77
  f3 75 f3 86 f3 78 f3 71 f3 63 f3 76
  f4 55 f4 66 f4 49 f4 64 f4 70 f4 68
```

To get the **same dataset** as our original fats data step, use the following program:

```
* using "infile" to read data from a file;

options ls=79 nodate;
data donut;
  infile 'ADA2_01_fats.dat';  * Infile with data file location (eg c:\fats.dat);
  input fat $ abs @@;                 * string $ ;
  label fat = 'type of fat'
        abs = 'amt of fat absorbed in grams';       * End of data step ;
run;
proc print;
run;
```

A separate " ; " line is not needed at the end of the **data step** when you use the **infile** statement.

It is important to distinguish two types of data files: one for which the data entries are separated by blanks, as above, and one for which the data entries are separated by **tab** characters. If you have **tab** delimited data, you must use the **expandtabs** option at the end of the **infile** statement.

The SET command The **set** command can be used in a **data step** to modify a previously created data set. For example, if a data set **d1** exists and you wish to select out the individuals with gender='M', you can specify:

```
data d2;           * new data set d2;
  set d1;          * first, copy data set d1;
  if gender='M';   * keep only those records with gender='M';
run;
```

The resulting data set **d2** will contain all records in **d1** for which gender='M'. You now have at least two data sets in your SAS program, so using the **data=DATASETNAME** option on **proc**s is advised.

The DROP and KEEP commands in a DATA STEP You can reduce the number of variables in a data set using the **drop** and **keep** commands. See the online help for details.

The DATASETS procedure The **datasets** procedure can be used to delete data sets from SAS or get information about a dataset. See the online help for details.

The FORMAT procedure The **format** procedure allows you to assign labels to the levels of a categorical or numerical variable.

The following example illustrates some of these features.

```
* turkey - format, drop/keep, output dataset;

title 'Further analysis of the turkey data set';
options ls=79 nodate nocenter;
data turkey;
  infile 'ADA2_01_turkey.dat';
  input age weight orig $;              * Note orig is character $ var ;
  if age gt 25
    then yt = 1; * older than 25;
    else yt = 0; * not older than 25;
  label age    = 'age at death'
        weight = 'weight at death'
        orig   = 'site of farm'
        yt     = 'coded age';
run;

proc sort data=turkey;
  by orig;
run;

* Illustrate the use of format - note form of output;
proc format;
  value ytfmt 0='young' 1='old';
  value weightfmt 1-12='light' 12.1-40='heavy';
run;

proc print data=turkey;
  var age weight orig yt;
  format yt ytfmt. weight weightfmt. ;
  * How would output look without format? ;
run;

proc freq data=turkey;
  tables weight;
  format weight weightfmt.;
  * What would happen without the format? ;
run;
```

```
* Create a new data set turkey2 from turkey ;
data turkey2;
  set turkey;
  if weight > 16 then delete;    * Delete heaviest turkey ;
  drop age yt;                   * Drop age and yt vars   ;
run;

proc print data = turkey2;
  format weight weightfmt. ;
run;

* Create output data set tsumm with mean and median weight  ;
* Also, suppress printing of all output with NOPRINT option ;
proc univariate data=turkey2 noprint;
  var weight;
  by orig;
  output out=tsumm mean=meanwt median=medwt;
run;

* Print out summarized data contained in tsumm ;
proc print data = tsumm;
run;
```

Abbreviated log window contents:

```
1    title 'Further analysis of the turkey data set';
2    options ls=79 nodate nocenter;
3    data turkey;
4      infile 'F:\Dropbox\UNM\teach\ADA2_stat528\sas\ADA2_01_turkey.dat';
5      input age weight orig $;             * Note orig is character $ var ;
6      if age gt 25
7        then yt = 1; * older than 25;
8        else yt = 0; * not older than 25;
9      label age    = 'age at death'
10           weight = 'weight at death'
11           orig   = 'site of farm'
12           yt     = 'coded age';
13   run;

NOTE: The infile 'F:\Dropbox\UNM\teach\ADA2_stat528\sas\ADA2_01_turkey.dat' is:
      File Name=F:\Dropbox\UNM\teach\ADA2_stat528\sas\ADA2_01_turkey.dat,
      RECFM=V,LRECL=256

NOTE: 15 records were read from the infile
      'F:\Dropbox\UNM\teach\ADA2_stat528\sas\ADA2_01_turkey.dat'.
      The minimum record length was 12.
      The maximum record length was 12.
NOTE: The data set WORK.TURKEY has 15 observations and 4 variables.
NOTE: DATA statement used (Total process time):
      real time           1.17 seconds
      cpu time            0.03 seconds


14
15   proc sort data=turkey;
16     by orig;
17   run;

NOTE: There were 15 observations read from the data set WORK.TURKEY.
NOTE: The data set WORK.TURKEY has 15 observations and 4 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time           0.20 seconds
      cpu time            0.01 seconds


18
19   * Illustrate the use of format - note form of output;
20   proc format;
21     value ytfmt 0='young' 1='old';
NOTE: Format YTFMT has been output.
```

```
22     value weightfmt 1-12='light' 12.1-40='heavy';
NOTE: Format WEIGHTFMT has been output.
23   run;

NOTE: PROCEDURE FORMAT used (Total process time):
      real time           0.18 seconds
      cpu time            0.01 seconds
```

Output window contents:

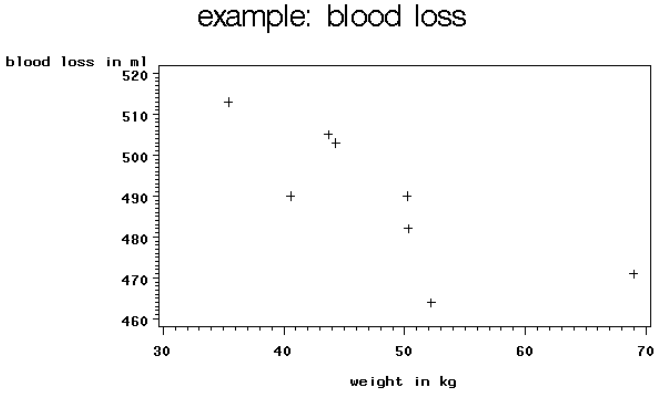
```
Further analysis of the turkey data set

Obs    age    weight    orig     yt
  1     28    heavy      va     old
  2     20    light      va     young
  3     32    heavy      va     old
  4     22    light      va     young
  5     29    heavy      va     old
  6     27    heavy      va     old
  7     28    heavy      va     old
  8     26    light      va     old
  9     25    heavy      wi     young
 10     23    heavy      wi     young
 11     21    light      wi     young
 12     31    heavy      wi     old
 13     27    heavy      wi     old
 14     29    heavy      wi     old
 15     30    heavy      wi     old


The FREQ Procedure
                      weight at death
                                          Cumulative    Cumulative
weight    Frequency     Percent     Frequency      Percent
-----------------------------------------------------------------
 light            4       26.67             4        26.67
 heavy           11       73.33            15       100.00


Obs    weight    orig
  1    heavy      va
  2    light      va
  3    heavy      va
  4    light      va
  5    heavy      va
  6    heavy      va
  7    heavy      va
  8    light      va
  9    heavy      wi
 10    heavy      wi
 11    light      wi
 12    heavy      wi
 13    heavy      wi
 14    heavy      wi


Obs    orig     meanwt    medwt
 1      va     12.2750    12.75
 2      wi     13.9833    14.00
```

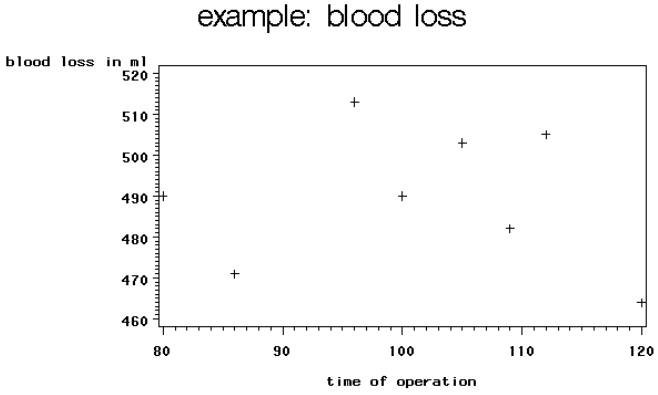
1.11 Correlation in SAS

The **corr** procedure gives Pearson and Spearman correlations and p-values for testing that the corresponding population correlations are zero. Here is SAS code for generating data plots and correlations for the blood loss data that we have come to know.

```
* bloodloss - correlation;

title 'example: blood loss';
options ls=79 nodate nocenter;
data bloodloss;
  input wt ti bl;
  label wt='weight in kg'
        ti='time of operation'
        bl='blood loss in ml';
  datalines;
  44.3 105 503
  40.6  80 490
  69.0  86 471
  43.7 112 505
  50.3 109 482
  50.2 100 490
  35.4  96 513
  52.2 120 464
;
run;

proc gplot data=bloodloss;
  plot bl*wt bl*ti;   * Multiple plots per statement, same as bl*(wt ti);
run;

proc corr data=bloodloss pearson spearman nosimple;
  var wt ti bl;
      * Pearson and Spearman corr for   ;
      *   each pair of vars in var list ;
      * nosimple suppresses simple summary statistic output. ;
run;
```

Output window contents:





example: blood loss 7

The CORR Procedure

3 Variables: wt ti bl

Pearson Correlation Coefficients, N = 8
Prob > |r| under H0: Rho=0

	wt	ti	bl
wt	1.00000	-0.06626	-0.77247
weight in kg		0.8761	0.0247
ti	-0.06626	1.00000	-0.10735
time of operation	0.8761		0.8003

```
bl                      -0.77247       -0.10735        1.00000
blood loss in ml          0.0247         0.8003


          Spearman Correlation Coefficients, N = 8
                Prob > |r| under H0: Rho=0

                                wt             ti             bl

wt                        1.00000        0.28571       -0.87427
weight in kg                             0.4927         0.0045

ti                        0.28571        1.00000       -0.15569
time of operation          0.4927                        0.7128

bl                      -0.87427       -0.15569        1.00000
blood loss in ml          0.0045         0.7128
```

Other common options on the **corr** statement are **noprob** (suppress printing of p-values), **kendall** (Kendall's τ coefficient), and **cov** (print pairwise covariances). The Pearson correlation is the default method if you do not specify the **pearson**, **spearman**, or **kendall** options. You can use a **by** statement with **corr**. See the online help for more details.

1.12 Regression in SAS

SAS has two main procedures for regression analysis: **reg** and **glm**. The **reg** procedure has several nice features for automated model selection and diagnostic analyses that are not available in **glm**, so I will initially focus on using **reg**. I will provide minimal details on **reg** (and **glm**) at this time because it has too many options. See the online help for more details.

As an example, the SAS program below fits a linear regression model to the blood loss data, using blood loss as the response variable and weight as the predictor. The form of the **model** statement is fairly natural, with the response and predictor variables appearing to the left and to the right of the equal sign, respectively. The **p**, **r**, **cli**, and **clm** options on the **model** statement (to the right of the /) give summary information on **p**= predicted values, **r**= residuals, **cli**= 95% CI for individual responses, and **clm**= 95% CI for the mean response at each weight in the data set.

The last two lines of the **data step** give WTs, but are missing TI and BL. These cases are not used in the calculation of the LS line, but the output will specify fitted values and summary information for these cases.

You should compare the output with our **Minitab** notes to see that they agree. I will point out several notable features.

```
* bloodloss - regression, fitting new observations, CIs and PIs;

options ls=79 nodate nocenter;
title 'blood loss regression bl=wt';
data bloodloss2;
  input wt ti bl;
  label wt='weight in kg'
        ti='time of operation'
        bl='blood loss in ml';
  datalines;
  44.3 105 503
  40.6  80 490
  69.0  86 471
```

```
  43.7 112 505
  50.3 109 482
  50.2 100 490
  35.4  96 513
  52.2 120 464
  60.0   .   .
  45.0   .   .
;
run;

proc reg data=bloodloss2;
  model bl = wt /p r cli clm; * pred, resid, PI new resp, CI mean resp;
run;
```

Output window contents:

```
blood loss regression bl=wt

The REG Procedure
Model: MODEL1
Dependent Variable: bl blood loss in ml

Number of Observations Read                          10
Number of Observations Used                           8
Number of Observations with Missing Values            2

                              Analysis of Variance

                                   Sum of           Mean
Source                     DF     Squares         Square    F Value    Pr > F
Model                       1  1207.45129     1207.45129       8.88    0.0247
Error                       6   816.04871      136.00812
Corrected Total             7  2023.50000

Root MSE              11.66225    R-Square     0.5967
Dependent Mean       489.75000    Adj R-Sq     0.5295
Coeff Var              2.38127

                           Parameter Estimates

                                   Parameter     Standard
Variable   Label             DF     Estimate        Error  t Value  Pr > |t|
Intercept  Intercept          1    552.44202     21.44088    25.77    <.0001
wt         weight in kg       1     -1.30033      0.43642    -2.98    0.0247

                         Output Statistics

     Dependent    Predicted       Std Error
Obs   Variable        Value    Mean Predict           95% CL Mean

  1   503.0000     494.8375          4.4628     483.9175     505.7576
  2   490.0000     499.6487          5.2951     486.6921     512.6054
  3   471.0000     462.7194          9.9650     438.3359     487.1030
  4   505.0000     495.6177          4.5694     484.4368     506.7986
  5   482.0000     487.0356          4.2227     476.7031     497.3681
  6   490.0000     487.1656          4.2135     476.8556     497.4756
  7   513.0000     506.4104          6.9474     489.4107     523.4102
  8   464.0000     484.5649          4.4754     473.6140     495.5159
  9          .     474.4224          6.5927     458.2905     490.5543
 10          .     493.9273          4.3551     483.2708     504.5838

                         Output Statistics

                                          Std Error   Student
Obs     95% CL Predict          Residual   Residual  Residual    -2-1 0 1 2

  1   464.2830   525.3921     8.1625     10.775     0.758  |      |*     |
  2   468.3086   530.9889    -9.6487     10.391    -0.929  |     *|      |
  3   425.1843   500.2546     8.2806      6.059     1.367  |      |**    |
```

```
  4   464.9690   526.2664     9.3823     10.730     0.874  |       |*      |
  5   456.6861   517.3851    -5.0356     10.871    -0.463  |       |       |
  6   456.8237   517.5075     2.8344     10.875     0.261  |       |       |
  7   473.1941   539.6267     6.5896      9.367     0.703  |       |*      |
  8   453.9994   515.1305   -20.5649     10.769    -1.910  |   ***|       |
  9   441.6418   507.2030           .          .         .
 10   463.4660   524.3886           .          .         .

Output Statistics

       Cook's
Obs         D

  1     0.049
  2     0.112
  3     2.527
  4     0.069
  5     0.016
  6     0.005
  7     0.136
  8     0.315
  9         .
 10         .

Sum of Residuals                           0
Sum of Squared Residuals           816.04871
Predicted Residual SS (PRESS)     2033.57316
```

1.12.1 Regression Diagnostics in SAS: Rocket Propellant Data

A rocket motor is manufactured by bonding an igniter propellant and a sustainer propellant together inside a metal housing. The shear strength of the bond between the two types of propellant is an important quality characteristic. It is suspected that shear strength is related to the age in weeks of the batch of sustainer propellant. Twenty observations on these two characteristics are given below, in a file named `ADA2_01_rocket.dat`. The first column is shear strength in psi, the second is age of propellant in weeks.

```
2158.70 15.50
1678.15 23.75
2316.00  8.00
2061.30 17.00
2207.50  5.50
1708.30 19.00
1784.70 24.00
2575.00  2.50
2357.90  7.50
2256.70 11.00
2165.20 13.00
2399.55  3.75
1779.80 25.00
2336.75  9.75
1765.30 22.00
2053.50 18.00
2414.40  6.00
2200.50 12.50
2654.20  2.00
1753.70 21.50
```

Most of the SAS program below should be clear. The **output** statement following the **model** statement creates an output data set called **rocket2** that contains the original data set (**d1**) input into **reg**, plus requested summaries from the linear regression fit. The SAS keywords **p=** and **student=** followed by the names I supplied indicate that we wish to save the predicted values and

studentized residuals, respectively. These summaries can be used in plots, but must be referenced with the user defined variable names (here: pv1 and str). Other case summaries from the regression can be saved. For example, **l95m=** and **u95m=** (followed by variable names) save lower and upper 95% CI for mean responses, **l95=** and **u95=** save lower and upper 95% prediction intervals for individual responses, **r=** saves residuals, **cookd=** saves Cook's D, and **h=** saves case leverages (which I have not defined).

The **rank** procedure step following the regression analysis computes the ranks for the variable **str** (the studentized residuals) listed on the **var** statement. The computed ranks are then stored in **rankstr**, specified in the **ranks** statement. The **normal=blom** option means that the ranks are the expected normal order statistics. At this point an internal SAS data set **rocket3** is created that contains the input data set **rocket2** plus the expected normal order statistics. From this information we can get a rankit plot or normal probability plot using the **plot** procedure.

The final **plot** procedure produces two standard diagnostic plots: a plot of r_i against $\hat{Y}_i$ (STR*PV1) and the rankit plot of the r_i (STR*RANKSTR).

This seems like a lot of work to get a rankit plot! A simple alternative is to bypass the **rank** procedure, and use the **univariate** procedure with **plot** and **normal** options to get a (lousy) rankit plot of the data in **str**. If you are using SAS for Windows there is a better way to get diagnostic plots which is shown by the **plot** statement in **proc reg**.

I **directly** generated high resolution diagnostic plots for a regression analysis using a **plot** statement in **reg**. I generated a plot of the studentized residuals versus the predicted values, a rankit plot of the studentized residuals, and an index plot of Cook's distance. Note the periods "." after the keywords **student** (for studentized residuals), **predicted** (for predicted values), **nqq** (for normal quantiles), and **obs** for observation number. As an aside, if you specify the **lineprinter** option at the end of the **reg** procedure statement, then the **plot** statement generates line printer quality plots in the output window, rather than high resolution plots in the graph window. This option does not work if you request a rankit plot. You might try the **lineprinter** option if you use the batch version on linux. Note the simplicity of generating the diagnostic plots relative to what we did earlier! In certain future analyses, I will generate the diagnostic plots the old fashioned way, which works regardless of which SAS version we are using.

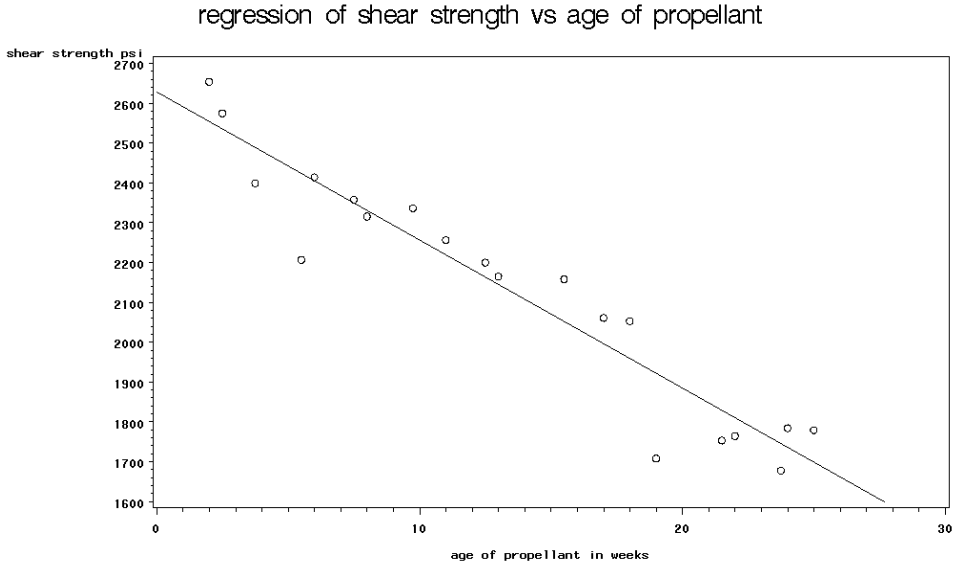
```
* rocket - infile, gplot, output residuals, plot residuals and QQ plot;

options ls=79 nodate nocenter;
title 'example: rocket';
data rocket;
  infile 'ADA2_01_rocket.dat';
  input v1 v2;
  label v1='shear strength psi' v2='age of propellant in weeks';
run;

title 'regression of shear strength vs age of propellant';
symbol1 v=circle c=black i=r;  * i=r makes a regression line;
proc gplot data=rocket;
  plot v1*v2;
run;

proc reg data=rocket;
  model v1=v2/p r;
    * diagnostic plots can be created directly in proc reg;
    plot student.*predicted. nqq.*student. cookd.*obs.;
  output out=rocket2 p=pv1 student=str;
run;

proc rank data=rocket2 out=rocket3 normal=blom;
```

```
   var str;
   ranks rankstr;
run;

symbol1 v=circle c=black i=none;
title 'std resid plot';
proc gplot data=rocket3;
   plot str*pv1;
run;
title 'QQ plot';
proc gplot data=rocket3;
   plot str*rankstr;
run;
quit;
```

Comments on the SAS Output The relationship between shear strength and age is fairly linear with predicted shear strength decreasing as the age of the propellant increases. The fitted LS line is

$$\text{Predicted shear strength} = 2627.8 - 37.2\text{Age}.$$

The test for $H_0 : \beta_1 = 0$ (zero slope for the population regression line) is highly significant: $p - value < 0.0001$. Also note that $R^2 = 0.9018$ so the linear relationship between shear strength and age explains about 90% of the variation in shear strength.



Output window contents:

```
example: rocket

The REG Procedure
Model: MODEL1
Dependent Variable: v1 shear strength psi

Number of Observations Read          20
Number of Observations Used          20

                          Analysis of Variance

                                   Sum of           Mean
Source                     DF      Squares         Square    F Value    Pr > F
```

Model	1	1527483	1527483	165.38 <.0001
Error	18	166255	9236.38100	
Corrected Total	19	1693738		

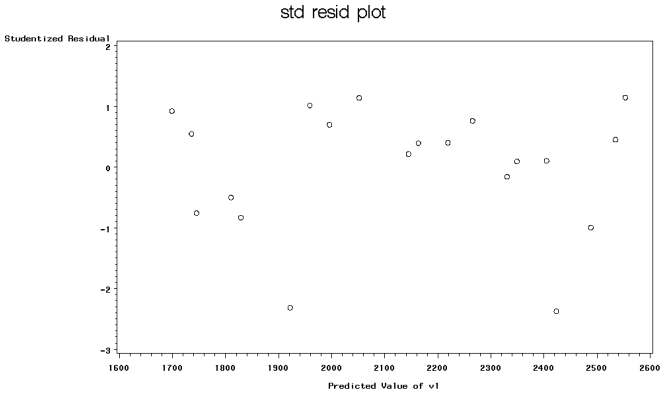
Root MSE	96.10609	R-Square	0.9018
Dependent Mean	2131.35750	Adj R-Sq	0.8964
Coeff Var	4.50915		

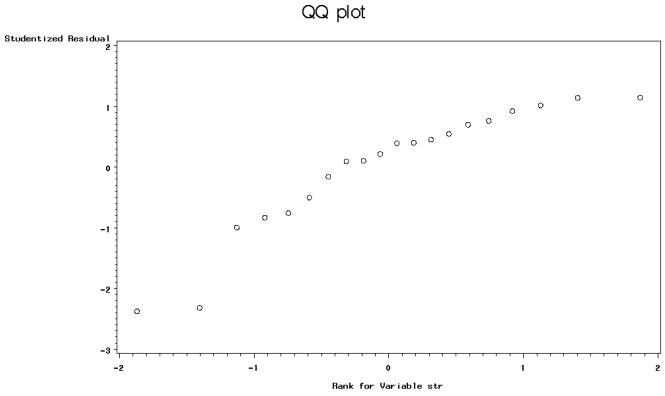
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value
Intercept	Intercept	1	2627.82236	44.18391	59.47
v2	age of propellant in weeks	1	-37.15359	2.88911	-12.86

Parameter Estimates

Variable	Label	DF	Pr > |t|
Intercept	Intercept	1	<.0001
v2	age of propellant in weeks	1	<.0001





Model Inadequacies The data plot and residual information identify observations 5 and 6 as potential outliers ($r_5 = -2.38, r_6 = -2.32$). The predicted values for these observations are much greater than the observed shear strengths. These same observations appear as potential outliers in the normal scores plot and the plot of r_i against $\hat{Y}_i$. Observations 5 and 6 also have the largest influence on the analysis; see the Cook's distance values.

A sensible next step would be to repeat the analysis holding out the most influential case, observation 5. It should be somewhat clear that the influence of case 6 would increase dramatically once case 5 is omitted from the analysis. Since both cases have essentially the same effect on the positioning of the LS line, I will assess the impact of omitting both simultaneously.

Before we hold out these cases, how do you think the LS line will change? My guess is these cases are pulling the LS line down, so the intercept of the LS line should increase once these cases are omitted. Holding out either case 5 or 6 would probably also affect the slope, but my guess is that when they are both omitted the slope will change little. (Is this my experience speaking, or have I already seen the output? Both.) What will happen to R^2 when we delete these points?

Output Statistics

Dependent	Predicted	Std Error		Std Error	Student

Obs	Variable	Value	Mean Predict	Residual	Residual	Residual
1	2159	2052	22.3597	106.7583	93.469	1.142
2	1678	1745	36.9114	-67.2746	88.735	-0.758
3	2316	2331	26.4924	-14.5936	92.383	-0.158
4	2061	1996	23.9220	65.0887	93.081	0.699
5	2208	2423	31.2701	-215.9776	90.877	-2.377
6	1708	1922	26.9647	-213.6041	92.246	-2.316
7	1785	1736	37.5010	48.5638	88.488	0.549
8	2575	2535	38.0356	40.0616	88.259	0.454
9	2358	2349	27.3623	8.7296	92.129	0.0948
10	2257	2219	22.5479	37.5671	93.424	0.402
11	2165	2145	21.5155	20.3743	93.667	0.218
12	2400	2488	35.1152	-88.9464	89.461	-0.994
13	1780	1699	39.9031	80.8174	87.431	0.924
14	2337	2266	23.8903	71.1752	93.089	0.765
15	1765	1810	32.9326	-45.1434	90.287	-0.500
16	2054	1959	25.3245	94.4423	92.709	1.019
17	2414	2405	30.2370	9.4992	91.226	0.104
18	2201	2163	21.6340	37.0975	93.639	0.396
19	2654	2554	39.2360	100.6848	87.732	1.148
20	1754	1829	31.8519	-75.3202	90.674	-0.831

Output Statistics

Obs	-2-1 0 1 2	Cook's D
1	|** |	0.037
2	*| |	0.050
3	| |	0.001
4	|* |	0.016
5	****| |	0.334
6	****| |	0.229
7	|* |	0.027
8	| |	0.019
9	| |	0.000
10	| |	0.005
11	| |	0.001
12	*| |	0.076
13	|* |	0.089
14	|* |	0.019
15	| |	0.017
16	|** |	0.039
17	| |	0.001
18	| |	0.004
19	|** |	0.132
20	*| |	0.043

Sum of Residuals	0
Sum of Squared Residuals	166255
Predicted Residual SS (PRESS)	205944

Further Analysis and Final Discussion: Some summaries for the complete analysis, and when cases 5 and 6 are held out, are given on the next page. To hold these cases out, place the following statement after the **input** statement in the original **data step**.

```
* rocket - delete two observations;
  if ( _n_ = 5 ) or ( _n_ = 6 ) then delete;
```

The summaries lead to the following conclusions:

1. Holding out cases 5 and 6 has little effect on on the estimated LS line. Predictions of shear

strength are slightly larger after holding out these two cases (recall that intercept increased, but slope was roughly the same!)

2. Holding out these two cases decreases $\hat{\sigma}$ considerably, and leads to a modest increase in R^2. The complete data set will give wider CI and prediction intervals than the analysis which deletes case 5 and 6 because $\hat{\sigma}$ decreases when these points are omitted.

3. Once these cases are held out, the normal scores plot and plot of the studentized residuals against fitted values shows no significant problems. (output not given). One observation has a large Cook's D but does not appear to be extremely influential.

Without any substantive reason to explain the low shear strengths for cases 5 and 6, I am hesitant to delete either case from the analysis. I feel relatively confident that including these cases will not seriously limit the use of the model.

Feature	Full data	Omit 5 and 6
b_0	2627.82	2658.97
b_1	−37.15	−37.69
R^2	0.9018	0.9578
$\hat{\sigma}$	96.10	62.96
p-val for $H_0 : \beta_1 = 0$	0.0001	0.0001

Here is a comparison of the predicted or fitted values for selected observations in the data set, based on the two fits.

Observation	Actual Shear Strength	Pred, full	Pred, omit 5,6
1	2159	2052	2075
2	1678	1745	1764
4	2061	1996	2018
8	2575	2535	2565
10	2257	2219	2244
15	1765	1810	1830
18	2201	2163	2188
20	1754	1829	1849

1.13 Features of SAS for Windows

I have focused on the structure of SAS code, at the expense of paying little attention to working in the Windows environment. We will hopefully become more proficient with all features of SAS as we progress through the semester. I have emphasized how certain statistical methods that we discussed last semester are implemented in SAS. The website has links to sites with other examples of standard statistical tests in SAS.

The SAS techniques that have been presented can be implemented in **any** of the SAS versions: linux batch, linux interactive, and Windows. There are glaring weaknesses of the non-Windows versions: there are few options for interacting with the data, the graphics are line-printer quality, and data analysis is not interactive. Here are some useful tools for doing data analysis in SAS for Windows that overcome some of these limitations.

1.13.1 Using the Import and Export Wizards in Windows

The **import wizard** can be used to access datasets created from a variety of spreadsheet packages, for example EXCEL, LOTUS, dBASE, ACCESS, or text data. To use the wizard, follow the sequence SAS MENU BAR > FILE > IMPORT DATA. The **import wizard** dialog box will pop up after you click on IMPORT DATA. From this point, using the wizard is straightforward. In order, you need to: (1) select the data source (i.e., EXCEL 97, etc), (2) identify where the file is located (i.e., browse for it), (3) choose the SAS destination: LIBRARY = WORK, MEMBER = d2, (4) finish. If the data is loaded successfully, you will see a note to this effect in the **log window**, or error messages describing problems SAS had creating the data set. I specified MEMBER = d2 in step (3), which implies that the data set is identified in SAS as **d2**. Any valid data set name is permissible.

It is important to note that the **import wizard** does not allow you to view or manipulate the data in a spreadsheet environment within **SAS**. The wizard provides an easy way to import data from a worksheet into SAS. Once entered in SAS, you still might need to specify another **data step** to place the imported data (using the **set** command) and then to manipulate it.

After inputting data with the wizard, I recommend that you list the contents of the data set using the **contents** procedure, as outlined below. This will tell you the types and names of the variables you read. This is important because EXCEL headers, for example, may be truncated by SAS.

To illustrate, I read an EXCEL 95 spreadsheet into SAS using the import wizard. The SAS program and its output below tells me that I read in 41 observations on 16 variables, 15 of which are numerical. At this time, I know what variable names and types of data I have in the data set **d2**.

```
options ls=79 nodate;
proc contents data=d2 position;
run;
```

Output window contents:

```
                          The CONTENTS Procedure

Data Set Name: WORK.D2                            Observations:          41
Member Type:   DATA                               Variables:             16
Engine:        V8                                 Indexes:               0
Created:       12:17 Thursday, February 22, 2001  Observation Length:    152
Last Modified: 12:17 Thursday, February 22, 2001  Deleted Observations:  0
```

```
Protection:                                          Compressed:           NO
Data Set Type:                                       Sorted:               NO
Label:

                  -----Engine/Host Dependent Information-----

   Data Set Page Size:         12288
   Number of Data Set Pages:   1
   First Data Page:            1
   Max Obs per Page:           80
   Obs in First Data Page:     41
   Number of Data Set Repairs: 0
   File Name:                  C:\TEMPSAS Temporary Files\_TD47\d2.sas7bdat
   Release Created:            8.0000M0
   Host Created:               WIN_NT

            -----Alphabetic List of Variables and Attributes-----

 #   Variable   Type   Len   Pos   Format     Informat   Label
-----------------------------------------------------------------
 3   AGE        Num      8     8   BEST8.     BEST8.     AGE
 5   AGEOFONS   Num      8    24   BEST11.    BEST11.    AGEOFONSET
 9   CHOLDCRE   Num      8    56   BEST8.     BEST8.     CHOLDCRE
 2   DATE       Num      8     0   MMDDYY8.   MMDDYY8.   DATE
 7   DEAD       Num      8    40   BEST11.    BEST11.    DEAD
14   DEGREEOF   Num      8    96   BEST17.    BEST17.    DEGREE OF DISABILITY
13   DISABILI   Num      8    88   BEST8.     BEST8.     DISABILITY
 4   DISEASED   Num      8    16   BEST15.    BEST15.    DISEASE DURATION
10   GLOMERUL   Num      8    64   BEST14.    BEST14.    GLOMERULONEPHRITIS
 8   NAADCRE    Num      8    48   BEST8.     BEST8.     NAADCRE
 1   NAME       Char    30   120   $30.       $30.       NAME
11   SLEDAI     Num      8    72   BEST11.    BEST11.    SLEDAI
12   SLICC      Num      8    80   BEST8.     BEST8.     SLICC
16   SSIDISAB   Num      8   112   BEST13.    BEST13.    SSI DISABILITY
15   WHOHANDI   Num      8   104   BEST16.    BEST16.    WHO HANDICAP
 6   YEARSOBS   Num      8    32   BEST13.    BEST13.    YEARS OBSERVED

                 -----Variables Ordered by Position-----

 #   Variable   Type   Len   Pos   Format     Informat   Label
-----------------------------------------------------------------
 1   NAME       Char    30   120   $30.       $30.       NAME
 2   DATE       Num      8     0   MMDDYY8.   MMDDYY8.   DATE
 3   AGE        Num      8     8   BEST8.     BEST8.     AGE
 4   DISEASED   Num      8    16   BEST15.    BEST15.    DISEASE DURATION
 5   AGEOFONS   Num      8    24   BEST11.    BEST11.    AGEOFONSET
 6   YEARSOBS   Num      8    32   BEST13.    BEST13.    YEARS OBSERVED
 7   DEAD       Num      8    40   BEST11.    BEST11.    DEAD
 8   NAADCRE    Num      8    48   BEST8.     BEST8.     NAADCRE
 9   CHOLDCRE   Num      8    56   BEST8.     BEST8.     CHOLDCRE
10   GLOMERUL   Num      8    64   BEST14.    BEST14.    GLOMERULONEPHRITIS
11   SLEDAI     Num      8    72   BEST11.    BEST11.    SLEDAI
12   SLICC      Num      8    80   BEST8.     BEST8.     SLICC
13   DISABILI   Num      8    88   BEST8.     BEST8.     DISABILITY
14   DEGREEOF   Num      8    96   BEST17.    BEST17.    DEGREE OF DISABILITY
15   WHOHANDI   Num      8   104   BEST16.    BEST16.    WHO HANDICAP
16   SSIDISAB   Num      8   112   BEST13.    BEST13.    SSI DISABILITY
```

The **output wizard** is used to save data sets to disk (in a variety of formats such as EXCEL, dBASE, LOTUS, text) that you have created in SAS. The process is started by following the steps SAS MENU BAR > FILE > EXPORT DATA. This is followed by a sequence of steps similar to that used in the **import wizard**.

1.13.2 SAS/Graph

SAS/Graph is a collection of procedures that provide **high resolution** graphs. For example, the **gplot** procedure generates high resolution 2D scatter plots. In SAS for Windows, these graphs pop up in the **graph window**, as described in the next section. **SAS/Graph** routines are also available under linux, but you must specify the format for the file(s) that will contain the graph(s). There is a link on the course WWW page (**Device Specifications for High Resolution Plots on Unix**) that provides some guidance on this issue. Further information is available through the SAS help facility: SAS MENU BAR > HELP > HELP ON SAS SOFTWARE PRODUCTS > SAS/GRAPH.

1.13.3 High Resolution Graphs in SAS for Windows

Several SAS for Windows procedures generate **high resolution** graphs that pop up in the **graph window**. In the simplest case, the graph window holds a single plot. This plot can be copied to the clipboard on the SAS Menu Bar, then pasted into, for example, **Microsoft Word**. Alternatively, the graph window (or catalog) can contain multiple plots, but unlike **Minitab**, only a single graph is shown at a time. and you can scroll through the graph catalog by using the slide bar on the side of the graph window, or by clicking on VIEW on the SAS Menu Bar and choosing the graph to look at. We saw examples earlier in this chapter of SAS programs where the procedures generate high resolution graphs.

1.13.4 SAS/Assist

SAS/Assist is a menu driven interface to many SAS tools, such as graphics and data analysis. For more information follow the path: SAS MENU BAR > HELP > HELP ON SAS SOFTWARE PRODUCTS > SAS/ASSIST. To use Assist, follow the steps: SAS MENU BAR > SOLUTIONS > ASSIST.

1.13.5 SAS/Analyst and SAS/Insight

These are two useful interactive data analysis environments that allow you to enter data into a spreadsheet, plot the data, and do a variety of simple analyses. For more information follow the path: SAS MENU BAR > HELP > USING SAS SOLUTIONS AND TOOLS (then go to **Statistical Analysis** and **Interactive Data Analysis**). To use Analyst or Insight, follow the steps: SAS MENU BAR > SOLUTIONS > ANALYSIS > ANALYST (for **Analyst**) or INTERACTIVE DATA ANALYSIS (for **Insight**).

Chapter 2

Introduction to Multiple Linear Regression

In **multiple linear regression**, a linear combination of two or more predictor variables (xs) is used to explain the variation in a response. In essence, the additional predictors are used to explain the variation in the response not explained by a simple linear regression fit.

2.1 Indian systolic blood pressure example

(The following problem was taken from the Minitab handbook.) Anthropologists conducted a study to determine the long-term effects of an environmental change on systolic blood pressure. They measured the blood pressure and several other characteristics of 39 Indians who migrated from a very primitive environment high in the Andes into the mainstream of Peruvian society at a lower altitude. All of the Indians were males at least 21 years of age, and were born at a high altitude.

The data are given below and in a file on the course website. The file contains 39 rows, one for each individual, and twelve columns corresponding to the variables, from left to right:

```
DATA:
    n = individual id
   x1 = age in years                x2 = years since migration
   x3 = weight in kilos             x4 = height in mm
   x5 = chin skin fold in mm        x6 = forearm skin fold in mm
   x7 = calf skin fold in mm        x8 = pulse rate-beats/min
   x9 = systolic bp                 x10 = diastolic bp
  x11 = years since migration divided by age

 1   21    1   71.0   1629    8.0    7.0   12.7   88   170    76   0.04762
 2   22    6   56.5   1569    3.3    5.0    8.0   64   120    60   0.27273
 3   24    5   56.0   1561    3.3    1.3    4.3   68   125    75   0.20833
 4   24    1   61.0   1619    3.7    3.0    4.3   52   148   120   0.04167
 5   25    1   65.0   1566    9.0   12.7   20.7   72   140    78   0.04000
 6   27   19   62.0   1639    3.0    3.3    5.7   72   106    72   0.70370
 7   28    5   53.0   1494    7.3    4.7    8.0   64   120    76   0.17857
 8   28   25   53.0   1568    3.7    4.3    0.0   80   108    62   0.89286
 9   31    6   65.0   1540   10.3    9.0   10.0   76   124    70   0.19355
10   32   13   57.0   1530    5.7    4.0    6.0   60   134    64   0.40625
11   33   13   66.5   1622    6.0    5.7    8.3   68   116    76   0.39394
12   33   10   59.1   1486    6.7    5.3   10.3   73   114    74   0.30303
13   34   15   64.0   1578    3.3    5.3    7.0   88   130    80   0.44118
14   35   18   69.5   1645    9.3    5.0    7.0   60   118    68   0.51429
15   35    2   64.0   1648    3.0    3.7    6.7   60   138    78   0.05714
16   36   12   56.5   1521    3.3    5.0   11.7   72   134    86   0.33333
17   36   15   57.0   1547    3.0    3.0    6.0   84   120    70   0.41667
18   37   16   55.0   1505    4.3    5.0    7.0   64   120    76   0.43243
```

19	37	17	57.0	1473	6.0	5.3	11.7	72	114	80	0.45946
20	38	10	58.0	1538	8.7	6.0	13.0	64	124	64	0.26316
21	38	18	59.5	1513	5.3	4.0	7.7	80	114	66	0.47368
22	38	11	61.0	1653	4.0	3.3	4.0	76	136	78	0.28947
23	38	11	57.0	1566	3.0	3.0	3.0	60	126	72	0.28947
24	39	21	57.5	1580	4.0	3.0	5.0	64	124	62	0.53846
25	39	24	74.0	1647	7.3	6.3	15.7	64	128	84	0.61538
26	39	14	72.0	1620	6.3	7.7	13.3	68	134	92	0.35897
27	41	25	62.5	1637	6.0	5.3	8.0	76	112	80	0.60976
28	41	32	68.0	1528	10.0	5.0	11.3	60	128	82	0.78049
29	41	5	63.4	1647	5.3	4.3	13.7	76	134	92	0.12195
30	42	12	68.0	1605	11.0	7.0	10.7	88	128	90	0.28571
31	43	25	69.0	1625	5.0	3.0	6.0	72	140	72	0.58140
32	43	26	73.0	1615	12.0	4.0	5.7	68	138	74	0.60465
33	43	10	64.0	1640	5.7	3.0	7.0	60	118	66	0.23256
34	44	19	65.0	1610	8.0	6.7	7.7	74	110	70	0.43182
35	44	18	71.0	1572	3.0	4.7	4.3	72	142	84	0.40909
36	45	10	60.2	1534	3.0	3.0	3.3	56	134	70	0.22222
37	47	1	55.0	1536	3.0	3.0	4.0	64	116	54	0.02128
38	50	43	70.0	1630	4.0	6.0	11.7	72	132	90	0.86000
39	54	40	87.0	1542	11.3	11.7	11.3	92	152	88	0.74074

A question we consider concerns the long term effects of an environmental change on the systolic blood pressure. In particular, is there a relationship between the systolic blood pressure and how long the Indians lived in their new environment as measured by the fraction of their life spent in the new environment: fraction = years since migration/age? Note that `sysbp=x9`=systolic bp and `yrage=x11`=fraction.

A plot of systolic blood pressure against `yrage` fraction given below suggests a weak linear relationship. Nonetheless, consider fitting the regression model

$$\text{sysbp} = \beta_0 + \beta_1 \text{yrage} + \varepsilon.$$

The least squares line is given by

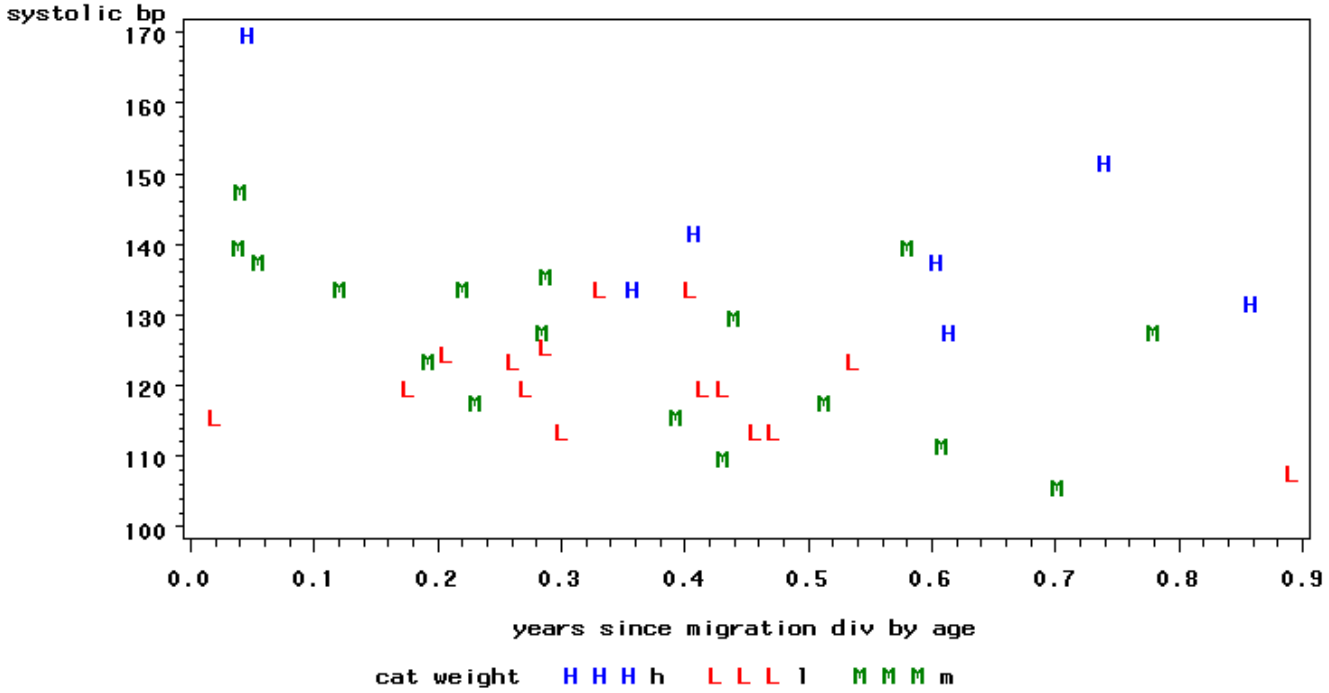
$$\widehat{\text{sysbp}} = 133.5 - 15.8\text{yrage},$$

and suggests that average systolic blood pressure decreases as the fraction of life spent in modern society increases. However, the t-test of $H_0 : \beta_1 = 0$ is not significant at the 5% level (p-value=0.089). That is, the weak linear relationship observed in the data is not atypical of a population where there is no linear relationship between systolic blood pressure and the fraction of life spent in a modern society.

Even if this test were significant, the small value of $R^2 = 0.076$ suggests that `yrage` fraction does not explain a *substantial* amount of the variation in the systolic blood pressures. If we omit the individual with the highest blood pressure (see the plot) then the relationship would be weaker. Program editor contents:

```
* Indian systolic blood pressure example;
* plot and simple linear regression;

options ls=79 nodate nocenter;

data indian;
  infile 'F:\Dropbox\UNM\teach\ADA2_stat528\sas\ADA2_02_indian.dat';
  input n x1-x11;
  if x3 lt 60 then x12='l';         * x12 (low, med, high weight) ;
    else if x3 ge 70 then x12='h'; *    for plotting by weight   ;
    else x12 ='m';
  label x1='age in years'          x2='years since migration'
    x3='weight in kilos'           x4='height in mm'
```

```
    x5='chin skin fold in mm'      x6='forearm skin fold in mm'
    x7='calf skin fold in mm'      x8='pulse rate-beats/min'
    x9='systolic bp'              x10='diastolic bp'
    x11='years since migration div by age'
    x12='cat weight';

  * give descriptive variable names to the variables we'll use;
  wt    = x3;
  sysbp = x9;
  yrage = x11;
  wtcat = x12;
  label wt='weight in kilos'
     sysbp='systolic bp'
     yrage='years since migration div by age'
     wtcat='cat weight';
run;

* define v=symbol, c=color, and i=interpolation;
symbol1 v="h" c=blue  i=none;
symbol2 v="l" c=red   i=none;
symbol3 v="m" c=green i=none;
proc gplot data=indian;
  plot sysbp*yrage=wtcat;
run;

proc reg data=indian;
  model sysbp = yrage;
run;
```



Output window contents:

```
The REG Procedure
Model: MODEL1
Dependent Variable: sysbp systolic bp
```

```
Number of Observations Read          39
Number of Observations Used          39

                              Analysis of Variance
                                   Sum of           Mean
Source                     DF       Squares         Square    F Value    Pr > F
Model                       1     498.06734      498.06734       3.05    0.0888
Error                      37    6033.36856      163.06402
Corrected Total            38    6531.43590

Root MSE              12.76965    R-Square     0.0763  <----
Dependent Mean       127.41026    Adj R-Sq     0.0513
Coeff Var             10.02247

                               Parameter Estimates
                                                        Parameter       Standard
Variable     Label                                DF      Estimate          Error
Intercept    Intercept                             1     133.49573        4.03801
yrage        years since migration div by age      1     -15.75187        9.01295

                            Parameter Estimates
Variable     Label                                DF    t Value    Pr > |t|
Intercept    Intercept                             1      33.06      <.0001
yrage        years since migration div by age      1      -1.75      0.0888 <----
```

2.1.1 Taking Weight Into Consideration

At best, there is a weak relationship between systolic blood pressure and the `yrage` fraction. However, it is usually accepted that systolic blood pressure and weight are related. A natural way to take weight into consideration is to include weight and `yrage` fraction as predictors of systolic blood pressure in the multiple regression model:

$$\text{sysbp} = \beta_0 + \beta_1 \text{yrage} + \beta_2 \text{wt} + \varepsilon.$$

As in simple linear regression, the model is written in the form:

$$\text{Response} = \text{Mean of Response} + \text{Residual},$$

so the model implies that that average systolic blood pressure is a linear combination of `yrage` fraction and weight. As in simple linear regression, the standard multiple regression analysis assumes that the responses are normally distributed with a constant variance σ^2. The parameters of the regression model β_0, β_1, β_2, and σ^2 are estimated by least squares (LS).

Here is the **reg** procedure and output for fitting the multiple regression model with `yrage` and `wt` (weight) as predictors. Just add `wt` to the right hand side of the previous **model** statement! Program editor contents:

```
* multiple regression;

proc reg data=indian;
  model sysbp = yrage wt;  * put both predictors to the right of "=" sign;
run;
```

Output window contents:

```
The REG Procedure
Model: MODEL1
Dependent Variable: sysbp systolic bp
```

```
Number of Observations Read          39
Number of Observations Used          39

                              Analysis of Variance
                                   Sum of           Mean
Source                     DF       Squares         Square    F Value    Pr > F
Model                       2    3090.07564    1545.03782      16.16    <.0001
Error                      36    3441.36025      95.59334
Corrected Total            38    6531.43590

Root MSE              9.77718    R-Square     0.4731 <----
Dependent Mean      127.41026    Adj R-Sq     0.4438
Coeff Var             7.67378

                                Parameter Estimates
                                                          Parameter       Standard
Variable     Label                                   DF     Estimate          Error
Intercept    Intercept                                1     60.89595       14.28088
yrage        years since migration div by age         1    -26.76721        7.21779
wt           weight in kilos                          1      1.21686        0.23369

                            Parameter Estimates
Variable     Label                                   DF    t Value    Pr > |t|
Intercept    Intercept                                1       4.26      0.0001
yrage        years since migration div by age         1      -3.71      0.0007 <----
wt           weight in kilos                          1       5.21      <.0001 <----
```

2.1.2 Important Points to Notice About the Regression Output

1. The LS estimates of the intercept and the regression coefficient for `yrage` fraction, and their standard errors, change from the simple linear model to the multiple regression model. For the simple linear regression:

   $$\widehat{\text{sysbp}} = 133.50 - 15.75\text{yrage}.$$

   For the multiple regression model:

   $$\widehat{\text{sysbp}} = 60.89 - 26.76\text{yrage} + 1.21\text{wt}.$$

2. Looking at the ANOVA tables for the simple linear and the multiple regression models we see that the Regression (model) *df* has increased from 1 to 2 (2=number of predictor variables) and the Residual (error) *df* has decreased from 37 to 36 ($=n-1-$ number of predictors). Adding a predictor increases the Regression *df* by 1 and decreases the Residual *df* by 1.

3. The Residual SS decreases by $6033.37 - 3441.36 = 2592.01$ upon adding the weight term. The Regression SS increased by 2592.01 upon adding the weight term term to the model. The Total SS does not depend on the number of predictors so it stays the same. The Residual SS, or the part of the variation in the response unexplained by the regression model, never increases when new predictors are added. (You can't add a predictor and explain less variation.)

4. The proportion of variation in the response explained by the regression model:

   $$R^2 = \text{Regression SS/Total SS}$$

   never decreases when new predictors are added to a model. The R^2 for the simple linear regression was 0.076, whereas $R^2 = 3090.08/6531.44 = 0.473$ for the multiple regression

model. Adding the weight variable to the model increases R^2 by 40%. That is, weight explains 40% of the variation in systolic blood pressure not already explained by fraction.

5. The estimated variability about the regression line

$$\text{Residual MS} = \hat{\sigma}^2$$

decreased dramatically after adding the weight effect. For the simple linear regression model $\hat{\sigma}^2 = 163.06$, whereas $\hat{\sigma}^2 = 95.59$ for the multiple regression model. This suggests that an important predictor has been added to model.

6. The F-statistic for the multiple regression model

$$F_{obs} = \text{Regression MS/Residual MS} = 1545.04/95.59 = 16.163$$

(which is compared to a F-table with 2 and 36 *df*) tests $H_0 : \beta_1 = \beta_2 = 0$ against $H_A :$ not H_0. This is a test of no relationship between the average systolic blood pressure and fraction and weight, assuming the relationship is linear. If this test is significant, then either fraction or weight, or both, are important for explaining the variation in systolic blood pressure.

7. Given the model

$$\text{sysbp} = \beta_0 + \beta_1 \text{yrage} + \beta_2 \text{wt} + \varepsilon,$$

we are interested in testing $H_0 : \beta_2 = 0$ against $H_A : \beta_2 \neq 0$. The t-statistic for this test

$$t_{obs} = \frac{b_2 - 0}{\text{SE}(b_2)} = \frac{1.217}{0.234} = 5.207$$

is compared to a t-critical value with Residual $df = 36$. **SAS** gives a p-value of < 0.0001, which suggests $\beta_2 \neq 0$. The t-test of $H_0 : \beta_2 = 0$ in the multiple regression model tests whether adding weight to the simple linear regression model explains a significant part of the variation in systolic blood pressure not explained by `yrage` fraction. In some sense, the t-test of $H_0 : \beta_1 = 0$ will be significant if the increase in R^2 (or decrease in Residual SS) obtained by adding weight to this simple linear regression model is substantial. We saw a big increase in R^2, which is deemed significant by the t-test. A similar interpretation is given to the t-test for $H_0 : \beta_1 = 0$.

8. The t-tests for $\beta_0 = 0$ and $\beta_1 = 0$ are conducted, assessed, and interpreted in the same manner. The p-value for testing $H_0 : \beta_0 = 0$ is 0.0001, whereas the p-value for testing $H_0 : \beta_1 = 0$ is 0.0007. This implies that fraction is important in explaining the variation in systolic blood pressure **after** weight is taken into consideration (by including weight in the model as a predictor).

9. We compute CIs for the regression parameters in the usual way: $b_i + t_{\text{crit}} SE(b_i)$, where t_{crit} is the t-critical value for the corresponding CI level with $df =$ Residual df.

2.1.3 Understanding the Model

The t-test for $H_0 : \beta_1 = 0$ is highly significant (p-value=0.0007), which implies that fraction is important in explaining the variation in systolic blood pressure **after weight is taken into consideration** (by including weight in the model as a predictor). Weight is called a **suppressor variable**. Ignoring weight suppresses the relationship between systolic blood pressure and `yrage` fraction.

The implications of this analysis are enormous! Essentially, the correlation between a predictor and a response says very little about the importance of the predictor in a regression model with one or more additional predictors. This conclusion also holds in situations where the correlation is high, in the sense that a predictor that is highly correlated with the response may be unimportant in a multiple regression model once other predictors are included in the model. In multiple regression *"everything depends on everything else."*

I will try to convince you that this was expected, given the plot of systolic blood pressure against fraction. This plot used a weight category variable `wtcat` L, M, or H as a plotting symbol. The relationship between systolic blood pressure and fraction is fairly linear within each weight category, and stronger than when we ignore weight. The slopes in the three groups are negative and roughly constant.

To see why `yrage` fraction is an important predictor after taking weight into consideration, let us return to the multiple regression model. The model implies that the average systolic blood pressure is a linear combination of `yrage` fraction and weight:

$$\widehat{\text{sysbp}} = \beta_0 + \beta_1 \text{yrage} + \beta_2 \text{wt}.$$

For each fixed weight, the average systolic blood pressure is linearly related to `yrage` fraction with a constant slope β_1, independent of weight. A similar interpretation holds if we switch the roles of `yrage` fraction and weight. That is, if we fix the value of fraction, then the average systolic blood pressure is linearly related to weight with a constant slope β_2, independent of `yrage` fraction.

To see this point, suppose that the LS estimates of the regression parameters are the true values

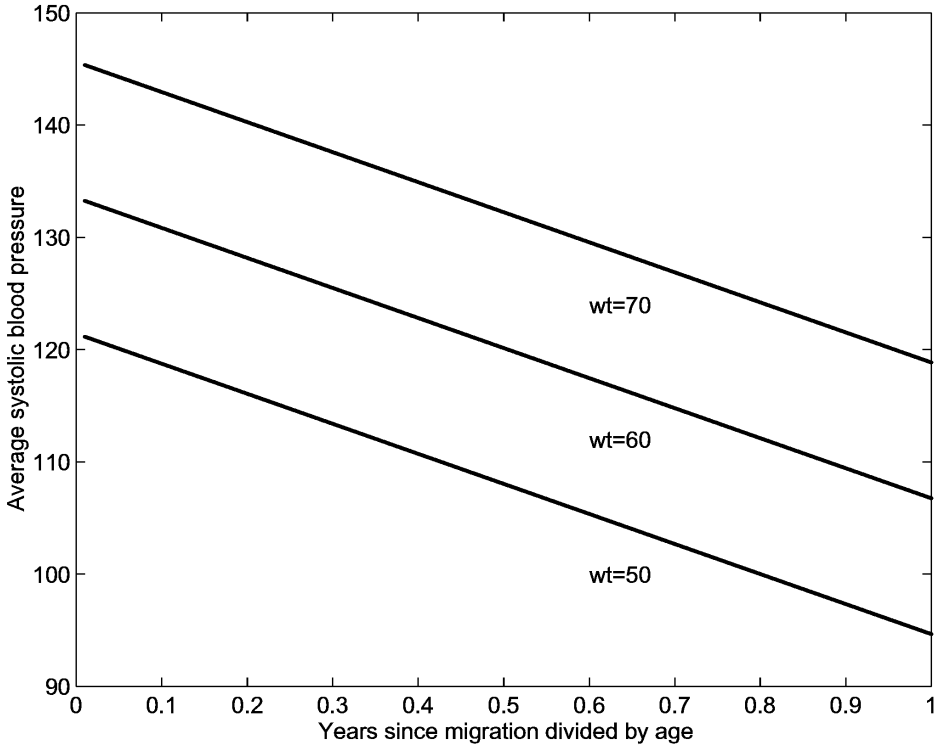
$$\widehat{\text{sysbp}} = 60.89 - 26.76\text{yrage} + 1.21\text{wt}.$$

If we restrict our attention to 50kg Indians, the average systolic blood pressure as a function of fraction is

$$\widehat{\text{sysbp}} = 60.89 - 26.76\text{yrage} + 1.21(50) = 121.39 - 26.76\text{yrage}.$$

For 60kg Indians,

$$\widehat{\text{sysbp}} = 60.89 - 26.76\text{yrage} + 1.21(60) = 133.49 - 26.76\text{yrage}.$$



Hopefully the pattern is clear: the average systolic blood pressure decreases by 26.76 for each increase of 1 on fraction, regardless of one's weight. If we vary weight over its range of values, we get a set of parallel lines (i.e., equal slopes) when we plot average systolic blood pressure as a function of `yrage` fraction. The intercept increases by 1.21 for each increase of 1kg in weight.

Similarly, if we plot the average systolic blood pressure as a function of weight, for several fixed values of fraction, we see a set of parallel lines with slope 1.21, and intercepts decreasing by 26.76 for each increase of 1 in fraction.

If we had more data we could check the model by plotting systolic blood pressure against fraction, broken down by individual weights. The plot should show a fairly linear relationship between systolic blood pressure and fraction, with a constant slope across weights. I grouped the weights into categories because of the limited number of observations. The same phenomenon should approximately hold, and it does. If the slopes for the different weight groups changed drastically with weight, but the relationships were linear, we would need to include an **interaction** or product variable wt $\times$ yrage in the model, in addition to weight and `yrage` fraction. This is probably not warranted here.

A final issue that I wish to address concerns the interpretation of the estimates of the regression coefficients in a multiple regression model. For the fitted model

$$\widehat{\text{sysbp}} = 60.89 - 26.76\text{yrage} + 1.21\text{wt}$$

our interpretation is consistent with the explanation of the regression model given above. For example, focus on the `yrage` fraction coefficient. The negative coefficient indicates that the predicted systolic blood pressure decreases as `yrage` fraction increases **holding weight constant**. In particular, the predicted systolic blood pressure decreases by 26.76 for each unit increase in fraction, holding weight constant at any value. Similarly, the predicted systolic blood pressure increases by 1.21 for each unit increase in weight, holding `yrage` fraction constant at any level.

This example was meant to illustrate multiple regression. A more complete analysis of the data, including diagnostics, will be given later.

2.2 GCE exam score example

The data below are selected from a larger collection of data referring to candidates for the General Certificate of Education (GCE) who were being considered for a special award. Here, Y denotes the candidate's total mark, out of 1000, in the GCE exam, while X_1 is the candidate's score in the compulsory part of the exam, which has a maximum score of 200 of the 1000 points on the exam. X_2 denotes the candidates' score, out of 100, in a School Certificate English Language paper taken on a previous occasion.

```
  Y   X1  X2
476  111  68
457   92  46
540   90  50
551  107  59
575   98  50
698  150  66
545  118  54
574  110  51
645  117  59
690  114  80
634  130  57
637  118  51
390   91  44
562  118  61
560  109  66
```

A goal here is to compute a multiple regression of Y on X_1 and X_2, and make the necessary tests to enable you to comment intelligently on the extent to which current performance in the compulsory test (X_1) may be used to predict aggregate performance on the GCE exam (Y), and on whether previous performance in the School Certificate English Language (X_2) has any predictive value independently of what has already emerged from the current performance in the compulsory papers.

I will lead you through a number of steps to help you answer this question. Let us answer the following straightforward questions based on the **SAS** output.

1. Plot Y against X_1 and X_2 individually, and comment on the form (i.e., linear, non-linear, logarithmic, etc.), strength, and direction of the relationships.

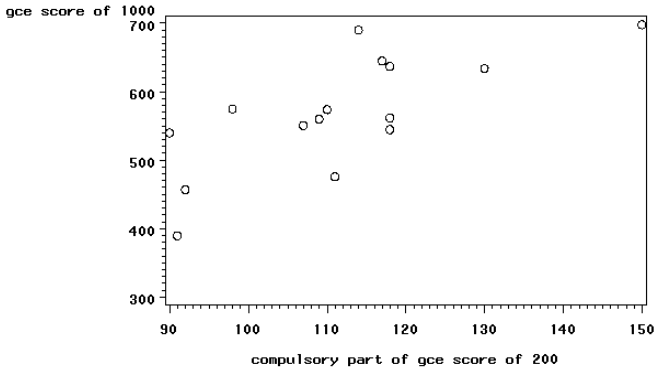
2. Plot X_1 against X_2 and comment on the form, strength, and direction of the relationship.

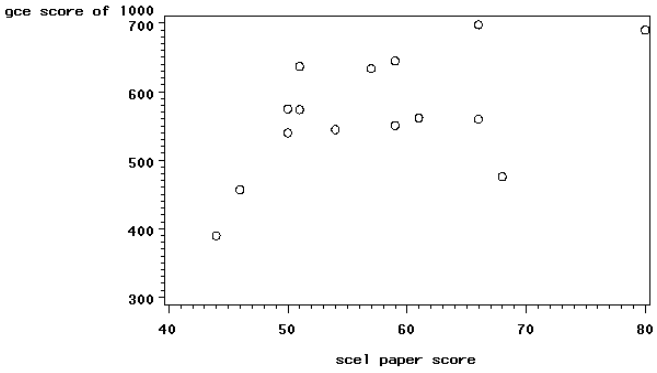
3. Compute the correlation between all pairs of variables. Do the correlations appear sensible, given the plots?

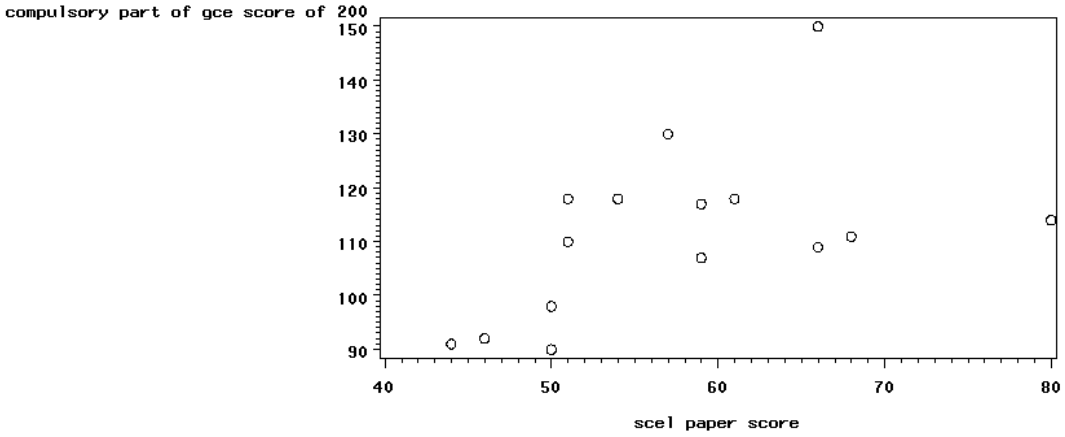
Program editor contents:

```
* GCE example - plots, correlation, individual simple linear regressions;

data gce;
  infile 'F:\Dropbox\UNM\teach\ADA2_stat528\sas\ADA2_02_gce.dat';
  input y x1 x2;
  label y='gce score of 1000'
        x1='compulsory part of gce score of 200'
```

```
       x2='scel paper score';
run;

* define v=symbol, c=color, and i=interpolation;
symbol1 v=circle c=black i=none;
proc gplot data=gce;
  plot y*(x1 x2);
  plot x1*x2;
run;

proc corr data=gce;
  var y x1 x2;
run;
```







Output window contents:
Correlation:

Pearson Correlation Coefficients, N = 15
Prob > |r| under H0: Rho=0

	y	x1	x2
y	1.00000	0.73073	0.54771
gce score of 1000		0.0020	0.0346
x1	0.73073	1.00000	0.50889
compulsory part of gce score of 200	0.0020		0.0527
x2	0.54771	0.50889	1.00000
scel paper score	0.0346	0.0527	

In parts 4 through 9, ignore the possibility that Y, X_1 or X_2 might ideally need to be transformed.

4. Which of X_1 and X_2 explains a larger proportion of the variation in Y? Which would appear to be a better predictor of Y? (Explain).

   Answer: R^2 is 0.53 for the model with X_1 and 0.30 with X_2. Equivilantly, the Model SS is larger for X_1 (53970) than for X_2 (30321). Thus, X_1 appears to be a better predictor of Y than X_2.

5. Consider 2 simple linear regression models for predicting Y, one with X_1 as a predictor, and the other with X_2 as the predictor. Do X_1 and X_2 individually appear to be important for explaining the variation in Y? (i.e., test that the slopes of the regression lines are zero). Which, if any, of the output, support, or contradicts, your answer to the previous question?

   Answer: The model with X_1 has a t-statistic of 3.86 with an associated p-value of 0.0020, while X_2 has a t-statistic of 2.36 with an associated p-value of 0.0346. Both predictors explain a significant amount of variability in Y. This is consistant with part (4).

Program editor contents:

```
* individual simple linear regression (SLR);

proc reg data=gce;
  model y=x1;
  model y=x2;
run;
```

Output window contents:
Model $y = \beta_0 + \beta_1 x_1 + \varepsilon$:

```
The REG Procedure
Model: MODEL1
Dependent Variable: y gce score of 1000

Number of Observations Read          15
Number of Observations Used          15

                            Analysis of Variance
                                   Sum of           Mean
Source                     DF        Squares         Square    F Value    Pr > F
Model                       1          53970          53970      14.90    0.0020
Error                      13          47103     3623.32355
Corrected Total            14         101073

Root MSE              60.19405    R-Square     0.5340
Dependent Mean       568.93333    Adj R-Sq     0.4981
Coeff Var             10.58016

                               Parameter Estimates
                                                        Parameter       Standard
Variable     Label                                DF      Estimate          Error
Intercept    Intercept                             1     128.54786       115.16037
x1           compulsory part of gce score          1       3.94847         1.02307
             of 200

                          Parameter Estimates
Variable     Label                                DF    t Value    Pr > |t|
Intercept    Intercept                             1       1.12      0.2845
x1           compulsory part of gce score          1       3.86      0.0020
             of 200
```

Model $y = \beta_0 + \beta_1 x_2 + \varepsilon$:

```
The REG Procedure
Model: MODEL2
Dependent Variable: y gce score of 1000

Number of Observations Read          15
Number of Observations Used          15

                          Analysis of Variance
                                   Sum of           Mean
Source                  DF        Squares         Square    F Value    Pr > F
Model                    1          30321          30321       5.57    0.0346
Error                   13          70752     5442.48412
Corrected Total         14         101073

Root MSE             73.77319    R-Square     0.3000
Dependent Mean      568.93333    Adj R-Sq     0.2461
Coeff Var            12.96693

                          Parameter Estimates
                                                    Parameter       Standard
Variable     Label                            DF      Estimate          Error
Intercept    Intercept                         1     291.58589      119.03815
x2           scel paper score                  1       4.82623        2.04474

                          Parameter Estimates
Variable     Label                            DF    t Value    Pr > |t|
Intercept    Intercept                         1       2.45      0.0292
x2           scel paper score                  1       2.36      0.0346
```

6. Fit the multiple regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$

Test $H_0 : \beta_1 = \beta_2 = 0$ at the 5% level. Describe in words what this test is doing, and what the results mean here.

Answer: The ANOVA table reports an F-statistic of 8.14 with associated p-value of 0.0058 indicating that the regression model with both X_1 and X_2 explains significantly more variability in Y than a model with the intercept, alone. That is, X_1 and X_2 explain variability in Y together. This does not tell us which of or whether X_1 or X_2 are individually important (recall the results of the Indian systolic blood pressure example).

7. In the multiple regression model, test $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$ individually. Describe in words what these tests are doing, and what the results mean here.

Answer: Each hypothesis is testing, conditional on all other predictors being in the model, whether the addition of the predictor being tested explains significantly more variability in Y than without it.

For $H_0 : \beta_1 = 0$, the t-statistic is 2.79 with an associated p-value of 0.0163. Thus, we reject H_0 in favor of the alternative that the slope is statistically significantly different from 0 conditional on X_2 being in the model. That is, X_1 explains significantly more variability in Y given that X_2 is already in the model.

For $H_0 : \beta_2 = 0$, the t-statistic is 1.09 with an associated p-value of 0.2987. Thus, we fail to reject H_0 concluding that there is insufficient evidence that the slope is different from 0 conditional on X_1 being in the model. That is, X_2 does not explain significantly more variability in Y given that X_1 is already in the model.

8. How does the R^2 from the multiple regression model compare to the R^2 from the individual simple linear regressions? Is what you are seeing here appear reasonable, given the tests on the individual coefficients?

   Answer: The R^2 for the model with only X_1 is 0.5340, only X_2 is 0.3000, and both X_1 and X_2 is 0.5757. There is only a very small increase in R^2 from the model with only X_1 when X_2 is added, which is consistent with X_2 not being important given that X_1 is already in the model.

9. Do your best to answer the question posed above, in the paragraph after the data "A goal ...". Provide an equation (LS) for predicting Y.

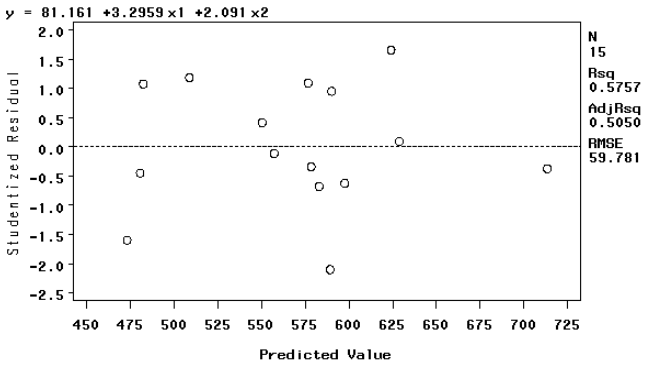
   Answer: Yes, we've seen that X_1 may be used to predict Y, and that X_2 does not explain significantly more variability in the model with X_1. Thus, the preferred model has only X_1:

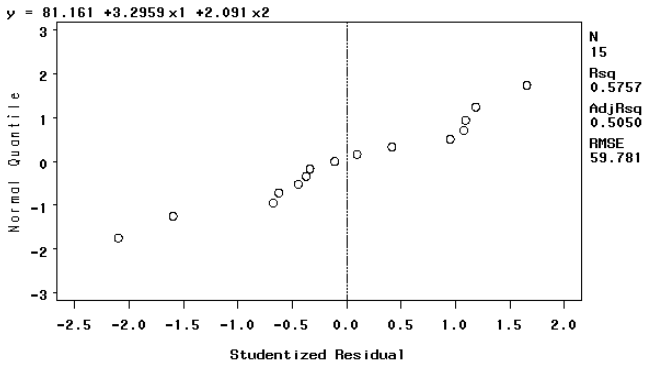
$$\hat{y} = 128.55 + 3.95x_1.$$

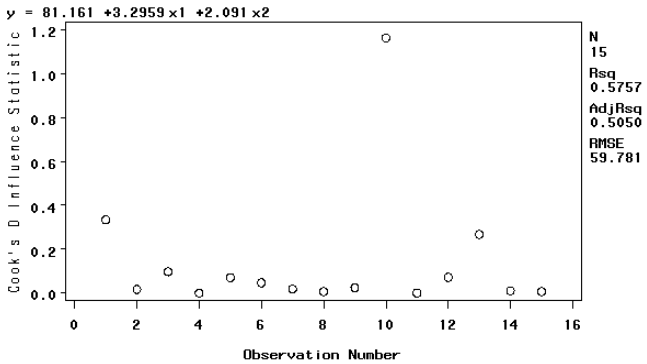
Program editor contents:

```
* multiple regression with diagnostic plots;

proc reg data=gce;
  model y=x1 x2/p r;
  plot student.*predicted. nqq.*student. cookd.*obs.; * diagnostic plots;
run;
```

Diagnostic plots suggest the residuals are roughly normal with no substantial outliers, though the Cook's distance is substantially larger for observation 10. We may wish to fit the model without observation 10 to see whether conclusions change.







Output window contents:

The REG Procedure
Model: MODEL1
Dependent Variable: y gce score of 1000

Number of Observations Read 15
Number of Observations Used 15

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	58188	29094	8.14	0.0058
Error	12	42885	3573.78575		
Corrected Total	14	101073			

Root MSE	59.78115	R-Square	0.5757	<---- not much improvement
Dependent Mean	568.93333	Adj R-Sq	0.5050	over X1 alone
Coeff Var	10.50758			

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error
Intercept	Intercept	1	81.16147	122.40594
x1	compulsory part of gce score of 200	1	3.29594	1.18032
x2	scel paper score	1	2.09104	1.92480

Parameter Estimates

Variable	Label	DF	t Value	Pr > |t|	
Intercept	Intercept	1	0.66	0.5198	
x1	compulsory part of gce score of 200	1	2.79	0.0163	
x2	scel paper score	1	1.09	0.2987	<---- not sig

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual
1	476.0000	589.2011	25.7429	-113.2011	53.955	-2.098
2	457.0000	480.5755	27.1852	-23.5755	53.242	-0.443
3	540.0000	482.3477	26.8144	57.6523	53.430	1.079
4	551.0000	557.1980	17.0782	-6.1980	57.290	-0.108
5	575.0000	508.7152	21.5941	66.2848	55.745	1.189
6	698.0000	713.5606	42.5495	-15.5606	41.992	-0.371
7	545.0000	582.9981	19.8210	-37.9981	56.400	-0.674
8	574.0000	550.3575	19.3269	23.6425	56.571	0.418
9	645.0000	590.1574	16.4078	54.8426	57.485	0.954
10	690.0000	624.1814	44.7140	65.8186	39.679	1.659
11	634.0000	628.8225	27.0938	5.1775	53.289	0.0972
12	637.0000	576.7250	23.4122	60.2750	55.006	1.096
13	390.0000	473.0974	29.2939	-83.0974	52.112	-1.595
14	562.0000	597.6354	17.0274	-35.6354	57.305	-0.622
15	560.0000	578.4272	23.8109	-18.4272	54.835	-0.336

Output Statistics

Obs	-2-1 0 1 2	Cook's D	
1	| ****| |	0.334	
2	| | |	0.017	
3	| |** |	0.098	
4	| | |	0.000	
5	| |** |	0.071	
6	| | |	0.047	
7	| *| |	0.019	
8	| | |	0.007	
9	| |* |	0.025	
10	| |*** |	1.165	<---- large Cook's D
11	| | |	0.001	
12	| |** |	0.073	

```
 13  |   ***|      |       0.268
 14  |     *|      |       0.011
 15  |      |      |       0.007

Sum of Residuals                            0
Sum of Squared Residuals                42885
Predicted Residual SS (PRESS)           79668
```

2.2.1 Some Comments on GCE Analysis

I will give you my thoughts on these data, and how I would attack this problem, keeping the ultimate goal in mind. I will examine whether transformations of the data are appropriate, and whether any important conclusions are dramatically influenced by individual observations. I will use some new tools to attack this problem, and will outline how they are used.

The plot of GCE (Y) against COMP (X_1) is fairly linear, but the trend in the plot of GCE (Y) against SCEL (X_2) is less clear. You might see a non-linear trend here, but the relationship is not very strong. When I assess plots I try to not allow a few observations affect my perception of trend, and with this in mind, I do not see any strong evidence at this point to transform any of the variables.

One difficulty that we must face when building a multiple regression model is that these two-dimensional (2D) plots of a response against individual predictors may have little information about the appropriate scales for a multiple regression analysis. In particular, the 2D plots only tell us whether we need to transform the data in a simple linear regression analysis. If a 2D plot shows a strong non-linear trend, I would do an analysis using the suggested transformations, including any other effects that are important. However, it might be that no variables need to be transformed in the multiple regression model.

The **partial regression residual plot** is a graphical tool that provides information about the need for transformations in a multiple regression model. The following **reg** procedure generates diagnostics and the partial residual plots for each predictor in the multiple regression model that has COMP and SCEL as predictors of GCE. **SAS** also gives a partial residual plot for the intercept but this does not help much so I omitted it. The partial residual plots are line-printer quality, and sent to the output window.

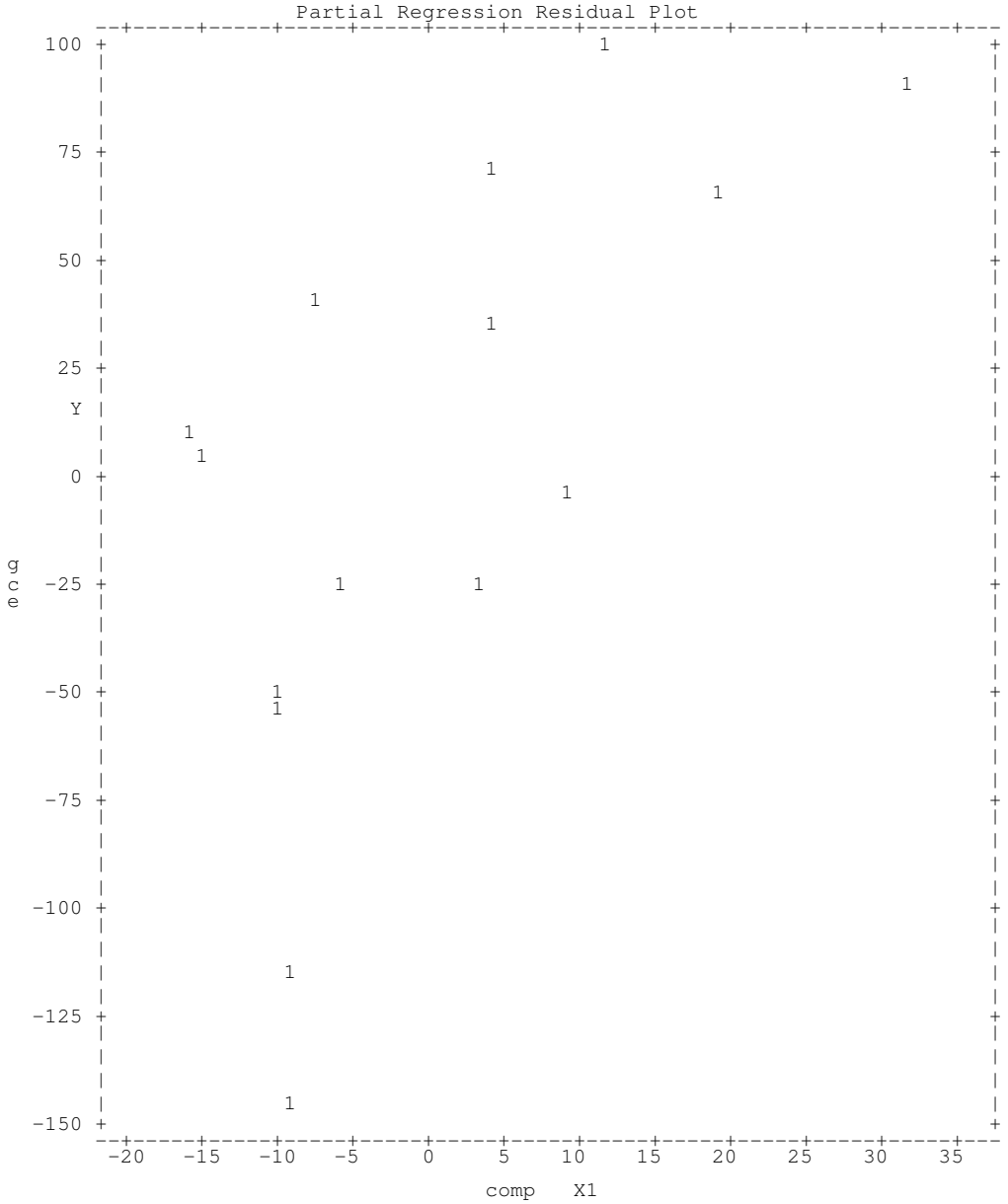
The first partial regression residual plot for COMP, given below, "**adjusts**" GCE and COMP for their common dependence on all the other predictors in the model (only SCEL here). This plot tells us whether we need to transform COMP in the multiple regression model, and whether any observations are influencing the significance of COMP in the fitted model. A roughly linear trend suggests that no transformation of COMP is warranted. The positive relationship seen here is consistent with the coefficient of COMP being positive in the multiple regression model. The partial residual plot for COMP shows little evidence of curvilinearity, and much less so than the original 2D plot of GCE against COMP. This indicates that there is no strong evidence for transforming COMP in a multiple regression model that includes SCEL.

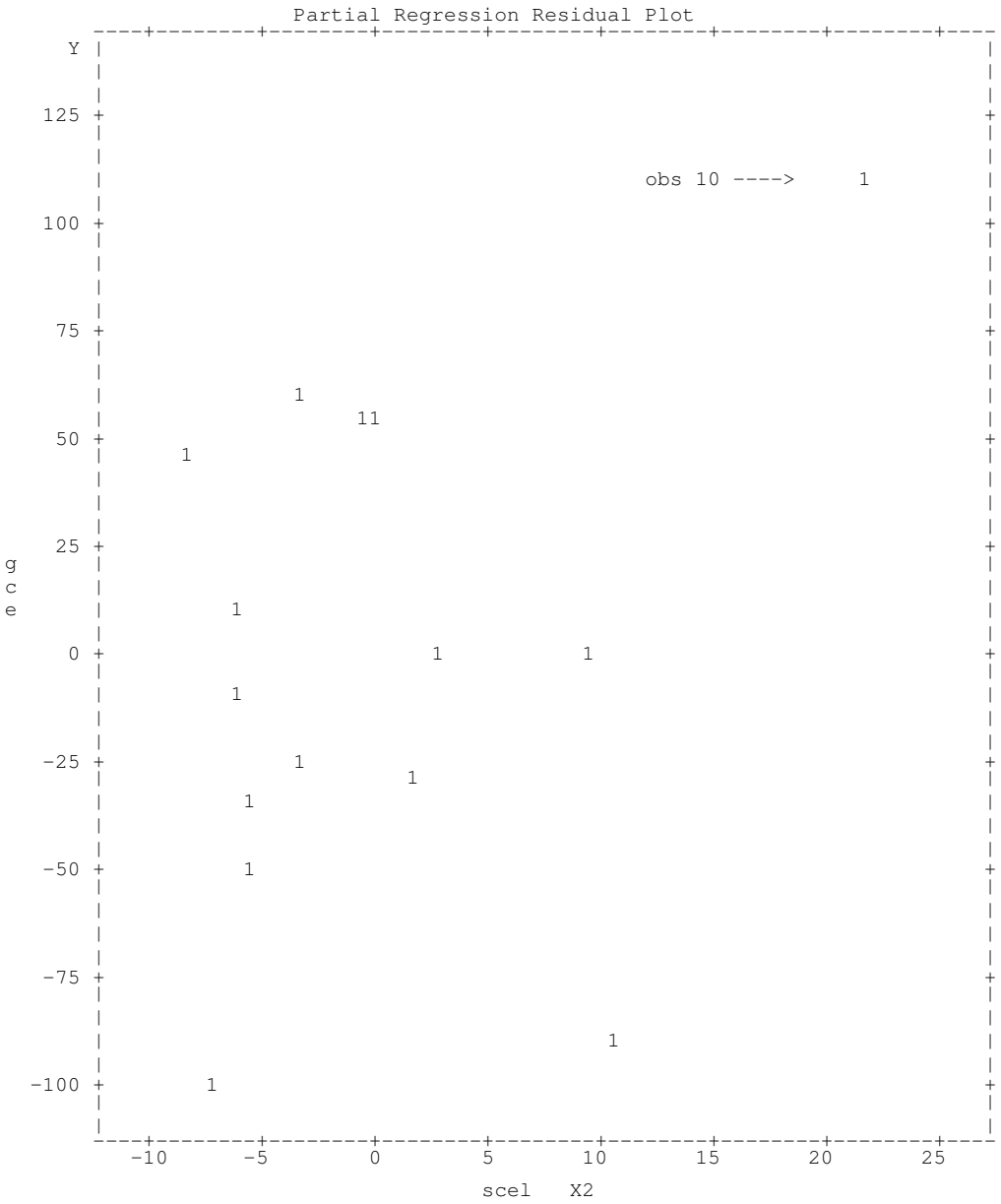
Although SCEL appears to somewhat useful as a predictor of GCE on it's own, the multiple regression output indicates that SCEL does not explain a significant amount of the variation in GCE, once the effect of COMP has been taken into account. Put another way, previous performance in the School Certificate English Language (X_2) has little predictive value independently of what has already emerged from the current performance in the compulsory papers (X_1 or COMP). This conclusion is is consistent with the fairly weak linear relationship between GCE against SCEL seen

in the second partial residual plot.

Program editor contents:

```
* partial residual plots;
*   to check relationship of X_i with Y after accounting for all other Xs in model;

  * note: uncomment ods lines to produce high-resolution plots in newer SAS versions;
*ods graphics on;
proc reg data=gce;
  model y=x1 x2/p r partial;  * /partial gives partial resid plots for each predictor;
run;
*ods graphics off;
```





Do diagnostics suggest any deficiencies associated with this conclusion? The partial residual plot of SCEL highlights observation 10, which has the largest value of Cook's distance in the multiple regression model. If we visually hold observation 10 out from this partial residual plot, it would appear that the relationship observed in this plot would weaken. This suggests that observation 10 is actually **enhancing** the significance of SCEL in the multiple regression model. That is, the p-value for testing the importance of SCEL in the multiple regression model would be inflated by holding out observation 10. The following output confirms this conjecture. The studentized residuals, Cook's distances and partial residual plots showed no serious deficiencies, so this output was omitted.

Program editor contents:

```
* run analysis excluding influential observation 10 ;

data gce2;
  set gce;
```

```
  if _N_ ~= 10;   * keep all observations that are not obs 10 (i.e., delete obs 10);
run;

proc reg data=gce2;
  model y=x1 x2/p r partial;  * /partial gives partial resid plots for each predictor;
run;
```

Output window contents:

```
The REG Procedure
Model: MODEL1
Dependent Variable: y gce score of 1000

Number of Observations Read          14
Number of Observations Used          14

                            Analysis of Variance
                                   Sum of           Mean
Source                     DF        Squares         Square    F Value    Pr > F
Model                       2          52317          26158       8.71    0.0054
Error                      11          33052     3004.73867
Corrected Total            13          85369

Root MSE              54.81550    R-Square     0.6128
Dependent Mean       560.28571    Adj R-Sq     0.5424
Coeff Var              9.78349

                              Parameter Estimates
                                                        Parameter       Standard
Variable     Label                                DF      Estimate          Error
Intercept    Intercept                             1     159.46077       120.29478
x1           compulsory part of gce score          1       4.24135         1.20185
             of 200
x2           scel paper score                      1      -1.27968         2.56646

                          Parameter Estimates
Variable     Label                                DF    t Value    Pr > |t|
Intercept    Intercept                             1       1.33      0.2118
x1           compulsory part of gce score          1       3.53      0.0047
             of 200
x2           scel paper score                      1      -0.50      0.6279 <----
```

What are my conclusions? It would appear that SCEL is not a useful predictor in the multiple regression model. For simplicity, I would likely use a simple linear regression model to predict GCE from COMP only. A diagnostic analysis of the model showed no serious deficiencies.

Chapter 3

A Taste of Model Selection for Multiple Regression

3.1 Model

Given data on a response variable Y and k predictor variables $X_1, X_2, \ldots, X_k$, we wish to develop a regression model to predict Y. Assuming that the collection of variables is measured on the correct scale, and that the candidate list of predictors includes all the important predictors, the most general model is

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon.$$

In most problems one or more of the predictors can be eliminated from this general or **full model** without (much) loss of information. We want to identify the important predictors, or equivalently, eliminate the predictors that are not very useful for explaining the variation in Y (conditional on the other predictors in the model).

We will study several **automated** methods for model selection. Given a specific criterion for selecting a model, **SAS** gives the best predictors. Before applying any of the methods, you should plot Y against each predictor $X_1, X_2, \ldots, X_k$ to see whether transformations are needed. If a transformation of X_i is suggested, include the transformation along with the original X_i in the candidate list. Note that you can transform the predictors differently, for example, $\log(X_1)$ and $\sqrt{X_2}$. However, if several transformations are suggested for the response, then you should consider doing one analysis for each suggested response scale before deciding on the final scale.

At this point, I will only consider the **backward elimination method**. Other approaches will be addressed later this semester.

3.2 Backward Elimination

The backward elimination procedure deletes unimportant variables, one at a time, starting from the full model. The steps is the procedure are:

1. Fit the full model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon.$$

2. Find the variable which when omitted from the full model (1) reduces R^2 the least, or equivalently, increases the Residual SS the least. This is the variable that gives the largest p-value for testing an individual regression coefficient $H_0 : \beta_i = 0$ for $i > 0$. Suppose this variable is X_k. If you reject H_0, stop and conclude that the full model is best. If you do not reject H_0, delete X_k from the full model, giving the **new full model**

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{k-1} X_{k-1} + \varepsilon.$$

Repeat steps 1 and 2 sequentially until no further predictors can be deleted.

In backward elimination we isolate the least important predictor left in the model, and check whether it is important. If not, delete it and repeat the process. Otherwise, stop. The default test level on the individual predictors in **SAS** is $\alpha = .10$.

Epidemiologists use a slightly different approach to building models. They argue strongly for the need to always include **confounding** variables in a model, regardless of their statistical significance. I will briefly discuss this issue, but you should recognize that there is no universally accepted best approach to building models. A related issue is that several sets of predictors might give nearly identical fits and predictions to those obtained using any model selection method. This should not be too surprising because predictors are often correlated with each other. However, this should make us question whether one could ever completely unravel which variables are important (and which are not) for predicting a response.

3.3 Example: Peru Indian blood pressure

I will illustrate backward elimination on the Peru Indian data, using systolic blood pressure ($x9$) as the response, and seven candidate predictors: $x3$ = weight in kilos; $x4$ = height in mm; $x5$ = chin skin fold in mm; $x6$ = forearm skin fold in mm; $x7$ = calf skin fold in mm; $x8$ = pulse rate-beats/min, and $x11$ = fraction.

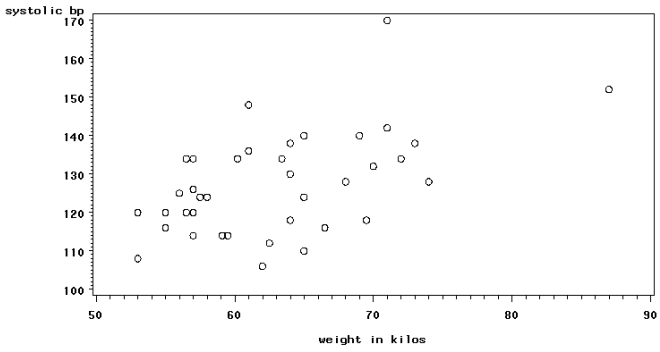
The **SAS** program given below generates simple summary statistics and plots. The plots do not suggest any apparent transformations of the response or the predictors, so we will analyze the data using the given scales. The **corr** procedure asks for correlations between the response and each potential predictor. I will note that the predictors are not highly correlated with each other.
Program editor contents:

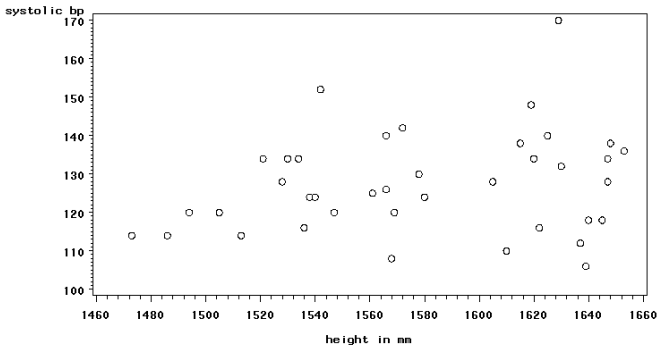
```
options ls=79 nodate nocenter;

data indian;
  infile "F:\Dropbox\UNM\teach\ADA2_stat528\sas\ADA2_02_indian.dat";
  input n x1-x11;
  label x1='age in years'          x2='years since migration'
    x3='weight in kilos'           x4='height in mm'
    x5='chin skin fold in mm'      x6='forearm skin fold in mm'
    x7='calf skin fold in mm'      x8='pulse rate-beats/min'
    x9='systolic bp'              x10='diastolic bp'
    x11='years since migration div by age'
    x12='cat weight';
  y=x9; * response variable;
  label x9='systolic bp';
run;

proc corr data=indian nosimple noprob;
  var y;                        * corr of y with these x variables ;
```

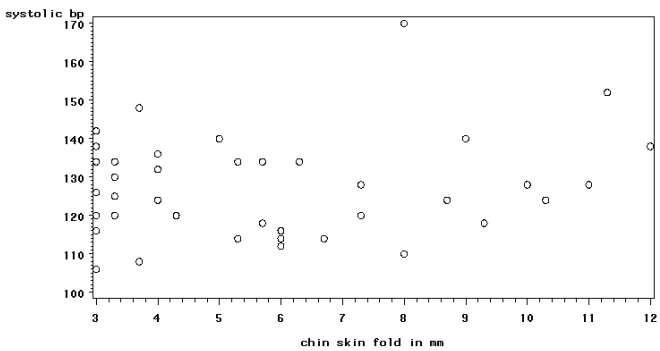
```
  with x3 x4 x5 x6 x7 x8 x11;  * (no xs with each other)          ;
run;

* define v=symbol, c=color, and i=interpolation;
symbol1 v=circle c=black i=none;
proc gplot data=indian;
  plot x9*(x3 x4 x5 x6 x7 x8 x11);
run;
```

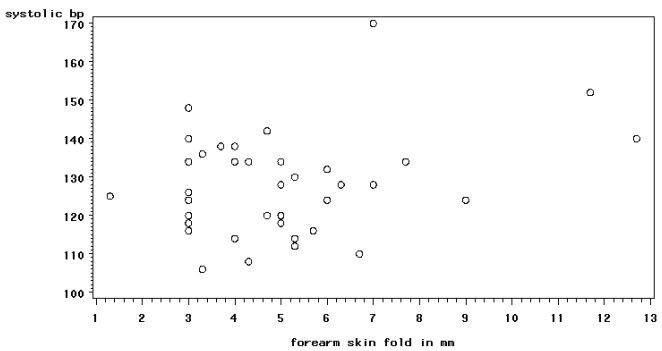
Output window contents: (reformatted a little to fit on page better)

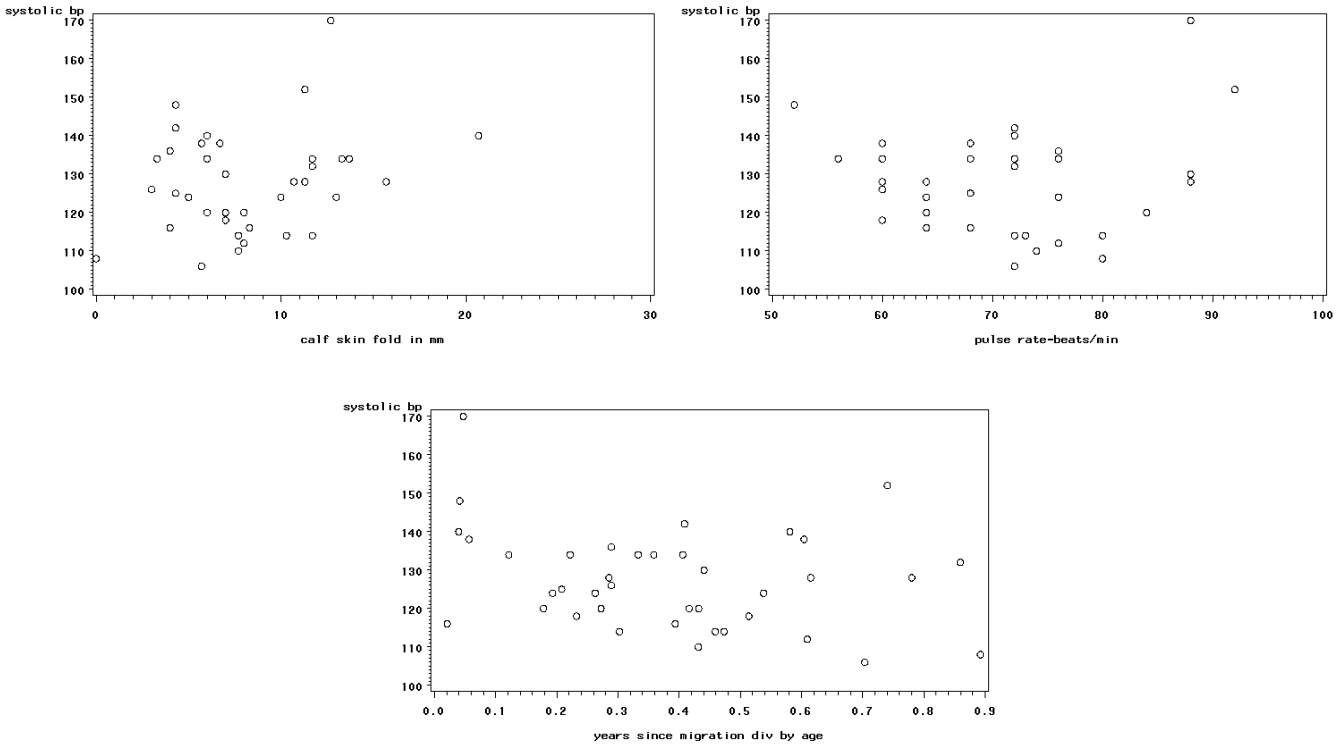
```
The CORR Procedure

 7 With Variables:   x3     x4     x5     x6     x7     x8     x11
 1      Variables:   y

   Pearson Correlation Coefficients, N = 39
                                          y
x3   weight in kilos                    0.52136
x4   height in mm                       0.21911
x5   chin skin fold in mm               0.17019
x6   forearm skin fold in mm            0.27228
x7   calf skin fold in mm               0.25079
x8   pulse rate-beats/min               0.13256
x11  years since migration div by age  -0.27615
```











To get output for the backward elimination method in **reg**, specify the following **SAS** statement in addition to the earlier **data step**:

Program editor contents:

```
proc reg data=indian;
  model y = x3 x4 x5 x6 x7 x8 x11 /selection = backward; * backward selection;
run;
```

Output window contents:

```
The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read          39
Number of Observations Used          39

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.5259 and C(p) = 8.0000

                              Analysis of Variance
                                   Sum of           Mean
Source                     DF        Squares         Square    F Value    Pr > F
Model                       7     3434.98916      490.71274       4.91    0.0008
Error                      31     3096.44673       99.88538
Corrected Total            38     6531.43590

              Parameter     Standard
Variable       Estimate        Error   Type II SS  F Value  Pr > F
Intercept    106.45745     53.91303    389.46328     3.90  0.0573
x3             1.71094      0.38659   1956.48532    19.59  0.0001
x4            -0.04533      0.03945    131.87622     1.32  0.2593
x5            -1.15724      0.84612    186.84503     1.87  0.1812
x6            -0.70182      1.34986     27.00094     0.27  0.6068
x7             0.10357      0.61170      2.86340     0.03  0.8667  <----
x8             0.07485      0.19570     14.61395     0.15  0.7047
```

```
x11          -29.31807      7.86839   1386.76004    13.88  0.0008

Bounds on condition number: 3.6493, 105.6
------------------------------------------------------------------------------
```

Remarks on Step 0: The full model has 7 predictors so REG df = 7. The F-test in the full model ANOVA table ($F = 4.91$ with p-value = 0.0008) tests the hypothesis that the regression coefficient for each predictor variable is zero. This test is highly significant, indicating that one or more of the predictors is important in the model.

In the parameter estimate table, the F-value column gives the square of the t-statistic for testing the significance of the individual predictors in the full model (conditional on all other predictors being in the model). The p-value is the same whether the t-statistic or F-value is shown.

The least important variable in the full model, as judged by the p-value, is $x7$ = calf skin fold. This variable, upon omission, reduces R^2 the least, or equivalently, increases the Residual SS the least. The p-value of 0.87 exceeds the default 0.10 cut-off, so $x7$ is omitted from the model.
Output window contents: (continued)

```
Backward Elimination: Step 1

Variable x7 Removed: R-Square = 0.5255 and C(p) = 6.0287

                              Analysis of Variance
                                    Sum of           Mean
Source                     DF       Squares         Square    F Value    Pr > F
Model                       6    3432.12577      572.02096       5.91    0.0003
Error                      32    3099.31013       96.85344
Corrected Total            38    6531.43590

              Parameter     Standard
Variable       Estimate        Error   Type II SS  F Value  Pr > F
Intercept    106.13720     53.05580    387.60072     4.00  0.0540
x3             1.70900      0.38051   1953.76360    20.17  <.0001
x4            -0.04478      0.03871    129.55602     1.34  0.2560
x5            -1.14165      0.82823    184.02391     1.90  0.1776
x6            -0.56731      1.07462     26.99288     0.28  0.6012
x8             0.07103      0.19142     13.33727     0.14  0.7130  <----
x11          -29.53996      7.63982   1448.00203    14.95  0.0005

Bounds on condition number: 2.8627, 68.302
------------------------------------------------------------------------------
```

Remarks on Step 1: After deleting $x7$, the six predictor model is fitted. At least one of the predictors left is important, as judged by the overall F-test p-value. The least important predictor left is $x8$ = pulse rate. This variable is omitted from the model because the p-value for including it exceeds the 0.10 threshold.
Output window contents: (continued)

```
Backward Elimination: Step 2

Variable x8 Removed: R-Square = 0.5234 and C(p) = 4.1622

                              Analysis of Variance
                                    Sum of           Mean
Source                     DF       Squares         Square    F Value    Pr > F
Model                       5    3418.78850      683.75770       7.25    0.0001
Error                      33    3112.64740       94.32265
Corrected Total            38    6531.43590

              Parameter     Standard
Variable       Estimate        Error   Type II SS  F Value  Pr > F
Intercept    110.27853     51.18665    437.80889     4.64  0.0386
x3             1.71825      0.37470   1983.50473    21.03  <.0001
```

x4	-0.04504	0.03820	131.12231	1.39	0.2468
x5	-1.17715	0.81187	198.29475	2.10	0.1565
x6	-0.43384	0.99933	17.77746	0.19	0.6670
x11	-28.98166	7.39170	1450.01807	15.37	0.0004

Bounds on condition number: 2.8504, 48.657

Backward Elimination: Step 3

Variable x6 Removed: R-Square = 0.5207 and C(p) = 2.3402

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	3401.01104	850.25276	9.23	<.0001
Error	34	3130.42486	92.07132		
Corrected Total	38	6531.43590			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	104.52276	48.84626	421.58326	4.58	0.0396
x3	1.64631	0.33202	2263.63282	24.59	<.0001
x4	-0.03957	0.03563	113.56435	1.23	0.2745
x5	-1.31083	0.74220	287.19721	3.12	0.0863
x11	-28.32576	7.14878	1445.51712	15.70	0.0004

Bounds on condition number: 2.2929, 26.188

Backward Elimination: Step 4

Variable x4 Removed: R-Square = 0.5033 and C(p) = 1.4771

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3287.44669	1095.81556	11.82	<.0001
Error	35	3243.98921	92.68541		
Corrected Total	38	6531.43590			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	52.90931	15.08952	1139.52747	12.29	0.0013
x3	1.44074	0.27656	2515.32927	27.14	<.0001
x5	-1.01353	0.69454	197.37104	2.13	0.1534
x11	-27.35223	7.11846	1368.43705	14.76	0.0005

Bounds on condition number: 1.5803, 12.43

Backward Elimination: Step 5

Variable x5 Removed: R-Square = 0.4731 and C(p) = 1.4531

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3090.07564	1545.03782	16.16	<.0001
Error	36	3441.36025	95.59334		
Corrected Total	38	6531.43590			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	60.89595	14.28088	1738.18032	18.18	0.0001
x3	1.21686	0.23369	2592.00831	27.11	<.0001
x11	-26.76721	7.21779	1314.69602	13.75	0.0007

Bounds on condition number: 1.094, 4.3759

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination

Step	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square
1	x7	calf skin fold in mm	6	0.0004	0.5255
2	x8	pulse rate-beats/min	5	0.0020	0.5234
3	x6	forearm skin fold in mm	4	0.0027	0.5207
4	x4	height in mm	3	0.0174	0.5033
5	x5	chin skin fold in mm	2	0.0302	0.4731

Summary of Backward Elimination

Step	C(p)	F Value	Pr > F
1	6.0287	0.03	0.8667
2	4.1622	0.14	0.7130
3	2.3402	0.19	0.6670
4	1.4771	1.23	0.2745
5	1.4531	2.13	0.1534

Remark on Summary Table: The partial R^2 is the reduction in R^2 achieved by omitting variables sequentially.

The backward elimination procedure eliminates five variables from the full model, in the following order: calf skin fold ($x7$), pulse rate ($x8$), forearm skin fold ($x6$), height ($x4$), and chin skin fold ($x5$). The model selected by backward elimination includes two predictors: weight ($x3$) and fraction ($x11$). As we progress from the full model to the selected model, R^2 decreases as follows: 0.53, 0.53, 0.52, 0.52, 0.50, and 0.47. The decrease is slight across this spectrum of models.

Using a mechanical approach, we are led to a model with weight and years by age fraction as predictors of systolic blood pressure. At this point we should closely examine this model.

3.3.1 Analysis for Selected Model

A **reg** proc for generating summaries for the selected model, followed by the output, is given on the next several pages.

Program editor contents:

```
proc reg data=indian;
  model y=x3 x11/p r partial;
  plot student.*predicted. nqq.*student. cookd.*obs.; * diagnostic plots;
run;
```

Output window contents:

```
The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read          39
Number of Observations Used          39
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3090.07564	1545.03782	16.16	<.0001
Error	36	3441.36025	95.59334		
Corrected Total	38	6531.43590			

Root MSE	9.77718	R-Square	0.4731
Dependent Mean	127.41026	Adj R-Sq	0.4438
Coeff Var	7.67378		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error

Intercept	Intercept	1	60.89595	14.28088
x3	weight in kilos	1	1.21686	0.23369
x11	years since migration div by age	1	-26.76721	7.21779

Parameter Estimates

Variable	Label	DF	t Value	Pr > |t|
Intercept	Intercept	1	4.26	0.0001
x3	weight in kilos	1	5.21	<.0001
x11	years since migration div by age	1	-3.71	0.0007

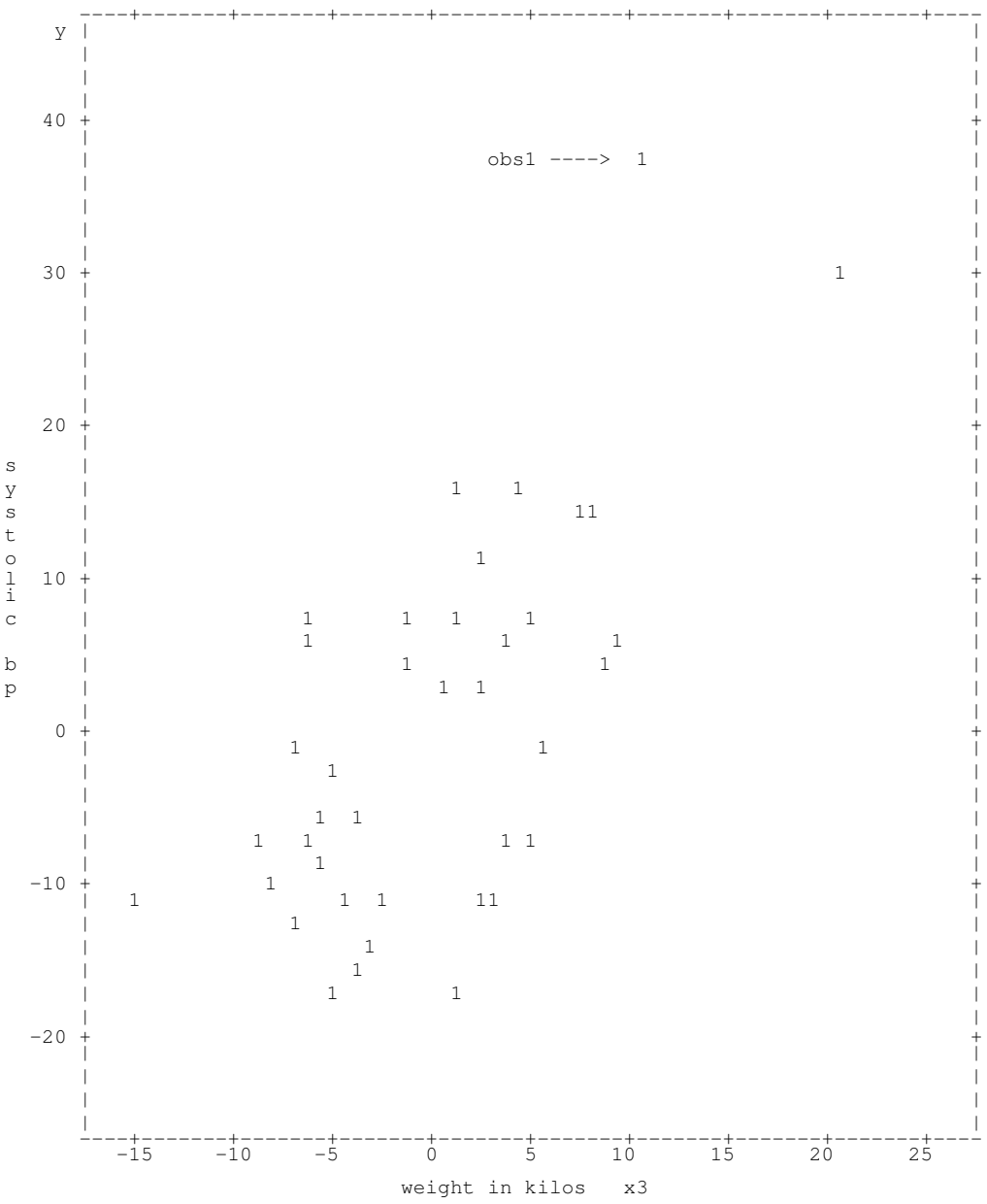
Output Statistics

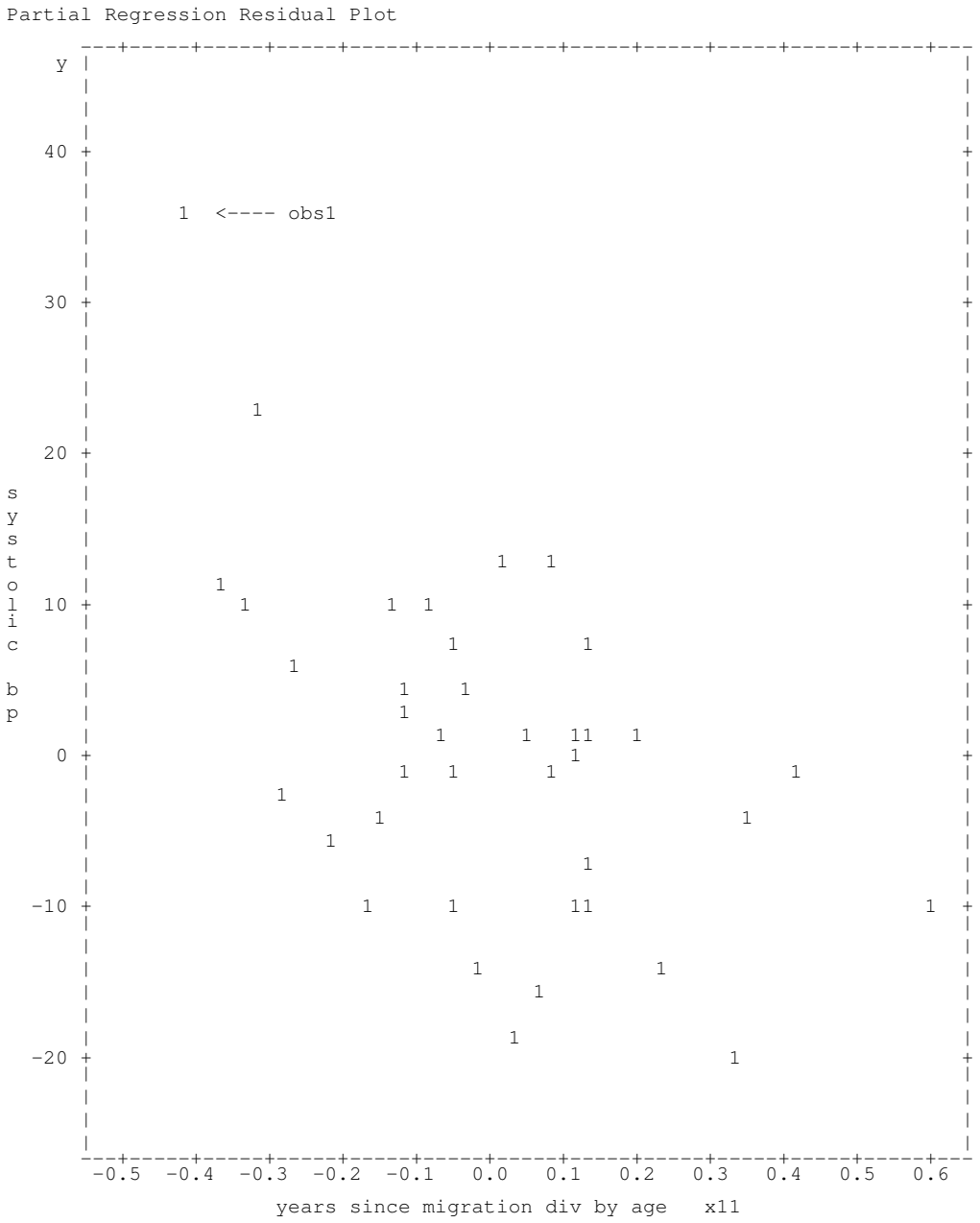
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D
1	170.0000	146.0181	3.7962	23.9819	9.010	2.662	|***** |	0.419
2	120.0000	122.3481	2.1902	-2.3481	9.529	-0.246	| |	0.001
3	125.0000	123.4635	2.3750	1.5365	9.484	0.162	| |	0.001
4	148.0000	134.0088	2.8563	13.9912	9.351	1.496	| ** |	0.070
5	140.0000	138.9209	3.0847	1.0791	9.278	0.116	| |	0.000
6	106.0000	117.5049	2.8523	-11.5049	9.352	-1.230	| ** |	0.047
7	120.0000	120.6095	2.8721	-0.6095	9.346	-0.0652	| |	0.000
8	108.0000	101.4899	5.1518	6.5101	8.310	0.783	| * |	0.079
9	124.0000	134.8108	2.2189	-10.8108	9.522	-1.135	| ** |	0.023
10	134.0000	119.3826	2.1598	14.6174	9.536	1.533	| |***	0.040
11	116.0000	131.2722	1.7432	-15.2722	9.621	-1.587	| *** |	0.028
12	114.0000	124.7009	1.8379	-10.7009	9.603	-1.114	| ** |	0.015
13	130.0000	126.9656	1.6127	3.0344	9.643	0.315	| |	0.001
14	118.0000	131.7013	2.1674	-13.7013	9.534	-1.437	| ** |	0.036
15	138.0000	137.2453	2.8998	0.7547	9.337	0.0808	| |	0.000
16	134.0000	120.7260	2.1610	13.2740	9.535	1.392	| ** |	0.033
17	120.0000	119.1037	2.1807	0.8963	9.531	0.0940	| |	0.000
18	120.0000	116.2481	2.5630	3.7519	9.435	0.398	| |	0.004
19	114.0000	117.9583	2.2905	-3.9583	9.505	-0.416	| |	0.003
20	124.0000	124.4295	2.0166	-0.4295	9.567	-0.0449	| |	0.000
21	114.0000	120.6198	1.9737	-6.6198	9.576	-0.691	| * |	0.007
22	136.0000	127.3759	1.7285	8.6241	9.623	0.896	| |*	0.009
23	126.0000	122.5084	2.1028	3.4916	9.548	0.366	| |	0.002
24	124.0000	116.4521	2.5013	7.5479	9.452	0.799	| |*	0.015
25	128.0000	134.4713	3.0245	-6.4713	9.298	-0.696	| * |	0.017
26	134.0000	138.9010	2.6453	-4.9010	9.413	-0.521	| * |	0.007
27	112.0000	120.6279	2.2850	-8.6279	9.506	-0.908	| * |	0.016
28	128.0000	122.7506	3.1525	5.2494	9.255	0.567	| |*	0.012
29	134.0000	134.7804	2.4817	-0.7804	9.457	-0.0825	| |	0.000
30	128.0000	135.9945	2.1772	-7.9945	9.532	-0.839	| * |	0.012
31	140.0000	129.2966	2.2738	10.7034	9.509	1.126	| |**	0.024
32	138.0000	133.5416	2.8458	4.4584	9.354	0.477	| |	0.007
33	118.0000	132.5498	1.9620	-14.5498	9.578	-1.519	| *** |	0.032
34	110.0000	128.4330	1.6313	-18.4330	9.640	-1.912	| *** |	0.035
35	142.0000	136.3425	2.3789	5.6575	9.483	0.597	| |*	0.007
36	134.0000	128.2025	1.9627	5.7975	9.578	0.605	| |*	0.005
37	116.0000	127.2534	3.1756	-11.2534	9.247	-1.217	| ** |	0.058
38	132.0000	123.0561	3.6731	8.9439	9.061	0.987	| |*	0.053
39	152.0000	146.9349	5.6286	5.0651	7.995	0.634	| |*	0.066

Sum of Residuals	0
Sum of Squared Residuals	3441.36025
Predicted Residual SS (PRESS)	4078.66721

Recall that the partial regression residual plot for weight, given below, **adjusts** systolic blood pressure and weight for their common dependence on all the other predictors in the model (only years by age fraction here). This plot tells us whether we need to transform weight in the multiple regression model, and whether any observations are influencing the significance of weight in the fitted model. A roughly linear trend, as seen here, suggests that no transformation of weight is warranted. The positive relationship seen here is consistent with the coefficient of weight being positive in the multiple regression model.

The partial residual plot for fraction exhibits a stronger relationship than is seen in the earlier 2D plot of systolic blood pressure against year by age fraction. This means that fraction is more useful as a predictor after taking an individual's weight into consideration.


Partial Regression Residual Plot
y
40
30
20
10
0
-10
-20
obs1 ----> 1
systolic bp
-15
-10
-5
0
5
10
15
20
25
weight in kilos   x3



3.3.2 Residual and Diagnostic Analysis for Selected Model

Comments:

1. The individual with the highest systolic blood pressure (case 1) has a large studentized residual r_i and the largest Cook's D_i.

2. Except for case 1, the rankit plot and the plot of the studentized residuals against the fitted values show no gross abnormalities.

3. The plots of studentized residuals against the individual predictors show no patterns. The partial residual plots show roughly linear trends. These plots collectively do not suggest the

need to transform either of the predictors. Although case 1 is prominent in the partial residual plots, it does not appear to be influencing the significance of these predictors.

Model selection methods can be highly influenced by outliers and influential cases. We should hold out case 1, and rerun the backward procedure to see whether case 1 unduly influenced the selection of the two predictor model. If we hold out case 1, we find that the model with weight and fraction as predictors is suggested again. After holding out case 1, there are no large residuals, no extremely influential points, or any gross abnormalities in plots. The R^2 for the selected model is now $R^2 = 0.408$. This decrease in R^2 should have been anticipated. Why?[1]

The two analyses suggest that the "best model" for predicting systolic blood pressure is

$$\text{sys bp} \quad = \quad \beta_0 + \beta_1 \text{weight} + \beta_2 \text{year by age fraction} + \varepsilon.$$

Should case 1 be deleted? I have not fully explored this issue, but I will note that eliminating this case does have a significant impact on the estimates of the regression coefficients, and on predicted values. What do you think?

3.4 Options with Backward Elimination

Here are three useful **options** when doing backward elimination in **SAS**. The **SLSTAY=** option allows you to specify the level used for testing the significance of predictors, with 0.10 being the default. The **details=summary** option eliminates the printing of details on the individual steps and just provides the summary table as output. The **include=n** option forces the regression model to always include the first n predictors listed on the **model** statement. For example, the following **model** statement requests that $x3$ and $x4$ be forced or included in the final model, that the test level be changed to 0.15, and that only the summary table be provided.
Program editor contents:

```
proc reg;
  model y = x3 x4 x5 x6 x7 x8 x11/selection=backward
                                  include=2  slstay=0.15  details=summary;
run;
```

3.5 Example: Dennis Cook's Rat Data

This example illustrates the importance of a careful diagnostic analysis.

An experiment was conducted to investigate the amount of a particular drug present in the liver of a rat. Nineteen rats were randomly selected, weighed, placed under light ether anesthesia and given an oral dose of the drug. Because it was felt that large livers would absorb more of a given dose than small livers, the actual dose an animal received was approximately determined as 40mg of the drug per kilogram of body weight. (Liver weight is known to be strongly related to body weight.) After a fixed length of time, each rat was sacrificed, the liver weighed, and the percent of the dose in the liver determined.

The experimental hypothesis was that, for the method of determining the dose, there is no relationship between the percentage of dose in the liver (Y) and the body weight, liver weight, and

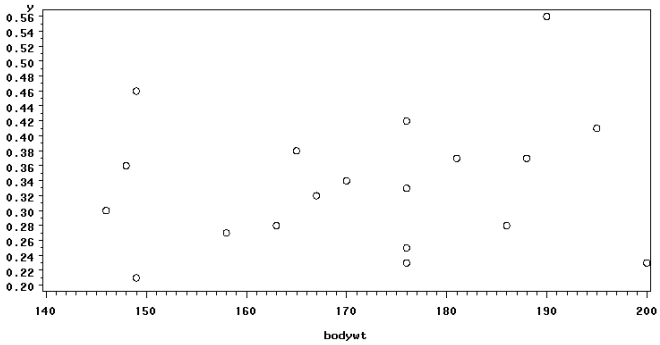
[1] Obs 1 increases the SST, but greatly increases model relationship so greatly increases SSR.

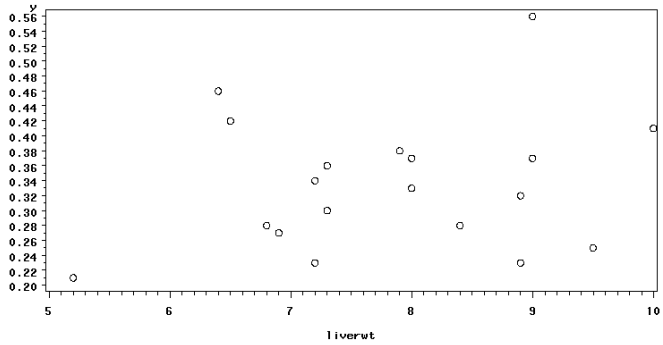
relative dose. A **SAS** program and output is given on the next several pages. Things to notice: (1) The correlation between Y and each predictor is small, as expected. (2) The backward elimination procedure selects weight and dose as predictors. The p-values for testing the importance of these variables, when added last to this two predictor model, are small, 0.019 and 0.015.

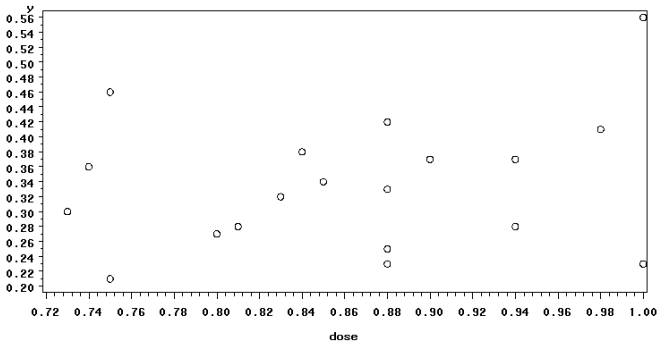
This cursory analysis leads to a conclusion that a combination of dose and body weight is associated with Y, but that neither of these predictors is important of its own. Although this commonly happens in regression problems, it is somewhat paradoxical here because dose was approximately a multiple of body weight, so to a first approximation, these predictors are linearly related and so only one of them should be needed in a linear regression model. Note that the correlation between dose and body weight is 0.99.

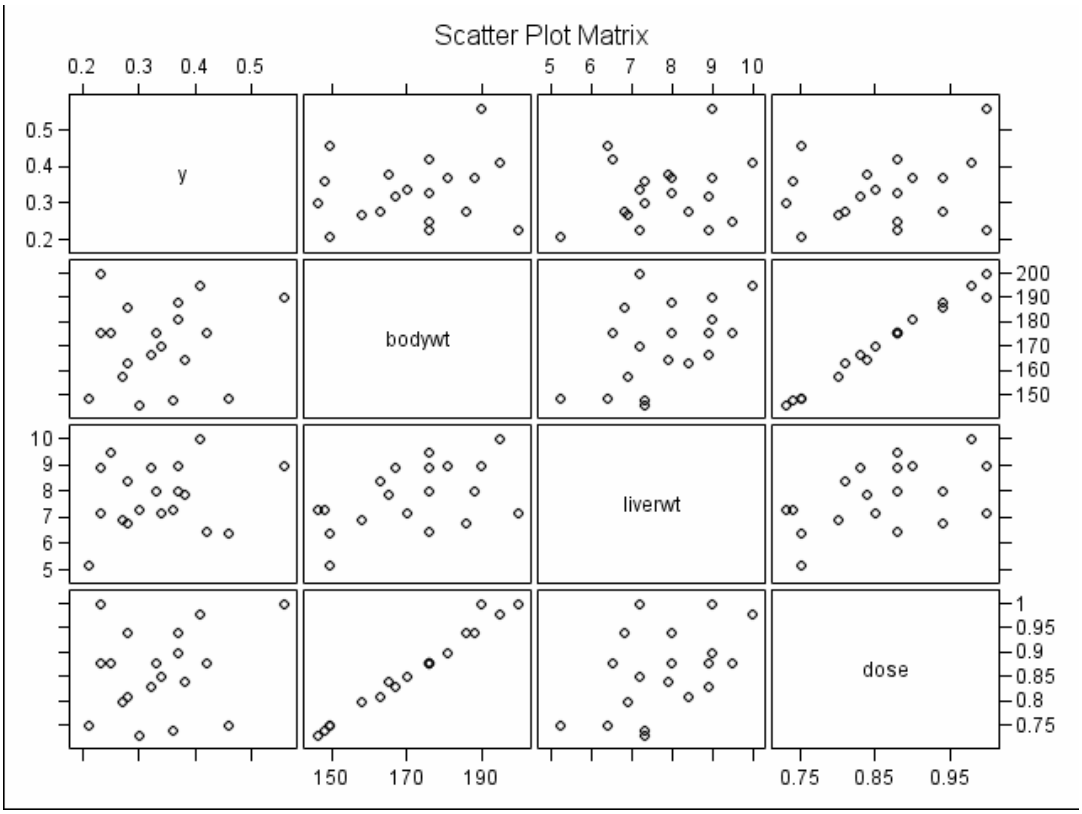
Program editor contents:

```
* read data;
data ratliver;
  infile "ADA2_03_ratliver.dat";
  input bodywt liverwt dose y;
run;

* correlation of y with each x;
proc corr data=ratliver nosimple noprob;
run;

* plot y with each x;
* define v=symbol, c=color, and i=interpolation;
symbol1 v=circle c=black i=none;
proc gplot data=ratliver;
  plot y*(bodywt liverwt dose);
run;

  ** Two ways to get scatterplot matrices in SAS;
  * use the "output delivery system" in SAS v9.1+ ;
  *    cite: http://www.ats.ucla.edu/stat/sas/faq/sas9_stat_plots.htm ;
  * scatterplot matrix;
  ods html style=journal;           * turn on html output;
  ods graphics on;                  * turn on ods graphics;
  proc corr data=ratliver plots;
    var y bodywt liverwt dose;
  run;
  ods graphics off;
  ods html close;

  * sgscatter is in v9.2+ ;
  *    cite: http://support.sas.com/documentation/cdl/en/grstatproc/62603/HTML/default/v
  * scatterplot matrix;
  proc sgscatter data=ratliver;
    matrix y bodywt liverwt dose;
  run;

* regression with backward selection;
proc reg data=ratliver;
  model y = bodywt liverwt dose /selection = backward; * backward selection;
run;
```









Output window contents:

```
The CORR Procedure

   4  Variables:    bodywt   liverwt  dose     y
```

```
          Pearson Correlation Coefficients, N = 19
              bodywt       liverwt          dose             y
bodywt        1.00000       0.50001       0.99021       0.15109
liverwt       0.50001       1.00000       0.49007       0.20333
dose          0.99021       0.49007       1.00000       0.22754
y             0.15109       0.20333       0.22754       1.00000

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read          19
Number of Observations Used          19

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.3639 and C(p) = 4.0000

                            Analysis of Variance
                                    Sum of           Mean
Source                    DF        Squares         Square    F Value    Pr > F
Model                      3        0.05126        0.01709       2.86    0.0720
Error                     15        0.08961        0.00597
Corrected Total           18        0.14087

            Parameter     Standard
Variable     Estimate        Error   Type II SS  F Value  Pr > F
Intercept     0.26592      0.19459      0.01116     1.87  0.1919
bodywt       -0.02125      0.00797      0.04241     7.10  0.0177
liverwt       0.01430      0.01722      0.00412     0.69  0.4193
dose          4.17811      1.52263      0.04498     7.53  0.0151

Bounds on condition number: 52.102, 314.59
------------------------------------------------------------------------------

Backward Elimination: Step 1

Variable liverwt Removed: R-Square = 0.3347 and C(p) = 2.6897

                            Analysis of Variance
                                    Sum of           Mean
Source                    DF        Squares         Square    F Value    Pr > F
Model                      2        0.04714        0.02357       4.02    0.0384
Error                     16        0.09373        0.00586
Corrected Total           18        0.14087

            Parameter     Standard
Variable     Estimate        Error   Type II SS  F Value  Pr > F
Intercept     0.28552      0.19127      0.01305     2.23  0.1550
bodywt       -0.02044      0.00784      0.03985     6.80  0.0190
dose          4.12533      1.50647      0.04393     7.50  0.0146

Bounds on condition number: 51.338, 205.35
------------------------------------------------------------------------------

All variables left in the model are significant at the 0.1000 level.

                    Summary of Backward Elimination
      Variable    Number    Partial     Model
Step  Removed     Vars In   R-Square   R-Square    C(p)     F Value   Pr > F
  1   liverwt         2      0.0292     0.3347     2.6897      0.69   0.4193
```

The apparent paradox can be resolved only with a careful diagnostic analysis! For the model with dose and body weight as predictors, there are no cases with large $|r_i|$ values, but case 3 has a relatively large Cook's D value. Further, the partial residual plot for body weight clearly highlights case 3. Without this case we would see roughly a random scatter of points, suggesting that body weight is unimportant after taking dose into consideration. The importance of body

weight as a predictor in the multiple regression model is due solely to the placement of case 3. The partial residual plot for dose gives the same message.

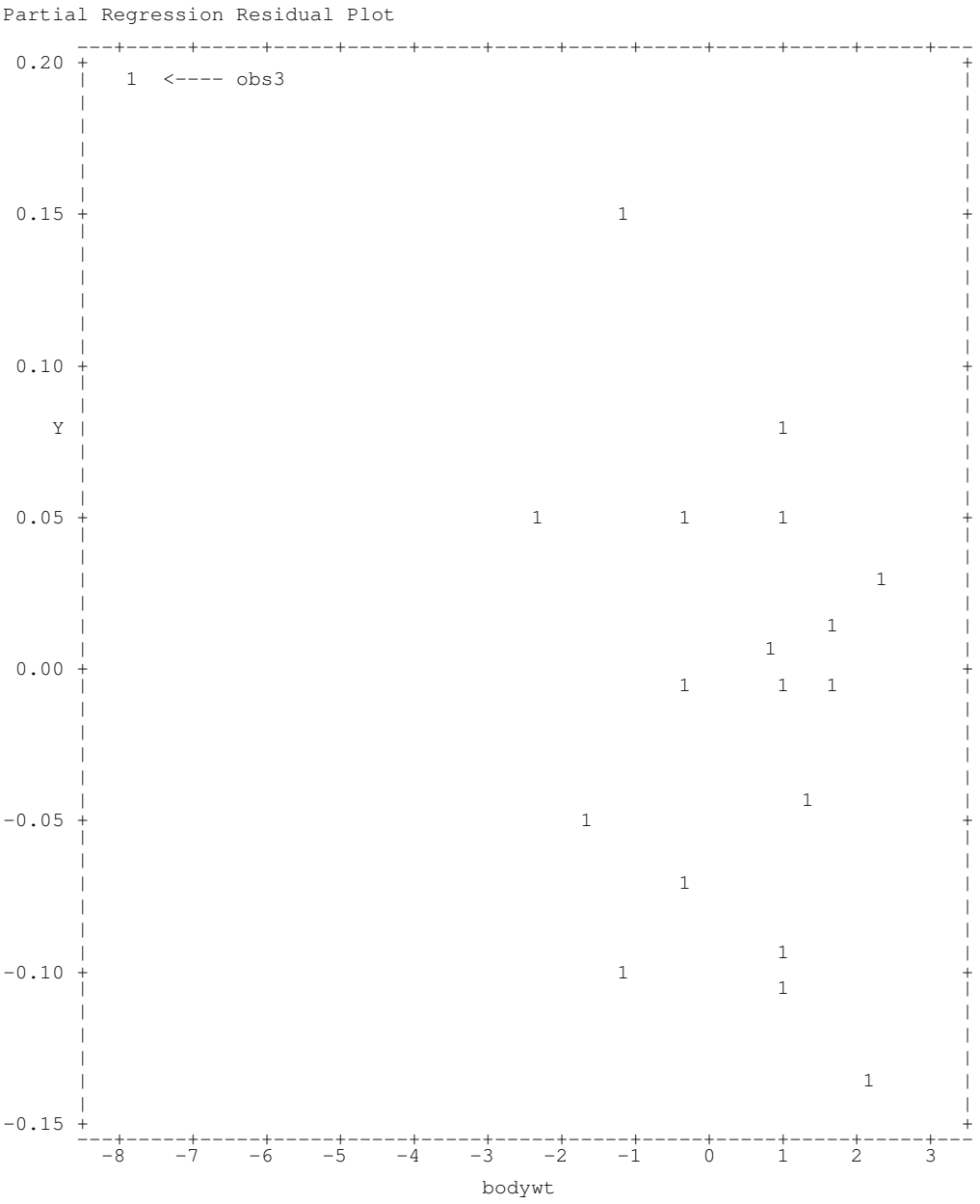
Program editor contents:

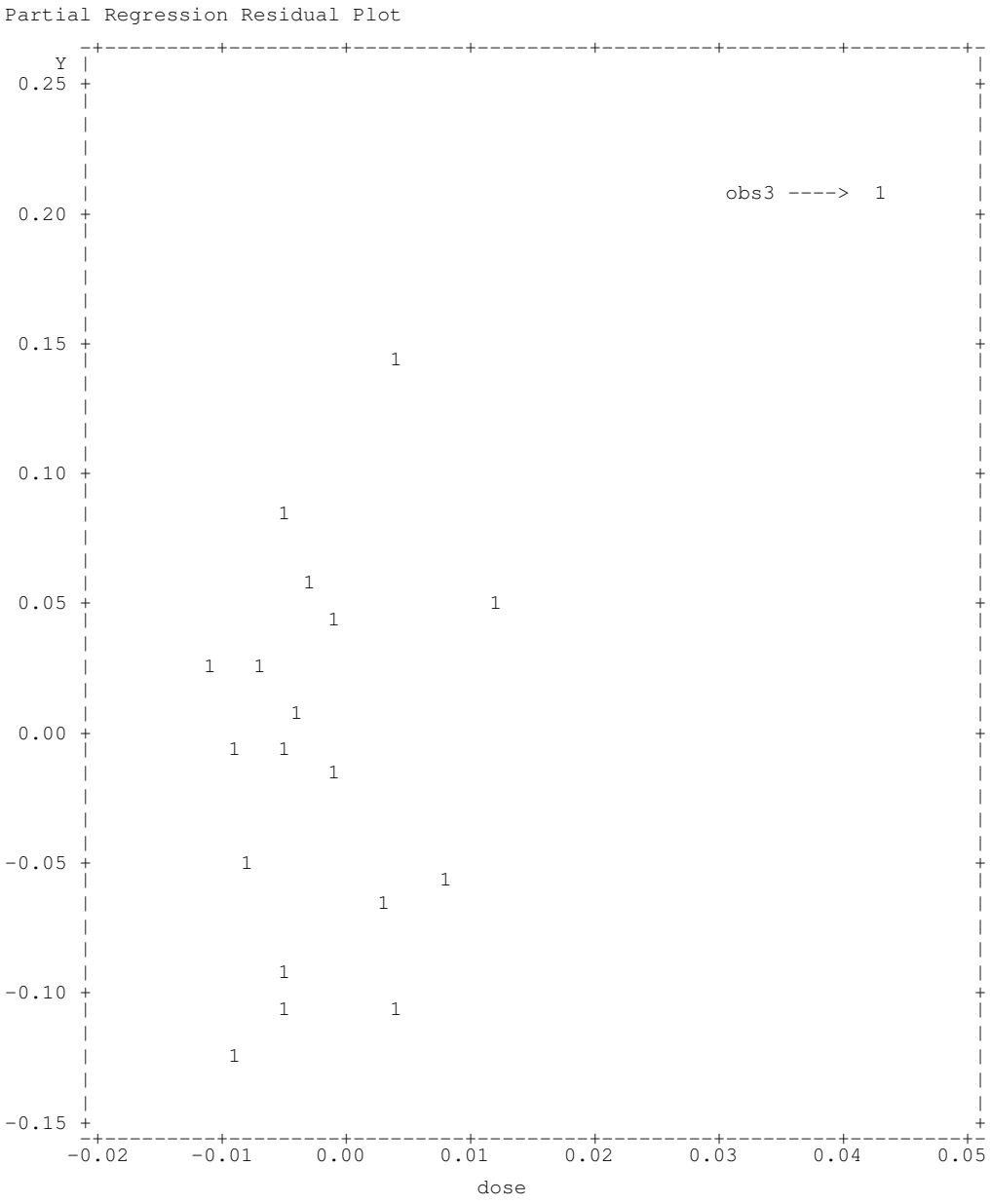
```
* Diagnostic Analysis for Model Selected by Backward Elimination;
proc reg data=ratliver;
  model y = bodywt dose /p r partial;
run;
```

Output window contents:

```
The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read          19
Number of Observations Used          19

                                Analysis of Variance
                                     Sum of           Mean
Source                     DF        Squares         Square    F Value    Pr > F
Model                       2        0.04714        0.02357       4.02    0.0384
Error                      16        0.09373        0.00586
Corrected Total            18        0.14087

Root MSE              0.07654    R-Square     0.3347
Dependent Mean        0.33526    Adj R-Sq     0.2515
Coeff Var            22.82924

                           Parameter Estimates
                        Parameter        Standard
Variable     DF          Estimate           Error    t Value    Pr > |t|
Intercept     1           0.28552         0.19127       1.49      0.1550
bodywt        1          -0.02044         0.00784      -2.61      0.0190
dose          1           4.12533         1.50647       2.74      0.0146
```

```
                                Output Statistics
     Dependent  Predicted     Std Error                 Std Error   Student                    Cook's
Obs   Variable      Value  Mean Predict   Residual   Residual  Residual    -2-1 0 1 2          D
  1     0.4200     0.3176        0.0198     0.1024     0.0739     1.385  |      |**    |  0.046
  2     0.2500     0.3176        0.0198    -0.0676     0.0739    -0.915  |     *|     |  0.020
  3     0.5600     0.5264        0.0697     0.0336     0.0316     1.061  |      |**   |  1.820 <-
  4     0.2300     0.3176        0.0198    -0.0876     0.0739    -1.185  |    **|     |  0.034
  5     0.2300     0.3220        0.0381    -0.0920     0.0664    -1.386  |    **|     |  0.212
  6     0.3200     0.2954        0.0225     0.0246     0.0731     0.337  |      |     |  0.004
  7     0.3700     0.3198        0.0273     0.0502     0.0715     0.702  |      |*    |  0.024
  8     0.4100     0.3417        0.0314     0.0683     0.0698     0.979  |      |*    |  0.065
  9     0.3300     0.3176        0.0198     0.0124     0.0739     0.167  |      |     |  0.001
 10     0.3800     0.3775        0.0257   0.002503     0.0721    0.0347  |      |     |  0.000
 11     0.2700     0.3556        0.0256    -0.0856     0.0721    -1.187  |    **|     |  0.059
 12     0.3600     0.3125        0.0312     0.0475     0.0699     0.679  |      |*    |  0.031
 13     0.2100     0.3333        0.0308    -0.1233     0.0700    -1.761  |   ***|     |  0.200
 14     0.2800     0.2946        0.0234    -0.0146     0.0729    -0.201  |      |     |  0.001
 15     0.3400     0.3165        0.0188     0.0235     0.0742     0.316  |      |     |  0.002
 16     0.2800     0.3607        0.0242    -0.0807     0.0726    -1.111  |    **|     |  0.046
 17     0.3000     0.3122        0.0330    -0.0122     0.0691    -0.176  |      |     |  0.002
 18     0.3700     0.2979        0.0262     0.0721     0.0719     1.002  |      |**   |  0.044
 19     0.4600     0.3333        0.0308     0.1267     0.0700     1.808  |      |***  |  0.211

Sum of Residuals                           0
Sum of Squared Residuals             0.09373
Predicted Residual SS (PRESS)        0.16353
```


Partial Regression Residual Plot
1 <---- obs3
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
Y
-8
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
bodywt



If we delete this case and redo the analysis we find, as expected, no important predictors of Y. The output below shows that the backward elimination removes each predictor from the model. Thus, the apparent relationship between Y and body weight and dose in the initial analysis can be ascribed to Case 3 alone. Can you see this case in the plots?

Program editor contents:

```
* Redo model selection after holing out case 3;
data ratliver2;
  set ratliver;
  if _N_ = 3 then delete;
run;

* regression with backward selection;
proc reg data=ratliver2;
  model y = bodywt liverwt dose /selection = backward; * backward selection;
run;
```

Output window contents:

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	dose	2	0.0112	0.0099	2.1599	0.16	0.6953
2	bodywt	1	0.0041	0.0058	0.2180	0.06	0.8075
3	liverwt	0	0.0058	0.0000	-1.6988	0.09	0.7636

All variables are omitted!

In his text[2], Weisberg writes: The careful analyst must now try to understand exactly why the third case is so influential. Inspection of the data indicates that this rat with weight 190g, was reported to have received a full dose of 1.000, which was a larger dose than it should have received according to the rule for assigning doses, see scatterplot above (e.g., rat 8 with a weight of 195g got a lower dose of 0.98). A number of causes for the result found in the first analysis are possible: (1) the dose or weight recorded for case 3 was in error, so the case should probably be deleted from the analysis, or (2) the regression fit in the second analysis is not appropriate except in the region defined by the 18 points excluding case 3. It is possible that the combination of dose and rat weight chosen was fortuitous, and that the lack of relationship found would not persist for any other combinations of them, since inclusion of a data point apparently taken under different conditions leads to a different conclusion. This suggests the need for collection of additional data, with dose determined by some rule other than a constant proportion of weight.

I hope the point of this analysis is clear! What have we learned from this analysis?

[2]Applied Linear Regression, 3rd Ed. by Sanford Weisberg, published by Wiley/Interscience in 2005 (ISBN 0-471-66379-4)

Chapter 4

One Factor Designs and Extensions

This section describes an experimental design to compare the effectiveness of four insecticides to eradicate beetles. The primary interest is determining which treatment is most effective, in the sense of providing the lowest typical survival time.

In a **completely randomized design** (CRD), the scientist might select a sample of genetically identical beetles for the experiment, and then randomly assign a predetermined number of beetles to the treatment groups (insecticides). The sample sizes for the groups need not be equal. A power analysis is often conducted to determine sample sizes for the treatments. For simplicity, assume that 48 beetles will be used in the experiment, with 12 beetles assigned to each group.

After assigning the beetles to the four groups, the insecticide is applied (uniformly to all experimental units or beetles), and the individual survival times recorded. A natural analysis of the data collected from this **one factor** design would be to compare the survival times using a one-way ANOVA.

There are several important controls that should be built into this experiment. The same strain of beetles should be used to ensure that the four treatment groups are alike as possible, so that differences in survival times are attributable to the insecticides, and not due to genetic differences among beetles. Other factors that may influence the survival time, say the concentration of the insecticide or the age of the beetles, would be held constant, or fixed by the experimenter, if possible. Thus, the same concentration would be used with the four insecticides.

In complex experiments, there are always potential influences that are not realized or are thought to be unimportant that you do not or can not control. The **randomization** of beetles to groups ensures that there is no systematic dependence of the observed treatment differences on the uncontrolled influences. This is extremely important in studies where genetic and environmental influences can not be easily controlled (as in humans, more so than in bugs or mice). The randomization of beetles to insecticides tends to diffuse or greatly reduce the effect of the uncontrolled influences on the comparison of insecticides, in the sense that these effects become part of the uncontrolled or error variation of the experiment.

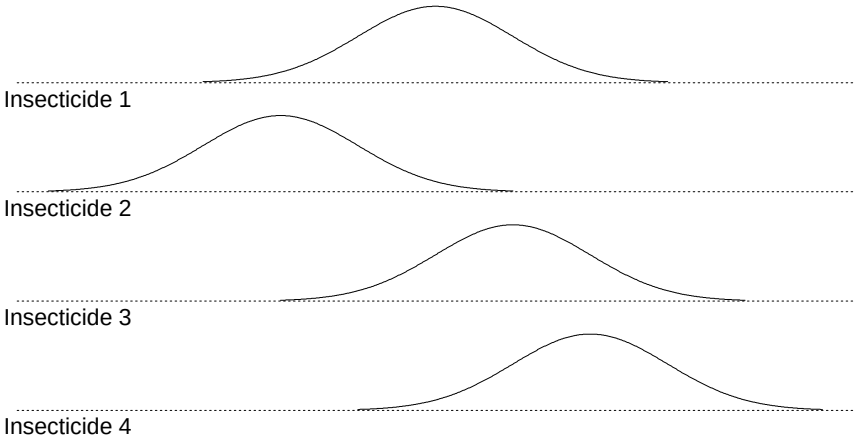
In summary, an **experiment** is to impose a treatment on experimental units to observe a response. Randomization and carefully controlling factors are important considerations.

Suppose y_{ij} is the response for the j^{th} experimental unit in the i^{th} treatment group, where $i = 1, 2, \ldots, I$. The statistical model for a **completely randomized one-factor design** that leads to a one-way ANOVA is given by:

$$y_{ij} = \mu_i + e_{ij},$$

where μ_i is the (unknown) population mean for all potential responses to the i^{th} treatment, and e_{ij} is the residual or deviation of the response from the population mean. The responses within and across treatments are assumed to be independent, normal random variables with constant variance.

For the insecticide experiment, y_{ij} is the survival time for the j^{th} beetle given the i^{th} insecticide, where $i = 1, 2, 3, 4$ and $j = 1, 2, \ldots, 12$. The random selection of beetles coupled with the randomization of beetles to groups ensures the independence assumptions. The assumed population distributions of responses for the $I = 4$ insecticides can be represented as follows:



Let

$$\mu = \frac{1}{I}\sum_i \mu_i$$

be the grand mean, or average of the population means. Let

$$\alpha_i = \mu_i - \mu$$

be the i^{th} group **treatment effect**. The treatment effects are constrained to add to zero, $\alpha_1 + \alpha_2 + \cdots + \alpha_I = 0$, and measure the difference between the treatment population means and the grand mean. Given this notation, the one-way ANOVA model is

$$y_{ij} = \mu + \alpha_i + e_{ij}.$$

The model specifies that the

$$\text{Response} = \text{Grand Mean} + \text{Treatment Effect} + \text{Residual}.$$

An hypothesis of interest is whether the population means are equal: $H_0 : \mu_1 = \cdots = \mu_I$, which is equivalent to the hypothesis of no treatment effects: $H_0 : \alpha_1 = \cdots = \alpha_I = 0$. If H_0 is true, then the one-way model is

$$y_{ij} = \mu + e_{ij},$$

where μ is the common population mean. You know how to test H_0 and do multiple comparisons of the treatments, so I will not review this material.

Most texts use treatment effects to specify ANOVA models, a convention that I will also follow. A difficulty with this approach is that the treatment effects must be constrained to be uniquely estimable from the data (because the I population means μ_i are modeled in terms of $I+1$ parameters: $\mu_i = \mu + \alpha_i$). An infinite number of constraints can be considered each of which gives the same structure on the population means. The standard constraint where the treatment effects sum to zero was used above, but many statistical packages, including **SAS**, impose the constraint $\alpha_I = 0$ (or sometimes $\alpha_1 = 0$). Although estimates of treatment effects depend on which constraint is chosen, the null and alternative models used with the ANOVA F-test, and pairwise comparisons of treatment effects, do not. I will downplay the discussion of estimating treatment effects to minimize problems.

Chapter 5

Paired Experiments and Randomized Block Experiments

A **randomized block design** is often used instead of a completely randomized design in studies where there is extraneous variation among the experimental units that may influence the response. A significant amount of the extraneous variation may be removed from the comparison of treatments by partitioning the experimental units into fairly **homogeneous subgroups** or **blocks**.

For example, suppose you are interested in comparing the effectiveness of four antibiotics for a bacterial infection. The recovery time after administering an antibiotic may be influenced by the patients general health, the extent of their infection, or their age. Randomly allocating experimental subjects to the treatments (and then comparing them using a one-way ANOVA) may produce one treatment having a "favorable" sample of patients with features that naturally lead to a speedy recovery. Alternatively, if the characteristics that affect the recovery time are spread across treatments, then the variation within samples due to these uncontrolled features can dominate the effects of the treatment, leading to an inconclusive result.

A better way to design this experiment would be to **block** the subjects into groups of four patients who are alike as possible on factors other than the treatment that influence the recovery time. The four treatments are then randomly assigned to the patients (one per patient) within a block, and the recovery time measured. The blocking of patients usually produces a more sensitive comparison of treatments than does a completely randomized design because the variation in recovery times due to the blocks is eliminated from the comparison of treatments.

A randomized block design is a **paired experiment** when two treatments are compared. The usual analysis for a paired experiment is a parametric or non-parametric paired comparison.

Randomized block (RB) designs were developed to account for soil fertility gradients in agricultural experiments. The experimental field would be separated into strips (blocks) of fairly constant fertility. Each strip is partitioned into equal size plots. The treatments, say varieties of corn, are randomly assigned to the plots, with each treatment occurring the same number of times (usually once) per block. All other factors that are known to influence the response would be controlled or fixed by the experimenter. For example, when comparing the mean yields, each plot would receive the same type and amount of fertilizer and the same irrigation plan.

The discussion will be limited to randomized block experiments with one factor. Two or more factors can be used with a randomized block design. For example, the agricultural experiment could be modified to compare four combinations of two corn varieties and two levels of fertilizer in each

block instead of the original four varieties. In certain experiments, each experimental unit receives each treatment. The experimental units are "natural" blocks for the analysis.

Example: Comparison of Treatments for Itching Ten[1] male volunteers between 20 and 30 years old were used as a study group to compare seven treatments (5 drugs, a placebo, and no drug) to relieve itching. Each subject was given a different treatment on seven study days. The time ordering of the treatments was randomized across days. Except on the no-drug day, the subjects were given the treatment intravenously, and then itching was induced on their forearms using an effective itch stimulus called cowage. The subjects recorded the duration of itching, in seconds. The data are given in the table below. From left to right the drugs are: papaverine, morphine, aminophylline, pentobarbitol, tripelenamine.

The volunteers in the study were treated as blocks in the analysis. At best, the volunteers might be considered a representative sample of males between the ages of 20 and 30. This limits the extent of inferences from the experiment. The scientists can not, without sound medical justification, extrapolate the results to children or to senior citizens.

Patient	Nodrug	Placebo	Papv	Morp	Amino	Pento	Tripel
1	174	263	105	199	141	108	141
2	224	213	103	143	168	341	184
3	260	231	145	113	78	159	125
4	255	291	103	225	164	135	227
5	165	168	144	176	127	239	194
6	237	121	94	144	114	136	155
7	191	137	35	87	96	140	121
8	100	102	133	120	222	134	129
9	115	89	83	100	165	185	79
10	189	433	237	173	168	188	317

5.1 Analysis of a Randomized Block Design

Assume that you designed a randomized block experiment with I blocks and J treatments, where each treatment occurs once in each block. Let y_{ij} be the response for the j^{th} treatment within the i^{th} block. The model for the experiment is

$$y_{ij} = \mu_{ij} + e_{ij},$$

where μ_{ij} is the population mean response for the j^{th} treatment in the i^{th} block and e_{ij} is the deviation of the response from the mean. The population means are assumed to satisfy the additive model

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$

where μ is a grand mean, α_i is the effect for the i^{th} block, and β_j is the effect for the j^{th} treatment. The responses are assumed to be independent across blocks, normally distributed and with constant variance. The randomized block model does not require the observations within a block to be

[1] Beecher, 1959

independent, but does assume that the correlation between responses within a block is identical for each pair of treatments.

The model is sometimes written as

$$\text{Response} = \text{Grand Mean} + \text{Treatment Effect} + \text{Block Effect} + \text{Residual}.$$

Given the data, let $\bar{y}_{i\cdot}$ be the i^{th} block sample mean (the average of the responses in the i^{th} block), $\bar{y}_{\cdot j}$ be the j^{th} treatment sample mean (the average of the responses on the j^{th} treatment), and $\bar{y}_{\cdot\cdot}$ be the average response of all IJ observations in the experiment.

An ANOVA table for the randomized block experiment partitions the Model SS into SS for Blocks and Treatments.

Source	*df*	SS	MS
Blocks	$I-1$	$J\sum_i(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2$	
Treats	$J-1$	$I\sum_j(\bar{y}_{\cdot j}-\bar{y}_{\cdot\cdot})^2$	
Error	$(I-1)(J-1)$	$\sum_{ij}(\bar{y}_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y}_{\cdot\cdot})^2$	
Total	$IJ-1$	$\sum_{ij}(y_{ij}-\bar{y}_{\cdot\cdot})^2$.	

A primary interest is testing whether the treatment effects are zero: $H_0 : \beta_1 = \cdots = \beta_J = 0$. The treatment effects are zero if in each block the population mean responses are identical for each treatment. A formal test of no treatment effects is based on the p-value from the F-statistic F_{obs} = MS Treat/MS Error. The p-value is evaluated in the usual way (i.e., as an upper tail area from an F-distribution with $J-1$ and $(I-1)(J-1)$ df.) This H_0 is rejected when the treatment averages $\bar{y}_{\cdot j}$ vary significantly relative to the error variation.

A test for no block effects ($H_0 : \alpha_1 = \cdots = \alpha_I = 0$) is often a secondary interest, because, if the experiment is designed well, the blocks will be, by construction, noticeably different. There are no block effects if the population mean response for an arbitrary treatment is identical across blocks. A formal test of no block effects is based on the p-value from the the F-statistic F_{obs} = MS Blocks/MS Error. This H_0 is rejected when the block averages $\bar{y}_{i\cdot}$ vary significantly relative to the error variation.

The randomized block model is easily fitted in the **glm** procedure. Before illustrating the analysis on the itching data, let me mention five important points about randomized block analyses:

1. The F-test p-value for comparing $J = 2$ treatments is identical to the p-value for comparing the two treatments using a paired t-test.

2. The Block SS plus the Error SS is the Error SS from a one-way ANOVA comparing the J treatments. If the Block SS is large relative to the Error SS from the two-factor model, then the experimenter has eliminated a substantial portion of the variation that is used to assess the differences among the treatments. This leads to a more sensitive comparison of treatments than would have been obtained using a one-way ANOVA.

3. The RB model is equivalent to an additive or no interaction model for a two-factor experiment, where the blocks are levels of one of the factors. The analysis of a randomized block experiment under this model is the same analysis used for a two-factor experiment with no replication (one observation per cell). We will discuss the two-factor design soon.

4. Under the sum constraint on the parameters (i.e., $\sum_i \alpha_i = \sum_j \beta_j = 0$), the estimates of the grand mean, block effects, and treatment effects are $\hat{\mu} = \bar{y}_{..}$, $\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$, and $\hat{\beta}_j = \bar{y}_{.j} - \bar{y}_{..}$, respectively. The estimated mean response for the $(i,j)^{th}$ cell is $\hat{\mu}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j = \bar{y}_{i.} + \bar{y}_{.j} - \bar{y}_{..}$.

5. The F-test for comparing treatments is appropriate when the responses within a block have the same correlation. This is a reasonable working assumption in many analyses. A multivariate repeated measures model can be used to compare treatments when the constant correlation assumption is unrealistic, for example when the same treatment is given to an individual over time.

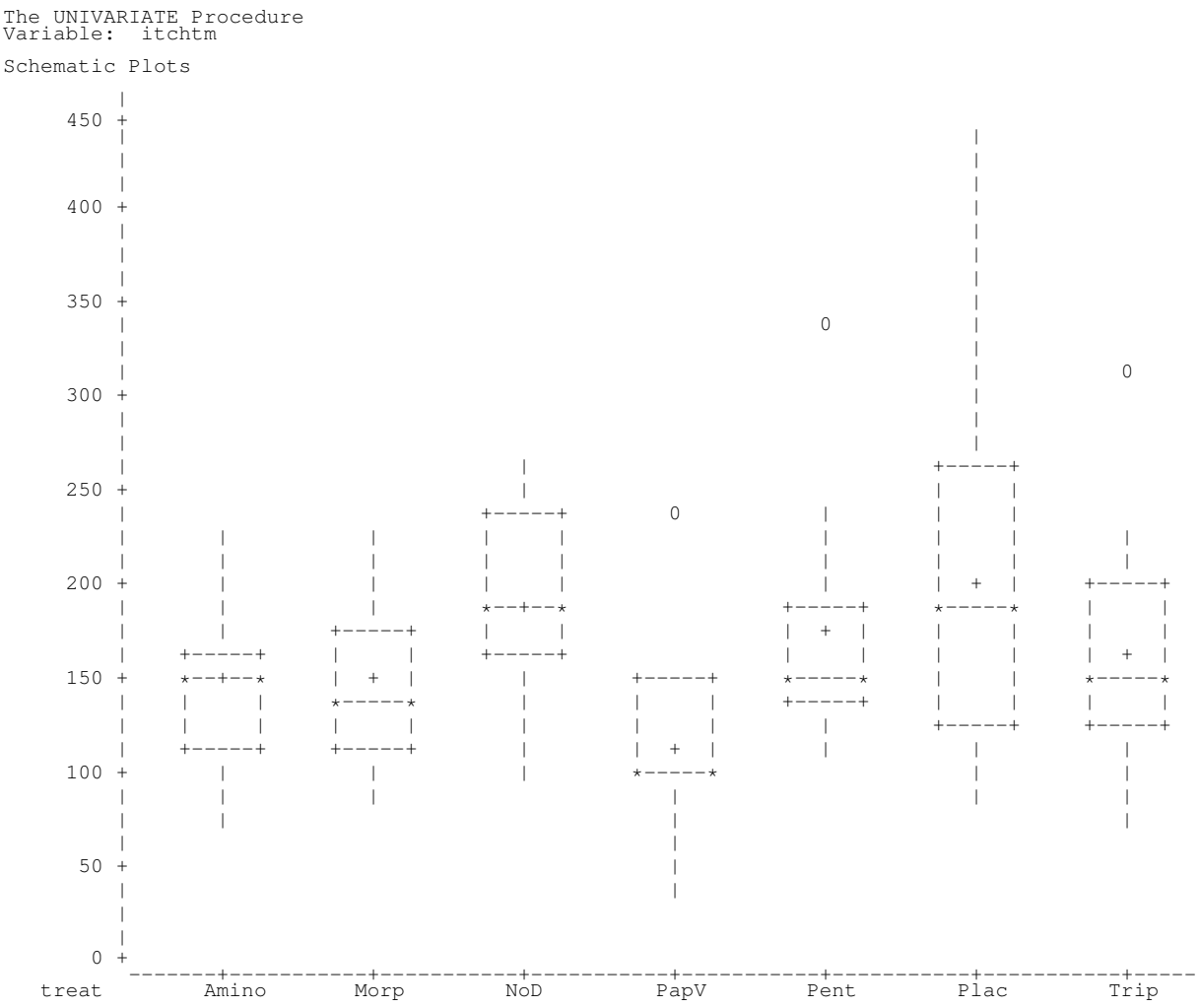
RB Analysis of the Itching Data A **SAS** program for analyzing the itching data is given below. As a first step, I made side-by-side boxplots of the itching durations across treatments. The boxplots are helpful for **informally** comparing treatments and visualizing the data. The differences in the level of the boxplots will usually be magnified by the F-test for comparing treatments because the variability within the boxplots includes block differences which are eliminated from the Error SS.

Program editor contents:

```
* itching data;
options ls=79 nodate nocenter;

data itch;
  infile "F:\Dropbox\UNM\teach\ADA2_stat528\sas\ADA2_05_itch.dat";
  input person treat $ itchtm;   * treatments are strings $ ;
  obs = _n_;  * for use later in plotting in glm;
run;

proc sort data=itch;
  by treat;
run;

proc univariate data=itch plot;
  var itchtm;
  by treat;
run;
```

Each of the five drugs appears to have an effect, compared to the placebo and to no drug. Papaverine appears to be the most effective drug. The placebo and no drug have similar medians. The relatively large spread in the placebo group suggests that some patients responded adversely to the placebo, whereas others responded positively.

Output window contents:



To fit the RB model in **glm**, you need to specify blocks (**person**) and treatments (**treat**) as **class** variables, and include each to the right of the equal sign on the **model** statement. The response variable **itchm** appears to the left of the equal sign. The **means** statement gives block and treatment means, with Bonferroni multiple comparisons.

The initial ANOVA table combines the Block and Treatment SS together to give the Model SS. The F-test in the overall ANOVA table tests for no block effects and no treatment effects. If there are no block effects and no tretment effects then the mean itching time is independent of treatment and individuals. The p-value of 0.0005 strongly suggests that the population mean itching times are not all equal.

The next summary gives two partitionings of the Model SS into the SS for Blocks and Treatments. The Mean Squares, F-statistics, and p-values for testing these effects are given. For a RB design with the same number of responses per block (i.e., no missing data), the Type I and Type III SS are identical, and correspond to the formulas given earlier. The distinction between Type I and Type III SS is important for unbalanced problems, an issue we discuss later.

The F-tests show significant differences among the treatments (p-value=0.017) and among patients (p-value=0.001). The Bonferroni comparisons suggest that papaverine induces a lower mean itching time than placebo. All the other comparisons of treatments are insignificant. The comparison of blocks is of less interest.

Program editor contents:

```
proc glm data=itch;
  class treat person;           * categorical treatment and person (blocks);
  model itchtm = treat person;  * itch time related to treatment and block;
  means treat person /bon;      * give treatment and block means ;
                                * with bonferroni multiple comparisons;
  output out=outglm student=sres predicted=pred cookd=cookd; * for diagnostics below;
run;
```

Output window contents:

```
The GLM Procedure

                Class Level Information
Class         Levels    Values
treat              7    Amino Morp NoD PapV Pent Plac Trip
person            10    1 2 3 4 5 6 7 8 9 10

Number of Observations Read          70
Number of Observations Used          70

Dependent Variable: itchtm

                                  Sum of
Source                     DF        Squares    Mean Square   F Value   Pr > F
Model                      15    156292.6000     10419.5067      3.37   0.0005
Error                      54    167129.6857      3094.9942
Corrected Total            69    323422.2857

R-Square     Coeff Var      Root MSE    itchtm Mean
0.483246      33.86337      55.63267       164.2857

Source                     DF      Type I SS    Mean Square   F Value   Pr > F
treat                       6     53012.8857      8835.4810      2.85   0.0173
person                      9    103279.7143     11475.5238      3.71   0.0011

Source                     DF    Type III SS    Mean Square   F Value   Pr > F
treat                       6     53012.8857      8835.4810      2.85   0.0173
person                      9    103279.7143     11475.5238      3.71   0.0011
```

Treatment multiple comparisons:

```
Bonferroni (Dunn) t Tests for itchtm

NOTE: This test controls the Type I experimentwise error rate, but it
generally has a higher Type II error rate than REGWQ.

Alpha                              0.05
Error Degrees of Freedom             54
Error Mean Square               3094.994
Critical Value of t              3.18843
Minimum Significant Difference    79.327

Means with the same letter are not significantly different.

 Bon
Groupi
  ng           Mean      N    treat
     A        204.80     10    Plac
B    A        191.00     10    NoD
B    A        176.50     10    Pent
B    A        167.20     10    Trip
B    A        148.00     10    Morp
B    A        144.30     10    Amino
B             118.20     10    PapV
```

Person multiple comparisons:

```
Bonferroni (Dunn) t Tests for itchtm

NOTE: This test controls the Type I experimentwise error rate, but it
generally has a higher Type II error rate than REGWQ.

Alpha                                      0.05
Error Degrees of Freedom                     54
Error Mean Square                      3094.994
Critical Value of t                     3.44530
Minimum Significant Difference           102.45

Means with the same letter are not significantly different.

 Bon
Groupi
  ng            Mean       N    person
     A         243.57       7    10
B    A         200.00       7    4
B    A         196.57       7    2
B    A         173.29       7    5
B    A         161.57       7    1
B    A         158.71       7    3
B    A         143.00       7    6
B              134.29       7    8
B              116.57       7    9
B              115.29       7    7
```

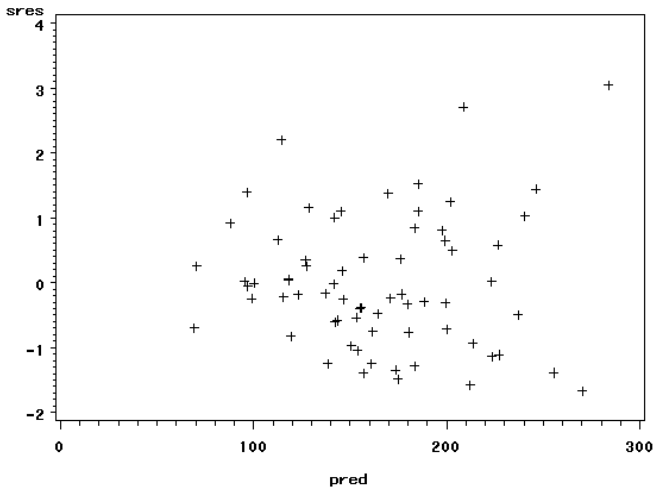
Diagnostic Analysis for the RB Analysis A diagnostic analysis of ANOVA-type models, including the RB model, is easily performed using **glm** output. As with the **reg** procedure, standardized residuals and Cook's distances can be saved in an output data set and then plotted. Unlike the **reg** procedure, **glm** does not generate high resolution diagnostic plots, nor does **glm** print the studentized residuals or distances as an option on the **model** statement.

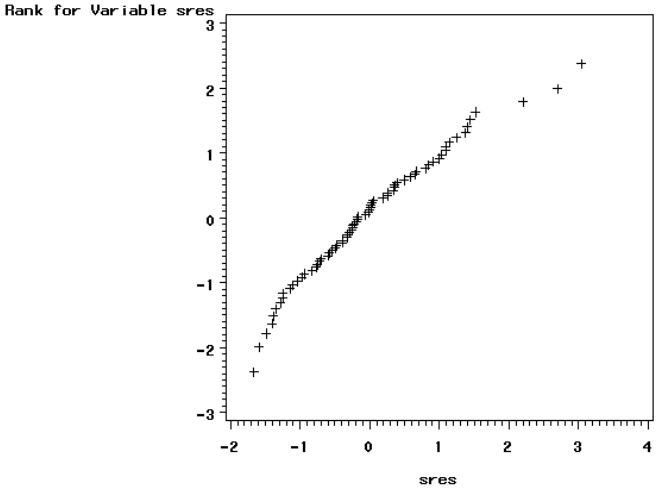
The program above did the RB analysis then saved and plots the studentized residuals and Cook's distances. In the initial **data step** I created an observation number variable to use when plotting the distances.

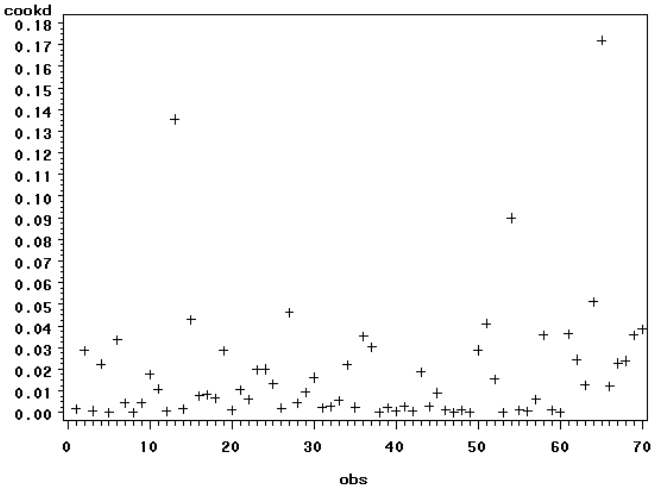
The rankit plot (normal quantile or QQ-plot) shows the residual distribution is slightly skewed to the right, in part, due to three cases that are not fitted well by the model (the outliers in the boxplots). Except for these cases, which are also the most influential cases, the plot of the studentized residuals against predicted values shows no gross abnormalities.

Program editor contents:

```
* diagnostics;
proc rank data=outglm out=outglm normal=blom;
  var sres;
  ranks rankstr;
run;

proc gplot data=outglm;
  plot rankstr*sres sres*pred cookd*obs;
run;
```







Although the F-test for comparing treatments is not overly sensitive to modest deviations from normality, I will present a non-parametric analysis as a backup, to see whether similar conclusions are reached about the treatments.

Non-parametric Analysis of a RB Experiment Milton Friedman developed a non-parametric test for comparing treatments in a randomized block design. Friedman's test can be approximated in **SAS** by comparing treatments in a one-way ANOVA of the **ranked responses** within each block. The block effect is not needed in the model because the Block SS is zero on the ranked data (the block means are equal).

The following **SAS** program was appended to the data step for the itching experiment to approximate Friedman's test. The output suggests significant differences among treatments, which supports the earlier conclusion.

Program editor contents:

```
* itching nonparametric analysis;
proc sort data=itch;
```

```
  by person;
run;

proc rank data=itch out=itch;
  var itchtm;
  ranks rankitch;
  by person;
run;

proc glm data=itch;
  class treat;
  model rankitch = treat;
  means treat /bon;
run;
```

Output window contents:

```
The GLM Procedure

                 Class Level Information
Class         Levels    Values
treat              7    Amino Morp NoD PapV Pent Plac Trip

Number of Observations Read          70
Number of Observations Used          70

Dependent Variable: rankitch   Rank for Variable itchtm
                                        Sum of
Source                     DF         Squares    Mean Square   F Value   Pr > F
Model                       6     69.3500000     11.5583333      3.47   0.0051
Error                      63    210.1500000      3.3357143
Corrected Total            69    279.5000000

R-Square     Coeff Var      Root MSE    rankitch Mean
0.248122      45.65984      1.826394          4.000000

Source                     DF      Type I SS    Mean Square   F Value   Pr > F
treat                       6    69.35000000    11.55833333      3.47   0.0051

Source                     DF    Type III SS    Mean Square   F Value   Pr > F
treat                       6    69.35000000    11.55833333      3.47   0.0051

Bonferroni (Dunn) t Tests for rankitch

NOTE: This test controls the Type I experimentwise error rate, but it
generally has a higher Type II error rate than REGWQ.

Alpha                              0.05
Error Degrees of Freedom             63
Error Mean Square              3.335714
Critical Value of t             3.16614
Minimum Significant Difference   2.5861

Means with the same letter are not significantly different.

 Bon
Grouping        Mean      N    treat
     A        5.1000     10    NoD
     A        4.9000     10    Plac
     A        4.9000     10    Pent
B    A        4.2500     10    Trip
B    A        3.5000     10    Morp
B    A        3.0500     10    Amino
B             2.3000     10    PapV
```

5.2 Extending the One-Factor Design to Multiple Factors

The CRD (completely randomized design) for comparing insecticides varies the levels of one factor (insecticide), while controlling other factors that influence survival time. The inferences from the

one-way ANOVA apply to beetles with a given age from the selected strain that might be given the selected concentration of the insecticides. Any generalization of the conclusions to other situations must be justified scientifically, typically through further experimentation.

There are several ways to broaden the scope of the study. For example, several strains of beetles or several concentrations of the insecticide might be used. For simplicity, consider a simple two-factor experiment where three concentrations (Low, Medium, and High) are applied with each of the four insecticides. This is a completely crossed **two-factor experiment** where each of the $4 \times 3 = 12$ combinations of the two factors (insecticide and dose) are included in the comparison of survival times. With this experiment, the scientist can compare insecticides, compare concentrations, and check for an interaction between dose and insecticide.

Assuming that 48 beetles are available, the scientist would randomly assign them to the 12 experimental groups, giving prespecified numbers of beetles to the 12 groups. For simplicity, assume that the experiment is **balanced**, that is, the same number of beetles (4) is assigned to each group ($12 \times 4 = 48$). This is a CRD with two factors.

5.2.1 Example: Beetle insecticide two-factor design

The data below were collected using the experimental design just described. The table gives survival times of groups of four beetles randomly allocated to twelve treatment groups obtained by crossing the levels of four insecticides (A, B, C, D) at each of three concentrations of the insecticides (1=Low, 2=Medium, 3=High). This is a balanced 4-by-3 **factorial** design (two-factor design) that is replicated four times. The unit of measure for the survival times is 10 hours, that is, 0.3 is a survival time of 3 hours.

Dose	Insecticide	Survival Times			
1	A	0.31	0.45	0.46	0.43
1	B	0.82	1.10	0.88	0.72
1	C	0.43	0.45	0.63	0.76
1	D	0.45	0.71	0.66	0.62
2	A	0.36	0.29	0.40	0.23
2	B	0.92	0.61	0.49	1.24
2	C	0.44	0.35	0.31	0.40
2	D	0.56	1.02	0.71	0.38
3	A	0.22	0.21	0.18	0.23
3	B	0.30	0.37	0.38	0.29
3	C	0.23	0.25	0.24	0.22
3	D	0.30	0.36	0.31	0.33

The basic unit of analysis is the **cell means**, which are the averages of the 4 observations in each of the 12 treatment combinations. For example, in the table below, the sample mean survival for the 4 beetles given a low dose (dose=1) of insecticide A is 0.413. From the cell means we obtain the dose and insecticide **marginal means** by averaging over the levels of the other factor. For example, the marginal mean for insecticide A is the average of the cell means for the 3 treatment combinations involving insecticide A: $0.314 = (0.413 + 0.320 + 0.210)/3$.

Cell Means	Dose			
Insecticide	1	2	3	Insect marg
A	0.413	0.320	0.210	0.314
B	0.880	0.815	0.335	0.677
C	0.568	0.375	0.235	0.393
D	0.610	0.668	0.325	0.534
Dose marg	0.618	0.544	0.277	0.480

Because the experiment is balanced, a marginal mean is the average of all observations that receive a given treatment. For example, the marginal mean for insecticide A is the average survival time for the 16 beetles given insecticide A.

Looking at the table of means, the insecticides have noticeably different mean survival times averaged over doses, with insecticide A having the lowest mean survival time averaged over doses. Similarly, higher doses tend to produce lower survival times. A more formal approach to analyzing the table of means is given in the next section.

5.2.2 The Interaction Model for a Two-Factor Experiment

Assume that you designed a **balanced** two-factor experiment with K responses at each combination of the I levels of factor 1 (F1) with the J levels of factor 2 (F2). The total number of responses is KIJ, or K times the IJ treatment combinations.

Let y_{ijk} be the k^{th} response at the i^{th} level of F1 and the j^{th} level of F2. A generic model for the experiment expresses y_{ijk} as a mean response plus a residual:

$$y_{ijk} = \mu_{ij} + e_{ijk},$$

where μ_{ij} is the population mean response for the treatment defined by the i^{th} level of F1 combined with the j^{th} level of F2. As in a one-way ANOVA, the responses within and across treatment groups are assumed to be independent, normally distributed, and have constant variance.

The **interaction model** expresses the population means as

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij},$$

where μ is a grand mean, α_i is the effect for the i^{th} level of F1, β_j is the effect for the j^{th} level of F2, and $(\alpha\beta)_{ij}$ is the interaction between the i^{th} level of F1 and the j^{th} level of F2. (Note that $(\alpha\beta)$ is an individual term distinct from α and β, $(\alpha\beta)$ is not their product.) The model is often written

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk},$$

meaning

$$\text{Response} = \text{Grand Mean} + \text{F1 effect} + \text{F2 effect} + \text{F1-by-F2 interaction} + \text{Residual}.$$

The model without the interaction term is $y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$, meaning

$$\text{Response} = \text{Grand Mean} + \text{F1 effect} + \text{F2 effect} + \text{Residual},$$

is called the **additive model** because the effects of F1 and F2 on the mean are additive.

Defining effects from cell means

The effects that define the population means and the usual hypotheses of interest can be formulated from the table of population means, given here.

	Level of F2				
Level of F1	1	2	$\cdots$	J	F1 marg
1	μ_{11}	μ_{12}	$\cdots$	μ_{1J}	$\bar{\mu}_{1\cdot}$
2	μ_{21}	μ_{22}	$\cdots$	μ_{2J}	$\bar{\mu}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	μ_{I1}	μ_{I2}	$\cdots$	μ_{IJ}	$\bar{\mu}_{I\cdot}$
F2 marg	$\bar{\mu}_{\cdot 1}$	$\bar{\mu}_{\cdot 2}$	$\cdots$	$\bar{\mu}_{\cdot J}$	$\bar{\mu}_{\cdot\cdot}$

The F1 **marginal population means** are averages within rows (over columns) of the table:

$$\bar{\mu}_{i\cdot} = \frac{1}{J}\sum_{c} \mu_{ic}.$$

The F2 marginal population means are averages within columns (over rows):

$$\bar{\mu}_{\cdot j} = \frac{1}{I}\sum_{r} \mu_{rj}.$$

The overall or **grand population mean** is the average of the cell means

$$\bar{\mu}_{\cdot\cdot} = \frac{1}{IJ}\sum_{rc} \mu_{rc} = \frac{1}{I}\sum_{i} \bar{\mu}_{i\cdot} = \frac{1}{J}\sum_{j} \bar{\mu}_{\cdot j}.$$

Using this notation, the effects in the interaction model are $\mu = \bar{\mu}_{\cdot\cdot}$, $\alpha_i = \bar{\mu}_{i\cdot} - \bar{\mu}_{\cdot\cdot}$, $\beta_j = \bar{\mu}_{\cdot j} - \bar{\mu}_{\cdot\cdot}$, and $(\alpha\beta)_{ij} = \mu_{ij} - \bar{\mu}_{i\cdot}) - \bar{\mu}_{\cdot j} + \bar{\mu}_{\cdot\cdot}$. The effects sum to zero:

$$\sum_{i} \alpha_i = \sum_{j} \beta_j = \sum_{ij} (\alpha\beta)_{ij} = 0,$$

and satisfy $\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$ (i.e., cell mean is sum of effects) required under the model.

The F1 and F2 effects are analogous to treatment effects in a one-factor experiment, except that here the treatment means are averaged over the levels of the other factor. The interaction effect will be interpreted later.

Estimating effects from the data

Let

$$\bar{y}_{ij} = \frac{1}{K}\sum_{k} y_{ijk} \qquad \text{and} \qquad s^2_{ij}$$

be the sample mean and variance, respectively, for the K responses at the i^{th} level of F1 and the j^{th} level of F2. Inferences about the population means are based on the table of sample means:

	Level of F2				
Level of F1	1	2	$\cdots$	J	F1 marg
1	$\bar{y}_{11}$	$\bar{y}_{12}$	$\cdots$	$\bar{y}_{1J}$	$\bar{y}_{1\cdot}$
2	$\bar{y}_{21}$	$\bar{y}_{22}$	$\cdots$	$\bar{y}_{2J}$	$\bar{y}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	$\bar{y}_{I1}$	$\bar{y}_{I2}$	$\cdots$	$\bar{y}_{IJ}$	$\bar{y}_{I\cdot}$
F2 marg	$\bar{y}_{\cdot 1}$	$\bar{y}_{\cdot 2}$	$\cdots$	$\bar{y}_{\cdot J}$	$\bar{y}_{\cdot\cdot}$

The F1 **marginal sample means** are averages within rows of the table:

$$\bar{y}_{i\cdot} = \frac{1}{J}\sum_{c} \bar{y}_{ic\cdot}$$

The F2 marginal sample means are averages within columns:

$$\bar{y}_{\cdot j} = \frac{1}{I}\sum_{r} \bar{y}_{rj\cdot}$$

Finally, $\bar{y}_{\cdot\cdot}$ is the average of the cell sample means:

$$\bar{y}_{\cdot\cdot} = \frac{1}{IJ}\sum_{ij} \bar{y}_{ij} = \frac{1}{I}\sum_{i} \bar{y}_{i\cdot} = \frac{1}{J}\sum_{j} \bar{y}_{\cdot j}.$$

The sample sizes in each of the IJ treatment groups are equal (K), so $\bar{y}_{i\cdot}$ is the sample average of all responses at the i^{th} level of F1, $\bar{y}_{\cdot j}$ is the sample average of all responses at the j^{th} level of F2, and $\bar{y}_{\cdot\cdot}$ is the average response in the experiment.

Under the interaction model, the estimated population mean for the $(i,j)^{th}$ cell is the observed cell mean: $\hat{\mu}_{ij} = \bar{y}_{ij}$. This can be partitioned into estimated effects

$$\begin{aligned}
\hat{\mu} &= \bar{y}_{\cdot\cdot} \quad \text{the estimated grand mean} \\
\hat{\alpha}_i &= \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot} \quad \text{the estimated F1 effect } i = 1, 2, ..., I \\
\hat{\beta}_j &= \bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot} \quad \text{the estimated F2 effect } j = 1, 2, ..., J \\
\widehat{(\alpha\beta)}_{ij} &= \bar{y}_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot} \quad \text{the estimated cell interaction}
\end{aligned} \tag{5.1}$$

that satisfy

$$\hat{\mu}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \widehat{(\alpha\beta)}_{ij}.$$

The ANOVA table

The ANOVA table for a balanced two-factor design decomposes the total variation in the data, as measured by the Total SS, into components that measure the variation of marginal sample means for F1 and F2 (the F1 SS and F2 SS), a component that measures the degree to which the factors interact (the F1-by-F2 Interaction SS), and a component that pools the sample variances across the IJ samples (the Error SS). Each SS has a df, given in the following ANOVA table. As usual, the MS for each source of variation is the corresponding SS divided by the df. The MS Error estimates the common population variance for the IJ treatments.

Source	*df*	SS	MS
F1	$I-1$	$KJ\sum_i(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2$	MS F1=SS/df
F2	$J-1$	$KI\sum_j(\bar{y}_{\cdot j}-\bar{y}_{\cdot\cdot})^2$	MS F2=SS/df
Interaction	$(I-1)(J-1)$	$K\sum_{ij}(\bar{y}_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y}_{\cdot\cdot})^2$	MS Interact=SS/df
Error	$IJ(K-1)$	$(K-1)\sum_{ij}s_{ij}^2$	MSE=SS/df
Total	$IJK-1$	$\sum_{ijk}(y_{ijk}-\bar{y}_{\cdot\cdot})^2$.	

The standard tests in the two-factor analysis, and their interpretations are:

The test of no F1 effect: $H_0:\alpha_1=\cdots=\alpha_I=0$ is equivalent to testing $H_0:\bar{\mu}_{1\cdot}=\bar{\mu}_{2\cdot}=\cdots=\bar{\mu}_{I\cdot}$. The absence of an F1 effect implies that each level of F1 has the same population mean response **when the means are averaged over levels of** F2. The test for no F1 effect is based on MS F1/MS Error, which is compared to the upper tail of an F-distribution with numerator and denominator df of $I-1$ and $IJ(K-1)$, respectively. H_0 is rejected when the F1 marginal means $\bar{y}_{i\cdot}$ vary significantly relative to the within sample variation. Equivalently, H_0 is rejected when the sum of squared F1 effects (between sample variation) is large relative to the within sample variation.

The test of no F2 effect: $H_0:\beta_1=\cdots=\beta_J=0$ is equivalent to testing $H_0:\bar{\mu}_{\cdot 1}=\bar{\mu}_{\cdot 2}=\cdots=\bar{\mu}_{\cdot J}$. The absence of a F2 effect implies that each level of F2 has the same population mean response **when the means are averaged over levels of** F1. The test for no F2 effect is based on MS F2/MS Error, which is compared to an F-distribution with numerator and denominator df of $J-1$ and $IJ(K-1)$, respectively. H_0 is rejected when the F2 marginal means $\bar{y}_{\cdot j}$ vary significantly relative to the within sample variation. Equivalently, H_0 is rejected when the sum of squared F2 effects (between sample variation) is large relative to the within sample variation.

The test of no interaction: $H_0:(\alpha\beta)_{ij}=0$ for all i and j is based on MS Interact/MS Error, which is compared to an F-distribution with numerator and denominator df of $(I-1)(J-1)$ and $IJ(K-1)$, respectively.

The interaction model places no restrictions on the populations means μ_{ij}. Since the population means can be arbitrary, the interaction model can be viewed as a one factor model with IJ treatments. One connection between the two ways of viewing the two-factor analysis is that the F1, F2, and Interaction SS for the two-way interaction model sum to the Treatment or Model SS for comparing the IJ treatments. The Error SS for the two-way interaction model is identical to the Error SS for a one-way ANOVA of the IJ treatments. **SAS** gives an overall test of no differences in the IJ population means as part of the two-way analysis.

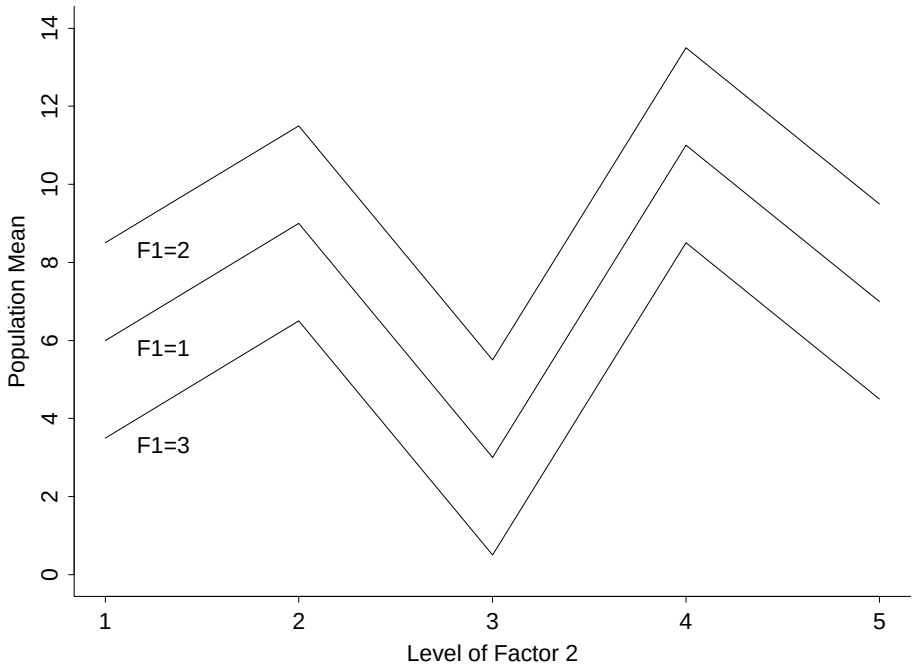
I always summarize the data using the cell and marginal means instead of the estimated effects, primarily because means are the basic building blocks for the analysis. My discussion of the model and tests emphasizes both approaches to help you make the connection with the two ways this material is often presented in texts.

Understanding interaction

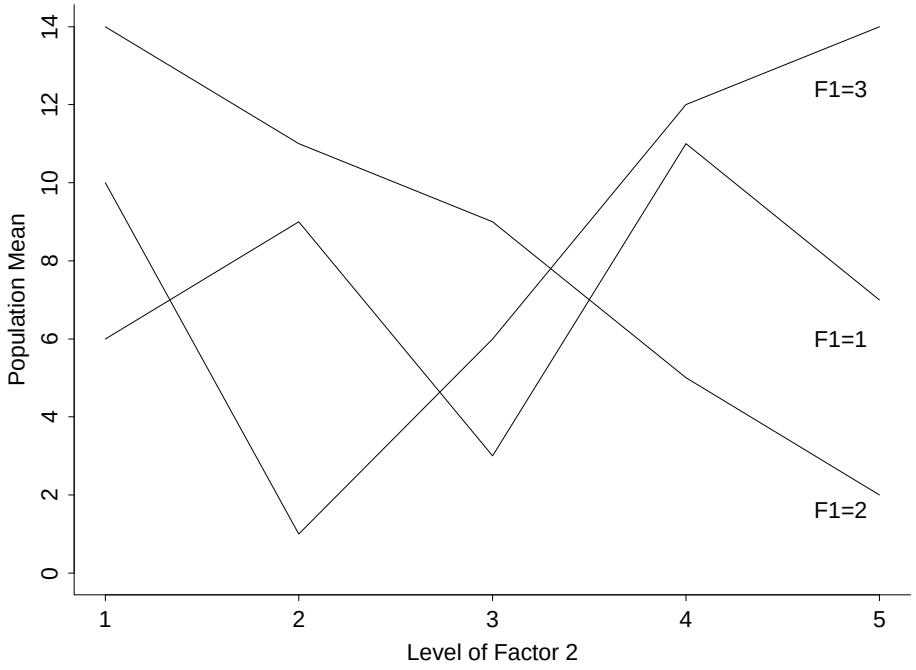
To understand interaction, suppose you (conceptually) plot the means in each row of the population table, giving what is known as the **population mean profile** plot. The F1 marginal population means average the population means within the F1 profiles. At each F2 level, the F2 marginal mean averages the population cell means across F1 profiles.

No interaction. If the plot has perfectly **parallel** F1 profiles, as the plot below for a 3×5 experiment, then the levels of F1 and F2 **do not interact**. That is,

parallel profiles $\Leftrightarrow$ $\mu_{ij} - \mu_{hj}$ is independent of j for each i and h
difference between levels of F1 does not depend on level of F2
$\Leftrightarrow$ $\mu_{ij} - \bar{\mu}_{i\cdot} = \mu_{hj} - \bar{\mu}_{h\cdot}$ for all i, j, h
difference between level of F2 j and F2 mean does not depend on level of F1
$\Leftrightarrow$ $\mu_{ij} - \bar{\mu}_{i\cdot} = \bar{\mu}_{\cdot j} - \bar{\mu}_{\cdot\cdot}$ for all i, j
difference between level of F2 j and F2 mean is the same for all levels of F1
$\Leftrightarrow$ $\mu_{ij} - \bar{\mu}_{i\cdot} - \bar{\mu}_{\cdot j} + \bar{\mu}_{\cdot\cdot} = 0$ for all i, j
interaction effect is 0
$\Leftrightarrow$ $(\alpha\beta)_{ij} = 0$ for all i, j
interaction effect is 0
$\Leftrightarrow$ no interaction term in model.



Interaction. **Interaction** is present if the profiles are not **perfectly parallel**. An example of a profile plot for two-factor experiment (3×5) with interaction is given below.



The roles of F1 and F2 can be reversed in these plots without changing the assessment of a presence or absence of interaction.

A qualitative check for interaction can be based on the **sample means profile plot**, but keep in mind that profiles of sample means are never perfectly parallel even when the factors do not interact in the population. The Interaction SS measures the extent of non-parallelism in the sample mean profiles. In particular, the Interaction SS is zero when the sample mean profiles are perfectly parallel because $\widehat{(\alpha\beta)}_{ij} = 0$ for all i and j.

5.2.3 Example: Survival Times of Beetles

The **SAS** program below generates cell means and a sample means profile plot for the beetle experiment. The **means** procedure was used to obtain the 12 treatment cell means. Three variables were needed to represent each response in the data set: dose (1-3, categorical), insecticide (A-D, categorical), and time (the survival time).

The **sort** procedure was used prior to the **means** procedure to sort the data on the levels of the two **by** variables needed for the **means** analysis. The **noprint** option suppresses the printed output from the **means** procedure. An output data set **meansurv** was created in **means** to contain the 12 cell means, identified by dose and insecticide. A profile plot of the cell means is given after printing **meansurv**. **The interpolation=join connects the like symbols with a straight line** (i.e., B–B–B). The marginal means (generated from later output) are given on the plot.

As noted earlier, the insecticides have noticeably different mean survival times averaged over doses, with insecticide A having the lowest mean. Similarly, higher doses tend to produce lower survival times. The sample means profile plot shows some evidence of interaction.

Program editor contents:

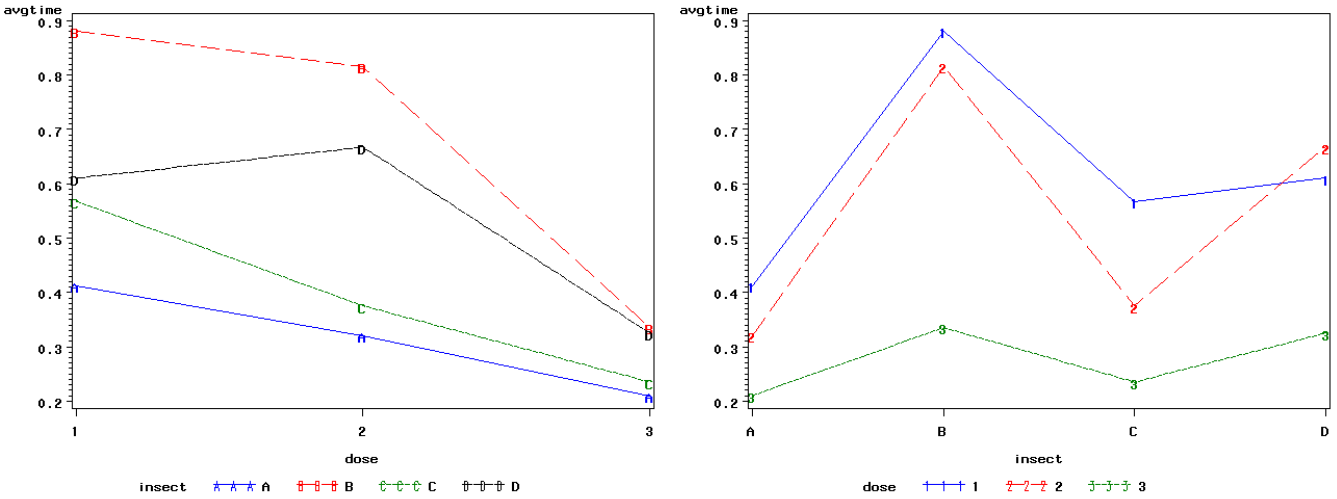
```
* Survival Times of Beetles;
```

```
options ls=79 nodate nocenter;

* input data;
data beetles;
  input dose $ insect $ t1-t4;  * dose and insect are both character data;
  time = t1; output; * output one observation line per time;
  time = t2; output;
  time = t3; output;
  time = t4; output;
  drop t1-t4;        * at end of reading data, remove variables t1, t2, t3, t4;
datalines;
1 A 0.31 0.45 0.46 0.43
1 B 0.82 1.10 0.88 0.72
1 C 0.43 0.45 0.63 0.76
1 D 0.45 0.71 0.66 0.62
2 A 0.36 0.29 0.40 0.23
2 B 0.92 0.61 0.49 1.24
2 C 0.44 0.35 0.31 0.40
2 D 0.56 1.02 0.71 0.38
3 A 0.22 0.21 0.18 0.23
3 B 0.30 0.37 0.38 0.29
3 C 0.23 0.25 0.24 0.22
3 D 0.30 0.36 0.31 0.33
;
run;

* need to sort for the means procedure;
*   note that the data are already in this order                        ;
*   but it is good practice to "sort by" when a following procedure   ;
*   uses a "by" statement                                               ;
proc sort data=beetles;
  by dose insect;
run;

* calculate means and output means to dataset meansurv;
proc means data=beetles noprint;    * do not print means;
  var time;
  by dose insect;
  output out=meansurv mean=avgtime; * output mean times to dataset meansurv;
run;

* print means;
proc print data=meansurv;
run;


* plot means in two profile plots;

* insecticide lines over doses;
* define v=symbol, c=color, i=interpolation, l=line style;
symbol1 v="A" c=blue  i=join l=1;
symbol2 v="B" c=red   i=join l=2;
symbol3 v="C" c=green i=join l=3;
symbol4 v="D" c=black i=join l=4;
proc gplot data=meansurv;
  plot avgtime*dose=insect;
run;

* dose lines over insecticides;
* define v=symbol, c=color, i=interpolation, l=line style;
symbol1 v="1" c=blue  i=join l=1;
symbol2 v="2" c=red   i=join l=2;
symbol3 v="3" c=green i=join l=3;
proc gplot data=meansurv;
  plot avgtime*insect=dose;
run;
```

Output window contents:

Obs	dose	insect	_TYPE_	_FREQ_	avgtime
1	1	A	0	4	0.4125
2	1	B	0	4	0.8800
3	1	C	0	4	0.5675
4	1	D	0	4	0.6100
5	2	A	0	4	0.3200
6	2	B	0	4	0.8150
7	2	C	0	4	0.3750
8	2	D	0	4	0.6675
9	3	A	0	4	0.2100
10	3	B	0	4	0.3350
11	3	C	0	4	0.2350
12	3	D	0	4	0.3250

Profile plots:



The **glm** procedure was used to fit the interaction model. The two factors, dose and insecticide, are identified on the **class** statement. The **model** statement identifies the response variable **time** to the left of the equal size. The dose, insecticide, and interaction effects are specified to the right of the equal sign, with either **dose*insect** or **insect*dose** as valid specifications for the interaction term. The **means** statement gives the marginal means (by specifying **dose insect**) and cell means (by specifying **dose*insect**), with Bonferroni multiple comparisons provided as an option.
Program editor contents:

```
proc glm data=beetles;
  class dose insect;                          * factor variables          ;
  model time = dose insect dose*insect;       * interaction model         ;
  means dose insect dose*insect /bon;         * marginal and cell means   ;
                                              *   with Bonferroni multiple comparisons;
run;
```

The initial ANOVA table tests for no differences among the population mean survival times for the 12 dose and insecticide combinations. The p-value of 0.0001 strongly suggests that the population mean survival times are not all equal.

The next summary gives two partitionings of the one-way ANOVA Treatment SS into the SS for Dose, Insecticide, and the Dose by Insecticide interaction. The Mean Squares, F-statistics and p-values for testing these effects are given. As noted in the RB analysis, the Type I and Type III SS are identical for a balanced design, and correspond to the formulas given earlier.

The p-values for the F-statistics indicate that the dose and insecticide effects are significant at the 0.01 level. The F-test for no dose by insecticide interaction is not significant at the 0.10 level (p-value=0.112). Thus, the interaction seen in the profile plot of the sample means might be due solely to chance or sampling variability.

Output window contents:

```
The GLM Procedure

    Class Level Information
Class         Levels    Values
dose               3    1 2 3
insect             4    A B C D

Number of Observations Read          48
Number of Observations Used          48

Dependent Variable: time
                                   Sum of
Source                     DF       Squares    Mean Square   F Value   Pr > F
Model                      11    2.20435625     0.20039602      9.01   <.0001
Error                      36    0.80072500     0.02224236
Corrected Total            47    3.00508125

R-Square     Coeff Var      Root MSE     time Mean
0.733543      31.11108      0.149139      0.479375

Source                     DF      Type I SS    Mean Square   F Value   Pr > F
dose                        2    1.03301250     0.51650625     23.22   <.0001
insect                      3    0.92120625     0.30706875     13.81   <.0001
dose*insect                 6    0.25013750     0.04168958      1.87   0.1123

Source                     DF    Type III SS    Mean Square   F Value   Pr > F
dose                        2    1.03301250     0.51650625     23.22   <.0001
insect                      3    0.92120625     0.30706875     13.81   <.0001
dose*insect                 6    0.25013750     0.04168958      1.87   0.1123

Bonferroni (Dunn) t Tests for time

NOTE: This test controls the Type I experimentwise error rate, but it
generally has a higher Type II error rate than REGWQ.

Alpha                              0.05
Error Degrees of Freedom             36
Error Mean Square              0.022242
Critical Value of t             2.51104
Minimum Significant Difference   0.1324

Means with the same letter are not significantly different.

Bon
Grouping   Mean      N    dose
A       0.61750     16    1
A       0.54438     16    2
B       0.27625     16    3
```

```
Bonferroni (Dunn) t Tests for time

NOTE: This test controls the Type I experimentwise error rate, but it
generally has a higher Type II error rate than REGWQ.

Alpha                              0.05
Error Degrees of Freedom             36
Error Mean Square              0.022242
Critical Value of t             2.79197
Minimum Significant Difference     0.17

Means with the same letter are not significantly different.

Bon
Grouping        Mean      N    insect
     A       0.67667     12    B
B    A       0.53417     12    D
B    C       0.39250     12    C
     C       0.31417     12    A

Level of     Level of           -------------time------------
dose         insect       N             Mean          Std Dev
1            A            4       0.41250000       0.06946222
1            B            4       0.88000000       0.16083117
1            C            4       0.56750000       0.15671099
1            D            4       0.61000000       0.11284207
2            A            4       0.32000000       0.07527727
2            B            4       0.81500000       0.33630343
2            C            4       0.37500000       0.05686241
2            D            4       0.66750000       0.27097048
3            A            4       0.21000000       0.02160247
3            B            4       0.33500000       0.04654747
3            C            4       0.23500000       0.01290994
3            D            4       0.32500000       0.02645751
```

Interpretation of the Dose and Insecticide Effects

The interpretation of the dose and insecticide **main effects** depends on whether interaction is present. The distinction is important, so I will give both interpretations to emphasize the differences. Given the test for interaction, I would likely summarize the main effects assuming no interaction.

The average survival time decreases as the dose increases, with estimated mean survival times of 0.618, 0.544, and 0.276, respectively. A Bonferroni comparison shows that the population mean survival time for the high dose (averaged over insecticides) is significantly less than the population mean survival times for the low and medium doses (averaged over insecticides). The two lower doses are not significantly different from each other. This leads to two dose groups:

```
Dose:          1=Low  2=Med  3=Hig
Marg Mean:     0.618  0.544  0.276
Groups:        ------------  -----
```

If dose and insecticide **interact**, you can conclude that beetles given a high dose of the insecticide typically survive for shorter periods of time **averaged over insecticides**. You can not, in general, conclude that the highest dose yields the lowest survival time **regardless** of insecticide. For example, the difference in the medium and high dose marginal means (0.544 - 0.276 = 0.268) estimates the typical decrease in survival time achieved by using the high dose instead of the medium dose, averaged over insecticides. If the two factors interact, then the difference in mean times between the medium and high doses on a given insecticide may be significantly greater than 0.268, significantly less than 0.268, or even negative. In the latter case the medium dose would be **better** than the high dose for the given insecticide, even though the high dose gives better performance averaged over insecticides. An interaction forces you to use the cell means to decide which combination of dose and insecticide gives the best results (and the **mean** command does not give multiple comparisons

of cell means). Of course, our profile plot tells us that this hypothetical situation is probably not tenable here, but it could be so when a significant interaction is present.

If dose and insecticide **do not interact**, then the difference in marginal dose means averaged over insecticides also estimates the difference in population mean survival times between two doses, **regardless of the insecticide**. This follows from the parallel profiles definition of no interaction. Thus, the difference in the medium and high dose marginal means (0.544 - 0.276 = 0.268) estimates the expected decrease in survival time anticipated from using the high dose instead of the medium dose, **regardless of the insecticide** (and hence also when averaged over insecticides). A practical implication of no interaction is that you can conclude that the high dose is best, regardless of the insecticide used. The difference in marginal means for two doses estimates the difference in average survival expected, regardless of the insecticide.

An ordering of the mean survival times on the four insecticides (averaged over the three doses) is given below. Three groups are obtained from the Bonferroni comparisons, with any two insecticides separated by one or more other insecticides in the ordered string having significantly different mean survival times averaged over doses.

If interaction is present, you can conclude that insecticide A is no better than C, but significantly better than B or D, when performance is **averaged over doses**. If the interaction is absent, then A is not significantly better than C, but is significantly better than B or D, **regardless of the dose**. Furthermore, for example, the difference in marginal means for insecticides B and A of 0.677 - 0.314 = 0.363 is the expected decrease in survival time from using A instead of B, regardless of dose. This is also the expected decrease in survival times when averaged over doses.

```
Insect:          B      D      C      A
Marg Mean:   0.677  0.534  0.393  0.314
Groups:      ------------
                    ------------
                           ------------
```

5.2.4 Example: Output voltage for batteries

The maximum output voltage for storage batteries is thought to be influenced by the temperature in the location at which the battery is operated and the material used in the plates. A scientist designed a two-factor study to examine this hypothesis, using three temperatures (50, 65, 80), and three materials for the plates (1, 2, 3). Four batteries were tested at each of the 9 combinations of temperature and material type. The maximum output voltage was recorded for each battery. This is a balanced 3-by-3 factorial experiment with four observations per treatment. The data are given with the **SAS** program below. The **SAS** program is similar to the insecticide program.

The two-way ANOVA table indicates that the main effects (temperature and material) and the interaction are significant at the 0.05 level. The p-values for testing no interaction, no material effect, and no temperature effect (individually) are 0.019, 0.002, and 0.0001, respectively. The cell means plots of the material profiles have different slopes, which is consistent with the presence of a temperature-by-material interaction.

The Bonferroni comparisons indicate that the population mean output for the three temperatures **averaged over material types** decreases as the temperature increases:

```
      Temp:    80      65      50
 Marg mean:  64.17  107.58  144.83
     Group: ------  ------  ------
```

Similarly, the population mean output for material 1 is lower than the output for material 3 **averaged over temperatures**:

```
 Material:     1      2      3
Marg mean:  83.17 108.33 125.08
    Group  -------------
           -------------
```

The significant interaction between temperature and material implies that you can not directly conclude that batteries stored at 50 degrees have the highest average output regardless of the material. Nor can you directly conclude that material 3 has a higher average output than material 1 regardless of the temperature. You can only conclude that the differences are significant when averaged over the levels of the other factor.

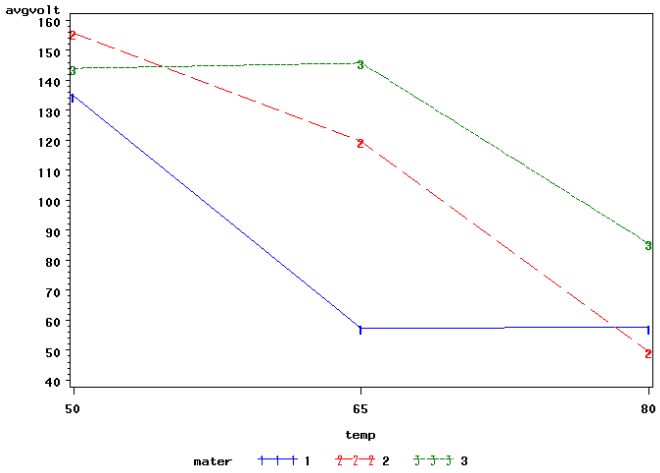
However, you can compare materials at each temperature, and you can compare temperatures for each material. At individual temperatures, material 2 and 3 (or 1 and 2) might be significantly different even though they are not significantly different when averaged over temperatures. For example, material 2 might produce a significantly higher average output than the other two material types at 50 degrees. This comparison of cell means is relevant if you are interested in using the batteries at 50 degrees! Comparing cell means is possible using the **lsmeans** command in **glm**, a point I will return to later.

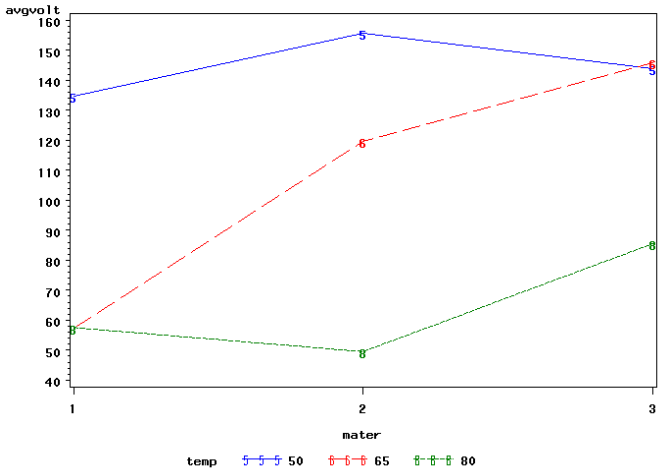
Program editor contents:

```
* Output voltage for batteries;
options ls=79 nodate nocenter;

* input data;
data battvolt;
  input mater $ temp $ v1-v4;  * mater and temp are both character data;
  voltage = v1; output; * output one observation line per voltage;
  voltage = v2; output;
  voltage = v3; output;
  voltage = v4; output;
  drop v1-v4;         * at end of reading data, remove variables t1, t2, t3, t4;
datalines;
1 50  130 155  74 180
1 65   34  40  80  75
1 80   20  70  82  58
2 50  150 188 159 126
2 65  136 122 106 115
2 80   25  70  58  45
3 50  138 110 168 160
3 65  174 120 150 139
3 80   96 104  82  60
;

* need to sort for the means procedure;
*   note that the data are already in this order                        ;
*   but it is good practice to "sort by" when a following procedure     ;
*   uses a "by" statement                                               ;
proc sort data=battvolt;
  by mater temp;
run;

* calculate means and output means to dataset meanvolt;
proc means data=battvolt noprint;    * do not print means;
  var voltage;
  by mater temp;
  output out=meanvolt mean=avgvolt; * output mean voltage to dataset meanvolt;
run;


* plot means in two profile plots;

* mater lines over temp;
* define v=symbol, c=color, i=interpolation, l=line style;
```

```
symbol1 v="1" c=blue  i=join l=1;
symbol2 v="2" c=red   i=join l=2;
symbol3 v="3" c=green i=join l=3;
proc gplot data=meanvolt;
  plot avgvolt*temp=mater;
run;

* temp lines over mater;
* define v=symbol, c=color, i=interpolation, l=line style;
symbol1 v="50" c=blue  i=join l=1;
symbol2 v="65" c=red   i=join l=2;
symbol3 v="80" c=green i=join l=3;
proc gplot data=meanvolt;
  plot avgvolt*mater=temp;
run;


proc glm data=battvolt;
  class mater temp;                          * factor variables              ;
  model voltage = mater temp mater*temp;     * interaction model             ;
  means mater temp mater*temp /bon;          * marginal and cell means       ;
                                             *   with Bonferroni comparisons;
run;
```

Profile plots:





Output window contents:

The GLM Procedure

Class Level Information

Class	Levels	Values
mater	3	1 2 3
temp	3	50 65 80

Number of Observations Read 36
Number of Observations Used 36

Dependent Variable: voltage

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	59416.22222	7427.02778	11.00	<.0001
Error	27	18230.75000	675.21296		
Corrected Total	35	77646.97222			

R-Square	Coeff Var	Root MSE	voltage Mean
0.765210	24.62372	25.98486	105.5278

Source	DF	Type I SS	Mean Square	F Value	Pr > F
mater	2	10683.72222	5341.86111	7.91	0.0020
temp	2	39118.72222	19559.36111	28.97	<.0001
mater*temp	4	9613.77778	2403.44444	3.56	0.0186

Source	DF	Type III SS	Mean Square	F Value	Pr > F
mater	2	10683.72222	5341.86111	7.91	0.0020
temp	2	39118.72222	19559.36111	28.97	<.0001
mater*temp	4	9613.77778	2403.44444	3.56	0.0186

Bonferroni (Dunn) t Tests for voltage

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	27
Error Mean Square	675.213
Critical Value of t	2.55246
Minimum Significant Difference	27.077

Means with the same letter are not significantly different.

Bon Grouping	Mean	N	mater
A	125.08	12	3
B A	108.33	12	2
B	83.17	12	1

Bonferroni (Dunn) t Tests for voltage

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	27
Error Mean Square	675.213
Critical Value of t	2.55246
Minimum Significant Difference	27.077

Means with the same letter are not significantly different.

Bon Grouping	Mean	N	temp
A	144.83	12	50
B	107.58	12	65
C	64.17	12	80

Level of mater	Level of temp	N	voltage Mean	voltage Std Dev
1	50	4	134.750000	45.3532432
1	65	4	57.250000	23.5990819
1	80	4	57.500000	26.8514432
2	50	4	155.750000	25.6173769
2	65	4	119.750000	12.6589889
2	80	4	49.500000	19.2613603
3	50	4	144.000000	25.9743463
3	65	4	145.750000	22.5444006
3	80	4	85.500000	19.2786583

5.2.5 Checking assumptions in a two-factor experiment

The normality and constant variance assumptions for a two-factor design can be visually checked using side-by-side boxplots and residual plots. Another useful tool for checking constant variances is to plot the sample deviations for each group against the group means.

Let us check the distributional assumptions for the insecticide experiment. The sampling design suggests that the independence assumptions are reasonable. The group sample sizes are small, so the residual plots are likely to be more informative than the side-by-side boxplots and the plot of the standard deviations.

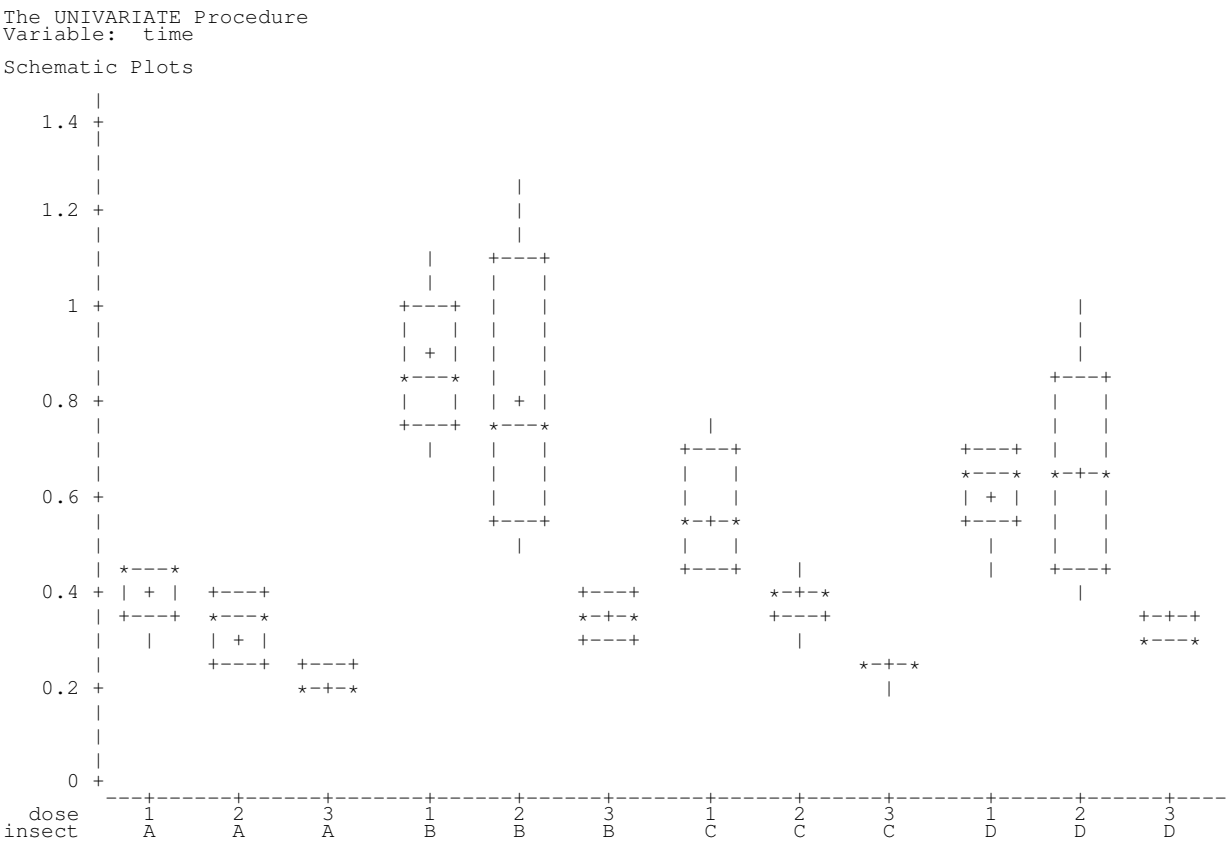
The **SAS** program given below generates plots and summary statistics for the survival times. I used the **univariate** procedure to get boxplots and to also compute and store the means $\bar{y}_{ij}$ and standard deviations s_{ij} for the 12 treatment combinations. The **univariate** procedure is more convenient than the **means** procedure that I used in earlier analyses to save group means because it also gives the boxplots. The **glm** procedure was used to generate output for the residual plots. Only the relevant output is presented.

Program editor contents:

```
* Checking assumptions in beetles example;
options ls=79 nodate nocenter;

* input data;
data beetles;
  input dose $ insect $ t1-t4 @@;  * dose and insect are both character data;
  time = t1; obs = 4 * _N_ - 3; output; * output one observation line per time;
  time = t2; obs = 4 * _N_ - 2; output; * obs is observation number;
  time = t3; obs = 4 * _N_ - 1; output;
  time = t4; obs = 4 * _N_ - 0; output;
  drop t1-t4;        * at end of reading data, remove variables t1, t2, t3, t4;
datalines;
1 A 0.31 0.45 0.46 0.43     1 B 0.82 1.10 0.88 0.72
1 C 0.43 0.45 0.63 0.76     1 D 0.45 0.71 0.66 0.62
2 A 0.36 0.29 0.40 0.23     2 B 0.92 0.61 0.49 1.24
2 C 0.44 0.35 0.31 0.40     2 D 0.56 1.02 0.71 0.38
3 A 0.22 0.21 0.18 0.23     3 B 0.30 0.37 0.38 0.29
3 C 0.23 0.25 0.24 0.22     3 D 0.30 0.36 0.31 0.33
;
run;

* need to sort for procedures using "by" statement;
proc sort data=beetles;
  by insect dose;
run;

* Use for boxplots and cell means and stddev;
proc univariate data=beetles plot;  * create boxplots;
  var time;
  by insect dose;
  output out=meanstdsurv mean=avgtime std=stdtime; * output mean and stddev;
```

```
                                                    *   to dataset meanstdsurv;
run;

* print means and standard deviations;
proc print data=meanstdsurv;
run;

* plot to check whether means and standard deviations are related;
symbol1 v=plus c=black i=none;
proc gplot data=meanstdsurv;
  plot stdtime*avgtime;
run;

proc glm data=beetles;
  class dose insect;                    * factor variables  ;
  model time = dose insect dose*insect; * interaction model ;

  * output to dataset glmout the studentized residuals,
                               predicted values, and cook's D ;
  output out=glmout student=str predicted=pred cookd=cookd;
run;

* create normal quantiles for QQ plot;
proc rank data=glmout out=glmout normal=blom;
  var str;
  ranks rankstr;
run;

symbol1 v=plus c=black i=none;
proc gplot data=glmout;
  plot rankstr*str str*pred cookd*obs;
run;
```

Output window contents:
The set of box plots (each representing 4 points) for each insecticide/dose combination indicates both that means and standard deviations of treatments seem different. Also, there appears to be less variability for dose=3 than for doses 1 and 2; the model assumes that variability is the same and does not depend on treatment.



Dataset with means and standard deviations:
When ordered by dose, the smaller standard deviations are for dose=3.

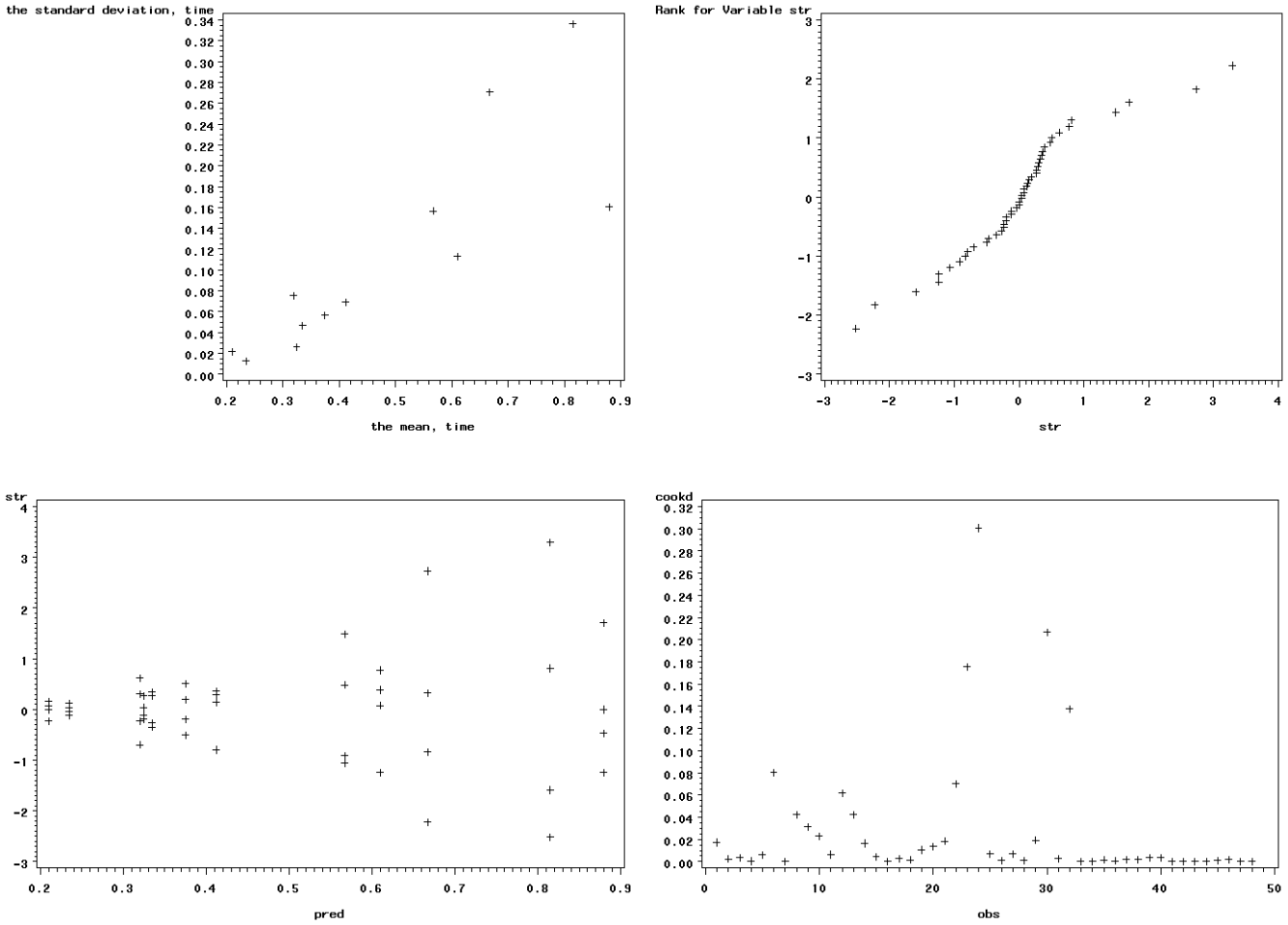
Ordered by insecticide

Obs	insect	dose	avgtime	stdtime
1	A	1	0.4125	0.06946
2	A	2	0.3200	0.07528
3	A	3	0.2100	0.02160
4	B	1	0.8800	0.16083
5	B	2	0.8150	0.33630
6	B	3	0.3350	0.04655
7	C	1	0.5675	0.15671
8	C	2	0.3750	0.05686
9	C	3	0.2350	0.01291
10	D	1	0.6100	0.11284
11	D	2	0.6675	0.27097
12	D	3	0.3250	0.02646

Ordered by dose

Obs	insect	dose	avgtime	stdtime
1	A	1	0.4125	0.06946
4	B	1	0.8800	0.16083
7	C	1	0.5675	0.15671
10	D	1	0.6100	0.11284
2	A	2	0.3200	0.07528
5	B	2	0.8150	0.33630
8	C	2	0.3750	0.05686
11	D	2	0.6675	0.27097
3	A	3	0.2100	0.02160
6	B	3	0.3350	0.04655
9	C	3	0.2350	0.01291
12	D	3	0.3250	0.02646

Diagnostic plots:
1. The plot of the mean vs standard deviation shows an increasing trend.
2. The normal quantile plot shows an “S” shape rather than a straight line, suggesting the residuals are not normal, but have higher kurtosis (more peaky) than a normal distribution. 3. The residuals vs the predicted value show that the higher the predicted value the more variability (horn shaped).
4. The plot of the Cook’s distances indicate a few influential observations.



Survival times are usually right skewed, with the spread or variability in the distribution increasing as the mean or median increases. Ideally, the distributions should be symmetric, normal, and the standard deviation should be fairly constant across groups.

The boxplots (note the ordering) and the plot of the s_{ij} against $\bar{y}_{ij}$ show the tendency for the spread to increase with the mean. This is reinforced by the residual plot, where the variability increases as the predicted values (the cell means under the two-factor interaction model) increase.

As noted earlier, the rankit plot of the studentized residuals is better suited to examine normality here than the boxplots which are constructed from 4 observations. Not surprisingly, the boxplots do not suggest non-normality. Looking at the rankit plot we clearly see evidence of non-normality.

5.2.6 A Remedy for Non-Constant Variance

A plot of cell standard deviations against the cell means is sometimes used as a diagnostic tool for suggesting transformations of the data. Here are some suggestions for transforming non-negative measurements to make the variability independent of the mean (i.e., stabilize the variance). The transformations also tend to reduce skewness, if present (and may induce skewness if absent). As an aside, I note that some statisticians prefer to plot the IQR against the median to get a more robust view of the dependence of spread on typical level because s_{ij} and $\bar{y}_{ij}$ are sensitive to outliers.

1. If s_{ij} increases linearly with $\bar{y}_{ij}$, use a **log** transformation of the response.

2. If s_{ij} increases as a quadratic function of $\bar{y}_{ij}$, use a reciprocal (**inverse**) transformation of the response.

3. If s_{ij} increases as a square root function of $\bar{y}_{ij}$, use a **square root** transformation of the response.

4. If s_{ij} is roughly independent of $\bar{y}_{ij}$, do not transform the response. This idea does not require the response to be non-negative!

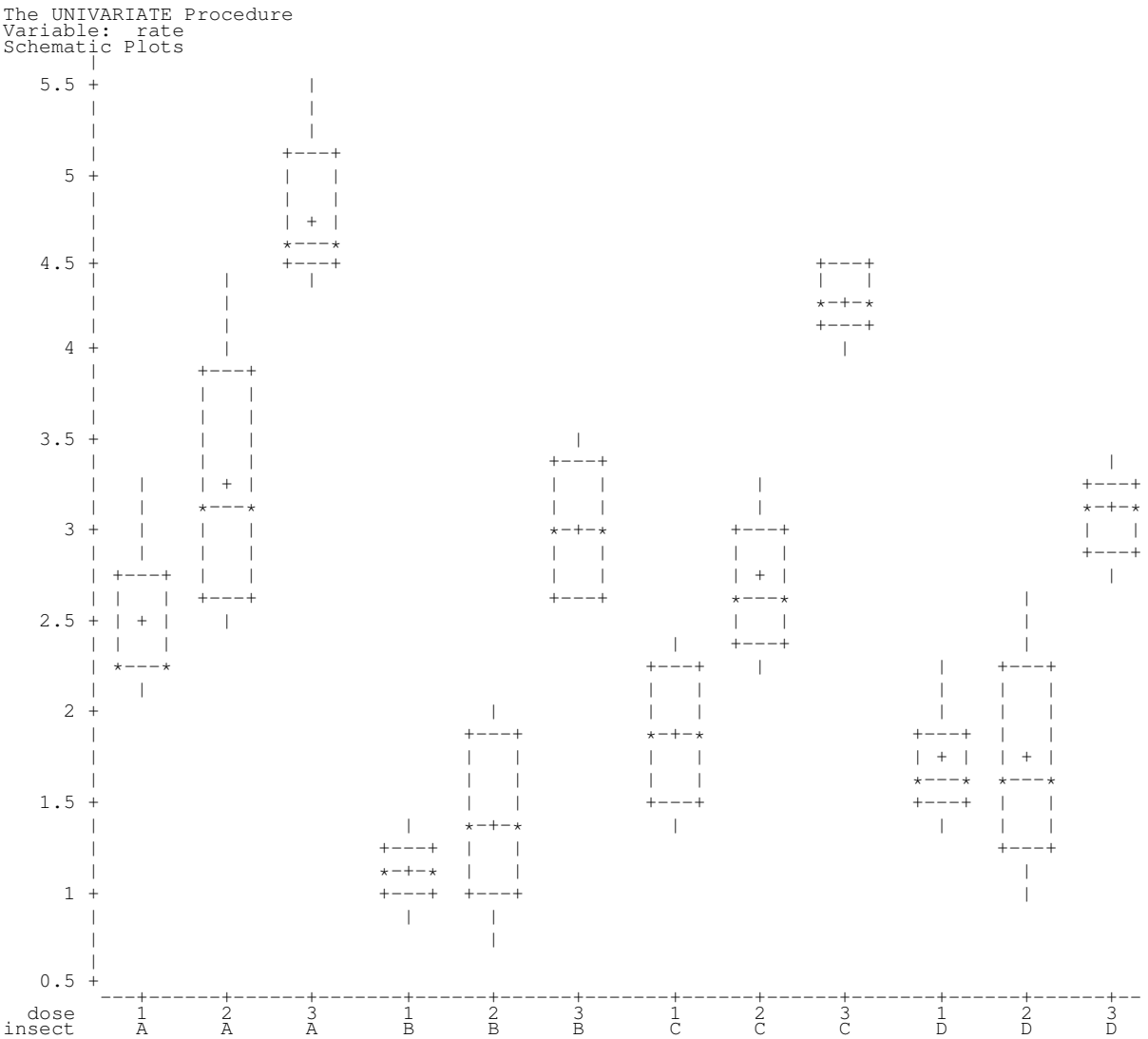
A logarithmic transformation or a reciprocal (inverse) transformation of the survival times might help to stabilize the variance. The survival time distributions are fairly symmetric, so these non-linear transformations may destroy the symmetry. As a first pass, I will consider the reciprocal transformation because the inverse survival time has a natural interpretation as the dying rate. For example, if you survive 2 hours, then 1/2 is the proportion of your remaining lifetime expired in the next hour. The unit of time is actually 10 hours, so 0.1/time is the actual rate. The 0.1 scaling factor has no effect on the analysis provided you appropriately rescale the results on the mean responses.

Three simple modifications of the **SAS** program are needed to do the analysis on the rates: create the rates (**rate**) in the **data step**, change **time** to **rate** in subsequent **procs**, and add a **means** statement for multiple comparisons.

The distributions of the rates are reasonably symmetric and have no outliers. Further, the cell standard deviations are more nearly constant than in the original analysis. Thus, the distributions of rates appear to satisfy the ANOVA assumptions.

Program editor contents:

```
* input data;
data beetles;     * replace dataset beetles   ;
  set beetles;    * start with beetles dataset;
  rate = 1/time;  * add rate column to dataset;
run;

** same code with substitutions of rate for time, etc. **;
** see ADA2_05.sas on website for complete listing     **;
```

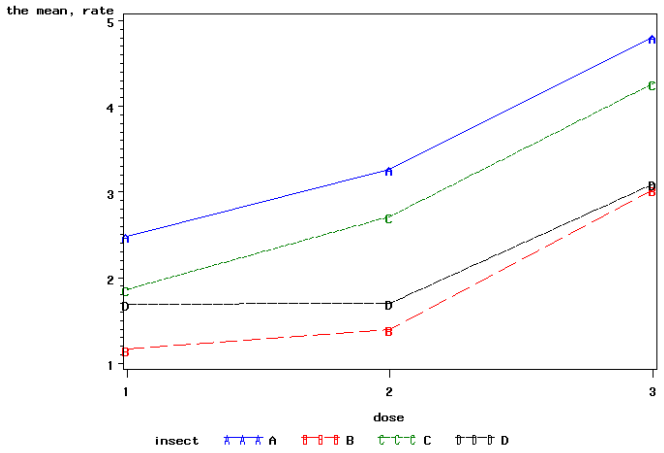


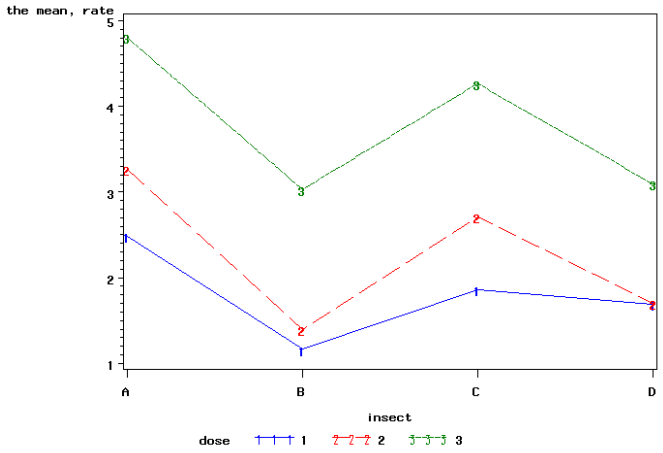
The standard deviations of rate appear much more similar than those of time did.

Obs	insect	dose	avgrate	stdrate
1	A	1	2.48688	0.49666
2	A	2	3.26847	0.82233
3	A	3	4.80269	0.52964
4	B	1	1.16346	0.19950
5	B	2	1.39339	0.55319
6	B	3	3.02897	0.42144
7	C	1	1.86272	0.48938
8	C	2	2.71392	0.41751
9	C	3	4.26499	0.23481
10	D	1	1.68968	0.36471
11	D	2	1.70153	0.70191
12	D	3	3.09181	0.24405

A profile plot, the ANOVA table with multiple comparisons, and diagnostic plots are given below. As before, the main effects are significant but the interaction is not. Unlike the original analysis, the residual plots do not show any gross deviations from assumptions. Also, no case seems relatively influential. Let us discuss the multiple comparisons and the implications about the mean rates.

Profile plots:





The GLM Procedure
Dependent Variable: rate

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	56.86218143	5.16928922	21.53	<.0001
Error	36	8.64308307	0.24008564		
Corrected Total	47	65.50526450			

R-Square	Coeff Var	Root MSE	rate Mean
0.868055	18.68478	0.489985	2.622376

Source	DF	Type I SS	Mean Square	F Value	Pr > F
dose	2	34.87711982	17.43855991	72.63	<.0001
insect	3	20.41428935	6.80476312	28.34	<.0001
dose*insect	6	1.57077226	0.26179538	1.09	0.3867

Source	DF	Type III SS	Mean Square	F Value	Pr > F
dose	2	34.87711982	17.43855991	72.63	<.0001
insect	3	20.41428935	6.80476312	28.34	<.0001
dose*insect	6	1.57077226	0.26179538	1.09	0.3867

Bonferroni (Dunn) t Tests for rate

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

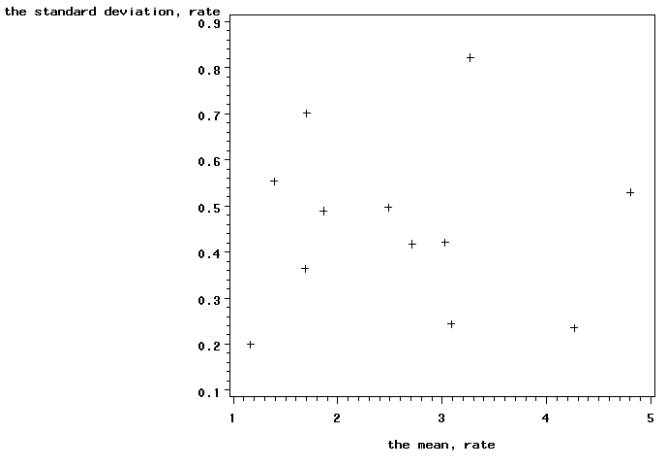
Alpha	0.05
Error Degrees of Freedom	36
Error Mean Square	0.240086
Critical Value of t	2.51104
Minimum Significant Difference	0.435

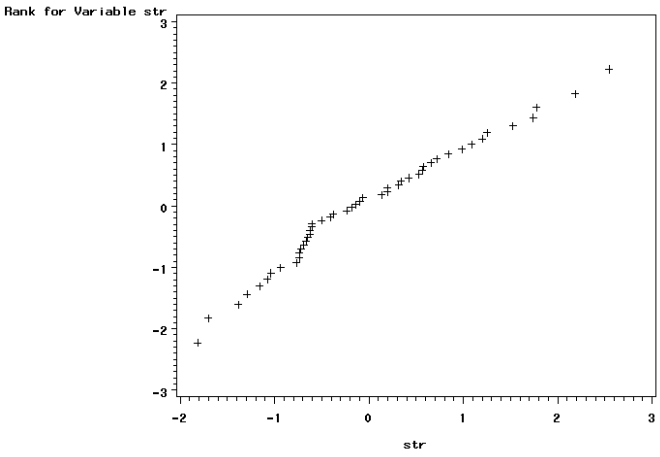
Means with the same letter are not significantly different.

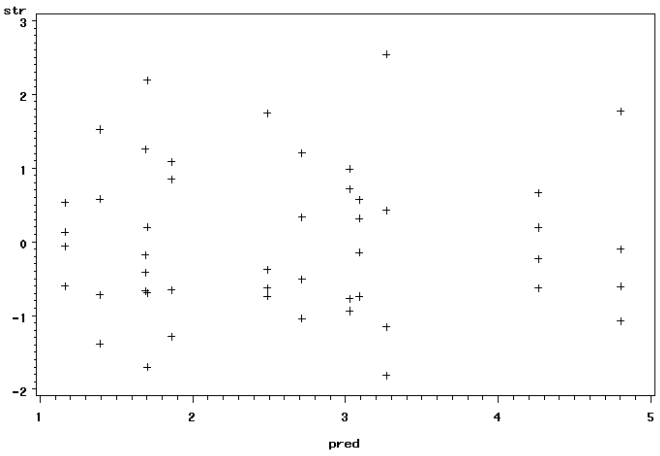
Bon Grouping	Mean	N	dose
A	3.7971	16	3
B	2.2693	16	2
C	1.8007	16	1

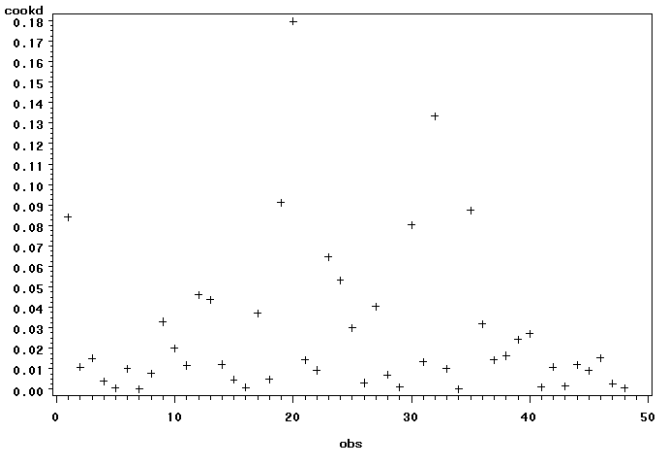
```
Bonferroni (Dunn) t Tests for rate

NOTE: This test controls the Type I experimentwise error rate, but it
generally has a higher Type II error rate than REGWQ.

Alpha                                   0.05
Error Degrees of Freedom                  36
Error Mean Square                   0.240086
Critical Value of t                  2.79197
Minimum Significant Difference         0.5585

Means with the same letter are not significantly different.

Bon
Grouping    Mean      N    insect
A         3.5193     12    A
B         2.9472     12    C
C         2.1610     12    D
C         1.8619     12    B
```

Diagnostic plots:









Comments on the Two Analyses of Survival Times

The effects of the transformation are noticeable. For example, the comparisons among doses and insecticides are less sensitive (differences harder to distinguish) on the original scale (look at the Bonferroni groupings). A comparison of the interaction p-values and profile plots for the two analyses suggests that the transformation eliminates much of the observed interaction between the main effects. Although the interaction in the original analysis was not significant at the 10% level (p-value=0.112), the small sample sizes suggest that power for detecting interaction might be low. To be on the safe side, one might interpret the main effects in the original analysis as if an interaction were present. This need appears to be less pressing with the rates.

The statistical assumptions are reasonable for an analysis of the rates. I think that the simplicity of the main effects interpretation is a strong motivating factor for preferring the analysis of the transformed data to the original analysis. You might disagree, especially if you believe that the original time scale is most relevant for analysis.

Given the suitability of the inverse transformation, I did not consider the logarithmic transformation.

5.2.7 glm: MEANS versus LSMEANS

In **glm**, either the **means** or **lsmeans** statement can be used to get multiple comparisons of the levels of the factors in a design. The **means** procedure is more convenient because the output gives the groupings of levels (for balanced designs and one-way ANOVA). At most, the **lsmeans** output provides p-values for each pairwise comparison, from which the groupings are constructed manually. However, the **lsmeans** statement provides a way to compare cell means, something that is not possible with the **means** statement.

The **SAS** program and output below uses **means** and **lsmeans** to get multiple comparisons for the battery data. The **pdiff** option on the **lsmeans** statement gives p-values for pairwise comparisons. To construct groupings based on the p-values, let us recall how the Bonferroni correction works. A comparison of 3 pairs of materials (or temperatures) having a family error rate of 0.05 or less is attained by comparing pairs of materials at the 0.05/3=0.0167 level. Using this criteria, the population mean voltages for two material types (averaged over temperatures) are significantly different if the p-value for the test is 0.0167 or less. In general, the cut-off level for statistical significance with a 0.05 family error rate is $.05/c$, where c is the number of pairwise comparisons. Using these ideas, hopefully you can show that **means** and **lsmeans** give the same groupings of materials and temperatures.

An important point is that the cell and marginal averages given by the **means** and **lsmeans** statements agree here because the design is balanced. For unbalanced designs with two or more factors, **lsmeans** and **means** compute different averages. I will argue that **lsmeans** are the appropriate averages for unbalanced analyses. You should use the **means** statement with caution — it is OK for balanced or unbalanced one-factor designs, and for the balanced two-factor designs (including the RB) that we have discussed.

Program editor contents:

```
* means vs lsmeans in battery voltage example;
proc glm data=battvolt;
  class mater temp;                        * factor variables            ;
  model voltage = mater temp mater*temp; * interaction model           ;
  means   mater temp mater*temp /bon  ;  * marginal and cell means      ;
```

```
                                          *   with Bonferroni comparisons;
  lsmeans mater temp mater*temp /pdiff;   * pdiff = p-values of the differences;
run;
```

MEANS output:

```
The GLM Procedure
Bonferroni (Dunn) t Tests for voltage

NOTE: This test controls the Type I experimentwise error rate, but it
generally has a higher Type II error rate than REGWQ.

Alpha                              0.05
Error Degrees of Freedom             27
Error Mean Square                675.213
Critical Value of t              2.55246
Minimum Significant Difference    27.077

Means with the same letter are not significantly different.

Bon Grouping    Mean      N    mater
     A        125.08     12    3
B    A        108.33     12    2
B             83.17     12    1
```

```
Bonferroni (Dunn) t Tests for voltage

NOTE: This test controls the Type I experimentwise error rate, but it
generally has a higher Type II error rate than REGWQ.

Alpha                              0.05
Error Degrees of Freedom             27
Error Mean Square                675.213
Critical Value of t              2.55246
Minimum Significant Difference    27.077

Means with the same letter are not significantly different.

Bon
Grouping   Mean      N    temp
A        144.83     12    50
B        107.58     12    65
C         64.17     12    80


Level of     Level of          -----------voltage-----------
mater        temp         N             Mean          Std Dev

1            50           4       134.750000       45.3532432
1            65           4        57.250000       23.5990819
1            80           4        57.500000       26.8514432
2            50           4       155.750000       25.6173769
2            65           4       119.750000       12.6589889
2            80           4        49.500000       19.2613603
3            50           4       144.000000       25.9743463
3            65           4       145.750000       22.5444006
3            80           4        85.500000       19.2786583
```

LSMEANS output:
differences for main effect mater

```
The GLM Procedure
Least Squares Means

                voltage      LSMEAN
mater            LSMEAN      Number
1             83.166667           1
2            108.333333           2
3            125.083333           3

     Least Squares Means for effect mater
     Pr > |t| for H0: LSMean(i)=LSMean(j)

          Dependent Variable: voltage
i/j                 1              2              3
   1                          0.0251         0.0005
   2          0.0251                         0.1260
   3          0.0005         0.1260

NOTE: To ensure overall protection level, only probabilities associated with
      pre-planned comparisons should be used.
```

differences for main effect temp

```
               voltage      LSMEAN
temp            LSMEAN      Number
50          144.833333           1
65          107.583333           2
80           64.166667           3

     Least Squares Means for effect temp
     Pr > |t| for H0: LSMean(i)=LSMean(j)

          Dependent Variable: voltage
i/j                 1              2              3
   1                          0.0016         <.0001
   2          0.0016                         0.0003
   3          <.0001         0.0003

NOTE: To ensure overall protection level, only probabilities associated with
      pre-planned comparisons should be used.
```

differences for any pair of cell means

```
                          voltage      LSMEAN
mater     temp             LSMEAN      Number

1         50           134.750000           1
1         65            57.250000           2
1         80            57.500000           3
2         50           155.750000           4
2         65           119.750000           5
2         80            49.500000           6
3         50           144.000000           7
3         65           145.750000           8
3         80            85.500000           9

                Least Squares Means for effect mater*temp
                   Pr > |t| for H0: LSMean(i)=LSMean(j)

                            Dependent Variable: voltage

i/j         1         2         3         4         5         6         7         8         9
    1             0.0002    0.0003    0.2631    0.4214    <.0001    0.6187    0.5544    0.0124
    2    0.0002             0.9892    <.0001    0.0021    0.6765    <.0001    <.0001    0.1358
    3    0.0003    0.9892             <.0001    0.0022    0.6667    <.0001    <.0001    0.1392
    4    0.2631    <.0001    <.0001              0.0605    <.0001    0.5279    0.5907    0.0007
    5    0.4214    0.0021    0.0022    0.0605              0.0007    0.1980    0.1685    0.0732
    6    <.0001    0.6765    0.6667    <.0001    0.0007              <.0001    <.0001    0.0605
    7    0.6187    <.0001    <.0001    0.5279    0.1980    <.0001              0.9248    0.0036
    8    0.5544    <.0001    <.0001    0.5907    0.1685    <.0001    0.9248              0.0029
    9    0.0124    0.1358    0.1392    0.0007    0.0732    0.0605    0.0036    0.0029

NOTE: To ensure overall protection level, only probabilities associated with
      pre-planned comparisons should be used.
```

5.2.8 Unbalanced Two-Factor Designs and Analysis

The sample sizes are usually unequal, or **unbalanced**, for the different treatment groups in an experiment. Although this has no consequence on the specification of a model, you have to be more careful with the **analysis of unbalanced experiments** that have two or more factors. With unbalanced designs, the Type I and Type III SS differ, as do the main effect averages given by **means** and **lsmeans**. Inferences might critically depend on which summaries are used. I will use the following example to emphasize the differences between the types of SS and averages.

Example: rat insulin

The **SAS** data step below gives the insulin levels in rats a certain length of time after a fixed dose of insulin was injected into their jugular or portal veins. This is a two-factor study with two vein types (jugular, portal) and three time levels (0, 30, and 60). An unusual feature of this experiment is that the rats used in the six vein and time combinations are distinct. I will fit a two-factor interaction model, which assumes that the responses are independent within and across treatments. The design is **unbalanced**, with sample sizes varying from 3 to 12.

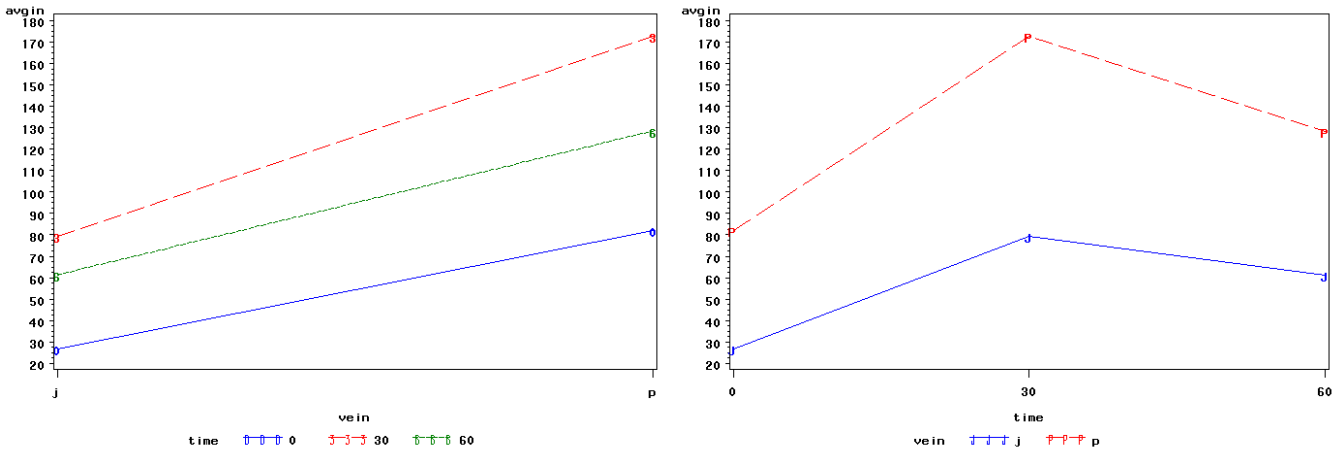
An alternative experimental design might randomly assign rats to the two vein groups, and then measure the insulin levels of each rat at the three time points. Depending on the questions of interest, you could compare veins using a one-way MANOVA, or use other multivariate techniques that allow correlated responses within rats.

Looking at the output, we see the Type I and Type III SS are different, except for the interaction term, and that the vein **means** differ from the vein **lsmeans**. The **means** output for time does not give individual averages but rather CIs for all pairwise differences in means (**SAS**'s method for

unequal sample sizes when a factor has 3+ levels). Although the differences between the types of SS and means are small I will show you an example where serious discrepancies are found.
Program editor contents:

```
* Rat insulin example;
options ls=79 nodate nocenter;

data ratinsulin;
  input vein $ time $ insulin @@ ;
datalines;
j  0  18   j  0  36   j  0  12   j  0  24   j  0  43
j 30  61   j 30 116   j 30  63   j 30 132   j 30  68   j 30  37
j 60  18   j 60 133   j 60  33
p  0  96   p  0  72   p  0  34   p  0  41   p  0  98   p  0  77
p  0 120   p  0  49   p  0  92   p  0 111   p  0  99   p  0  94
p 30 146   p 30 193   p 30  78   p 30 127   p 30 136   p 30 144
p 30 115   p 30 199   p 30 253   p 30 338
p 60 132   p 60 110   p 60 141   p 60 204   p 60  69   p 60 152
p 60 196   p 60 195   p 60  84   p 60 105   p 60  71   p 60  83
;
run;

* need to sort for the means procedure;
proc sort data=ratinsulin;
  by vein time;
run;

proc means data=ratinsulin noprint;  * do not print means;
  var insulin;
  by vein time;
  output out=meanins mean=avgin; * output means dataset meanins;
run;

proc print data=meanins;
run;

* plot means in two profile plots;

* mater lines over temp;
* define v=symbol, c=color, i=interpolation, l=line style;
symbol1 v="j" c=blue  i=join l=1;
symbol2 v="p" c=red   i=join l=2;
proc gplot data=meanins;
  plot avgin*time=vein;
run;

* temp lines over mater;
* define v=symbol, c=color, i=interpolation, l=line style;
symbol1 v="0"  c=blue  i=join l=1;
symbol2 v="30" c=red   i=join l=2;
symbol3 v="60" c=green i=join l=3;
proc gplot data=meanins;
  plot avgin*vein=time;
run;

proc glm data=ratinsulin;
  class vein time;                    * factor variables             ;
  model insulin=vein time vein*time;  * interaction model            ;
  means   vein time vein*time /bon;   * marginal and cell means      ;
                                      *   with Bonferroni comparisons;
  lsmeans vein time vein*time /pdiff; * pdiff = p-values of the differences;
run;
```

Profile plots:
Because the lines all seem parallel, it appears there is not sufficient evidence for a vein-by-time interaction.



Output window contents:
Treatment means

Obs	vein	time	_TYPE_	_FREQ_	avgin
1	j	0	0	5	26.600
2	j	30	0	6	79.500
3	j	60	0	3	61.333
4	p	0	0	12	81.917
5	p	30	0	10	172.900
6	p	60	0	12	128.500

Type I and Type III SS

glm gives Type I and Type III SS for each effect in a model. The Type I SS is the **sequential reduction** in Error SS achieved when an effect is added to a model that includes only the prior effects listed in the **model** statement. The Type III SS are more difficult to define explicitly, but they roughly correspond to the reduction in Error SS achieved when an effect is added **last** to the model.

Type I SS and Type III SS are equal for balanced designs and for one-way ANOVA, but are typically different for unbalanced designs, where there is no unique way to define the SS for an effect. The problem here is similar to multiple regression, where the SS for a predictor X is the decrease in Residual SS when X is added to a model. This SS is not unique because the change in the Residual SS depends on which predictors are included in the model prior to X. In a regression analysis, the standard tests for effects in a model are based on **glm** Type III SS and not on the Type I SS.

For the insulin analysis, the Type I and Type III interaction SS are identical because this effect was added last to the **model** statement. The Type I and III SS for the main effects are not equal. Also note that the Type I SS for the main effects and interaction add to the model SS, but the Type III SS do not.

```
The GLM Procedure

    Class Level Information
Class          Levels    Values
vein                2    j p
time                3    0 30 60

Number of Observations Read          48
Number of Observations Used          48

Dependent Variable: insulin
                                    Sum of
Source                      DF        Squares    Mean Square   F Value   Pr > F
Model                        5     99478.4833     19895.6967      8.12   <.0001
Error                       42    102914.1833      2450.3377
Corrected Total             47    202392.6667

R-Square     Coeff Var      Root MSE    insulin Mean
0.491512      47.06899      49.50089        105.1667

Source                      DF      Type I SS    Mean Square   F Value   Pr > F
vein                         1    46400.28011    46400.28011     18.94   <.0001
time                         2    50332.28932    25166.14466     10.27   0.0002
vein*time                    2     2745.91390     1372.95695      0.56   0.5752

Source                      DF     Type III SS    Mean Square   F Value   Pr > F
vein                         1    48212.70374    48212.70374     19.68   <.0001
time                         2    37734.18798    18867.09399      7.70   0.0014
vein*time                    2     2745.91390     1372.95695      0.56   0.5752
```

MEANS versus LSMEANS

The **lsmeans** (sometimes called **adjusted means**) for a single factor is an arithmetic average of cell means. For example, the mean responses in the jugular vein at times 0, 30, and 60 are 26.6, 79.5, and 61.33, respectively. The **lsmeans** for the jugular vein is thus $55.81 = (26.6 + 79.5 + 61.33)/3$. This average gives equal weight to the 3 times even though the sample sizes at these times differ (5, 6, and 3). The **means** of 56.71 for the jugular is the average of the 14 jugular responses, ignoring time. If the cell sample sizes were equal, the **lsmeans** and **means** averages would agree.

The **means** and **lsmeans** for individual cells (i.e., for the 6 **vein*time** combinations) are identical, and equal to cell means.

Which to use? Regardless of whether the design is balanced, the basic building blocks for a two-factor analysis are cell means, and the marginal means, defined as the average of the cell means over the levels of the other factor.

The F-statistics based on Type III SSs are appropriate for unbalanced two-factor designs because they test the same hypotheses that were considered in balanced designs. That is, the Type III F-tests on the main effects check for equality in population means averaged over levels of the other factor. The Type III F-test for no interaction checks for parallel profiles. Given that the Type III F-tests for the main effects check for equal population cell means averaged over the levels of the other factor, multiple comparisons for main effects should be based on **lsmeans**.

The Type I SS and F-tests and the multiple comparisons based on **means** should be ignored because they do not, in general, test meaningful hypotheses. The problem with using the **means** output is that the experimenter has fixed the sample sizes for a two-factor experiment, so comparisons of **means**, which ignore the second factor, introduces a potential bias due to choice of sample sizes. Put another way, any differences seen in the **means** in the jugular and portal could be solely due to the sample sizes used in the experiment and not due to differences in the veins.

Focusing on the Type III SS, the F-tests indicate that the vein and time effects are significant, but that the interaction is not significant. The jugular and portal profiles are reasonably parallel, which is consistent with a lack of interaction. What can you conclude from the **lsmeans** comparisons of veins and times?

MEANS output: do not use when design is unbalanced, or, to be safe, always use LSMEANS below.

```
Bonferroni (Dunn) t Tests for insulin

NOTE: This test controls the Type I experimentwise error rate, but it
generally has a higher Type II error rate than REGWQ.

Alpha                              0.05
Error Degrees of Freedom             42
Error Mean Square              2450.338
Critical Value of t             2.01808
Minimum Significant Difference   31.723
Harmonic Mean of Cell Sizes    19.83333

NOTE: Cell sizes are not equal.

Means with the same letter are not significantly different.

Bon
Grouping   Mean      N    vein
A        125.12     34    p
B         56.71     14    j



Bonferroni (Dunn) t Tests for insulin

NOTE: This test controls the Type I experimentwise error rate, but it
generally has a higher Type II error rate than Tukey's for all pairwise
comparisons.

Alpha                          0.05
Error Degrees of Freedom         42
Error Mean Square          2450.338
Critical Value of t         2.49367

Comparisons significant at the 0.05 level are indicated by ***.

              Difference      Simultaneous
   time          Between     95% Confidence
Comparison         Means         Limits
 30 - 60           22.81     -21.56   67.17
 30 - 0            72.23      29.23  115.22  ***
 60 - 30          -22.81     -67.17   21.56
 60 - 0            49.42       5.69   93.15  ***
 0  - 30          -72.23    -115.22  -29.23  ***
 0  - 60          -49.42     -93.15   -5.69  ***


Level of     Level of           ----------insulin----------
vein         time          N            Mean         Std Dev
j            0             5        26.600000      12.7593103
j            30            6        79.500000      36.4458502
j            60            3        61.333333      62.5166644
p            0            12        81.916667      27.7470992
p            30           10       172.900000      76.1175261
p            60           12       128.500000      49.7182973
```

LSMEANS output: significant differences between veins, and between times 0 and 30.

```
The GLM Procedure
Least Squares Means

                          H0:LSMean1=
              insulin       LSMean2
vein           LSMEAN       Pr > |t|
j           55.811111          <.0001
p          127.772222

              insulin       LSMEAN
time           LSMEAN       Number
0           54.258333             1
30         126.200000             2
60          94.916667             3

     Least Squares Means for effect time
     Pr > |t| for H0: LSMean(i)=LSMean(j)

          Dependent Variable: insulin
i/j               1              2              3
   1                        0.0003         0.0562
   2         0.0003                        0.1338
   3         0.0562         0.1338

NOTE: To ensure overall protection level, only probabilities associated with
      pre-planned comparisons should be used.
```

The table below is useful for finding differences when an interaction is present, but not that informative for a model with only main effects.

```
                        insulin       LSMEAN
vein    time             LSMEAN       Number

j       0             26.600000             1
j       30            79.500000             2
j       60            61.333333             3
p       0             81.916667             4
p       30           172.900000             5
p       60           128.500000             6

                  Least Squares Means for effect vein*time
                    Pr > |t| for H0: LSMean(i)=LSMean(j)

                          Dependent Variable: insulin
i/j            1              2              3              4              5              6
   1                      0.0849         0.3422         0.0418         <.0001         0.0004
   2         0.0849                      0.6065         0.9227         0.0007         0.0543
   3         0.3422         0.6065                      0.5230         0.0014         0.0416
   4         0.0418         0.9227         0.5230                      0.0001         0.0262
   5         <.0001         0.0007         0.0014         0.0001                      0.0423
   6         0.0004         0.0543         0.0416         0.0262         0.0423

NOTE: To ensure overall protection level, only probabilities associated with
      pre-planned comparisons should be used.
```

Chapter 6

A Short Discussion of Observational Studies

"Thou shall adjust for what thou can not control."

In most scientific studies, the groups being compared do not consist of identical experimental units that have been randomly assigned to receive a treatment. Instead, the groups might be extremely heterogeneous on factors that might be related to a specific response on which you wish to compare the groups. Inferences about the nature of differences among groups in such **observational studies** can be flawed if this heterogeneity is ignored in the statistical analysis.

The following problem emphasizes the care that is needed when analyzing **observational studies**, and highlights the distinction between the **means** and **lsmeans** output for a two-way table. The **data are artificial**, but the conclusions are consistent with an interesting analysis conducted by researchers at Sandia laboratories.

A representative sample of 550 high school seniors was selected in 1970. A similar sample of 550 was selected in 1990. The final SAT scores (on a 1600 point scale) were obtained for each student. A simple analysis might compare the average SAT scores for the two years, to see whether students are scoring higher, lower, or about the same, over time. The one-way **lsmeans** and **means** breakdowns of the SAT scores are identical (except that **means** output also gives standard deviations). The average SAT scores for 1970 and 1990 are 892.8 and 882.2, respectively. The one-way ANOVA, combined with the observed averages, indicates that the typical SAT score has decreased significantly over the 20 year period.

The boxplots for the two samples (not given) show heavy tailed distributions with similar spreads. Given the large sample sizes, the F-test comparing populations is approximately valid even though the population distributions are non-normal.

Program editor contents:

```
* SAT example;
options ls=79 nodate nocenter;

data sat;
   infile "F:\Dropbox\UNM\teach\ADA2_stat528\sas\ADA2_06_sat.dat";
   input year eth grade;
run;

proc glm data=sat;
  class year;
  model grade=year;
  means year;
```

```
  lsmeans year;
run;
```

Output window contents:

```
The GLM Procedure

     Class Level Information
Class         Levels    Values
year               2    1970 1990

Number of Observations Read        1100
Number of Observations Used        1100

Dependent Variable: grade
                                    Sum of
Source                     DF       Squares    Mean Square   F Value   Pr > F
Model                       1     31409.8945     31409.8945     61.90   <.0001
Error                    1098    557117.1018       507.3926
Corrected Total          1099    588526.9964

R-Square     Coeff Var      Root MSE    grade Mean
0.053370      2.538076      22.52538      887.4982

Source                     DF      Type I SS    Mean Square   F Value   Pr > F
year                        1    31409.89455    31409.89455     61.90   <.0001

Source                     DF    Type III SS    Mean Square   F Value   Pr > F
year                        1    31409.89455    31409.89455     61.90   <.0001

MEANS OUTPUT:
Level of              ------------grade------------
year            N             Mean          Std Dev
1970          550       892.841818       22.9617262
1990          550       882.154545       22.0804071

LSMEANS OUTPUT:
Least Squares Means

year    grade LSMEAN
1970      892.841818
1990      882.154545
```

Should we be alarmed? Should we be concerned that students entering college have fewer skills than students 20 years ago? Should we be pumping billions of dollars into the bloated bureaucracies of our public school systems with the hope that a few of these dollars might be put to good use in programs to enhance performance? This is the consensus among some people in the know, all of whom wax eloquently about the impending inability of the U.S. to compete in the new global economy.

The SAT study is not a well-designed experiment, where a scientist has controlled all the factors that might affect the response (the SAT score) other than the treatment (the year). Even without these controls, there is no randomization of treatments to students selected from a target population.

The SAT study is an **observational study** of two distinct populations. The observed differences in SAT scores may indeed be due to a decrease in performance. The differences might also be due to factors that make the two populations incomparable for assessing changes in performance over time.

My hypothetical populations have students from two ethnic groups (1 and 2). If you construct box-plots of the SAT scores for the four combinations of ethnicity and year, you see that the typical SAT score **within** each ethnic group has **increased** over time, whereas the typical SAT score **ignoring** ethnicity decreased over time. Is this a paradox, and what are appropriate conclusions in the analysis?

I fit a two-factor model with year and ethnicity effects plus an interaction. The two-factor model gives a method to compare the SAT scores over time, after **adjusting** for the effect of ethnicity on performance. The F-test for comparing years adjusts for ethnicity because it is based on comparing the average SAT scores across years after averaging the cell means over ethnicities, thereby eliminating from the comparison of years any effects due to changes in the ethnic composition of the populations. The two-way analysis is preferable to the unadjusted one-way analysis which **ignores** ethnicity.

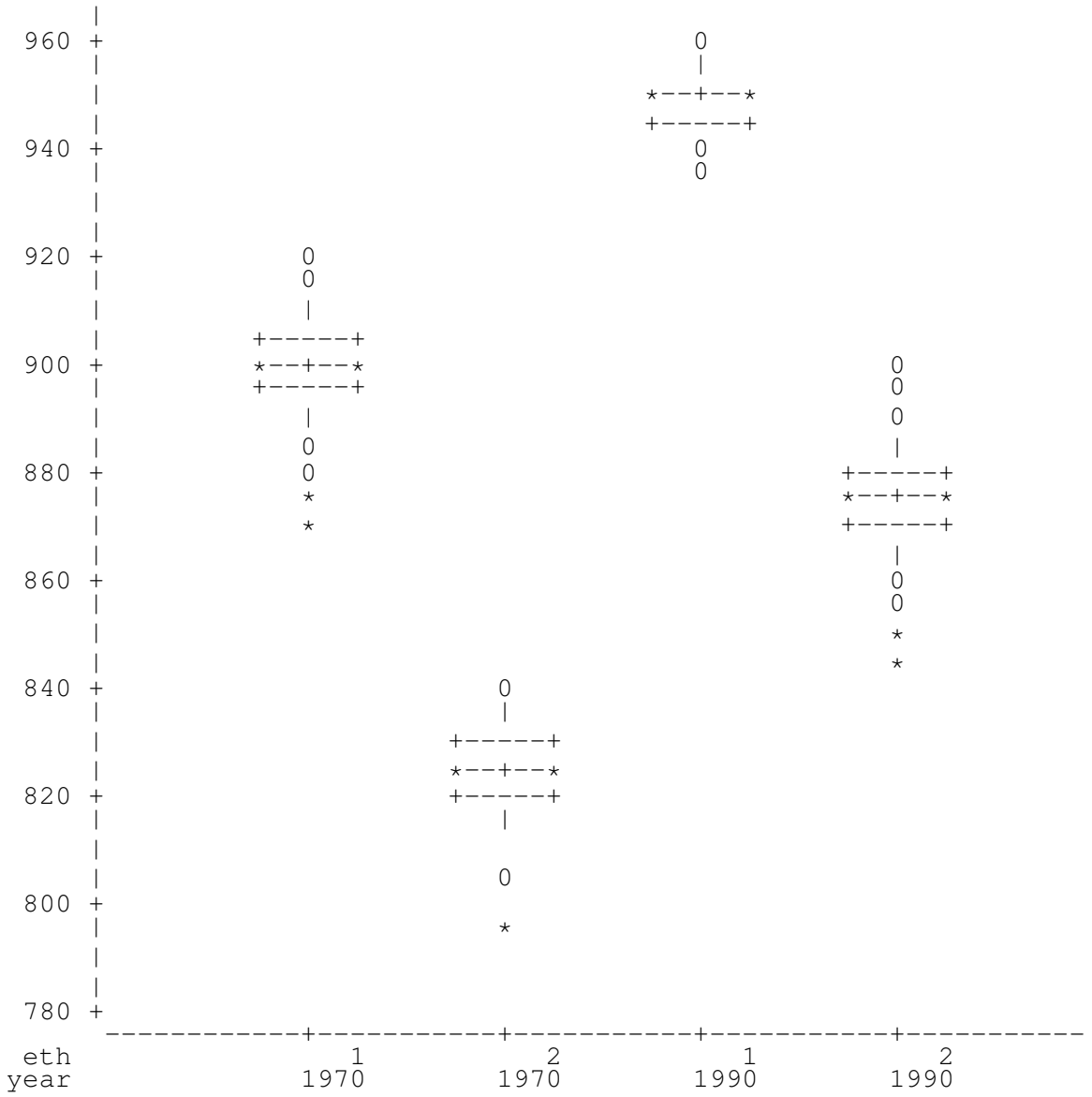
The year and ethnicity main effects are significant in the two factor model, but the interaction is not. The marginal **lsmeans** indicate that the average SAT score increased significantly over time when averaged over ethnicities. This is consistent with the cell mean SAT scores increasing over time **within** each ethnic group. Given the lack of a significant interaction, the expected increase in SAT scores from 1970 to 1990 **within each** ethnic group is the difference in marginal averages: 912.0 - 861.9 = 50.1.

Program editor contents:

```
* SAT example adjusting for ethnicity;
proc sort data=sat;
  by year eth;
run;

proc univariate data=sat plot;
  var grade;
  by year eth;
run;

proc glm data=sat;
  class year eth;
  model grade = year eth year*eth;
  means   year eth year*eth;
  lsmeans year eth year*eth;
run;
```

Output window contents:

```
The UNIVARIATE Procedure
Variable:  grade

Schematic Plots
```



```
The GLM Procedure

     Class Level Information
Class         Levels    Values
year               2    1970 1990
eth                2    1 2

Number of Observations Read        1100
Number of Observations Used        1100

Dependent Variable: grade
                                    Sum of
Source                     DF       Squares    Mean Square   F Value   Pr > F
Model                       3    533539.2264    177846.4088   3544.78   <.0001
Error                    1096     54987.7700        50.1713
Corrected Total          1099    588526.9964

R-Square     Coeff Var      Root MSE    grade Mean
0.906567      0.798106      7.083172      887.4982

Source                     DF      Type I SS    Mean Square   F Value   Pr > F
year                        1     31409.8945     31409.8945    626.05   <.0001
eth                         1    501984.3165    501984.3165  10005.4   <.0001
year*eth                    1       145.0154       145.0154      2.89   0.0894

Source                     DF    Type III SS    Mean Square   F Value   Pr > F
year                        1    228282.9383    228282.9383   4550.07   <.0001
eth                         1    501984.3165    501984.3165  10005.4   <.0001
year*eth                    1       145.0154       145.0154      2.89   0.0894

MEANS OUTPUT:
Level of              ------------grade------------
year             N            Mean          Std Dev
1970           550      892.841818       22.9617262
1990           550      882.154545       22.0804071

Level of              ------------grade------------
eth              N            Mean          Std Dev
1              550      904.152727       15.7643837
2              550      870.843636       16.3622780

Level of      Level of              ------------grade------------
year          eth             N            Mean          Std Dev
1970          1             500      899.712000       7.28719854
1970          2              50      824.140000       8.28081136
1990          1              50      948.560000       5.48898712
1990          2             500      875.514000       6.88480974

LSMEANS OUTPUT:
Least Squares Means

year    grade LSMEAN
1970      861.926000
1990      912.037000

eth    grade LSMEAN
1        924.136000
2        849.827000

year    eth    grade LSMEAN
1970    1        899.712000
1970    2        824.140000
1990    1        948.560000
1990    2        875.514000
```

As noted in the insulin analysis, the marginal **lsmeans** and **means** are different for unbalanced two-factor analyses. The marginal **means ignore** the levels of the other factors when averaging responses. The marginal **lsmeans** are averages of cell means over the levels of the other factor. Thus, for example, the 1970 **mean** SAT score of 892.8 is the average of the 550 scores selected that year. The 1970 **lsmeans** SAT score of 861.9 is midway between the average 1970 SAT scores for

the two ethnic groups: $861.9 = (899.7 + 824.1)/2$. Hopefully, this discussion also clarifies why the year marginal **means** are identical in the one and two-factor analyses, but the year **lsmeans** are not.

The 1970 and 1990 marginal **means** estimate the typical SAT score ignoring all factors that may influence performance. These marginal averages are not relevant for **understanding** any trends in performance over time because they do not account for changes in the composition of the population that may be related to performance.

The average SAT scores (ignoring ethnicity) decreased from 1970 to 1990 because the ethnic composition of the student population changed. Ten out of every eleven students sampled in 1970 were from the first ethnic group. Only one out of eleven students sampled in 1990 was from this group. Students in the second ethnic group are underachievers, but they are becoming a larger portion of the population over time. The decrease in average (**means**) performance inferred from comparing 1970 to 1990 is **confounded** with the increased representation of the underachievers over time. Once ethnicity was taken into consideration, the typical SAT scores were shown to have increased, rather than decreased.

In summary, the one-way analysis ignoring ethnicity is valid, and allows you to conclude that the typical SAT score has decreased over time, but it does not provide any insight into the nature of the changes that have occurred. A two-factor analysis backed up with a comparison of the marginal **lsmeans** is needed to compare performances over time, adjusting for the changes in ethnic composition.

The Sandia study reached the same conclusion. The Sandia team showed that the widely reported decreases in SAT scores over time are due to changes in the ethnic distribution of the student population over time, with individuals in historically underachieving ethnic groups becoming a larger portion of the student population over time.

A more complete analysis of the SAT study would adjust the SAT scores to account for other potential confounding factors, such as sex, and differences due to the number of times the exam was taken. These confounding effects are taken into consideration by including them as effects in the model. I will discuss this issue more completely in the coming week.

The interpretation of the results from an observational study with several effects of interest, and several confounding variables, is greatly simplified by eliminating the insignificant effects from the model. For example, the year by ethnicity interaction in the SAT study might be omitted from the model to simplify interpretation. The year effects would then be estimated after fitting a two-way additive model with year and ethnicity effects only. The same approach is sometimes used with designed experiments, say the insulin study that we analyzed earlier.

An important caveat

The ideas that we discussed on the design and analysis of experiments and observational studies are universal. They apply regardless of whether you are analyzing categorical data, counts, or measurements.

Chapter 7

Analysis of Covariance: Comparing Regression Lines

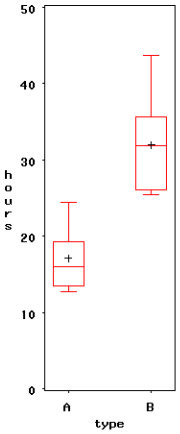
Suppose that you are interested in comparing the typical lifetime of two tool types (A = 0, B = 1). A simple analysis of the data given below would consist of making side-by-side boxplots followed by a two-sample test of equal means (or medians). The standard two-sample test using the pooled variance estimator is a special case of the one-way ANOVA with two groups. The summaries suggest that the distribution of lifetimes for the tool types are different. In the output below, μ_i is population mean lifetime for tool type i ($i = 0, 1$).

```
                        0 = type A
                        1 = type B
Lifetime(Hours)  Rpm     Type
          18.73    610      A
          14.52    950      A
          17.43    720      A
          14.54    840      A
          13.44    980      A
          24.39    530      A
          13.34    680      A
          22.71    540      A
          12.68    890      A
          19.32    730      A
          30.16    670      B
          27.09    770      B
          25.40    880      B
          26.05   1000      B
          33.49    760      B
          35.62    590      B
          26.07    910      B
          36.78    650      B
          34.95    810      B
          43.67    500      B
```

Program editor contents:

```
* tool data;
options ls=79 nodate nocenter;

data tools;
  infile "F:\Dropbox\UNM\teach\ADA2_stat528\sas\ADA2_07_tools.dat";
  input hours rpm type $;     * type is character data $ ;
  if type = 'A' then toolb=0;
  if type = 'B' then toolb=1;
  obs = _n_;  * for use later in plotting in glm;
run;
```

```
title "";
proc boxplot data=tools;
  plot hours * type / vzero;  * vzero includes 0 in axis;
run;

* 2-sample t-test on hours by type ;
proc ttest data=tools;
  var hours;
  class type;
run;
```



Output window contents:
A two sample t-test comparing mean lifetimes of tool types:

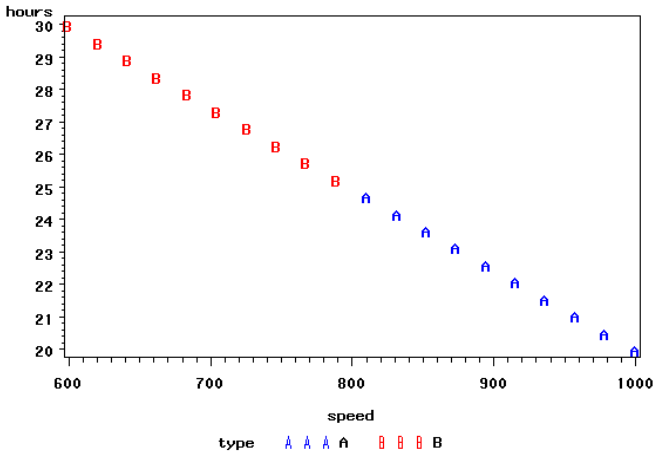
```
The TTEST Procedure
                              Statistics
                          Lower CL            Upper CL
Variable  type         N      Mean     Mean      Mean  Std Dev  Std Err
hours     A           10    14.164   17.11    20.056   4.1176   1.3021
hours     B           10    27.632  31.928    36.224   6.0059   1.8992
hours     Diff (1-2)        -19.66  -14.82     -9.98    5.149   2.3027

                                T-Tests
Variable    Method              Variances      DF    t Value    Pr > |t|
hours       Pooled              Equal          18      -6.44      <.0001
hours       Satterthwaite       Unequal      15.9      -6.44      <.0001

                    Equality of Variances
Variable    Method      Num DF    Den DF    F Value    Pr > F
hours       Folded F         9         9       2.13    0.2761
```

This comparison is potentially misleading because the samples are not comparable. A one-way ANOVA is most appropriate for designed experiments where all the factors influencing the response, other than the treatment (tool type), are fixed by the experimenter. The tools were operated at different speeds. If speed influences lifetime, then the observed differences in lifetimes could be due to differences in speeds at which the two tool types were operated.

For example, suppose speed is inversely related to lifetime of the tool. Then, the differences seen in the boxplots above could be due to tool type B (1) being operated at lower speeds than tool type A (0). To see how this is possible, consider the data plot given below, where the relationship between lifetime and speed is identical in each sample. A simple linear regression model relating hours to speed, ignoring tool type, fits the data exactly, yet the lifetime distributions for the tool types, ignoring speed, differ dramatically. (The data were generated to fall exactly on a straight line). The regression model indicates that you would expect identical mean lifetimes for tool types

A and B, if they were, or could be, operated at identical speeds. This is not exactly what happens in the actual data. However, I hope the point is clear.
Program editor contents:

```
* FAKE tools data;
data toolsfake;
  infile "F:\Dropbox\UNM\teach\ADA2_stat528\sas\ADA2_07_toolsfake.dat";
  input speed hours type $;    * type is string $ ;
run;

* define v=symbol, c=color, i=interpolation, l=line style;
symbol1 v="A" c=blue  i=none;
symbol2 v="B" c=red   i=none;
proc gplot data=toolsfake;
  plot hours*speed=type;
run;
```



As noted in the SAT example, you should be wary of group comparisons where important factors that influence the response have not been accounted for or controlled. In the SAT example, the differences in scores were affected by a change in the ethnic composition over time. A two-way ANOVA with two factors, time and ethnicity, gave the most sensible analysis.

For the tool lifetime problem, you should compare groups (tools) after adjusting the lifetimes to account for the influence of a measurement variable, speed. The appropriate statistical technique for handling this problem is called **analysis of covariance** (ANCOVA).

7.1 ANCOVA

A natural way to account for the effect of speed is through a multiple regression model with lifetime as the response and two predictors, speed and tool type. A binary **categorical variable**, here tool type, is included in the model as a **dummy variable** or **indicator variable** (a (0,1) variable).

Consider the model

$$\text{Tool lifetime} = \beta_0 + \beta_1 \text{ Toolb} + \beta_2 \text{ Speed} + e,$$

where Toolb is 0 for type A tools, and 1 for type B tools. For type A tools, the model simplifies to:

$$\begin{aligned} \text{Tool lifetime} &= \beta_0 + \beta_1(0) + \beta_2 \text{ Speed} + e \\ &= \beta_0 + \beta_2 \text{ Speed} + e. \end{aligned}$$

For type B tools, the model simplifies to:

$$\begin{aligned} \text{Tool lifetime} &= \beta_0 + \beta_1(1) + \beta_2 \text{ Speed} + e \\ &= (\beta_0 + \beta_1) + \beta_2 \text{ Speed} + e. \end{aligned}$$

This ANCOVA model fits two regression lines, one for each tool type, but restricts the slopes of the regression lines to be identical. To see this, let us focus on the interpretation of the regression coefficients. For the ANCOVA model,

β_2 = slope of population regression lines for tool types A and B.

and

β_0 = intercept of population regression line for tool A (called the reference group).

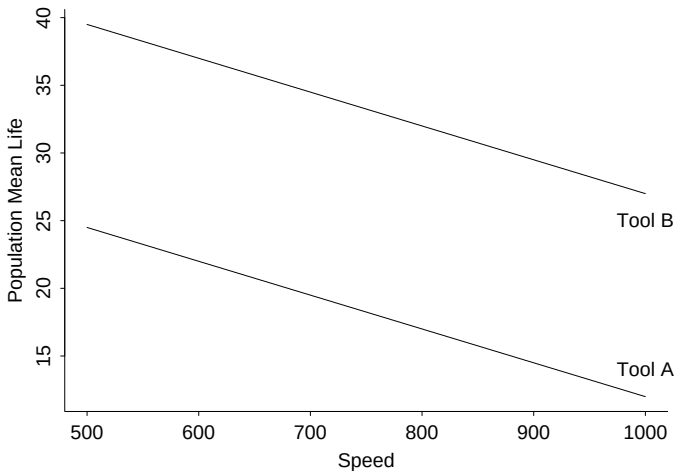
Given that

$\beta_0 + \beta_1$ = intercept of population regression line for tool B,

it follows that

β_1 = difference between tool B and tool A intercepts.

A picture of the population regression lines for one version of the model is given below.



An important feature of the ANCOVA model is that β_1 measures the difference in mean response for the tool types, regardless of the speed. A test of $H_0 : \beta_1 = 0$ is the primary interest, and is interpreted as a **comparison of the tool types, after adjusting or allowing for the speeds at which the tools were operated**.

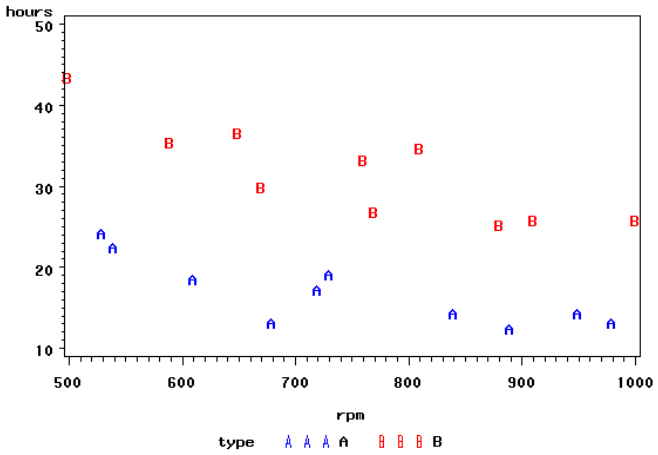
The following **SAS** program fits the ANCOVA model to the tool data.
Program editor contents:

```
proc print data=tools;
run;

* define v=symbol, c=color, i=interpolation, l=line style;
symbol1 v="A" c=blue  i=none; * i=r l=1;
symbol2 v="B" c=red   i=none; * i=r l=2;
proc gplot data=tools;
  plot hours*rpm=type;
run;

proc reg data=tools;
  model hours = toolb rpm;                  * Fit equal slopes model;
  output out=outglm p=p student=student;
run;

proc rank data=outglm out=outglm normal=blom;
  var student;
  ranks rankstr;
run;

* define v=symbol, c=color, i=interpolation, l=line style;
symbol1 v="A" c=blue  i=none;
symbol2 v="B" c=red   i=none;
proc gplot data=outglm;
  plot student*p=type rankstr*student=type;
run;
```

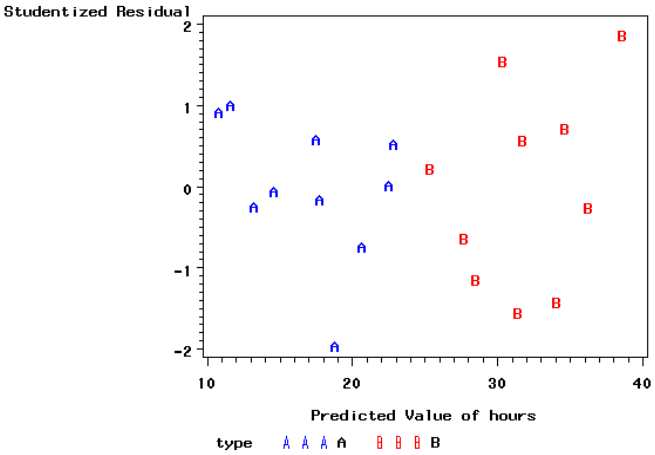
The ANCOVA model is palusible. The relationship between lifetime and speed is roughly linear within tool types, with similar slopes but unequal intercepts across groups. The plot of the studentized residuals against the fitted values shows no gross abnormalities, but suggests that the variability about the regression line for tool type A is somewhat smaller than the variability for tool type B. The model assumes that the variability of the responses is the same for each group. The rankit plot does not show any gross deviations from a straight line.

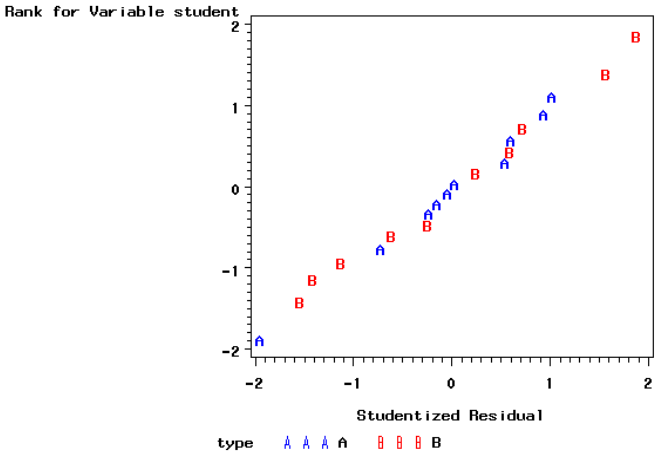
Output window contents:

Obs	hours	rpm	type	toolb	obs
1	18.73	610	A	0	1
2	14.52	950	A	0	2
3	17.43	720	A	0	3
4	14.54	840	A	0	4
5	13.44	980	A	0	5
6	24.39	530	A	0	6
7	13.34	680	A	0	7
8	22.71	540	A	0	8
9	12.68	890	A	0	9
10	19.32	730	A	0	10
11	30.16	670	B	1	11
12	27.09	770	B	1	12
13	25.40	880	B	1	13
14	26.05	1000	B	1	14
15	33.49	760	B	1	15
16	35.62	590	B	1	16
17	26.07	910	B	1	17
18	36.78	650	B	1	18
19	34.95	810	B	1	19
20	43.67	500	B	1	20




```
The REG Procedure
Dependent Variable: hours

Number of Observations Read          20
Number of Observations Used          20

                              Analysis of Variance
                                         Sum of           Mean
Source                     DF           Squares         Square    F Value    Pr > F
Model                       2     1418.03362       709.01681      76.75    <.0001
Error                      17      157.05456         9.23850
Corrected Total            19     1575.08818

Root MSE               3.03949    R-Square     0.9003
Dependent Mean        24.51900    Adj R-Sq     0.8886
Coeff Var             12.39647

                           Parameter Estimates
                                    Parameter       Standard
Variable      DF                     Estimate          Error    t Value    Pr > |t|
Intercept      1                     36.98560        3.51038      10.54      <.0001
toolb          1                     15.00425        1.35967      11.04      <.0001
rpm            1                     -0.02661        0.00452      -5.89      <.0001
```





The fitted relationship for the combined data set is

$$\text{Predicted Lifetime} \;=\; 36.99 + 15.00\ \text{Toolb} - 0.0266\ \text{Speed}.$$

Assigning the LS estimates to the appropriate parameters, the fitted relationships for the two tool types must be, for tool type B:

$$\begin{aligned} \text{Predicted Lifetime} &= (36.99 + 15.00) - 0.0266 \text{ Speed} \\ &= 51.99 - 0.0266 \text{ Speed}, \end{aligned}$$

and for tool type A:

$$\text{Predicted Lifetime} = 36.99 - 0.0266 \text{ Speed}.$$

The t-test of $H_0 : \beta_1 = 0$ checks whether the intercepts for the population regression lines are equal, assuming equal slopes. The t-test p-value < 0.0001 suggests that the population regression lines for tools A and B have unequal intercepts. The LS lines indicate that the average lifetime of either type tool decreases by 0.0266 hours for each increase in 1 RPM. Regardless of the lathe speed, the model predicts that type B tools will last 15 hours longer (i.e., the regression coefficient for the Toolb predictor) than type A tools. Summarizing this result another way, the $t-$test suggests that there is a significant difference between the lifetimes of the two tool types, after adjusting for the effect of the speeds at which the tools were operated. The estimated difference in average lifetime is 15 hours, regardless of the lathe speed.

7.2 Generalizing the ANCOVA Model to Allow Unequal Slopes

I will present a flexible approach for checking equal slopes and equal intercepts in ANCOVA-type models. The algorithm also provides a way to build regression models in studies where the primary interest is comparing the regression lines across groups rather than comparing groups after adjusting for a regression effect. The approach can be applied to an arbitrary number of groups and predictors. For simplicity, I will consider a problem with three groups and a single regression effect.

The data given in the **SAS** program on the following pages are the IQ scores of identical twins, one raised in a foster home (IQF) and the other raised by natural parents (IQN). The 27 pairs are divided into three groups by social status of the natural parents (H=high, M=medium, L=low). I will examine the regression of IQF on IQN for each of the three social classes[1].

There is no **a priori** reason to assume that the regression lines for the three groups have equal slopes or equal interepts. These are, however, reasonable hypotheses to examine. The easiest way to check these hypotheses is to fit a multiple regression model to the combined data set, and check whether certain carefully defined regression effects are zero. The most general model has six parameters, and corresponds to fitting a simple linear regression model to the three groups separately ($3 \times 2 = 6$).

Two indicator variables are needed to uniquely identify each observation by social class. For example, let $I1 = 1$ for H status families and $I1 = 0$ otherwise, and let $I2 = 1$ for M status families and $I2 = 0$ otherwise. The indicators $I1$ and $I2$ jointly assume 3 values:

Status	$I1$	$I2$
L	0	0
M	0	1
H	1	0

[1]The data were originally analyzed by Sir Cyril Burt.

Given the indicators $I1$ and $I2$ and the predictor IQN define two **interaction** or **product effects**: $P4 = I1 * IQN$ and $P5 = I2 * IQN$.

The most general model allows separate slopes and intercepts for each group:

$$IQF = \beta_0 + \beta_1 I1 + \beta_2 I2 + \beta_3 IQN + \beta_4 P4 + \beta_5 P5 + e. \tag{7.1}$$

Note that Ij $(j = 1, 2)$ is the **I**ndicator variable aligned with β_j whereas Pk $(k = 4, 5)$ is the **P**roduct effect aligned with β_k.

This model is best understood by considering the three status classes separately. If status = L, then $I1 = I2 = P4 = P5 = 0$. For these families

$$IQF = \beta_0 + \beta_3 IQN + e.$$

If status = M, then $I1 = P4 = 0$, whereas $I2 = 1$ and $P5 = IQN$. For these families

$$\begin{aligned} IQF &= \beta_0 + \beta_2(1) + \beta_3 IQN + \beta_5 IQN + e \\ &= (\beta_0 + \beta_2) + (\beta_3 + \beta_5) IQN + e. \end{aligned}$$

Finally, if status = H, then $I2 = P5 = 0$, whereas $I1 = 1$ and $P4 = IQN$. For these families

$$\begin{aligned} IQF &= \beta_0 + \beta_1(1) + \beta_3 IQN + \beta_4 IQN + e \\ &= (\beta_0 + \beta_1) + (\beta_3 + \beta_4) IQN + e. \end{aligned}$$

The regression coefficients β_0 and β_3 are the intercept and slope for the L status population regression line. The other parameters measure differences in intercepts and slopes across the three groups, using L status families as a **baseline** or **reference group**. In particular:

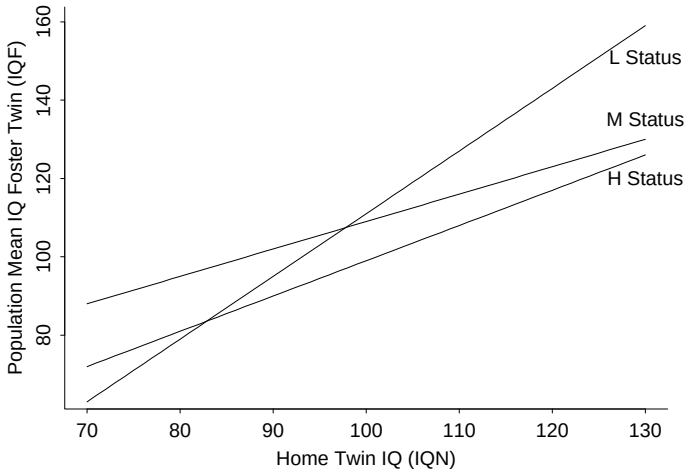
β_1 = difference between the intercepts of the H and L population regression lines.

β_2 = difference between the intercepts of the M and L population regression lines.

β_4 = difference between the slopes of the H and L population regression lines.

β_5 = difference between the slopes of the M and L population regression lines.

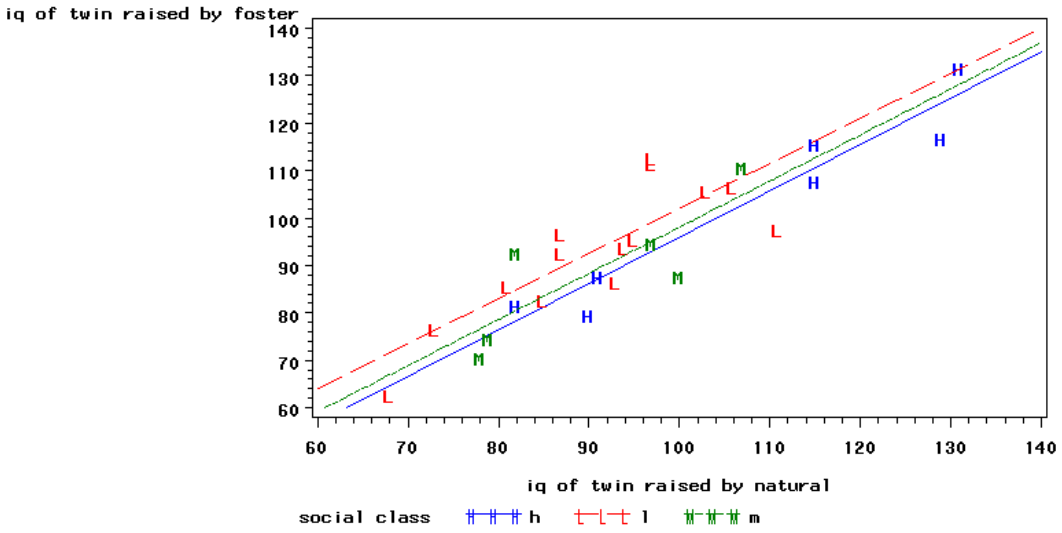
The plot gives a possible picture of the population regression lines corresponding to the general model (7.1).



A **SAS** program and output from fitting the general model to the twins data is given below. For now, ignore the **test** statement.

Program editor contents:

```
* twins data;
options ls=79 nodate nocenter;

data twins;
  infile "F:\Dropbox\UNM\teach\ADA2_stat528\sas\ADA2_07_twins.dat";
  input IQF IQN status $;      * char $ ;
  if status='h' then i1=1; else i1=0;   * Create indicators ;
  if status='m' then i2=1; else i2=0;
  P4=i1*IQN;                            * Create interaction variables ;
  P5=i2*IQN;
  label IQF='iq foster'
        IQN='iq natural'
        status='social class';
run;

proc print data=twins;
run;

* define v=symbol, c=color, i=interpolation, l=line style;
symbol1 v="h" c=blue  i=r l=1;
symbol2 v="l" c=red   i=r l=2;
symbol3 v="m" c=green i=r l=3;
proc gplot data=twins;
  plot IQF*IQN=status;
run;

proc reg data=twins;
  model IQF = IQN i1 i2 P4 P5;    * Fit general model ;
  test P4=0, P5=0;              * Test that P4 and P5 can be omitted ;
run;
```



Output window contents:

Obs	IQF	IQN	status	i1	i2	P4	P5
1	82	82	h	1	0	82	0
2	80	90	h	1	0	90	0
3	88	91	h	1	0	91	0
4	108	115	h	1	0	115	0
5	116	115	h	1	0	115	0
6	117	129	h	1	0	129	0
7	132	131	h	1	0	131	0
8	71	78	m	0	1	0	78
9	75	79	m	0	1	0	79
10	93	82	m	0	1	0	82
11	95	97	m	0	1	0	97
12	88	100	m	0	1	0	100
13	111	107	m	0	1	0	107
14	63	68	l	0	0	0	0
15	77	73	l	0	0	0	0
16	86	81	l	0	0	0	0
17	83	85	l	0	0	0	0
18	93	87	l	0	0	0	0
19	97	87	l	0	0	0	0
20	87	93	l	0	0	0	0
21	94	94	l	0	0	0	0
22	96	95	l	0	0	0	0
23	112	97	l	0	0	0	0
24	113	97	l	0	0	0	0
25	106	103	l	0	0	0	0
26	107	106	l	0	0	0	0
27	98	111	l	0	0	0	0

```
The REG Procedure
Dependent Variable: IQF iq of twin raised by foster

Number of Observations Read          27
Number of Observations Used          27

                              Analysis of Variance
                                     Sum of           Mean
Source                    DF        Squares         Square    F Value    Pr > F
Model                      5     5407.19567     1081.43913      17.24    <.0001
Error                     21     1317.47100       62.73671
Corrected Total           26     6724.66667

Root MSE              7.92065    R-Square     0.8041
Dependent Mean       95.11111    Adj R-Sq     0.7574
Coeff Var             8.32779

                              Parameter Estimates
                              Parameter     Standard
Variable   Label        DF     Estimate        Error  t Value  Pr > |t|
Intercept  Intercept     1      7.20461     16.75126     0.43    0.6715
IQN        iq natural    1      0.94842      0.18218     5.21    <.0001
i1                       1     -9.07665     24.44870    -0.37    0.7142
i2                       1     -6.38859     31.02087    -0.21    0.8388
P4                       1      0.02914      0.24458     0.12    0.9063
P5                       1      0.02414      0.33933     0.07    0.9439

        Test 1 Results for Dependent Variable IQF
                                  Mean
Source              DF           Square    F Value    Pr > F
Numerator            2          0.46591       0.01    0.9926
Denominator         21         62.73671
```

The natural way to express the fitted model is to give separate prediction equations for the three status groups. Here is an easy way to get the separate fits. For the general model (7.1), the predicted IQF satisfies

Predicted IQF = (Intercept + Coeff for Status Indicator) + (Coeff for IQN + Coeff for Status Product

For the baseline group, use 0 as the coefficients for the status indicator and product effect.

Thus, for the baseline group with status = L,

$$\begin{aligned} \text{Predicted IQF} &= 7.20 + 0 + (0.948 + 0)\text{IQN} \\ &= 7.20 + 0.948\text{IQN}. \end{aligned}$$

For the M status group with indicator $I2$ and product effect $P5$:

$$\begin{aligned} \text{Predicted IQF} &= 7.20 - 6.39 + (0.948 + 0.024)\text{IQN} \\ &= 0.81 + 0.972\text{IQN}. \end{aligned}$$

For the H status group with indicator $I1$ and product effect $P4$:

$$\begin{aligned} \text{Predicted IQF} &= 7.20 - 9.08 + (0.948 + 0.029)\text{IQN} \\ &= -1.88 + 0.977\text{IQN}. \end{aligned}$$

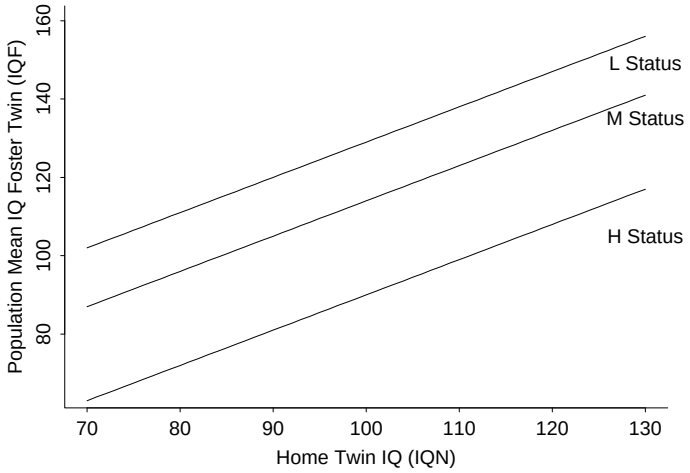
The LS lines are identical to separately fitting simple linear regressions to the three groups.

Equal slopes ANCOVA model

There are three other models of potential interest besides the general model. The **equal slopes** or ANCOVA model

$$IQF = \beta_0 + \beta_1 I1 + \beta_2 I2 + \beta_3 IQN + e$$

is a special case of (7.1) with $\beta_4 = \beta_5 = 0$. In the ANCOVA model, β_3 is the slope for all three regression lines. The other parameters have the same interpretation as in the general model (7.1), see the plot above. Output from the ANCOVA model is given below.



Program editor contents:

```
* twins equal slopes model;
proc reg data=twins;
  model IQF = IQN i1 i2;  * Fit equal slopes model ;
  test i1=0, i2=0;      * Test that I1 and I2 can be omitted ;
run;
```

Output window contents:

```
The REG Procedure
Model: MODEL1
Dependent Variable: IQF iq foster

Number of Observations Read          27
Number of Observations Used          27

                          Analysis of Variance
                                   Sum of           Mean
Source                 DF         Squares         Square    F Value    Pr > F
Model                   3      5406.26386     1802.08795      31.44    <.0001
Error                  23      1318.40281       57.32186
Corrected Total        26      6724.66667

Root MSE              7.57112    R-Square     0.8039
Dependent Mean       95.11111    Adj R-Sq     0.7784
Coeff Var             7.96029

                          Parameter Estimates
                             Parameter      Standard
Variable    Label        DF      Estimate         Error   t Value   Pr > |t|
Intercept   Intercept     1       5.61883       9.96285      0.56     0.5782
IQN         iq natural    1       0.96581       0.10695      9.03     <.0001
i1                        1      -6.22643       3.91708     -1.59     0.1256
i2                        1      -4.19109       3.69512     -1.13     0.2684
```

F-test for equal intercepts, assuming equal slopes:

```
        Test 1 Results for Dependent Variable IQF
                                   Mean
Source              DF         Square    F Value    Pr > F
Numerator            2       87.56529       1.53    0.2383
Denominator         23       57.32186
```

For the ANCOVA model, the predicted IQF for the three groups satisfies

$$\text{Predicted IQF} \;=\; (\text{Intercept} + \text{Coeff for Status Indicator}) + (\text{Coeff for IQN}) \times \text{IQN}.$$

As with the general model, use 0 as the coefficients for the status indicator and product effect for the baseline group.

For L status families:

$$\text{Predicted IQF} \;=\; 5.62 + 0.966\text{IQN},$$

for M status:

$$\begin{aligned}\text{Predicted IQF} &= 5.62 - 4.19 + 0.966 * \text{IQN} \\ &= 1.43 + 0.966 * \text{IQN},\end{aligned}$$

and for H status:

$$\begin{aligned}\text{Predicted IQF} &= 5.62 - 6.23 + 0.966\text{IQN} \\ &= -0.61 + 0.966 * \text{IQN}.\end{aligned}$$

Equal slopes and equal intercepts ANCOVA model

The model with **equal slopes** and **equal intercepts**

$$\text{IQF} \;=\; \beta_0 + \beta_3\text{IQN} + e$$

is a special case of the ANCOVA model with $\beta_1 = \beta_2 = 0$. This model does not distinguish among social classes. The common intercept and slope for the social classes are β_0 and β_3, respectively.

The predicted IQF for this model is

$$\text{IQF} \quad = \quad 9.21 + 0.901 * \text{IQN}$$

for each social class. Output for this model is given below.

Program editor contents:

```
* twins equal intercepts and slopes model;
proc reg data=twins;
  model IQF = IQN;  * fit equal intercepts and slopes model ;
run;
```

Output window contents:

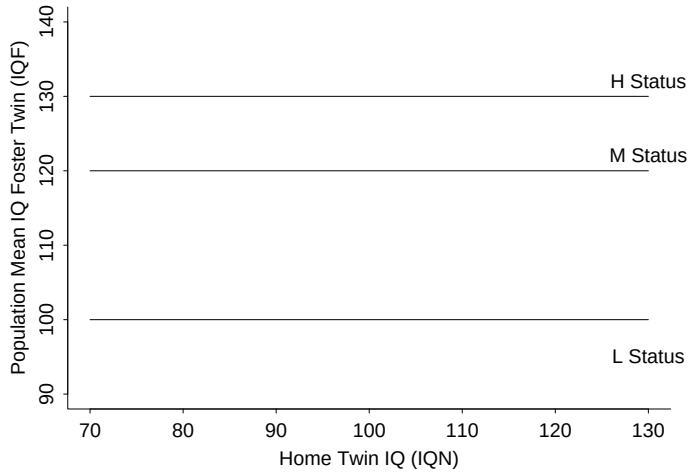
```
The REG Procedure
Model: MODEL1
Dependent Variable: IQF iq foster

Number of Observations Read          27
Number of Observations Used          27

                            Analysis of Variance
                                    Sum of           Mean
Source                     DF      Squares         Square    F Value    Pr > F
Model                       1   5231.13328   5231.13328      87.56    <.0001
Error                      25   1493.53339     59.74134
Corrected Total            26   6724.66667

Root MSE              7.72925    R-Square     0.7779
Dependent Mean       95.11111    Adj R-Sq     0.7690
Coeff Var             8.12655

                           Parameter Estimates
                               Parameter      Standard
Variable    Label        DF     Estimate         Error   t Value   Pr > |t|
Intercept   Intercept     1      9.20760       9.29990      0.99     0.3316
IQN         iq natural    1      0.90144       0.09633      9.36     <.0001
```

No slopes, but intercepts ANCOVA model

The model with **no predictor** (IQN) effects

$$\text{IQF} \quad = \quad IQF = \beta_0 + \beta_1 I1 + \beta_2 I2 + e$$

is a special case of the ANCOVA model with $\beta_3 = 0$. In this model, social status has an effect on IQF but IQN does not. This model of **parallel regression lines** with **zero slopes** is identical to a one-way ANOVA model for the three social classes, where the intercepts play the role of the population means, see the plot below.



For the ANOVA model, the predicted IQF for the three groups satisfies

$$\text{Predicted IQF} \quad = \quad \text{Intercept} + \text{Coeff for Status Indicator}$$

Again, use 0 as the coefficients for the baseline status indicator.

For L status families:

$$\text{Predicted IQF} \quad = \quad 93.71,$$

for M status:

$$\begin{aligned} \text{Predicted IQF} &= 93.71 - 4.88 \\ &= 88.83, \end{aligned}$$

and for H status:

$$\begin{aligned} \text{Predicted IQF} &= 93.71 + 9.57 \\ &= 103.28. \end{aligned}$$

The predicted IQFs are the mean IQFs for the three groups.

Program editor contents:

```
* twins no slope, different intercept model ;
proc reg data=twins;
  model IQF = i1 i2;
  test i1=0, i2=0;
run;
```

Output window contents:

```
The REG Procedure
Model: MODEL1
Dependent Variable: IQF iq foster

Number of Observations Read          27
Number of Observations Used          27

                          Analysis of Variance
                                   Sum of           Mean
Source                  DF       Squares         Square    F Value    Pr > F
Model                    2     731.54762      365.77381       1.46    0.2511
Error                   24    5993.11905      249.71329
Corrected Total         26    6724.66667

Root MSE              15.80232    R-Square     0.1088
Dependent Mean        95.11111    Adj R-Sq     0.0345
Coeff Var             16.61459

                       Parameter Estimates
                          Parameter       Standard
Variable    Label       DF      Estimate         Error   t Value   Pr > |t|
Intercept   Intercept    1      93.71429       4.22335     22.19     <.0001
i1                       1       9.57143       7.31505      1.31     0.2031
i2                       1      -4.88095       7.71074     -0.63     0.5327

        Test 1 Results for Dependent Variable IQF
                                  Mean
Source              DF          Square    F Value    Pr > F
Numerator            2      365.77381       1.46    0.2511
Denominator         24      249.71329
```

7.2.1 Relating Models to Two-Factor ANOVA

Recall the multiple regression formulation of the general model:

$$IQF = \beta_0 + \beta_1 I1 + \beta_2 I2 + \beta_3 IQN + \beta_4 P4 + \beta_5 P5 + e.$$

If you think of β_0 as a grand mean, $\beta_1 I1 + \beta_2 I2$ as the status effect (i.e., the two indicators $I1$ and $I2$ allow you to differentiate among social classes), $\beta_3 IQN$ as the IQN effect and $\beta_4 P4 + \beta_5 P5$ as the status by IQN interaction, then you can represent the model as

IQF = Grand Mean + Status Effect + IQN effect + Status*IQN interaction + Residual.

This representation has the same form as a two-factor ANOVA model with interaction, except that IQN is a quantitative effect rather than a qualitative (i.e., categorical) effect. The general model has the same structure as a two-factor interaction ANOVA model because the plot of the population means allows non-parallel profiles. However, the general model is a special case of the two-factor interaction ANOVA model because it restricts the means to change linearly with IQN.

The ANCOVA model (7.2) has main effects for status and IQN but no interaction:

$$\text{IQF} \;=\; \text{Grand Mean} + \text{Status Effect} + \text{IQN effect} + \text{Residual}. \tag{7.2}$$

The ANCOVA model is a special case of the additive two-factor ANOVA model because the plot of the population means has parallel profiles, but is not equivalent to the additive two-factor ANOVA model.

The model with equal slopes and intercepts (7.3) has no main effect for status nor an interaction between status and IQN:

$$\text{IQF} \quad = \quad \text{Grand Mean} + \text{IQN effect} + \text{Residual}. \tag{7.3}$$

The one-way ANOVA model (7.4) has no main effect for IQN nor an interaction between status and IQN:

$$\text{IQF} \quad = \quad \text{Grand Mean} + \text{Status Effect} + \text{Residual}. \tag{7.4}$$

I will expand on these ideas later, as they are useful for understanding the connections between regression and ANOVA models.

7.2.2 Choosing Among Models

I will suggest a backward sequential method to select which of models (7.1), (7.2), and (7.3) fits best. You would typically be interested in the one-way ANOVA model (7.4) only when the effect of IQN was negligible.

Step 1: Fit the full model (7.1) and test the hypothesis of equal slopes $H_0 : \beta_4 = \beta_5 = 0$. (aside: t-tests are used to test **either** $\beta_4 = 0$ or $\beta_5 = 0$.) To test H_0, eliminate the predictor variables $P4$ and $P5$ associated with β_4 and β_5 from the full model (7.1). Then fit the reduced model (7.2) with equal slopes. Reject $H_0 : \beta_4 = \beta_5 = 0$ if the increase in the Residual SS obtained by deleting $P4$ and $P5$ from the full model is significant. Formally, compute the $F-$statistic:

$$F_{obs} = \frac{\text{(ERROR SS for reduced model - ERROR SS for full model)}/2}{\text{ERROR MS for full model}}$$

and compare it to an upper-tail critical value for an $F-$distribution with 2 and *df* degrees of freedom, where *df* is the Residual *df* for the full model. The F-test is a direct extension of the single degree-of-freedom $F-$tests in the stepwise fits. A p-value for F-test is obtained in **SAS** using the `test P4=0, P5=0;` statement with the fit to the full model; see code starting on on page 141. If H_0 is rejected, stop and conclude that the population regression lines have different slopes (and then I do not care whether the intercepts are equal). Otherwise, proceed to

Step 2: Fit the equal slopes or ANCOVA model (7.2) and test for equal intercepts $H_0 : \beta_1 = \beta_2 = 0$. Follow the procedure outlined in Step 1, treating the ANCOVA model as the full model and the model $IQF = \beta_0 + \beta_3 IQN + e$ with equal slopes and intercepts as the reduced model. In **SAS**, the `test I1=0, I2=0;` statement with the ANCOVA **reg** proc gives a p-value for the test. If H_0 is rejected, conclude that that population regression lines are parallel with unequal intercepts. Otherwise, conclude that regression lines are identical.

Step 3: Estimate the parameters under the appropriate model, and conduct a diagnostic analysis. Summarize the fitted model by status class.

A comparison of regression lines across $k > 3$ groups requires $k - 1$ indicator variables to define the groups, and $k - 1$ interaction variables, assuming the model has a single predictor. The

comparison of models mimics the discussion above, except that the numerator of the $F-$statistic is divided by $k-1$ instead of 2, and the numerator *df* for the $F-$test is $k-1$ instead of 2. If $k=2$, the $F-$tests for comparing the three models are equivalent to $t-$tests given with the parameter estimates summary. For example, recall how you tested for equal intercepts in the tools problems.

The plot of the twins data shows fairly linear relationships within each social class. The linear relationships appear to have similar slopes and similar intercepts. The p-value for testing the hypothesis that the slopes of the population regression lines are equal is essentially 1. The observed data are consistent with the reduced model of equal slopes.

The p-value for comparing the model of equal slopes and equal intercepts to the ANCOVA model is 0.238, so there is insufficient evidence to reject the reduced model with equal slopes and intercepts. The estimated regression line, regardless of social class, is:

$$\text{Predicted IQF} = 9.21 + 0.901\text{*IQN}.$$

There are no serious inadequacies with this model, based on a diagnostic analysis (not shown).

An interpretation of this analysis is that the natural parents' social class has no impact on the relationship between the IQ scores of identical twins raised apart. What other interesting features of the data would be interesting to explore? For example, what values of the intercept and slope of the population regression line are of intrinsic interest?

7.2.3 Comments on Comparing Regression Lines

In the twins example, I defined two indicator variables (plus two interaction variables) from an ordinal categorical variable: status (H, M, L). Many researchers would assign numerical codes to the status groups and use the coding as a predictor in a regression model. For status, a "natural" coding might be to define NSTAT=0 for L, 1 for M, and 2 for H status families. This suggests building a multiple regression model with a single status variable (i.e., single df):

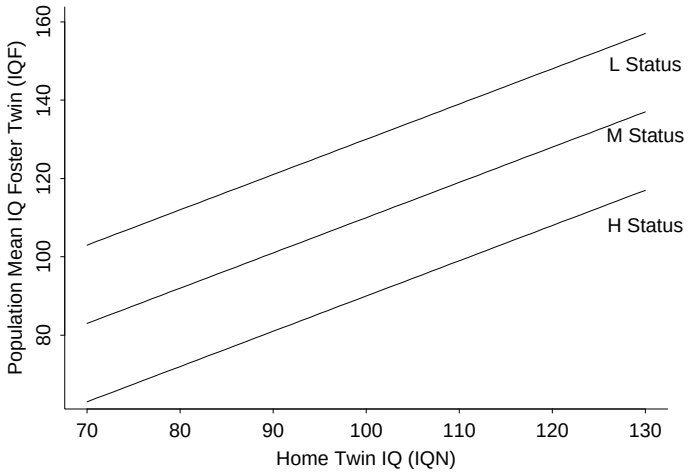
$$IQF = \beta_0 + \beta_1 IQN + \beta_2 \text{NSTAT} + e.$$

If you consider the status classes separately, the model implies that

$$\begin{aligned}
IQF &= \beta_0 + \beta_1 IQN + \beta_2(0) + e = \beta_0 + \beta_1 IQN + e && \text{for L status families,}\\
IQF &= \beta_0 + \beta_1 IQN + \beta_2(1) + e = (\beta_0 + \beta_2) + \beta_1 IQN + e && \text{for M status families,}\\
IQF &= \beta_0 + \beta_1 IQN + \beta_2(2) + e = (\beta_0 + 2\beta_2) + \beta_1 IQN + e && \text{for H status families.}
\end{aligned}$$

The model assumes that the IQF by IQN regression lines are parallel for the three groups, and are separated by a constant β_2. This model is more restrictive (and less reasonable) than the ANCOVA model with equal slopes but arbitrary intercepts. Of course, this model is a easier to work with because it requires keeping track of only one status variable instead of two status indicators.

A plot of the population regression lines under this model is given above, assuming $\beta_2 < 0$.



7.3 Using GLM to Fit Models

This section illustrates how to do what was done above in proc **glm** *and using LSMEANS. Please read on your own.*

glm can be used to fit regression models with **factors** (qualitative or categorical characteristics) as effects, measurements as **predictors**, or mixtures of the two types of effects. Variables to be treated as factors are specified in the **class** statement. In the twins study, **IQN** is a **predictor** and **status** is a **factor**. The twins ANCOVA model is a regression model with one predictor (IQN) and one factor, (status). The unequal slopes or general model is a regression model with one predictor (IQN), one factor (status), plus an interaction between the IQN predictor and the status factor.

glm is easier to use than **reg** when fitting models with factors because you do not need to create the indicators and the interaction effects. However, the output is a bit more difficult to interpret, and **glm** produces less diagnostic information than **reg**.

A factor is included in a model by adding one indicator variable for each level of the factor. This induces perfect collinearity among the variables whenever the model includes an intercept (assumed by **SAS**). To eliminate this problem without changing the **interpretation** of the model, you can exclude any one of the indicators from the model. To get parameter estimates, **glm** orders the levels of the factor alphanumerically, then creates the indicators internally and omits the indicator corresponding to the last ordered level. The last ordered level then plays the role of a **baseline** or **reference** group.

In the **SAS** data step below, I created the **sta** factor for the twins data so that the last ordered level (**clow**) identified the L status families. Any **glm** model with **sta** as a factor will then include indicators for the H and M status groups, as in our earlier **reg** analyses. If I used **status** as a factor, instead of **sta**, then the M status families would be the baseline group. The choice of a baseline groups has no important impact on the analysis.

SAS glm programs and output from fitting models (7.1) – (7.4) to the twins data are given below. The first model considered is a one-way ANOVA comparing the three social classes (identified

by the levels of **sta**) on the response variable IQF. The ANOVA model is not a primary interest here, but it is given to emphasize that a regression model with a single factor **sta** is a one-way ANOVA model.

ANOVA model

The **lsmeans** output for the one-way ANOVA model gives the average IQFs for the three status groups. The **stderr** option gives standard errors. You should compare these to the summaries given on page 147.

Program editor contents:

```
* twins data, anova model using class statement and lsmeans ;
options ls=79 nodate nocenter;

data twins;
  infile "F:\Dropbox\UNM\teach\ADA2_stat528\sas\ADA2_07_twins.dat";
  input IQF IQN status $;      * char $ ;
  if status = 'h' then sta = 'ahi';
    else if status = 'm' then sta = 'bmed';
    else sta = 'clow';
run;

proc glm data=twins;
  class sta;
  model IQF = sta;          * ANOVA model ;
  lsmeans sta/stderr pdiff;
run;
```

Output window contents:

```
The GLM Procedure
      Class Level Information
Class         Levels    Values
sta                3    ahi bme clo

Number of Observations Read          27
Number of Observations Used          27

Dependent Variable: IQF
                                    Sum of
Source                      DF        Squares    Mean Square   F Value   Pr > F
Model                        2     731.547619     365.773810      1.46   0.2511
Error                       24    5993.119048     249.713294
Corrected Total             26    6724.666667

R-Square     Coeff Var      Root MSE      IQF Mean
0.108786      16.61459      15.80232      95.11111

Source                      DF      Type I SS    Mean Square   F Value   Pr > F
sta                          2    731.5476190    365.7738095      1.46   0.2511

Source                      DF    Type III SS    Mean Square   F Value   Pr > F
sta                          2    731.5476190    365.7738095      1.46   0.2511
```

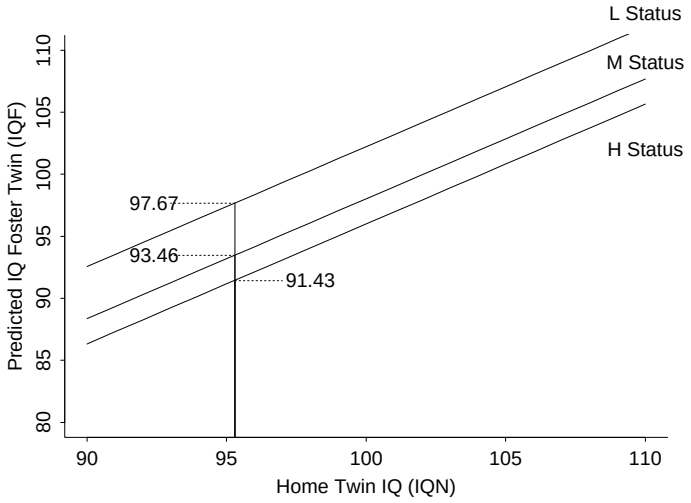
```
Least Squares Means
                              Standard                      LSMEAN
sta       IQF LSMEAN             Error    Pr > |t|         Number
ahi       103.285714          5.972715      <.0001              1
bme        88.833333          6.451270      <.0001              2
clo        93.714286          4.223347      <.0001              3

      Least Squares Means for effect sta
     Pr > |t| for H0: LSMean(i)=LSMean(j)
           Dependent Variable: IQF
i/j              1              2              3
   1                        0.1132         0.2031
   2         0.1132                        0.5327
   3         0.2031         0.5327

NOTE: To ensure overall protection level, only probabilities associated with
      pre-planned comparisons should be used.
```

Equal slopes model

The **equal slopes** or ANCOVA model is obtained by adding the predictor IQN to the ANOVA model. Output for the ANCOVA model is given below. The **solution** option on the **model** statement gives the parameter estimates. I omitted this option with the ANOVA model.

The **glm parameter estimates table** gives the estimated intercept (for the reference group), the coefficient for IQN and the coefficients for the H and M status indicators. A zero is reported as the coefficient for the baseline, or L status group. These summaries agree with the **reg** output for the ANCOVA model given on page 144. The note following the parameter estimates table means that the ANCOVA model is overparameterized (has one or more collinearities), and that a specific set of constraints was used to eliminate the problem. **glm**'s constraint is to zero out the L status indicator, as described above.

The F-statistic for the **sta** factor based on the Type III SS tests the simultaneous importance of the two status indicators in the ANCOVA model. The reduced model with no **sta** factor fits equal slopes and intercepts to the three groups. Thus, the F-test on **sta** gives the p-value or significance level for testing equal intercepts across the three regression lines, assuming equal slopes. To see this, compare the output to page 144.



lsmeans gives the estimated average response IQF for the three social classes when IQN is at the average IQN-value in the data set: $\overline{IQN} = 95.3$. The **lsmeans** are the predicted responses for the three regression lines at $\overline{IQN} = 95.3$. To better visualize the **lsmeans** output for the ANCOVA model, recall that for L status families:

$$\text{Predicted IQF} = 5.62 + 0.966\text{IQN},$$

for M status:

$$\begin{aligned}\text{Predicted IQF} &= 5.62 - 4.19 + 0.966 * \text{IQN} \\ &= 1.43 + 0.966 * \text{IQN},\end{aligned}$$

and for H status:

$$\begin{aligned}\text{Predicted IQF} &= 5.62 - 6.23 + 0.966\text{IQN} \\ &= -0.61 + 0.966 * \text{IQN}.\end{aligned}$$

At $\overline{IQN} = 95.3$, the estimated mean IQFs are 97.656, 93.465, and 91.430 for the L, M, and H status groups.

Because the fitted lines are parallel, the difference in any pair of **lsmeans** is the estimated mean difference in IQFs at any IQN for which the ANCOVA model is reasonable. Furthermore, the difference in **lsmeans** is the difference between the intercepts for the two regression lines in the comparison. For example, the H minus L **lsmeans** of $91.430 - 97.656 = -6.226$ is the difference between the intercepts for the H and L status regression lines. This is the estimated coefficient of the H indicator variable.

Program editor contents:

```
* twins data, equal slopes model using class statement and lsmeans ;
proc glm data=twins;
  class sta;
  model IQF = IQN sta/solution;  * equal slopes model ;
  lsmeans sta/stderr pdiff;
run;
```

Output window contents:

```
The GLM Procedure
      Class Level Information
Class          Levels    Values
sta                 3    ahi bme clo

Number of Observations Read          27
Number of Observations Used          27

Dependent Variable: IQF
                                        Sum of
Source                      DF         Squares    Mean Square   F Value   Pr > F
Model                        3    5406.263856    1802.087952     31.44   <.0001
Error                       23    1318.402810      57.321861
Corrected Total             26    6724.666667

R-Square     Coeff Var      Root MSE      IQF Mean
0.803945      7.960290      7.571120      95.11111

Source                      DF       Type I SS    Mean Square   F Value   Pr > F
IQN                          1    5231.133275    5231.133275     91.26   <.0001
sta                          2     175.130581      87.565291      1.53   0.2383

Source                      DF     Type III SS    Mean Square   F Value   Pr > F
IQN                          1    4674.716237    4674.716237     81.55   <.0001
sta                          2     175.130581      87.565291      1.53   0.2383

                                                    Standard
Parameter                 Estimate                     Error    t Value    Pr > |t|
Intercept           5.618826023 B          9.96284527          0.56      0.5782
IQN                 0.965807702            0.10694813          9.03      <.0001
sta       ahi      -6.226425986 B          3.91708267         -1.59      0.1256
sta       bme      -4.191089737 B          3.69511811         -1.13      0.2684
sta       clo       0.000000000 B                   .             .           .

NOTE: The X'X matrix has been found to be singular, and a generalized inverse
      was used to solve the normal equations.  Terms whose estimates are
      followed by the letter 'B' are not uniquely estimable.

The GLM Procedure
Least Squares Means
                              Standard                      LSMEAN
sta       IQF LSMEAN             Error    Pr > |t|         Number
ahi       91.4302970        3.1483785      <.0001              1
bme       93.4656332        3.1331719      <.0001              2
clo       97.6567230        2.0700257      <.0001              3

      Least Squares Means for effect sta
     Pr > |t| for H0: LSMean(i)=LSMean(j)
           Dependent Variable: IQF
i/j              1             2             3
   1                       0.6617        0.1256
   2         0.6617                      0.2684
   3         0.1256        0.2684

NOTE: To ensure overall protection level, only probabilities associated with
      pre-planned comparisons should be used.
```

Interaction model

The third model considered is the **interaction model**, which fits unequal slopes and intercepts for the social groups. The **unequal slopes** model is obtained by adding the **IQN** $*$ **sta** interaction to the ANCOVA model. The unequal slopes model has H and M status indicators, and their interactions or products with IQN. The L status indicator and its product with IQN are omitted from the model to eliminate collinearity.

The effects for the L status or baseline group are zeroed out in the parameter estimates table. Except for this point, the parameter estimates table gives the same information as the **reg** fit to this model. See page 142.

The F-statistic for testing the **IQN*sta** interaction based on the Type III SS is the test for equal slopes, allowing for unequal intercepts. The p-value of 0.9926 agrees with the **reg** p-value for this test.

The **lsmeans** are the predicted responses for the three regression lines at $\overline{IQN}$. The comparison of groups using **lsmeans** applies only at $\overline{IQN}$, and not at other values of IQN. For the unequal slopes model, the size and sign of the status group effects depend on the value of IQN. Consequently, if this model were the best summary for the data, there would be no natural way to define the main effect of social status on IQF!

Program editor contents:

```
* twins data, unequal slopes model using class statement and lsmeans ;
proc glm data=twins;
  class sta;
  model IQF = IQN sta IQN*sta/solution;  * Unequal slopes model ;
  lsmeans sta/stderr pdiff;
run;
```

Output window contents:

```
The GLM Procedure
      Class Level Information
Class          Levels    Values
sta                 3    ahi bme clo

Number of Observations Read          27
Number of Observations Used          27

Dependent Variable: IQF
                                        Sum of
Source                      DF         Squares    Mean Square   F Value   Pr > F
Model                        5    5407.195668    1081.439134     17.24   <.0001
Error                       21    1317.470999      62.736714
Corrected Total             26    6724.666667

R-Square     Coeff Var      Root MSE      IQF Mean
0.804084      8.327787      7.920651      95.11111

Source                      DF       Type I SS    Mean Square   F Value   Pr > F
IQN                          1    5231.133275    5231.133275     83.38   <.0001
sta                          2     175.130581      87.565291      1.40   0.2697
IQN*sta                      2       0.931811       0.465906      0.01   0.9926

Source                      DF     Type III SS    Mean Square   F Value   Pr > F
IQN                          1    3717.792414    3717.792414     59.26   <.0001
sta                          2       8.988848       4.494424      0.07   0.9311
IQN*sta                      2       0.931811       0.465906      0.01   0.9926

                                                    Standard
Parameter                Estimate                      Error    t Value    Pr > |t|
Intercept          7.204609862 B     16.75126441        0.43     0.6715
IQN                0.948422445 B      0.18217507        5.21     <.0001
sta       ahi     -9.076653525 B     24.44870426       -0.37     0.7142
sta       bme     -6.388585477 B     31.02087438       -0.21     0.8388
sta       clo      0.000000000 B              .           .          .
IQN*sta   ahi      0.029139714 B      0.24458020        0.12     0.9063
IQN*sta   bme      0.024144505 B      0.33932723        0.07     0.9439
IQN*sta   clo      0.000000000 B              .           .          .

NOTE: The X'X matrix has been found to be singular, and a generalized inverse
      was used to solve the normal equations.  Terms whose estimates are
      followed by the letter 'B' are not uniquely estimable.
```

```
Least Squares Means
                              Standard                      LSMEAN
sta      IQF LSMEAN             Error    Pr > |t|          Number
ahi      91.2860095         3.6021139      <.0001               1
bme      93.4980526         3.5130406      <.0001               2
clo      97.5857562         2.2437012      <.0001               3

      Least Squares Means for effect sta
     Pr > |t| for H0: LSMean(i)=LSMean(j)
           Dependent Variable: IQF
i/j              1              2              3
   1                        0.6647         0.1525
   2         0.6647                        0.3379
   3         0.1525         0.3379

NOTE: To ensure overall protection level, only probabilities associated with
      pre-planned comparisons should be used.
```

Slope (SLR) model

The final model fitted to the data is the simple linear regression model we selected earlier, with equal intercepts and slopes for the status groups. The model does not include a status effect or an interaction between status and IQN. The output agrees with the **reg** output for this model given on page 146.

Program editor contents:

```
* twins data, slope model, no status effect ;
proc glm data=twins;
  model IQF = IQN /solution;  * model with no status effect ;
run;
```

Output window contents:

```
The GLM Procedure
Number of Observations Read          27
Number of Observations Used          27

Dependent Variable: IQF
                                    Sum of
Source                      DF        Squares    Mean Square   F Value   Pr > F
Model                        1    5231.133275    5231.133275     87.56   <.0001
Error                       25    1493.533392      59.741336
Corrected Total             26    6724.666667

R-Square     Coeff Var      Root MSE      IQF Mean
0.777902      8.126550      7.729252      95.11111

Source                      DF      Type I SS    Mean Square   F Value   Pr > F
IQN                          1    5231.133275    5231.133275     87.56   <.0001

Source                      DF    Type III SS    Mean Square   F Value   Pr > F
IQN                          1    5231.133275    5231.133275     87.56   <.0001

                                   Standard
Parameter         Estimate            Error    t Value    Pr > |t|
Intercept      9.207598840       9.29989643       0.99      0.3316
IQN            0.901436001       0.09633286       9.36      <.0001
```

7.3.1 GLM Output for the Tool Lifetime ANCOVA Model

glm output for the ANCOVA or equal slopes model fitted to the tool data is obtained by appending the following **SAS** procedure to the data step on page 133. Tool type B is the reference group here, whereas the earlier analysis had tool type A as the reference group. What aspects of the fit should be the same, and what features should change? A question for your spare time.

Program editor contents:

```
* tools data, equal slopes, different intercepts, using class statement and lsmeans ;
proc glm data=tools;
  class type;
  model hours = rpm type /solution;
  lsmeans type /stderr pdiff;
run;
```

Output window contents:

```
The GLM Procedure
   Class Level Information
Class         Levels    Values
type               2    A B

Number of Observations Read          20
Number of Observations Used          20

Dependent Variable: hours
                                    Sum of
Source                      DF        Squares    Mean Square   F Value   Pr > F
Model                        2    1418.033618     709.016809     76.75   <.0001
Error                       17     157.054562       9.238504
Corrected Total             19    1575.088180

R-Square     Coeff Var      Root MSE    hours Mean
0.900288      12.39647      3.039491      24.51900

Source                      DF      Type I SS    Mean Square   F Value   Pr > F
rpm                          1     293.005403     293.005403     31.72   <.0001
type                         1    1125.028214    1125.028214    121.78   <.0001

Source                      DF    Type III SS    Mean Square   F Value   Pr > F
rpm                          1     320.167998     320.167998     34.66   <.0001
type                         1    1125.028214    1125.028214    121.78   <.0001

                                          Standard
Parameter           Estimate                 Error    t Value    Pr > |t|
Intercept        51.98985181 B      3.54082203      14.68      <.0001
rpm              -0.02660723        0.00451972      -5.89      <.0001
type      A     -15.00425061 B      1.35966970     -11.04      <.0001
type      B       0.00000000 B      .                .         .

NOTE: The X'X matrix has been found to be singular, and a generalized inverse
      was used to solve the normal equations.  Terms whose estimates are
      followed by the letter 'B' are not uniquely estimable.

Least Squares Means
                                                              H0:LSMean1=
                             Standard    H0:LSMEAN=0          LSMean2
type    hours LSMEAN            Error       Pr > |t|          Pr > |t|
A         17.0168747       0.9613015         <.0001            <.0001
B         32.0211253       0.9613015         <.0001
```

Chapter 8

Polynomial Regression

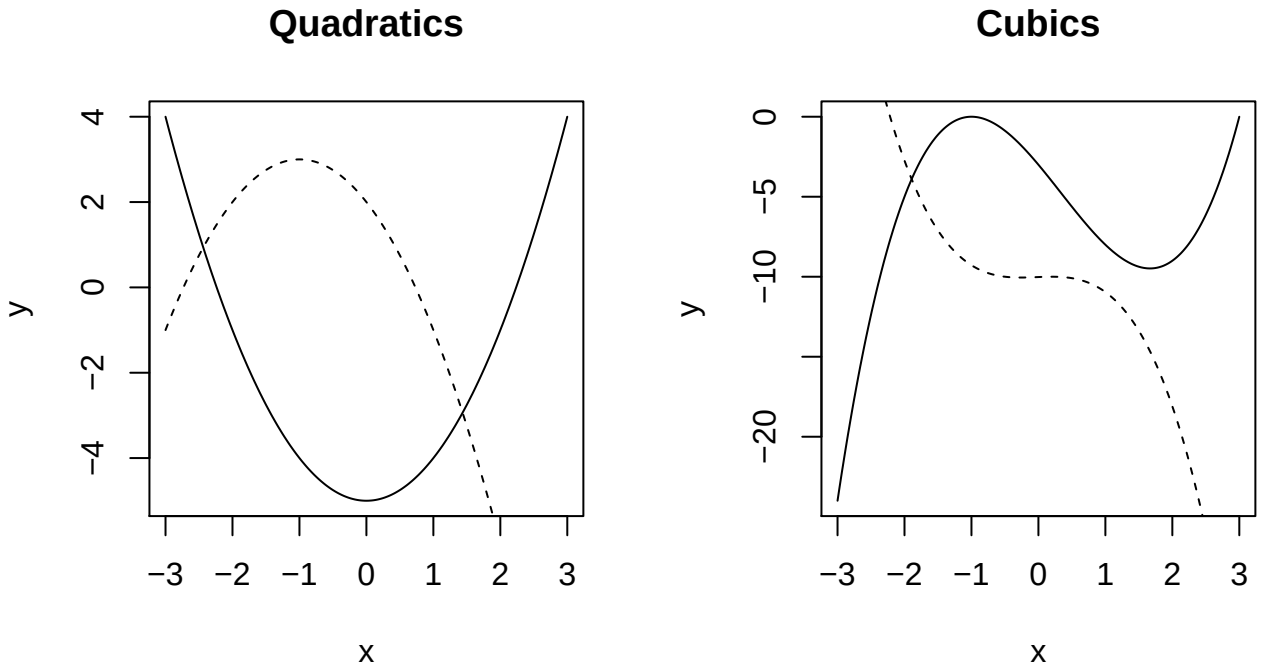
8.1 Polynomial Models with One Predictor

A p^{th} order polynomial model relating a dependent variable Y to a predictor X is given by

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_p X^p + \varepsilon.$$

This is a multiple regression model with predictors $X, X^2, \ldots, X^p$. For $p = 2$, 3, 4, and 5 we have quadratic, cubic, quartic and quintic relationships, respectively.

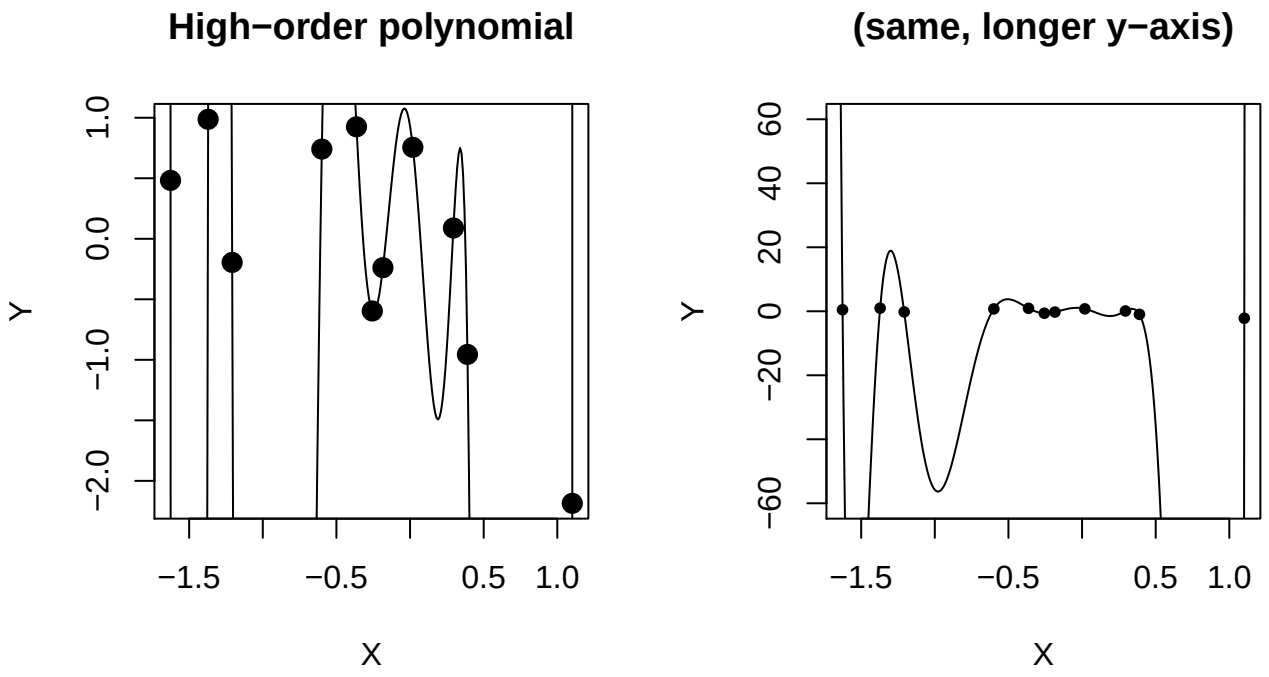
A second order polynomial (quadratic) allows at most one local maximum or minimum (i.e., a point where trend changes direction from increasing to decreasing, or from decreasing to increasing). A third order polynomial (cubic) allows at most two local maxima or minima. In general, a p^{th} order polynomial allows at most $p - 1$ local maxima or minima. The two panels below illustrate different quadratic and cubic relationships.



It is important to recognize that not all, or even none, of the "turning-points" in a polynomial may be observed if the range of X is suitably restricted.

Although polynomial models allow a rich class of non-linear relationships between Y and X (by virtue of Taylor's Theorem in calculus), some caution is needed when fitting polynomials. In particular, the extreme X-values can be highly influential, numerical instabilities occur when fitting high order models, and predictions based on high order polynomial models can be woeful.

To illustrate the third concern, consider a data set (Y_i, X_i) for $i = 1, 2, \ldots, n$ where the X_is are distinct. One can show mathematically that an $(n-1)^{st}$ degree polynomial will fit the observed data exactly. However, for a high order polynomial to fit exactly, the fitted curve must oscillate wildly between data points. In the picture below, I show the 10^{th} degree polynomial that fits exactly the 11 distinct data points. Although $R^2 = 1$, I would not use the fitted model to make predictions with new data. (If possible, models should always be validated using new data.) Intuitively, a quadratic or a lower order polynomial would likely be significantly better. In essence, the 10^{th} degree polynomial is modelling the variability in the data, rather than the trend.



Another concern is that the importance of lower order terms (i.e., $X, X^2, \ldots, X^{p-1}$) depends on the scale in which we measure X. For example, suppose for some chemical reaction,

$$\text{Time to reaction} = \beta_0 + \beta_1 \text{Temp} + \beta_2 \text{Temp}^2 + \varepsilon.$$

The significance level for the estimate of the Temp coefficient depends on whether we measure temperature in degrees Celsius or Fahrenheit.

To avoid these problems, I recommend

1. Center the X data at $\bar{X}$ and fit the model

$$Y = \beta_0 + \beta_1(X - \bar{X}) + \beta_2(X - \bar{X})^2 + \cdots + \beta_p(X - \bar{X})^p + \varepsilon.$$

This is usually important only for cubic and higher order models.

2. Restrict attention to low order models, say quartic or less. If a fourth order polynomial does not fit, a transformation may provide a more succinct summary.

3. Pay careful attention to diagnostics.

4. Add or delete variables using the natural hierarchy among powers of X and include all lower order terms if a higher order term is needed. For example, in a forward-selection type algorithm, add terms X, X^2, ..., sequentially until no additional term is significant, but do not delete powers that were entered earlier. Similarly, with a backward-elimination type algorithm, start with the model of maximum acceptable order (for example a fourth or third order polynomial) and consider deleting terms in the order X^p, X^{p-1}, ..., until no further terms can be omitted. The **select=backward** option in the **reg** procedure **does not allow** you to invoke the hierarchy principle with backward elimination. The backward option sequentially eliminates the least significant effect in the model, regardless of what other effects are included.

8.1.1 Example: Cloud point and percent I-8

The cloud point of a liquid is a measure of the degree of crystallization in a stock, and is measured by the refractive index [1]. It has been suggested that the percent of I-8 (variable "i8") in the base stock is an excellent predictor of the cloud point using a second order (quadratic) model:

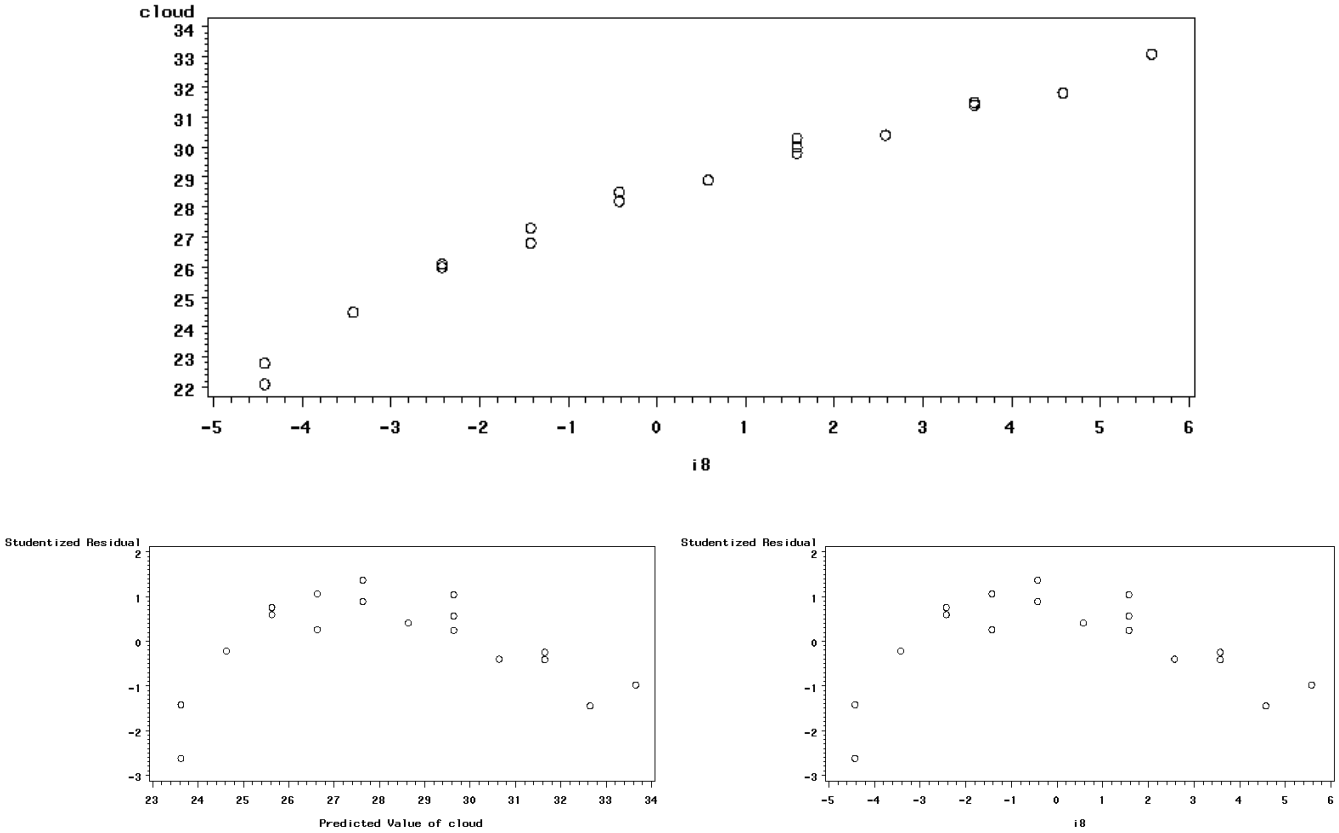
$$\text{Cloud point} = \beta_0 + \beta_1 \text{I8} + \beta_2 \text{I8}^2 + \varepsilon.$$

Data were collected to examine this model.
Program editor contents:

```
* Cloud point example;
options ls=79 nodate nocenter;
data cloudpoint;
  input i8 cloud @@;
  i8 = i8 - 4.42;       * center I8 using pre-calculated mean of I8 = 4.42 ;
  datalines;
0 22.1  1 24.5  2 26.0  3 26.8  4 28.2  5 28.9  6 30.0
7 30.4  8 31.4  2 26.1  4 28.5  6 30.3  8 31.5  0 22.8
3 27.3  6 29.8  9 31.8 10 33.1
;
run;

proc reg data=cloudpoint noprint;          * noprint - suppress printed output ;
  model cloud=i8;                          * simple linear regression
;
  output out=outreg p=p student=student;   * create predictions and residuals
;
run;

* define v=symbol, c=color, i=interpolation, l=line style;
symbol1 v=circle c=black  i=none;
proc gplot data=outreg;
   plot cloud*i8;
   plot student*p;
   plot student*i8;
run;
```

[1] Draper and Smith 1966, p. 162

Below are plots of the data and the studentized residuals.



The data plot is clearly nonlinear, suggesting that a simple linear regression model is inadequate. This is confirmed by a plot of the studentized residuals against the fitted values from a simple linear regression of Cloud point on i8. Also by the residuals against the i8 values. We do not see any local maxima or minima, so a second order model is likely to be adequate. To be sure, we will first fit a cubic model, and see whether the third order term is important.

The output below shows that the cubic term improves the fit of the quadratic model (i.e., the cubic term is important when added last to the model). The plot of the studentized residuals against the fitted values does not show any extreme abnormalities. Furthermore, no individual point is poorly fitted by the model. Case 1 has the largest studentized residual: $r_1 = -1.997$.

The first and last observations have the lowest and highest values of I8, given by 0 and 10, respectively. These cases are also the most influential points in the data set (largest Cook's D). If we delete these cases and redo the analysis (not shown) we find that the cubic term is no longer important (p-value=0.20) when added after the quadratic term. One may reasonably conclude that the significance of the cubic term in the original analysis is solely due to the two extreme I8 values, and that the quadratic model appears to fit well over the smaller range of $1 \leq \text{I8} \leq 9$.

Program editor contents:

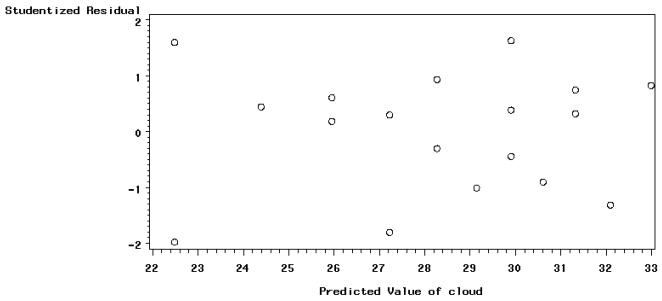
```
data cloudpoint;
  set cloudpoint;
  i8sq = i8**2;    * squared term ;
  i8cu = i8**3;    * cubed   term ;
;
```

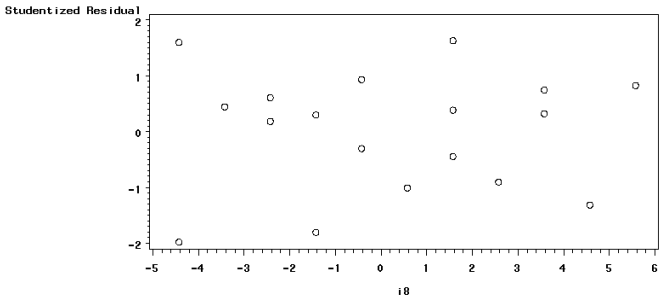
```
run;

proc reg data=cloudpoint;
  model cloud = i8 i8sq i8cu /r;          * cubic regression                  ;
  output out=outreg2 p=pcu student=stcu; * create predictions and residuals ;
run;

* define v=symbol, c=color, i=interpolation, l=line style;
symbol1 v=circle c=black  i=none;
proc gplot data=outreg2;
   plot stcu*pcu;
   plot stcu*i8;
run;
```

Output window contents:

```
The REG Procedure
Model: MODEL1
Dependent Variable: cloud

Number of Observations Read          18
Number of Observations Used          18

                                Analysis of Variance
                                           Sum of           Mean
Source                     DF         Squares         Square    F Value    Pr > F
Model                       3       164.66411       54.88804     812.87    <.0001
Error                      14         0.94533        0.06752
Corrected Total            17       165.60944

Root MSE              0.25985    R-Square     0.9943
Dependent Mean       28.30556    Adj R-Sq     0.9931
Coeff Var             0.91803

                           Parameter Estimates
                      Parameter       Standard
Variable     DF        Estimate          Error    t Value    Pr > |t|
Intercept     1        28.65714        0.08959     319.88      <.0001
i8            1         0.88222        0.04774      18.48      <.0001
i8sq          1        -0.07320        0.00790      -9.26      <.0001
i8cu          1         0.00974        0.00259       3.76      0.0021


                                Output Statistics
     Dependent  Predicted     Std Error                 Std Error   Student                    Cook's
Obs   Variable      Value  Mean Predict    Residual      Residual  Residual   -2-1 0 1 2          D
  1    22.1000    22.4869        0.1709     -0.3869         0.196    -1.977 |   ***|      |  0.745
  2    24.5000    24.3943        0.1072      0.1057         0.237     0.447 |      |      |  0.010
  3    26.0000    25.9555        0.1062      0.0445         0.237     0.188 |      |      |  0.002
  4    26.8000    27.2289        0.1044     -0.4289         0.238    -1.803 |   ***|      |  0.157
  5    28.2000    28.2730        0.0934     -0.0730         0.242    -0.301 |      |      |  0.003
  6    28.9000    29.1461        0.0883     -0.2461         0.244    -1.007 |    **|      |  0.033
  7    30.0000    29.9067        0.0970      0.0933         0.241     0.387 |      |      |  0.006
  8    30.4000    30.6132        0.1075     -0.2132         0.237    -0.901 |     *|      |  0.042
  9    31.4000    31.3240        0.1094      0.0760         0.236     0.322 |      |      |  0.006
 10    26.1000    25.9555        0.1062      0.1445         0.237     0.609 |      |*     |  0.019
 11    28.5000    28.2730        0.0934      0.2270         0.242     0.936 |      |*     |  0.033
 12    30.3000    29.9067        0.0970      0.3933         0.241     1.631 |      |***   |  0.108
 13    31.5000    31.3240        0.1094      0.1760         0.236     0.747 |      |*     |  0.030
 14    22.8000    22.4869        0.1709      0.3131         0.196     1.599 |      |***   |  0.488
 15    27.3000    27.2289        0.1044      0.0711         0.238     0.299 |      |      |  0.004
 16    29.8000    29.9067        0.0970     -0.1067         0.241    -0.443 |      |      |  0.008
 17    31.8000    32.0975        0.1274     -0.2975         0.226    -1.314 |    **|      |  0.137
 18    33.1000    32.9921        0.2250      0.1079         0.130     0.830 |      |*     |  0.515

Sum of Residuals                           0
Sum of Squared Residuals             0.94533
Predicted Residual SS (PRESS)        1.93560
```

Below are plots of the data and the studentized residuals.





8.2 Polynomial Models with Two Predictors

Polynomial models are sometimes fit to data collected from experiments with two or more predictors. For simplicity, consider the general quadratic model, with two predictors:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + \varepsilon.$$

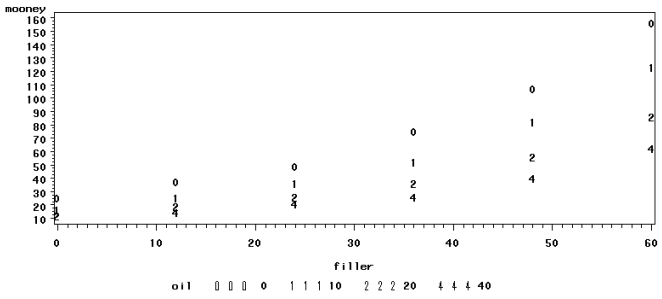
The model, which can be justified as a second order approximation to a smooth trend, includes quadratic terms in X_1 and X_2 and the product or interaction of X_1 and X_2.

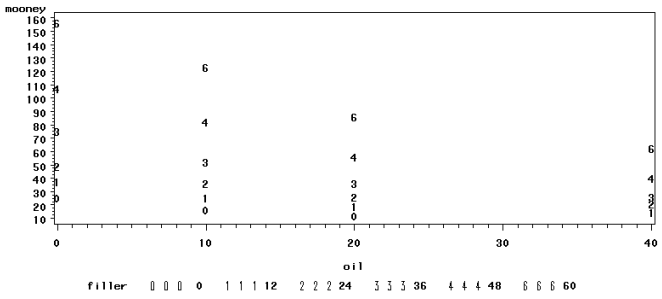
The data in the **SAS** program below give the Mooney viscosity at 100 degrees Celsius (Y) as a function of the filler level (X_1) and the naphthenic oil (X_2) level for an experiment involving filled and plasticized elastomer compounds.

8.2.1 Example: Mooney viscosity

Program editor contents:

```
* Mooney viscosity example;
options ls=79 nodate nocenter;
data viscosity;
  input oil filler mooney @@;
  datalines;
   0 0 26   0 12 38   0 24 50   0 36 76   0 48 108   0 60 157
  10 0 17  10 12 26  10 24 37  10 36 53  10 48  83  10 60 124
  20 0 13  20 12 20  20 24 27  20 36 37  20 48  57  20 60  87
  40 0  .  40 12 15  40 24 22  40 36 27  40 48  41  40 60  63
;
run;

* define v=symbol, c=color, i=interpolation, l=line style;
symbol1 v="0" c=black  i=none l=.;
symbol2 v="1" c=black  i=none l=.;
symbol3 v="2" c=black  i=none l=.;
symbol4 v="4" c=black  i=none l=.;
proc gplot data=viscosity;
  plot mooney*filler=oil;
run;
symbol1 v="0" c=black  i=none l=.;
symbol2 v="1" c=black  i=none l=.;
symbol3 v="2" c=black  i=none l=.;
symbol4 v="3" c=black  i=none l=.;
symbol5 v="4" c=black  i=none l=.;
symbol6 v="6" c=black  i=none l=.;
proc gplot data=viscosity;
  plot mooney*oil=filler;
run;
```





At each of the 4 oil levels, the relationship between the Mooney viscosity and filler level (with 6 levels) appears to be quadratic. Similarly, the relationship between the Mooney viscosity and oil level appears quadratic for each filler level (with 4 levels). This supports fitting the general quadratic model as a first step in the analysis.

The output below shows that each term is needed in the model. Although there are potentially influential points (cases 6 and 20), deleting either or both cases does not change the significance of the effects in the model (not shown).

Program editor contents:

```
data viscosity;
  set viscosity;
  oil2 = oil**2;         * quad term;
  filler2 = filler**2;   * quad term;
  pof  = oil*filler;     * product (interaction) term;
;

proc reg data=viscosity;
  model mooney = oil oil2 filler filler2 pof /p r;
run;
```

Output window contents:

```
The REG Procedure
Model: MODEL1
Dependent Variable: mooney

Number of Observations Read                          24
Number of Observations Used                          23
Number of Observations with Missing Values            1

                          Analysis of Variance
                                   Sum of           Mean
Source                     DF       Squares         Square    F Value    Pr > F
Model                       5         31400     6279.95138     405.22    <.0001
Error                      17     263.46047       15.49767
Corrected Total            22         31663

Root MSE              3.93671    R-Square     0.9917
Dependent Mean       52.34783    Adj R-Sq     0.9892
Coeff Var             7.52029

                     Parameter Estimates
                    Parameter       Standard
Variable     DF      Estimate          Error    t Value    Pr > |t|
Intercept     1      27.14458        2.61678      10.37      <.0001
oil           1      -1.27144        0.21353      -5.95      <.0001
oil2          1       0.03361        0.00466       7.21      <.0001
filler        1       0.43698        0.15266       2.86      0.0108
filler2       1       0.02732        0.00241      11.34      <.0001
pof           1      -0.03866        0.00319     -12.13      <.0001
```

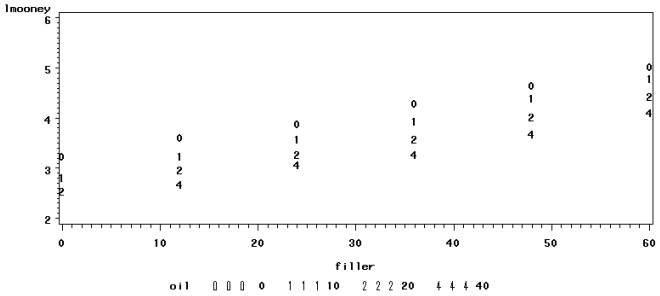
```
                          Output Statistics
     Dependent  Predicted     Std Error                Std Error   Student                    Cook's
Obs   Variable      Value  Mean Predict    Residual    Residual  Residual   -2-1 0 1 2           D

  1    26.0000    27.1446        2.6168    -1.1446       2.941    -0.389 |       |       |  0.020
  2    38.0000    36.3229        1.9255     1.6771       3.434     0.488 |       |       |  0.013
  3    50.0000    53.3700        1.8446    -3.3700       3.478    -0.969 |      *|       |  0.044
  4    76.0000    78.2861        1.8160    -2.2861       3.493    -0.655 |      *|       |  0.019
  5   108.0000   111.0712        1.9101    -3.0712       3.442    -0.892 |      *|       |  0.041
  6   157.0000   151.7251        2.7568     5.2749       2.810     1.877 |       |***    |  0.565
  7    17.0000    17.7912        2.0496    -0.7912       3.361    -0.235 |       |       |  0.003
  8    26.0000    22.3304        1.3266     3.6696       3.706     0.990 |       |*      |  0.021
  9    37.0000    34.7385        1.4283     2.2615       3.668     0.616 |       |*      |  0.010
 10    53.0000    55.0155        1.4419    -2.0155       3.663    -0.550 |      *|       |  0.008
 11    83.0000    83.1615        1.3431    -0.1615       3.701   -0.0436 |       |       |  0.000
 12   124.0000   119.1764        2.0078     4.8236       3.386     1.424 |       |**     |  0.119
 13    13.0000    15.1600        2.3648    -2.1600       3.147    -0.686 |      *|       |  0.044
 14    20.0000    15.0601        1.5564     4.9399       3.616     1.366 |       |**     |  0.058
 15    27.0000    22.8291        1.5996     4.1709       3.597     1.160 |       |**     |  0.044
 16    37.0000    38.4671        1.6461    -1.4671       3.576    -0.410 |       |       |  0.006
 17    57.0000    61.9739        1.5731    -4.9739       3.609    -1.378 |    **|       |  0.060
 18    87.0000    93.3497        2.0836    -6.3497       3.340    -1.901 |   ***|       |  0.234
 19           .    30.0638        4.2267           .           .        .           .  .
 20    15.0000    20.6857        2.8003    -5.6857       2.767    -2.055 |  ****|       |  0.721
 21    22.0000    19.1766        2.0945     2.8234       3.333     0.847 |       |*      |  0.047
 22    27.0000    25.5364        1.8849     1.4636       3.456     0.423 |       |       |  0.009
 23    41.0000    39.7651        2.0692     1.2349       3.349     0.369 |       |       |  0.009
 24    63.0000    61.8628        2.8991     1.1372       2.663     0.427 |       |       |  0.036

Sum of Residuals                           0
Sum of Squared Residuals           263.46047
Predicted Residual SS (PRESS)      579.81524
```

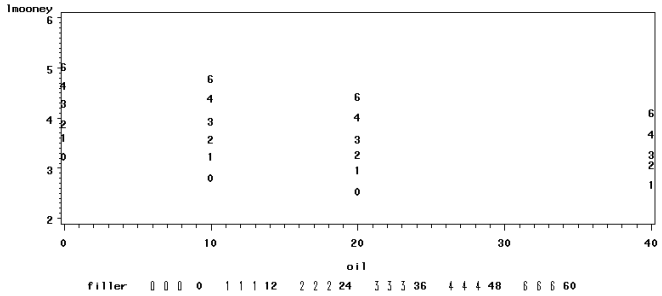
8.2.2 Example: Mooney viscosity on log scale

As noted earlier, transformations can often be used instead of polynomials. For example, the original data plots suggest transforming the Moody viscosity to a log scale. If we make this transformation and replot the data, we see that the log Mooney viscosity is roughly linearly related to the filler level at each oil level, but is a quadratic function of oil at each filler level. The plots of the transformed data suggest that a simpler model will be appropriate. To see this, we fit the full quadratic model. The interaction term can be omitted here, without much loss of predictive ability (R-squared is similar). The p-value for the interaction term in the quadratic model is 0.34.
Program editor contents:

```
* Viscosity data on log scale ;
data viscosity;
  set viscosity;
  lmooney = log(mooney);
;
run;

proc reg data=viscosity;
  model lmooney = oil oil2 filler filler2 pof;
run;

* define v=symbol, c=color, i=interpolation, l=line style;
symbol1 v="0" c=black  i=none l=.;
symbol2 v="1" c=black  i=none l=.;
symbol3 v="2" c=black  i=none l=.;
symbol4 v="4" c=black  i=none l=.;
proc gplot data=viscosity;
  plot lmooney*filler=oil;
run;
symbol1 v="0" c=black  i=none l=.;
symbol2 v="1" c=black  i=none l=.;
symbol3 v="2" c=black  i=none l=.;
symbol4 v="3" c=black  i=none l=.;
```

```
symbol5 v="4" c=black  i=none l=.;
symbol6 v="6" c=black  i=none l=.;
proc gplot data=viscosity;
  plot lmooney*oil=filler;
run;
```





Output window contents:

```
The REG Procedure
Model: MODEL1
Dependent Variable: lmooney

Number of Observations Read                          24
Number of Observations Used                          23
Number of Observations with Missing Values            1

                              Analysis of Variance
                                   Sum of           Mean
Source                     DF      Squares         Square    F Value    Pr > F
Model                       5     10.55468        2.11094     737.01    <.0001
Error                      17      0.04869        0.00286
Corrected Total            22     10.60338

Root MSE              0.05352    R-Square     0.9954
Dependent Mean        3.72680    Adj R-Sq     0.9941
Coeff Var             1.43603

                        Parameter Estimates
                     Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|
Intercept     1        3.23620        0.03557      90.97      <.0001
oil           1       -0.03921        0.00290     -13.51      <.0001
oil2          1     0.00042267     0.00006339       6.67      <.0001
filler        1        0.02860        0.00208      13.78      <.0001
filler2       1     0.00004657     0.00003276       1.42      0.1733
pof           1    -0.00004231     0.00004332      -0.98      0.3424
```

After omitting the interaction term, the quadratic effect in filler is not needed in the model (output not given). Once these two effects are removed, each of the remaining effects is significant.
Program editor contents:

```
* final model on log scale;
proc reg data=viscosity;
  model lmooney = oil oil2 filler /p r;
run;
```

Output window contents:

```
Root MSE              0.05423    R-Square     0.9947
Dependent Mean        3.72680    Adj R-Sq     0.9939
Coeff Var             1.45513

                    Parameter Estimates
                  Parameter       Standard
Variable     DF     Estimate          Error    t Value    Pr > |t|
Intercept     1      3.22966        0.02734     118.14      <.0001
oil           1     -0.04024        0.00270     -14.89      <.0001
oil2          1   0.00040974     0.00006356       6.45      <.0001
filler        1      0.03086     0.00057157      53.99      <.0001
```

I did not carefully examine this model for inadequacies. Assuming no difficulties, an important effect of the transformation is that the resulting model is simpler than the model selected on the original scale. This is satisfying to me, but you may not agree. Note that the selected model, with linear effects due to the oil and filler levels and a quadratic effect due to the oil level, agrees with our visual assessment of the data. Assuming no inadequacies, the predicted log Moody viscosity is given by

$$\widehat{\log \text{Moody viscosity}} = 3.2297 - 0.0402\text{Oil} + 0.0004\text{Oil}^2 + 0.0309\text{Filler}.$$

Quadratic models with two or more predictors are often used in industrial experiments to estimate the optimal combination of predictor values to maximize or minimize the response, over the range of predictor variable values where the model is reasonable. (This strategy is called "response surface methodology".) For example, we might wish to know what combination of oil level between 0 and 40 and filler level between 0 and 60 provides the lowest predicted Mooney viscosity (on the original or log scale). We can visually approximate the minimizer using the data plots, but one can do a more careful job of analysis using standard tools from calculus.

8.3 Using glm to Fit Polynomial Models

glm may be used to fit polynomial models. Although **glm** provides less diagnostic information than **reg**, polynomial models are easier to specify because they do not require you to explicitly define polynomial terms or interactions in the **data** step. The higher order effects are defined in the **model** statement as interactions. For example, in the Mooney viscosity analysis, a quadratic oil term is included as the product effect oil*oil.

Here is a **SAS** program for reproducing the initial fit to the data, using **glm** instead of **reg**. Compare the output to that on page 165. If you wish to do the analysis on the log of the Mooney viscosity, you must create the transformed response in a **data** step (you can't just put log(mooney) in the model statement)!

Program editor contents:

```
* glm for viscosity data;
proc glm data=viscosity;
  model mooney = oil filler oil*oil filler*filler oil*filler;
run;
```

Output window contents:

```
The GLM Procedure

Number of Observations Read          24
Number of Observations Used          23

Dependent Variable: mooney
                                        Sum of
Source                      DF         Squares     Mean Square   F Value   Pr > F
Model                        5     31399.75692      6279.95138    405.22   <.0001
Error                       17       263.46047        15.49767
Corrected Total             22     31663.21739

R-Square     Coeff Var      Root MSE    mooney Mean
0.991679      7.520291      3.936709       52.34783

Source                      DF       Type I SS     Mean Square   F Value   Pr > F
oil                          1     5193.85459      5193.85459    335.14   <.0001
filler                       1    22253.75983     22253.75983   1435.94   <.0001
oil*oil                      1      407.12188       407.12188     26.27   <.0001
filler*filler                1     1264.18453      1264.18453     81.57   <.0001
oil*filler                   1     2280.83608      2280.83608    147.17   <.0001

Source                      DF     Type III SS     Mean Square   F Value   Pr > F
oil                          1      549.452553      549.452553     35.45   <.0001
filler                       1      126.986365      126.986365      8.19   0.0108
oil*oil                      1      805.195432      805.195432     51.96   <.0001
filler*filler                1     1992.525521     1992.525521    128.57   <.0001
oil*filler                   1     2280.836085     2280.836085    147.17   <.0001

                                        Standard
Parameter                Estimate          Error    t Value    Pr > |t|
Intercept          27.14458151      2.61677902      10.37      <.0001
oil                -1.27144250      0.21353295      -5.95      <.0001
filler              0.43698421      0.15265832       2.86      0.0108
oil*oil             0.03361057      0.00466293       7.21      <.0001
filler*filler       0.02732264      0.00240965      11.34      <.0001
oil*filler         -0.03865898      0.00318667     -12.13      <.0001
```

Chapter 9

Discussion of Response Models with Factors and Predictors

We have considered simple models for designed experiments and observational studies where a response variable is modeled as a linear combination of effects due to **factors** or **predictors**, or both. With designed experiments, where only qualitative factors are considered, we get a "pure ANOVA" model. For example, in the experiment comparing survival times of beetles, the potential effects of **insecticide** (with levels A, B, C, and D) and **dose** (with levels 1=low, 2=medium, and 3=high) are included in the model as **factors** because these variables are qualitative. The natural model to consider is a two-way ANOVA with effects for dose and insecticide and a dose-by-insecticide interaction. If, however, the dose given to each beetle was recorded on a measurement scale, then the dosages can be used to define a predictor variable which can be used as a "regression effect" in a model. That is, the dose or some function of dose can be used as a (quantitative) predictor instead of as a qualitative effect.

For simplicity, assume that the doses are 10, 20, and 30, but the actual levels are irrelevant to the discussion. The simple additive model, or ANCOVA model, assumes that there is a linear relationship between mean survival time and dose, with different intercepts for the four insecticides. If data set includes the survival time (times) for each beetle, the insecticide (insect: an alphanumeric variable, with values A, B, C, and D), and dose, you would fit the ANCOVA model in **glm** using the command:

```
proc glm ;
  class insect;
  model times = insect dose;
run;
```

A more complex model that allows separate regression lines for each insecticide is specified as follows:

```
proc glm;
  class insect;
  model times = insect dose insect*dose;
run;
```

It is important to recognize that the **class** statement in the **glm** command defines which variables in the model are treated as **factors**. Each effect included in the class statement is treated as a factor. Effects in the model statement that are omitted from the class statement are treated as **predictors**. There is nothing that precludes a measurement variable from being included as a factor (with one level for each distinct observed value of the variable) instead of a predictor, so you

need to be careful about how you specify the class variables and the model. Thus, in the survival time experiment, the **glm** sequences

```
proc glm ;
  class dose insect;
  model times = insect dose;              * call this (A) for additive    ;
run;

proc glm ;
  class dose insect;
  model times = insect dose insect*dose;  * call this (I) for interaction ;
run;
```

give the analysis for a two-way ANOVA model without interaction and with interaction, respectively, where both dose and insecticide are treated as factors (since dose is in the class statement), even though we just defined dose on a measurement scale!

Is there a basic connection between the ANCOVA and separate regression line models for dose and two-way ANOVA models where dose and insecticide are treated as factors? Yes — I mentioned a connection when discussing ANCOVA and I will try now to make the connection more explicit.

For the moment, let us simplify the discussion and assume that only one insecticide was used at three dose levels. The LS estimates of the mean responses from the quadratic model

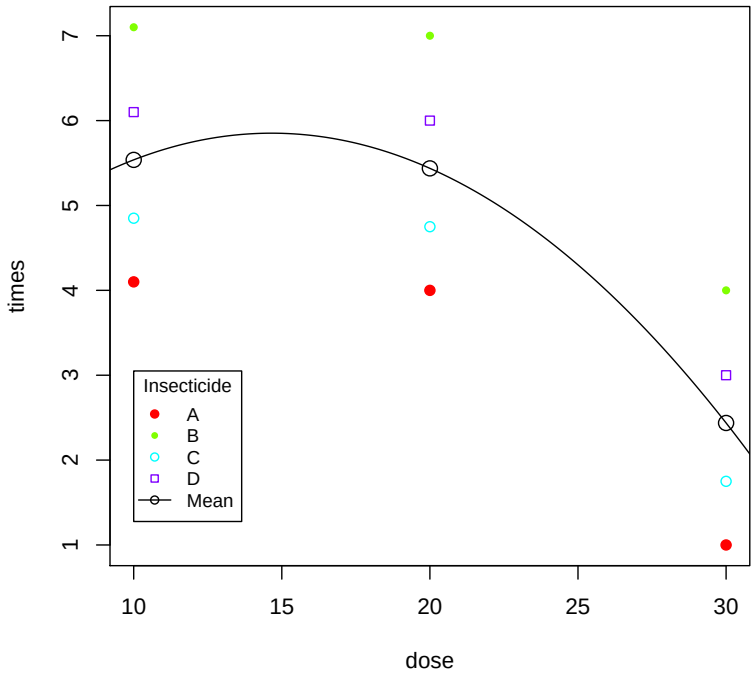
$$\text{Times} = \beta_0 + \beta_1 \text{Dose} + \beta_2 \text{Dose}^2 + \varepsilon$$

are the observed average survival times at the three dose levels. The LS curve goes through the mean survival time at each dose, as illustrated in the picture below.

If we treat dose as a **factor**, and fit the one-way ANOVA model

$$\text{Times} = \text{Grand Mean} + \text{Dose Effect} + \text{Residual},$$

then the LS estimates of the population mean survival times are the observed mean survival times. The two models are mathematically equivalent, but the parameters have different interpretations. In essence, the one-way ANOVA model places no restrictions on the values of the population means (no *a priori* relation between them) at the three doses, and neither does the quadratic model! (WHY?)



In a one-way ANOVA, the standard hypothesis of interest is that the dose effects are zero. This can be tested using the one-way ANOVA F-test, or by testing $H_0 : \beta_1 = \beta_2 = 0$ in the quadratic model (recall the **test** command in **reg**). With three dosages, the absence of a linear or quadratic effect implies that all the population mean survival times must be equal. An advantage of the polynomial model over the one-way ANOVA is that it provides an easy way to **quantify** how dose impacts the mean survival, and a convenient way to check whether a simple description such as a simple linear regression model is adequate to describe the effect.

More generally, if dose has p levels, then the one-way ANOVA model

$$\text{Times} = \text{Grand Mean} + \text{Dose Effect} + \text{Residual},$$

is equivalent to the $(p-1)^{st}$ degree polynomial

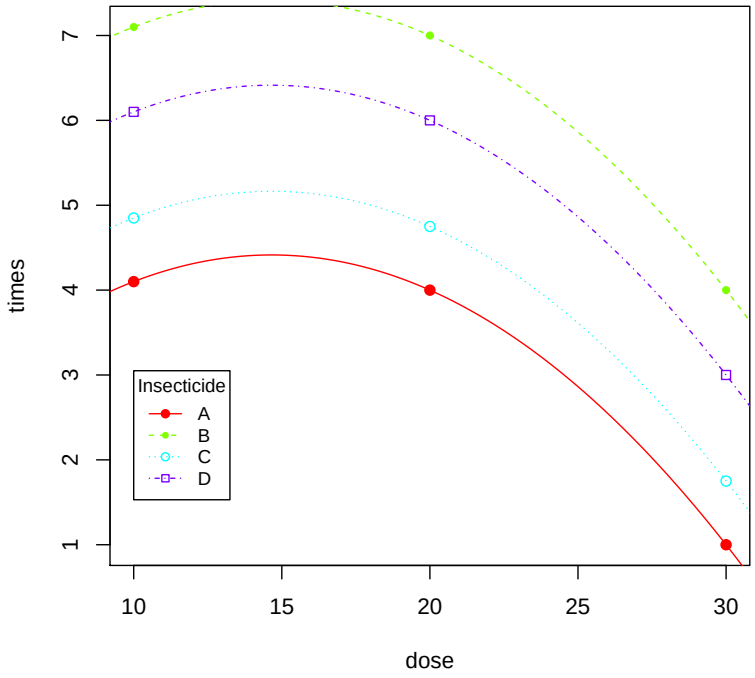
$$\text{Times} = \beta_0 + \beta_1 \text{Dose} + \beta_2 \text{Dose}^2 + \cdots + \beta_{p-1} \text{Dose}^{(p-1)} + \varepsilon$$

and the one-way ANOVA F-test for no treatment effects is equivalent to testing $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ in this polynomial.

Returning to the original experiment with 4 insecticides and 3 doses, I can show the following two equivalences. First, the two-way additive ANOVA model, with insecticide and dose as factors, i.e., model (A), is mathematically equivalent to an additive model with insecticide as a factor, and a quadratic effect in dose:

```
proc glm ;
  class insect;
  model times = insect dose dose*dose;
run;
```

Thinking of dose*dose as a quadratic term in dose, rather than as an interaction, this model has an additive insecticide effect, but the dose effect is not differentiated across insecticides. That is, the model assumes that the quadratic curves for the four insecticides differ only in level (i.e., different intercepts) and that the coefficients for the dose and dose*dose effects are identical across insecticides. This is an additive model, because the population means plot has parallel profiles. A possible pictorial representation of this model is given below.



Second, the two-way ANOVA interaction model, with insecticide and dose as factors, i.e., model (I), is mathematically equivalent to an interaction model with insecticide as a factor, and a quadratic effect in dose.

```
proc glm ;
  class insect;
  model times = insect dose dose*dose insect*dose insect*dose*dose;
run;
```

This model fits separate quadratic relationships for each of the four insecticides, by including interactions between insecticides and the linear and quadratic terms in dose. Because dose has three levels, this model places no restrictions on the mean responses.

To summarize, we have established that

- The additive two-way ANOVA model with insecticide and dose as factors is mathematically identical to an additive model with an insecticide factor and a quadratic effect in dose. The ANCOVA model with a linear effect in dose is a special case of these models, where the quadratic effect is omitted.

- The two-way ANOVA interaction model with insecticide and dose as factors is mathematically identical to a model with an insecticide factor, a quadratic effect in dose, and interactions

between the insecticide and the linear and quadratic dose effects. The separate regression lines model with a linear effect in dose is a special case of these models, where the quadratic dose effect and the interaction of the quadratic term with insecticide are omitted.

Recall that response models with **factors** and **predictors** as effects can be fit using the **reg** procedure, but each factor or interaction involving a factor must be represented in the model using indicator variables or product terms. The number of required indicators or product effects is one less than the number of distinct levels of the factor. For example, to fit the model with "parallel" quadratic curves in dose, you must define (in the **data** step!) three indicator variables for the insecticide effect, say $I1$, $I2$, and $I3$, and fit the model

$$\text{Times} = \beta_0 + \beta_1 I1 + \beta_2 I2 + \beta_3 I3 + \beta_4 \text{Dose} + \beta_5 \text{Dose}^2 + \varepsilon.$$

For the "quadratic interaction model", you must define 6 interaction or product terms between the 3 indicators and the 2 dose terms: $P6 = I1 * \text{Dose}$, $P7 = I2 * \text{Dose}$, $P8 = I3 * \text{Dose}$, $P9 = I1 * \text{Dose}^2$, and $P10 = I2 * \text{Dose}^2$, $P11 = I3 * \text{Dose}^2$, and fit

$$\begin{aligned} \text{Times} &= \beta_0 + \beta_1 I1 + \beta_2 I2 + \beta_3 I3 + \beta_4 \text{Dose} + \beta_5 \text{Dose}^2 \\ &\quad + \beta_6 P6 + \beta_7 P7 + \beta_8 P8 + \beta_9 P9 + \beta_{10} P10 + \beta_{11} P11 + \varepsilon. \end{aligned}$$

The $\beta_6 P6 + \beta_7 P7 + \beta_8 P8$ component in the model formally corresponds to the *insect* $*$ *dose* interaction in **glm**, whereas the $\beta_9 P9 + \beta_{10} P10 + \beta_{11} P11$ component is equivalent to the *insect* $*$ *dose* $*$ *dose* interaction (i.e., testing $H_0 : \beta_9 = \beta_{10} = \beta_{11} = 0$ in **reg** is equivalent to testing for the significance of the *insect* $*$ *dose* $*$ *dose* interaction in **glm**).

This discussion is not intended to confuse, but rather to impress upon you the intimate connection between regression and ANOVA, and to convince you of the care that is needed when modelling variation even in simple studies. Researchers are usually faced with more complex modelling problems than we have examined, where many variables might influence the response. If experimentation is possible, a scientist will often control the levels of variables that influence the response but that are not of primary interest. This can result in a manageable experiment with, say, four or fewer qualitative or quantitative variables that are systematically varied in a scientifically meaningful way. In observational studies, where experimentation is not possible, the scientist builds models to assess the effects of interest on the response, adjusting the response for all the uncontrolled variables that might be important. The uncontrolled variables are usually a mixture of factors and predictors. Ideally, the scientist knows what variables to control in an experiment and which to vary, and what variables are important to collect in an observational study.

The level of complexity that I am describing here might be intimidating, but certain basic principles can be applied to many of the studies you will see. Graduate students in statistics often take several courses (5+) in experimental design, regression analysis, and linear model theory to master the breadth of models, and the subtleties of modelling, that are needed to be a good data analyst. I can only scratch the surface here. I will discuss a reasonably complex study having multiple factors and multiple predictors. The example focuses on strategies for building models, with little attempt to do careful diagnostic analyses. Hopefully, the example will give you an appreciation for statistical modelling, but **please be careful — these tools are dangerous!**

9.1 Some Comments on Building Models

A primary goal in many statistical analyses is to build a model or models to understand the variation in a response. Fortunately, or unfortunately, there is no consensus on how this should be done. Ideally, theory would suggest models to compare, but in many studies the goal is to provide an initial model that will be refined, validated, or refuted, by further experimentation. An extreme view is that the selected model(s) should only include effects that are "**statistically important**", whereas another extreme suggests that all effects that might be "**scientifically important**" should be included.

A difficulty with implementing either approach is that importance is relative to specific goals (i.e., Why are you building the model and what do you plan to use the model for? Is the model a prescription or device to make predictions? Is the model a tool to understand the effect that one or more variables have on a response, after adjusting for uninteresting, but important effects that can not be controlled? etc.) Madigan and Raftery, in the 1994 edition of *The Journal of the American Statistical Association*, comment that "Science is an iterative process in which competing models of reality are compared on the basis of how well they predict what is observed; models that predict much less well than their competitors are discarded." They argue that models should be selected using Occum's razor, a widely accepted norm in scientific investigations whereby the simplest plausible model among all reasonable models, given the data, is preferred. Madigan and Raftery's ideas are fairly consistent with the first extreme, but can be implemented in a variety of ways, depending on how you measure prediction adequacy. They propose a Bayesian approach, based on model averaging and prior beliefs on the plausibility of different models. An alternative method using Mallow's C_p criterion will be discussed later.

A simple compromise between the two extremes might be to start the model building process with the most complex model that is scientifically reasonable, but still interpretable, and systematically eliminate effects using backward elimination. The initial or **maximal** model might include polynomial effects for predictors, main effects and interactions (2 factor, 3 factor, etc.) between factors, and products or interactions between predictors and factors. This approach might appear to be less than ideal because the importance of effects is assessed using hypothesis tests and no attempt is made to assess the effect of changes on predictions. However, one can show that the average squared error in predictions is essentially reduced by eliminating **insignificant** regression effects from the model, so this approach seems tenable.

In **glm**, it might be sensible to only assess significance of effects specified in the **model** statement. However, many of these effects consist of several degrees of freedom. That is, the effect corresponds to several regression coefficients in the model. (Refer to the discussion following the displayed equations on page 175). The individual regression variables that comprise an effect could also be tested individually. However, if the effect is a factor (with 3+ levels) or an interaction involving a factor, then the interpretation of tests on individual regression coefficients depends on the level of the factor that was selected to be the baseline category. The Type III F-test on the entire effect does not depend on the baseline category. In essence, two researchers can start with different representations of the same mathematical model (i.e., the parameters are defined differently for different choices of baseline categories), use the same algorithm for selecting a model, yet come to different final models for the data.

Statisticians often follow the **hierarchy** principle, which states that a lower order term (be it a factor or a predictor) may be considered for exclusion from a model only if no higher order effects

that include the term are present in the model. For example, given an initial model with effects A, B, C, and the $A*B$ interaction, the only candidates for omission at the first step are C and $A*B$. If you follow the hierarchy principle, and test an entire effect rather than test the single degrees of freedom components that comprise an effect, then the difficulty described above can not occur. The hierarchy principle is most appealing with **pure** ANOVA models (such as the three-factor model in the example below), where all the regression variables are indicators. In ANOVA models, the ANOVA effects are of interest because they imply certain structure on the means. The individual regression variables that define the effects are not usually a primary interest.

A non-hierarchical backward elimination algorithm where single degree of freedom effects are eliminated independently of the other effects in the model is implemented in the **reg** procedure. Recall our discussion of backwards elimination from earlier this semester.

9.2 Example: The Effect of Sex and Rank on Faculty Salary

The data in this example were collected from the personnel files of faculty at a small college in the 1970s. The data were collected to assess whether women were being discriminated against (consciously or unconsciously) in salary. The sample consists of tenured and tenure-stream faculty only. Temporary faculty were excluded from consideration (because they were already being discriminated against).

The variables in the **SAS** program below are id (individual identification numbers from 1 to 52), sex (coded 1 for female and 0 for male), rank (coded 1 for Asst. Professor, 2 for Assoc. Professor and 3 for Full Professor), year (number of years in current rank), degree (coded 1 for Doctorate, 0 else), yd (number of years since highest degree was earned), and salary (academic year salary in dollars).

Program editor contents:

```
* Faculty salary example ;
options ls=79 nodate nocenter;

data faculty;
  input id sex rank year degree yd salary @@;
  datalines;
 1 0 3 25 1 35 36350     2 0 3 13 1 22 35350
 3 0 3 10 1 23 28200     4 1 3  7 1 27 26775
 5 0 3 19 0 30 33696     6 0 3 16 1 21 28516
 7 1 3  0 0 32 24900     8 0 3 16 1 18 31909
 9 0 3 13 0 30 31850    10 0 3 13 0 31 32850
11 0 3 12 1 22 27025    12 0 2 15 1 19 24750
13 0 3  9 1 17 28200    14 0 2  9 0 27 23712
15 0 3  9 1 24 25748    16 0 3  7 1 15 29342
17 0 3 13 1 20 31114    18 0 2 11 0 14 24742
19 0 2 10 0 15 22906    20 0 3  6 0 21 24450
21 0 1 16 0 23 19175    22 0 2  8 0 31 20525
23 0 3  7 1 13 27959    24 1 3  8 1 24 38045
25 0 2  9 1 12 24832    26 0 3  5 1 18 25400
27 0 2 11 1 14 24800    28 1 3  5 1 16 25500
29 0 2  3 0  7 26182    30 0 2  3 0 17 23725
31 1 1 10 0 15 21600    32 0 2 11 0 31 23300
33 0 1  9 0 14 23713    34 1 2  4 0 33 20690
35 1 2  6 0 29 22450    36 0 2  1 1  9 20850
37 1 1  8 1 14 18304    38 0 1  4 1  4 17095
39 0 1  4 1  5 16700    40 0 1  4 1  4 17600
41 0 1  3 1  4 18075    42 0 1  3 0 11 18000
43 0 2  0 1  7 20999    44 1 1  3 1  3 17250
```

```
45 0 1  2 1  3 16500    46 0 1  2 1  1 16094
47 1 1  2 1  6 16150    48 1 1  2 1  2 15350
49 0 1  1 1  1 16244    50 1 1  1 1  1 16686
51 1 1  1 1  1 15000    52 1 1  0 1  2 20300
;
run;
```

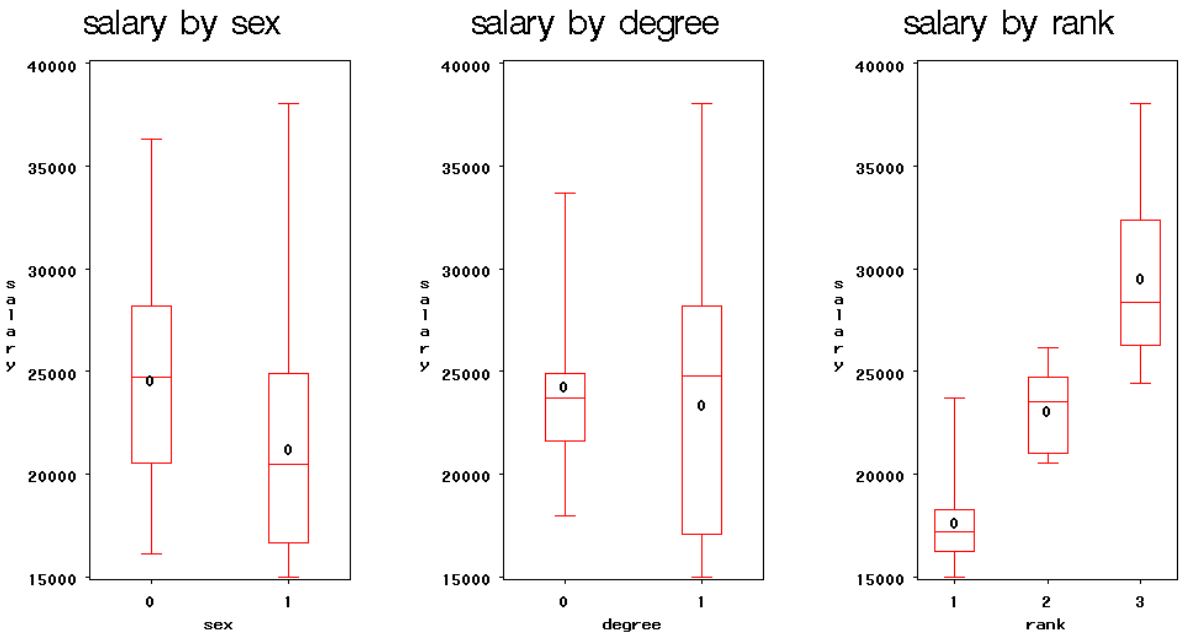
The data includes two potential predictors of salary, year and yd, and three factors, sex, rank, and degree. A primary statistical interest is whether males and females are compensated equally, on average, after adjusting salary for rank, years in rank, and the other given effects. Furthermore, we wish to know whether an effect due to sex is the same for each rank, or not.

Before answering these questions, let us look at the data. I will initially focus on the effect of the individual factors (sex, rank, and degree) on salary. A series of box-plots is given below, with the requisite **SAS** statements.

Looking at the boxplots, notice that women tend to earn less than men, that faculty with Doctorates tend to earn more than those without Doctorates (median), and that salary tends to increase with rank.

Program editor contents:

```
* boxplots of salary by categorical variables ;

title 'salary by sex';
proc sort data=faculty; by sex; run;
proc boxplot data=faculty; plot salary*sex; run;

title 'salary by degree';
proc sort data=faculty; by degree; run;
proc boxplot data=faculty; plot salary*degree; run;

title 'salary by rank';
proc sort data=faculty; by rank; run;
proc boxplot data=faculty; plot salary*rank; run;

title; * reset title to blank;
```



9.2.1 A Three-Way ANOVA on Salary Data

Hopefully, our earlier analyses have cured you of the desire to claim that a sex effect exists before considering whether the differences between male and female salaries might be due to other factors. The output below gives the sample sizes, means and standard deviations for the 11 combinations of sex, rank, and degree observed in the data. Side-by-side boxplots of the salaries for the 11 combinations are also provided. One combination of the three factors, female Associate Professors without Doctorates, was not observed.

Looking at the summaries, the differences between sexes **within** each combination of rank and degree appear to be fairly small. There is a big difference in the ranks of men and women, with a higher percentage of men in the more advanced ranks. This might explain the differences between male and female salaries, when other factors are ignored.

Program editor contents:

```
* calculate categorical summaries using glm ;
proc glm data=faculty;
  class sex degree rank;
  model salary = sex*rank*degree;
  means sex*rank*degree;  * this gives table of means and stds below ;
run;
```

Output window contents: (selected)

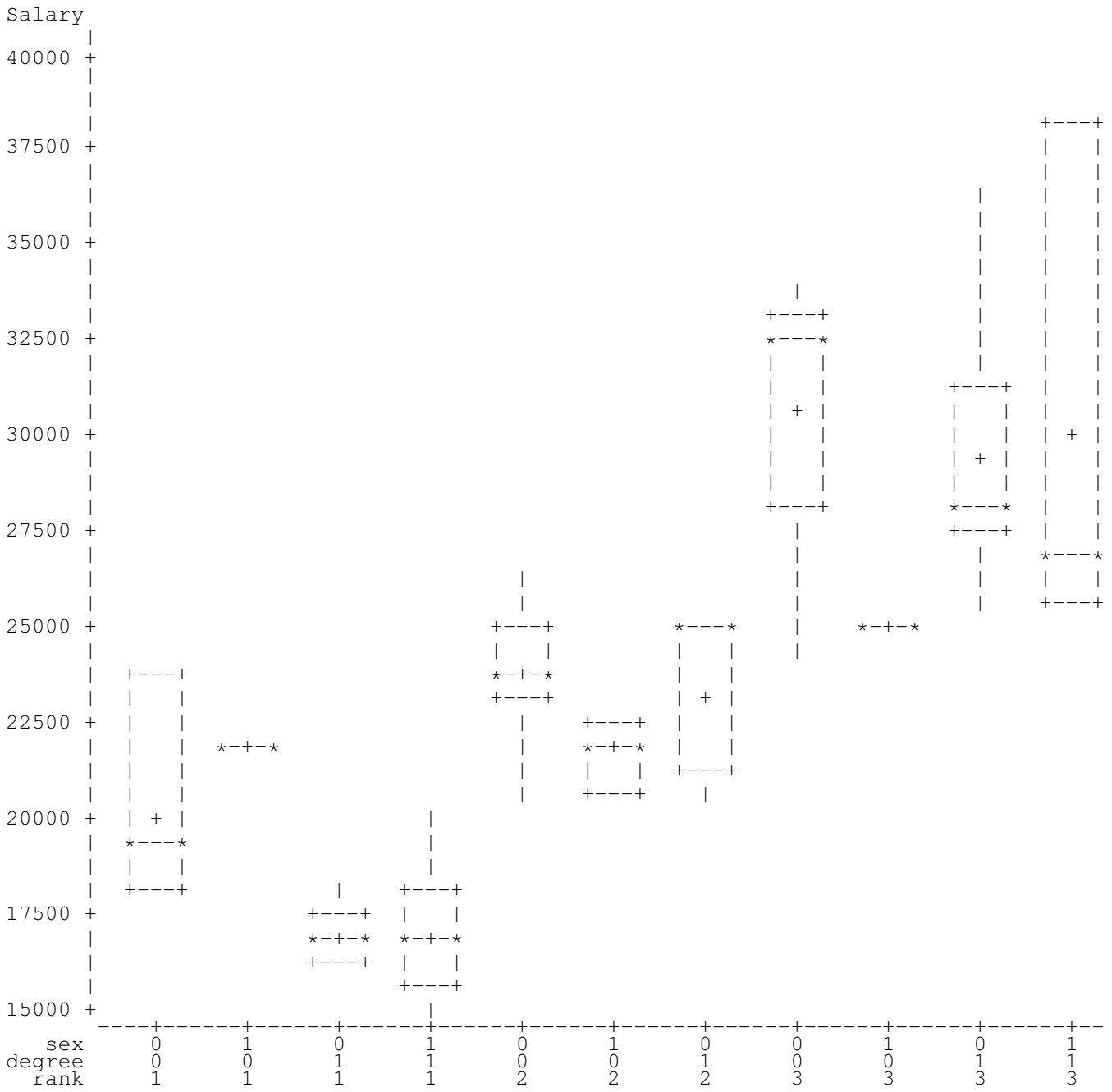
Level of sex	Level of rank	Level of degree	N	salary Mean	salary Std Dev
0	1	0	3	20296.0000	3016.96420
0	1	1	7	16901.1429	728.99621
0	2	0	7	23584.5714	1733.23646
0	2	1	5	23246.2000	2120.26843
0	3	0	4	30711.5000	4241.97230
0	3	1	12	29592.7500	3479.95659
1	1	0	1	21600.0000	.
1	1	1	7	17005.7143	1834.67906
1	2	0	2	21570.0000	1244.50793
1	3	0	1	24900.0000	.
1	3	1	3	30106.6667	6904.29275

I will consider two simple analyses of these data. The first analysis considers the effect of the three factors on salary. The second analysis considers the effect of the predictors. A complete analysis using both factors and predictors is then considered. I am doing the three factor analysis because the most complex pure ANOVA problem we considered this semester has two factors — the analysis is for illustration **only**!!

Program editor contents:

```
* boxplots by the three categorical variables ;
proc sort data=faculty;
  by rank degree sex;
run;
options ls=179;  * make wider linesize so all boxplots fit in single plot;
proc univariate plot; * for boxplots;
  var salary;
  by rank degree sex;
run;
options ls=79;  * reset linesize;
```

Output window contents:



The full model for a three-factor study includes the three main effects, the three possible two-factor interactions, plus the three-factor interaction. Identifying the factors by S (sex), D (degree) and R (rank), we write the full model as

Salary = Grand mean + S effect + D effect + R effect + S*D interaction + S*R interaction + R*D interaction + S*D*R interaction + Residual.

You should understand what main effects and two-factor interactions measure, but what about the three-factor term? If you look at the two levels of degree separately, then a three-factor interaction is needed if the interaction between sex and rank is different for the two degrees. (i.e., the profile plots are different for the two degrees). Not surprisingly, three-factor interactions are hard to interpret.

I considered a hierarchical backward elimination of effects using **glm**. Individual regression variables are not considered for deletion, unless they correspond to an effect in the model statement. All tests were performed at the 0.10 level, but this hardly matters here.

The first step in the elimination is to fit the full model and check whether the three-factor term is significant. The three-factor term was not significant. After eliminating this effect, I fit the model with all 3 two-factor terms, and then sequentially deleted the least important effects, one at a time, while still adhering to the hierarchy principle. The final model includes only an effect due to rank. I will lead you through the steps, given below.

Program editor contents:

```
* three-factor ANOVA backward selection ;
proc glm data=faculty;
  class sex degree rank;
  model salary = sex rank degree
                 sex*rank sex*degree rank*degree sex*rank*degree;
run;
```

Output window contents:

```
The GLM Procedure

   Class Level Information
Class         Levels    Values
sex                2    0 1
rank               3    1 2 3
degree             2    0 1

Number of Observations Read          52
Number of Observations Used          52

Dependent Variable: salary
                                    Sum of
Source                     DF       Squares    Mean Square   F Value   Pr > F
Model                      10    1424052631      142405263     16.14   <.0001
Error                      41     361677227        8821396
Corrected Total            51    1785729858

R-Square     Coeff Var      Root MSE    salary Mean
0.797463      12.48057      2970.083       23797.65

Source                     DF      Type I SS    Mean Square   F Value   Pr > F
sex                         1     114106220      114106220     12.94   0.0009
rank                        2    1239752324      619876162     70.27   <.0001
degree                      1      10855643       10855643      1.23   0.2738
sex*rank                    2       6171729        3085865      0.35   0.7069
sex*degree                  1       2500123        2500123      0.28   0.5973
rank*degree                 2      33433415       16716707      1.90   0.1632
sex*rank*degree             1      17233177       17233177      1.95   0.1697

Source                     DF    Type III SS    Mean Square   F Value   Pr > F
sex                         1      7895512.0      7895512.0      0.90   0.3497
rank                        2    483000187.3    241500093.7     27.38   <.0001
degree                      1      2078030.0      2078030.0      0.24   0.6300
sex*rank                    2     13768889.9      6884445.0      0.78   0.4649
sex*degree                  1      7996968.6      7996968.6      0.91   0.3466
rank*degree                 2     49434054.4     24717027.2      2.80   0.0723
sex*rank*degree             1     17233177.0     17233177.0      1.95   0.1697 *
```

omit SEX*RANK*DEGREE

Note that until I reach the final model, I will only give you the type III SS for each effect.

```
proc glm data=faculty;
  class sex degree rank;
  model salary = sex rank degree sex*rank sex*degree rank*degree;
run;
```

```
Source                     DF    Type III SS    Mean Square   F Value   Pr > F
sex                         1      7282954.1      7282954.1      0.81   0.3740
rank                        2    622453021.8    311226510.9     34.50   <.0001
degree                      1      1129300.6      1129300.6      0.13   0.7253
sex*rank                    2      2701492.6      1350746.3      0.15   0.8614 *
sex*degree                  1      7661925.5      7661925.5      0.85   0.3620
rank*degree                 2     33433415.0     16716707.5      1.85   0.1694
```

omit SEX*RANK

```
proc glm data=faculty;
  class sex degree rank;
  model salary = sex rank degree sex*degree rank*degree;
run;
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F	
sex	1	7501492.9	7501492.9	0.86	0.3574	
rank	2	801815852.4	400907926.2	46.22	<.0001	
degree	1	1555376.4	1555376.4	0.18	0.6740	
sex*degree	1	12335789.4	12335789.4	1.42	0.2394	*
rank*degree	2	32435968.1	16217984.0	1.87	0.1662	

omit SEX*DEGREE

```
proc glm data=faculty;
  class sex degree rank;
  model salary = sex rank degree rank*degree;
run;
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F	
sex	1	3009035.9	3009035.9	0.34	0.5606	*
rank	2	795017904.9	397508952.4	45.41	<.0001	
degree	1	12830901.9	12830901.9	1.47	0.2324	
rank*degree	2	27067985.3	13533992.7	1.55	0.2242	

omit SEX

```
proc glm data=faculty;
  class sex degree rank;
  model salary = rank degree rank*degree;
run;
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F	
rank	2	818613821.7	409306910.9	47.43	<.0001	
degree	1	12953723.7	12953723.7	1.50	0.2267	
rank*degree	2	31019254.7	15509627.3	1.80	0.1772	*

omit RANK*DEGREE

```
proc glm data=faculty;
  class sex degree rank;
  model salary = rank degree;
run;
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F	
rank	2	1349072233	674536116	75.65	<.0001	
degree	1	10970082	10970082	1.23	0.2729	*

omit DEGREE

```
proc glm data=faculty;
  class sex degree rank;
  model salary = rank;
run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1346783800	673391900	75.17	<.0001
Error	49	438946058	8958083		
Corrected Total	51	1785729858			

R-Square	Coeff Var	Root MSE	salary Mean
0.754192	12.57689	2993.006	23797.65

Source	DF	Type I SS	Mean Square	F Value	Pr > F
rank	2	1346783800	673391900	75.17	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
rank	2	1346783800	673391900	75.17	<.0001

This analysis suggests that sex is not predictive of salary, once other factors are taken into account. In particular, faculty rank appears to be the sole important effect, in the sense that once salaries are adjusted for rank no other factors explain a significant amount of the unexplained variation in salaries.

As noted earlier, *the analysis was meant to illustrate a three-factor ANOVA and backward selection*. The analysis is likely flawed, because it ignores the effects of year and year since degree on salary.

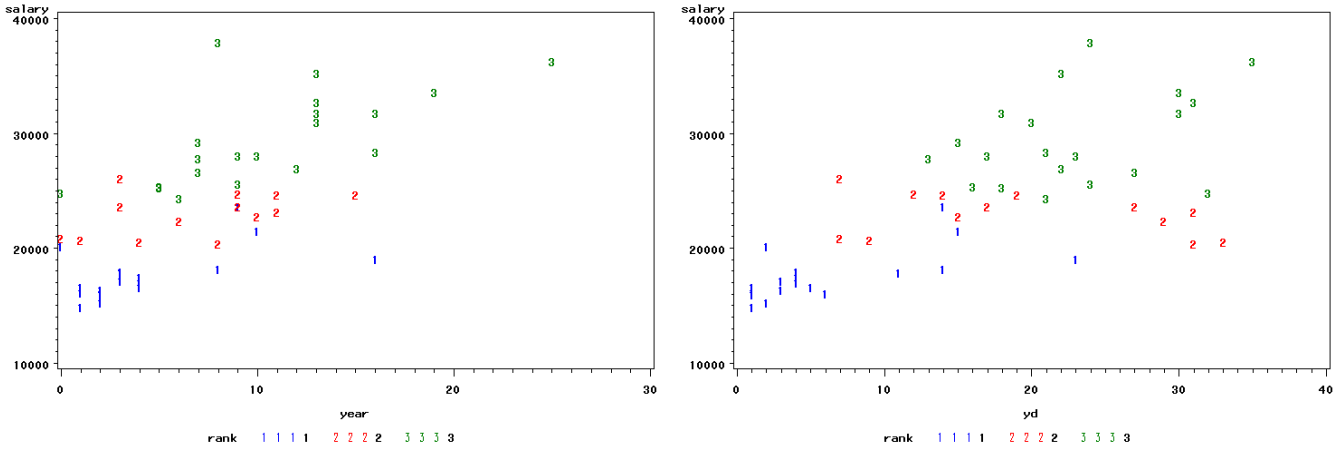
9.2.2 Using Year and Year Since Degree to Predict Salary

Plots of the salary against years in rank and years since degree show fairly strong associations with salary. The variability in salaries appears to be increasing with year and with year since degree, which might be expected. You might think to transform the salaries to a log scale to eliminate this effect, but doing so has little impact on the conclusions (not shown).

As a point of comparison with the three-factor ANOVA, I fit a multiple regression model with year and years since degree as predictors of salary. These two predictors are important for explaining the variation in salaries, but together they explain much less of the variation (58%) than rank does on its own (75%).

Program editor contents:

```
* prector variables year and yd plots and modeling ;

symbol1 v="1" c=blue  i=none;
symbol2 v="2" c=red   i=none;
symbol3 v="3" c=green i=none;
proc gplot data=faculty;
  plot salary*(year yd)=rank;
run;

proc reg data=faculty;
  model salary=year yd;
run;
```



Output window contents:

```
The REG Procedure
Model: MODEL1
Dependent Variable: salary

Number of Observations Read          52
Number of Observations Used          52

                              Analysis of Variance
                                   Sum of           Mean
Source                     DF       Squares         Square    F Value    Pr > F
Model                       2    1032512437      516256219      33.58    <.0001
Error                      49     753217421       15371784
Corrected Total            51    1785729858

Root MSE           3920.68669    R-Square     0.5782
Dependent Mean        23798      Adj R-Sq     0.5610
Coeff Var           16.47510

                         Parameter Estimates
                      Parameter       Standard
Variable     DF        Estimate          Error    t Value    Pr > |t|
Intercept     1           16556     1052.38960      15.73      <.0001
year          1       489.29446      129.56024       3.78      0.0004
yd            1       222.25130       69.80376       3.18      0.0025
```

9.2.3 Using Factors and Predictors to Model Salaries

The plots we looked at helped us to understand the data. In particular, the plot of salary against years in rank, using rank as a plotting symbol, suggests that a combination of predictors and factors will likely be better for modelling faculty salaries than either of the two models that we proposed up to this point.

There is no evidence of non-linearity in the plots of salary against the predictors, so I will not consider transforming years since degree, years in rank, or salary. Note that the increasing variability in salaries for increasing years in rank and increasing years since degree is partly due to differences in the relationships across ranks. The non-constant variance should be less of a concern in any model that includes rank and either years in rank or years since degree as effects.

I started the model building process with a **maximal** or full model with the five main effects plus the 10 possible interactions between two effects, regardless of whether the effects were factors or predictors. Notationally, this model is written as follows:

$$\begin{aligned}\text{Salary} \;=\; &\text{Grand mean} + \text{S effect} + \text{D effect} + \text{R effect} + \text{YEAR effect} + \text{YD effect}\\ &+\text{S*D interaction} + \text{S*R interaction} + \text{S*YEAR interaction} + \text{S*YD interaction}\\ &+\text{D*R interaction} + \text{D*YEAR interaction} + \text{D*YD interaction}\\ &+\text{R*YEAR interaction} + \text{R*YD interaction} + \text{YEAR*YD interaction} + \text{Residual},\end{aligned}$$

where the year and year since degree effects (YD) are linear terms (as in the multiple regression model we considered). To check whether any important effects might have been omitted, I added individual three-factor terms to this model. All of the three factor terms were insignificant (not shown), so I believe that my choice for the "maximal" model is sensible.

The output below gives the fit to the maximal model, and subsequent fits, using the hierarchy principle. Only selected summaries are provided.

```
* ANCOVA model with two-factor interactions, backward selection;
proc glm data=faculty;
  class sex degree rank;
  model salary = sex degree rank year yd
```

```
                  sex*degree sex*rank sex*year sex*yd
                  degree*rank degree*year degree*yd
                  rank*year rank*yd  year*yd;
run;

The GLM Procedure

   Class Level Information
Class          Levels    Values
sex                 2    0 1
rank                3    1 2 3
degree              2    0 1

Number of Observations Read          52
Number of Observations Used          52

                                 Sum of
Source                     DF       Squares    Mean Square   F Value   Pr > F
Model                      20    1595904404       79795220     13.03   <.0001
Error                      31     189825454        6123402
Corrected Total            51    1785729858

R-Square     Coeff Var      Root MSE    salary Mean
0.893699      10.39830      2474.551       23797.65

Source                     DF    Type III SS    Mean Square   F Value   Pr > F
sex                         1     4058489.47     4058489.47      0.66   0.4218
degree                      1     3352151.89     3352151.89      0.55   0.4649
rank                        2     6455271.26     3227635.63      0.53   0.5955
year                        1    18204335.68    18204335.68      2.97   0.0946
yd                          1     1489654.92     1489654.92      0.24   0.6253
sex*degree                  1     7164814.86     7164814.86      1.17   0.2877
sex*rank                    2      932236.83      466118.41      0.08   0.9269
year*sex                    1     7194387.81     7194387.81      1.17   0.2868
yd*sex                      1     2024209.60     2024209.60      0.33   0.5695
rank*degree                 2    13021265.02     6510632.51      1.06   0.3576
year*degree                 1     4510249.50     4510249.50      0.74   0.3974
yd*degree                   1     6407880.34     6407880.34      1.05   0.3142
year*rank                   2     1571932.58      785966.29      0.13   0.8800
yd*rank                     2     9822382.38     4911191.19      0.80   0.4575
year*yd                     1       50921.06       50921.06      0.01   0.9279 *
```

omit year*yd

```
Source                     DF    Type III SS    Mean Square   F Value   Pr > F
sex                         1     4484364.68     4484364.68      0.76   0.3911
degree                      1     3301400.94     3301400.94      0.56   0.4612
rank                        2     6594564.32     3297282.16      0.56   0.5791
year                        1    38545543.98    38545543.98      6.50   0.0158
yd                          1     1689072.95     1689072.95      0.28   0.5973
sex*degree                  1     7341234.98     7341234.98      1.24   0.2743
sex*rank                    2      907204.91      453602.45      0.08   0.9266 *
year*sex                    1     7178185.84     7178185.84      1.21   0.2796
yd*sex                      1     2152917.40     2152917.40      0.36   0.5512
rank*degree                 2    13240859.17     6620429.59      1.12   0.3401
year*degree                 1     4601975.55     4601975.55      0.78   0.3851
yd*degree                   1     6443383.32     6443383.32      1.09   0.3052
year*rank                   2     1930801.51      965400.76      0.16   0.8505
yd*rank                     2     9944911.41     4972455.70      0.84   0.4418
```

omit sex*rank

```
Source                     DF    Type III SS    Mean Square   F Value   Pr > F
sex                         1    10921090.89    10921090.89      1.95   0.1720
degree                      1     2832620.39     2832620.39      0.50   0.4822
rank                        2    25954361.29    12977180.64      2.31   0.1144
year                        1    41663883.26    41663883.26      7.43   0.0101
yd                          1     2080327.09     2080327.09      0.37   0.5466
sex*degree                  1    10640012.32    10640012.32      1.90   0.1775
year*sex                    1    10690025.93    10690025.93      1.91   0.1765
yd*sex                      1     3614220.95     3614220.95      0.64   0.4278
rank*degree                 2    16341405.47     8170702.73      1.46   0.2473
year*degree                 1     4894264.68     4894264.68      0.87   0.3569
yd*degree                   1     6719487.31     6719487.31      1.20   0.2815
year*rank                   2     5037089.08     2518544.54      0.45   0.6421 *
yd*rank                     2    15110673.14     7555336.57      1.35   0.2737
```

omit year*rank

Source	DF	Type III SS	Mean Square	F Value	Pr > F	
sex	1	10220393.83	10220393.83	1.88	0.1789	
degree	1	3045986.41	3045986.41	0.56	0.4591	
rank	2	24094177.36	12047088.68	2.21	0.1238	
year	1	87576116.47	87576116.47	16.10	0.0003	
yd	1	16576626.29	16576626.29	3.05	0.0894	
sex*degree	1	12831930.96	12831930.96	2.36	0.1333	
year*sex	1	13646798.56	13646798.56	2.51	0.1220	
yd*sex	1	2456465.51	2456465.51	0.45	0.5059	*
degree*rank	2	21836321.53	10918160.76	2.01	0.1491	
year*degree	1	7414065.54	7414065.54	1.36	0.2507	
yd*degree	1	9232871.95	9232871.95	1.70	0.2009	
yd*rank	2	41051000.17	20525500.09	3.77	0.0325	

omit yd*sex

Source	DF	Type III SS	Mean Square	F Value	Pr > F	
sex	1	7789865.58	7789865.58	1.45	0.2356	
degree	1	2561622.51	2561622.51	0.48	0.4936	
rank	2	27586945.64	13793472.82	2.57	0.0898	
year	1	86966219.66	86966219.66	16.23	0.0003	
yd	1	16417850.36	16417850.36	3.06	0.0883	
sex*degree	1	10394381.71	10394381.71	1.94	0.1720	
year*sex	1	22789419.22	22789419.22	4.25	0.0463	
degree*rank	2	21157939.11	10578969.56	1.97	0.1532	
year*degree	1	8497323.64	8497323.64	1.59	0.2158	*
yd*degree	1	9463399.52	9463399.52	1.77	0.1920	
yd*rank	2	42516601.53	21258300.76	3.97	0.0275	

omit year*degree

Source	DF	Type III SS	Mean Square	F Value	Pr > F	
sex	1	2244296.99	2244296.99	0.41	0.5246	
degree	1	49292.05	49292.05	0.01	0.9247	
rank	2	21035595.30	10517797.65	1.93	0.1587	
year	1	79881400.66	79881400.66	14.68	0.0005	
yd	1	11377658.77	11377658.77	2.09	0.1564	
sex*degree	1	3189135.99	3189135.99	0.59	0.4487	
year*sex	1	14489583.55	14489583.55	2.66	0.1110	
degree*rank	2	13515717.46	6757858.73	1.24	0.3003	
yd*degree	1	1695057.57	1695057.57	0.31	0.5800	*
yd*rank	2	34725538.51	17362769.25	3.19	0.0524	

omit yd*degree

Source	DF	Type III SS	Mean Square	F Value	Pr > F	
sex	1	1580594.72	1580594.72	0.30	0.5897	
degree	1	7150038.39	7150038.39	1.34	0.2545	
rank	2	19495122.38	9747561.19	1.82	0.1749	
year	1	83056284.75	83056284.75	15.54	0.0003	
yd	1	10460123.63	10460123.63	1.96	0.1698	
sex*degree	1	2505272.02	2505272.02	0.47	0.4976	*
year*sex	1	12832092.78	12832092.78	2.40	0.1294	
degree*rank	2	15741804.61	7870902.30	1.47	0.2418	
yd*rank	2	38135455.26	19067727.63	3.57	0.0378	

omit sex*degree

Source	DF	Type III SS	Mean Square	F Value	Pr > F	
sex	1	486920.72	486920.72	0.09	0.7628	
degree	1	14742150.99	14742150.99	2.80	0.1024	
rank	2	23161942.21	11580971.10	2.20	0.1245	
year	1	83953539.19	83953539.19	15.92	0.0003	
yd	1	12439599.15	12439599.15	2.36	0.1325	
year*sex	1	11377954.49	11377954.49	2.16	0.1497	
degree*rank	2	14519996.53	7259998.27	1.38	0.2642	*
yd*rank	2	38113372.51	19056686.26	3.61	0.0361	

omit rank*degree

Source	DF	Type III SS	Mean Square	F Value	Pr > F	
sex	1	936434.63	936434.63	0.17	0.6783	
degree	1	8902097.81	8902097.81	1.66	0.2049	*
rank	2	91805629.64	45902814.82	8.55	0.0008	
year	1	98692459.97	98692459.97	18.38	0.0001	
yd	1	9517559.47	9517559.47	1.77	0.1902	
year*sex	1	14134386.21	14134386.21	2.63	0.1122	
yd*rank	2	24905278.46	12452639.23	2.32	0.1108	

Remark: This model has no interactions involving **degree**, so I can consider omitting **degree** at this point, and in fact this is the least significant effect left in the model!

omit degree

Source	DF	Type III SS	Mean Square	F Value	Pr > F	
sex	1	1311737.31	1311737.31	0.24	0.6262	
rank	2	91215248.64	45607624.32	8.37	0.0009	
year	1	89906978.00	89906978.00	16.49	0.0002	
yd	1	1965670.97	1965670.97	0.36	0.5513	
year*sex	1	11545390.57	11545390.57	2.12	0.1528	*
yd*rank	2	27221003.18	13610501.59	2.50	0.0942	

omit year*sex

Source	DF	Type III SS	Mean Square	F Value	Pr > F	
sex	1	5552915.6	5552915.6	0.99	0.3244	*
rank	2	122529230.5	61264615.3	10.96	0.0001	
year	1	106510254.2	106510254.2	19.06	<.0001	
yd	1	899779.9	899779.9	0.16	0.6902	
yd*rank	2	23603682.1	11801841.1	2.11	0.1332	

omit sex

Source	DF	Type III SS	Mean Square	F Value	Pr > F	
rank	2	117957347.8	58978673.9	10.55	0.0002	
year	1	106004567.4	106004567.4	18.97	<.0001	
yd	1	23568.3	23568.3	0.00	0.9485	
yd*rank	2	23183132.0	11591566.0	2.07	0.1375	*

omit yd*rank

Source	DF	Type III SS	Mean Square	F Value	Pr > F	
rank	2	478539100.8	239269550.4	40.94	<.0001	
year	1	141105646.9	141105646.9	24.14	<.0001	
yd	1	2314414.0	2314414.0	0.40	0.5322	*

omit yd

```
* FINALLY! ;
proc glm data=faculty;
  class sex degree rank;
  model salary  = rank year /solution;  * get individual regress coefficients;
  lsmeans rank /stderr pdiff;
run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1508737124	502912375	87.15	<.0001
Error	48	276992734	5770682		
Corrected Total	51	1785729858			

R-Square	Coeff Var	Root MSE	salary Mean
0.844885	10.09437	2402.224	23797.65

Source	DF	Type I SS	Mean Square	F Value	Pr > F
rank	2	1346783800	673391900	116.69	<.0001
year	1	161953324	161953324	28.06	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
rank	2	632056216.9	316028108.5	54.76	<.0001
year	1	161953324.0	161953324.0	28.06	<.0001

Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept		25657.79140 B	926.8077292	27.68	<.0001
rank	1	-9454.52325 B	905.8301038	-10.44	<.0001
rank	2	-5192.23854 B	871.8328225	-5.96	<.0001
rank	3	0.00000 B	.	.	
year		375.69564	70.9177250	5.30	<.0001

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

```
Least Squares Means
                salary          Standard                      LSMEAN
rank            LSMEAN             Error     Pr > |t|        Number
1          19013.7606          613.0513       <.0001             1
2          23276.0453          642.2996       <.0001             2
3          28468.2838          582.2789       <.0001             3

     Least Squares Means for effect rank
     Pr > |t| for H0: LSMean(i)=LSMean(j)
          Dependent Variable: salary
i/j               1             2             3
   1                         <.0001        <.0001
   2         <.0001                        <.0001
   3         <.0001        <.0001

NOTE: To ensure overall protection level, only probabilities associated with
      pre-planned comparisons should be used.
```

9.2.4 Discussion of the Salary Analysis

The selected model is a simple ANCOVA model with a rank effect and a linear effect due to years in rank. Note that the maximal model has 20 single df effects with an $R^2 = 0.89$ while the selected model has 3 single df effects with $R^2 = 0.84$.

Looking at the parameter estimates table, all of the single df effects in the selected model are significant. The baseline group is Full Professors, with rank=3. Predicted salaries for the different ranks are given by:

$$\begin{aligned}
\widehat{\text{salary}} &= 25658 + 375.70\text{year} && \text{for Full} \\
\widehat{\text{salary}} &= 25658 - 5192 + 375.70\text{year} = 20466 + 375.70\text{year} && \text{for Assoc} \\
\widehat{\text{salary}} &= 25658 - 9454 + 375.70\text{year} = 16204 + 375.70\text{year} && \text{for Assis}
\end{aligned}$$

Do you remember how to interpret the **lsmeans**, and the p-values for comparing **lsmeans**?

You might be tempted to conclude that rank and years in rank are the only effects that are predictive of salaries, and that differences in salaries by sex are insignificant, once these effects have been taken into account. However, you must be careful because you have not done a diagnostic analysis. The following two issues are also important to consider.

A sex effect may exist even though there is insufficient evidence to support it based on these data. (**Lack of power corrupts; and absolute lack of power corrupts absolutely!**) If we are interested in the possibility of a sex effect, I think that we would do better by focusing on how large the effect might be, and whether it is important. A simple way to check is by constructing a confidence interval for the sex effect, based on a simple additive model that includes sex plus the effects that were selected as statistically significant, rank and year in rank. I am choosing this model because the omitted effects are hopefully small, and because the regression coefficient for a sex indicator is easy to interpret in an additive model. Other models might be considered for comparison. Summary output from this model is given below.

Women are the baseline group for the sex effect, so the predicted salaries for men are 524 dollars **less** than that for women, adjusting for rank and year. A rough 95% CI for the sex differential is the estimated sex coefficient plus or minus two standard errors, or $-524 \pm 2 * (835)$, or -2194 to 1146 dollars. The range of plausible values for the sex effect would appear to contain values of practical importance, so further analysis is warranted here.

Program editor contents:

```
* investigating sex effect using an additive model ;
proc glm data=faculty;
  class sex degree rank;
  model salary = sex rank year/solution;
run;
```

Output window contents:

```
                                        Sum of
Source                      DF         Squares     Mean Square   F Value   Pr > F
Model                        4      1511041772       377760443     64.64   <.0001
Error                       47       274688086         5844427
Corrected Total             51      1785729858

R-Square     Coeff Var      Root MSE    salary Mean
0.846176      10.15867      2417.525        23797.65

Source                      DF       Type I SS     Mean Square   F Value   Pr > F
sex                          1       114106220       114106220     19.52   <.0001
rank                         2      1239752324       619876162    106.06   <.0001
year                         1       157183229       157183229     26.89   <.0001

Source                      DF     Type III SS     Mean Square   F Value   Pr > F
sex                          1       2304648.3       2304648.3      0.39   0.5331
rank                         2     634005384.6     317002692.3     54.24   <.0001
year                         1     157183229.0     157183229.0     26.89   <.0001


                                                    Standard
Parameter              Estimate                        Error    t Value    Pr > |t|

Intercept          25914.80355 B                 1018.558367      25.44      <.0001

sex         0 M     -524.14921 B                  834.686879      -0.63      0.5331
sex         1 F        0.00000 B                        .            .           .

rank        1      -9483.84187 B                  912.794505     -10.39      <.0001
rank        2      -5109.92648 B                  887.123194      -5.76      <.0001
rank        3          0.00000 B                        .            .           .

year                 390.93576                   75.382975       5.19      <.0001

NOTE: The X'X matrix has been found to be singular, and a generalized inverse
      was used to solve the normal equations.  Terms whose estimates are
      followed by the letter 'B' are not uniquely estimable.
```

Another concern, and potentially a more important issue, was raised by M. O. Finkelstein in a 1980 discussion in the **Columbia Law Review** on the use of regression in discrimination cases: "... [a] **variable may reflect a position or status bestowed by the employer, in which case if there is discrimination in the award of the position or status, the variable may be 'tainted'.**" *Thus, if women are unfairly held back from promotion through the faculty ranks, then using faculty rank to adjust salary before comparing sexes may not be acceptable to the courts.* This suggests that an analysis comparing sexes but ignoring rank effects might be justifiable. What happens if this is done?

Program editor contents:

```
* investigating sex without rank using an additive model ;
proc glm data=faculty;
  class sex degree rank;
  model salary = sex year/solution;
run;
```

Output window contents:

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		18266.87220	B	1247.699941	14.64	<.0001
sex	0	-201.46685	B	1455.145030	-0.14	0.8905
sex	1	0.00000	B	.	.	.
year		759.01384		118.336295	6.41	<.0001

Similar result as before, insufficient evidence between sexes (due to large proportion of variability in salary explained by rank). However, rank and sex are (potentially) confounded. This data can not resolve this question. Instead, data on promotions would help resolve this issue.

Chapter 10

Automated Model Selection for Multiple Regression

Given data on a response variable Y and k predictors or binary variables $X_1, X_2, \ldots, X_k$, we wish to develop a regression model to predict Y. Assuming that the collection of variables is measured on the correct scale, and that the candidate list of effects includes all the important predictors or binary variables, the most general model is

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon.$$

In most problems one or more of the effects can be eliminated from the full model without loss of information. We want to identify the important effects, or equivalently, eliminate the effects that are not very useful for explaining the variation in Y.

We will study several automated non-hierarchical methods for model selection. Given a specific criterion for selecting a model, **SAS** gives the best model. *Before applying these methods*, plot Y against each predictor to see whether transformations are needed. Although transformations of binary variables are not necessary, side-by-side boxplots of the response across the levels of a factor give useful information on the predictive ability of the factor. If a transformation of X_i is suggested, include the transformation along with the original X_i in the candidate list. Note that we can transform the predictors differently, for example, $\log(X_1)$ and $\sqrt{X_2}$. However, if several transformations are suggested for the response, then one should consider doing one analysis for each suggested response scale before deciding on the final scale.

Different criteria for selecting models lead to different "best models." Given a collection of candidates for the best model, we make a choice of model on the basis of (1) a direct comparison of models, if possible (2) examination of model adequacy (residuals, influence, etc.) (3) simplicity — all things being equal, simpler models are preferred, and (4) scientific plausibility.

I view the various criteria as a means to generate interesting models for further consideration. I do not take any of them literally as best.

You should recognize that automated model selection methods should not replace scientific theory when building models! Automated methods are best suited for exploratory analyses, in situations where the researcher has little scientific information as a guide.

10.1 Forward Selection

In forward selection we add variables to the model one at a time. The steps in the procedure are:

1. Find the variable in the candidate list with the largest correlation (ignoring the sign) with Y. This variable gives a simple linear regression model with the largest R^2. Suppose this is X_1. Then fit the model

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \tag{10.1}$$

and test $H_0 : \beta_1 = 0$. If we reject H_0, go to step 2. Otherwise stop and conclude that no variables are important. A t-test can be used here, but **SAS** uses an equivalent ANOVA F-test.

2. Find the remaining variable which when added to model (10.1) increases R^2 the most (or equivalently decreases Residual SS the most). Suppose this is X_2. Fit the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \tag{10.2}$$

and test $H_0 : \beta_2 = 0$. If we do not reject H_0, stop and use model (10.1) to predict Y. If we reject H_0, replace model (10.1) with (10.2) and repeat step 2 sequentially until no further variables are added to the model.

In forward selection we sequentially isolate the most important effect left in the pool, and check whether it is needed in the model. If it is needed we continue the process. Otherwise we stop.

In **SAS**, the default level for the tests on the individual effects is $\alpha = 0.50$. This may seem needlessly high. However, in many problems certain variables may be important only in the presence of other variables. If we force the forward selection to test at standard levels then the process will never get "going" when none of the variables is important on its own.

10.2 Backward Elimination

The backward elimination procedure (discussed earlier this semester) deletes unimportant variables, one at a time, starting from the full model. The steps is the procedure are:

1. Fit the full model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon.$$

2. Find the variable which when omitted from the full model (10.3) reduces R^2 the least, or equivalently, increases the Residual SS the least. This is the variable that gives the largest p-value for testing an individual regression coefficient $H_0 : \beta_j = 0$ for $j > 0$. Suppose this variable is X_k. If you reject H_0, stop and conclude that the full model is best. If you do not reject H_0, delete X_k from the full model, giving the new full model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{k-1} X_{k-1} + \varepsilon$$

to replace (10.3). Repeat steps 1 and 2 sequentially until no further variables can be deleted.

In backward elimination we isolate the least important effect left in the model, and check whether it is important. If not, delete it and repeat the process. Otherwise, stop. The default test level on the individual variables in **SAS** is $\alpha = 0.10$.

10.3 Stepwise Regression

Stepwise regression combines features of forward selection and backward elimination. A deficiency of forward selection is that variables can not be omitted from the model once they are selected. This is problematic because many variables that are initially important are not important once several other variables are included in the model. In stepwise regression, we add variables to the model as in forward regression, but include a backward elimination step every time a new variable is added. That is, every time we add a variable to the model we ask whether any of the variables added earlier can be omitted. If variables can be omitted, they are placed back into the candidate pool for consideration at the next step of the process. The process continues until no additional variables can be added, and none of the variables in the model can be excluded.

The p-values used for including and excluding variables in stepwise regression are usually taken to be equal (Why is this reasonable?). In **SAS**, the default level is $\alpha = 0.15$.

10.3.1 Example: Indian systolic blood pressure

(from chapter 2)

Anthropologists conducted a study to determine the long-term effects of an environmental change on systolic blood pressure. They measured the blood pressure and several other characteristics of 39 Indians who migrated from a very primitive environment high in the Andes into the mainstream of Peruvian society at a lower altitude. All of the Indians were males at least 21 years of age, and were born at a high altitude.

Let us *illustrate* the three model selection methods to build a regression model, using systolic blood pressure ($v9$) as the response, and seven candidate predictors: $v3$ = weight in kilos; $v4$ = height in mm; $v5$ = chin skin fold in mm; $v6$ = forearm skin fold in mm; $v7$ = calf skin fold in mm; $v8$ = pulse rate-beats/min, and $v11$ = fraction, which is the proportion of each individual's lifetime spent in the new environment.

The **SAS** program given below generates simple summary statistics and plots. The plots do not suggest any apparent transformations of the response or the predictors, so we will analyze the data using the given scales.

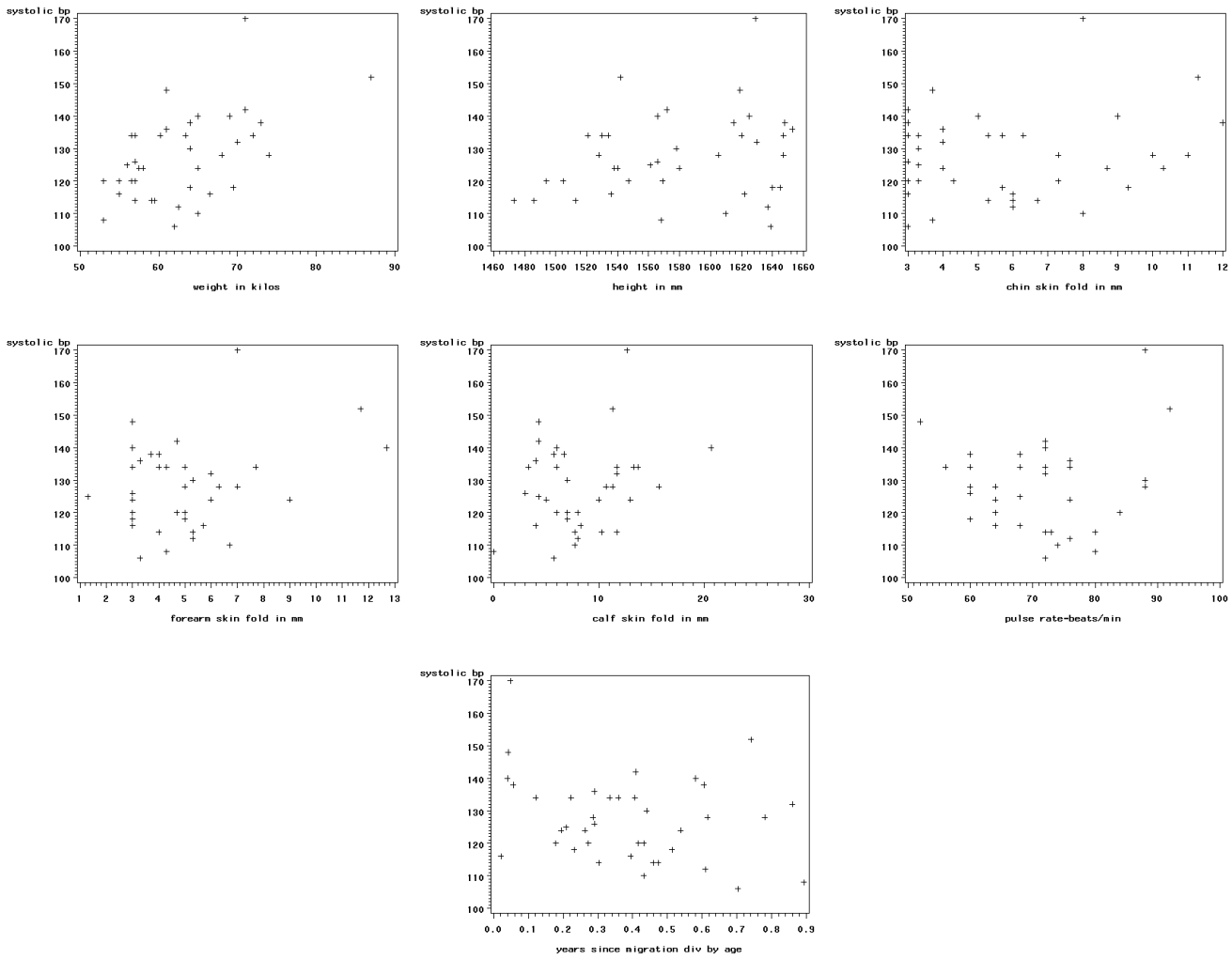
Program editor contents:

```
* Indian systolic blood pressure example;
options ls=79 nodate nocenter;

data indian;
  infile 'F:\Dropbox\UNM\teach\ADA2_stat528\sas\ADA2_02_indian.dat';
  input n v1-v11;
  label v1='age in years'          v2='years since migration'
    v3='weight in kilos'          v4='height in mm'
    v5='chin skin fold in mm'     v6='forearm skin fold in mm'
    v7='calf skin fold in mm'     v8='pulse rate-beats/min'
    v9='systolic bp'             v10='diastolic bp'
    v11='years since migration div by age';
run;

* look at Pearson correlations with response variable;
proc corr data=indian nosimple noprob;
  var v9;
  with v3 v4 v5 v6 v7 v8 v11;
run;

* plot variables with response;
```

```
proc gplot data=indian;
  plot v9*(v3 v4 v5 v6 v7 v8 v11);
run;
```

Output window contents:

```
The CORR Procedure
   7 With Variables:    v3     v4     v5     v6     v7     v8      v11
   1      Variables:    v9

   Pearson Correlation Coefficients, N = 39
                                          v9
v3  weight in kilos                    0.52136
v4  height in mm                       0.21911
v5  chin skin fold in mm               0.17019
v6  forearm skin fold in mm            0.27228
v7  calf skin fold in mm               0.25079
v8  pulse rate-beats/min               0.13256
v11 years since migration div by age -0.27615
```



To get output from the various model selection procedures, specify the following **SAS** statements in addition to the **data** step:
Program editor contents:

```
* selection methods: forward, backward, stepwise;
proc reg data=indian;
  model v9 = v3 v4 v5 v6 v7 v8 v11 / selection = forward;
  model v9 = v3 v4 v5 v6 v7 v8 v11 / selection = backward;
  model v9 = v3 v4 v5 v6 v7 v8 v11 / selection = stepwise;
run;
```

Forward selection output The output for the forward selection method is below.

```
The REG Procedure
Model: MODEL1
Dependent Variable: v9 systolic bp

Number of Observations Read          39
Number of Observations Used          39

Forward Selection: Step 1

Variable v3 Entered: R-Square = 0.2718 and C(p) = 12.6151

                          Analysis of Variance
                                   Sum of           Mean
Source                    DF        Squares         Square    F Value    Pr > F
Model                      1     1775.37963     1775.37963      13.81    0.0007
Error                     37     4756.05627      128.54206
Corrected Total           38     6531.43590

              Parameter     Standard
Variable       Estimate        Error   Type II SS  F Value  Pr > F
Intercept      66.59687     16.46390   2103.23243    16.36  0.0003
v3              0.96286      0.25908   1775.37963    13.81  0.0007

Bounds on condition number: 1, 1
```

Variable $v3$ =weight is entered first because it has the highest correlation with $v9$ =sys bp. The underlined F-value is the square of the t-statistic for testing the significance of the weight predictor in this simple linear regression model. The p-value is the same as the t-test p-value. The p-value is less than the default 0.50 level for inclusion, so we proceed to the next step.

```
Forward Selection: Step 2

Variable v11 Entered: R-Square = 0.4731 and C(p) = 1.4531

                          Analysis of Variance
                                   Sum of           Mean
Source                    DF        Squares         Square    F Value    Pr > F
Model                      2     3090.07564     1545.03782      16.16    <.0001
Error                     36     3441.36025       95.59334
Corrected Total           38     6531.43590

              Parameter     Standard
Variable       Estimate        Error   Type II SS  F Value  Pr > F
Intercept      60.89595     14.28088   1738.18032    18.18  0.0001
v3              1.21686      0.23369   2592.00831    27.11  <.0001
v11           -26.76721      7.21779   1314.69602    13.75  0.0007

Bounds on condition number: 1.094, 4.3759
```

Adding $v11$ =fraction to the simple linear regression model with weight as a predictor increases R^2 the most, or equivalently, decreases Residual SS the most. The underlined SS is the reduction in Residual SS for the simple linear regression model achieved by adding fraction as a predictor. The p-value for $v11$ is smaller than 0.50, so proceed to the next step.

```
Forward Selection: Step 3

Variable v5 Entered: R-Square = 0.5033 and C(p) = 1.4771

                              Analysis of Variance
                                       Sum of           Mean
Source                     DF         Squares         Square    F Value    Pr > F
Model                       3      3287.44669     1095.81556      11.82    <.0001
Error                      35      3243.98921       92.68541
Corrected Total            38      6531.43590

              Parameter     Standard
Variable       Estimate        Error   Type II SS  F Value  Pr > F
Intercept      52.90931     15.08952   1139.52747    12.29  0.0013
v3              1.44074      0.27656   2515.32927    27.14  <.0001
v5             -1.01353      0.69454    197.37104     2.13  0.1534
v11           -27.35223      7.11846   1368.43705    14.76  0.0005

Bounds on condition number: 1.5803, 12.43
```

Note that the Residual df decreases by one, and the Regression df increases by one for each additional predictor added to the model. At the third step, $v5$ =chin skin fold is added. The p-value for testing the importance of $v5$ when added last to the three predictor model is 0.153, which is less than the default 0.50 level, so proceed to the next step.

Regardless of which predictor was included at this step, the p-values in the far right column indicate the importance of the predictors when added last to the model.

```
Forward Selection: Step 4

Variable v4 Entered: R-Square = 0.5207 and C(p) = 2.3402

                              Analysis of Variance
                                       Sum of           Mean
Source                     DF         Squares         Square    F Value    Pr > F
Model                       4      3401.01104      850.25276       9.23    <.0001
Error                      34      3130.42486       92.07132
Corrected Total            38      6531.43590

              Parameter     Standard
Variable       Estimate        Error   Type II SS  F Value  Pr > F
Intercept     104.52276     48.84626    421.58326     4.58  0.0396
v3              1.64631      0.33202   2263.63282    24.59  <.0001
v4             -0.03957      0.03563    113.56435     1.23  0.2745
v5             -1.31083      0.74220    287.19721     3.12  0.0863
v11           -28.32576      7.14878   1445.51712    15.70  0.0004

Bounds on condition number: 2.2929, 26.188

No other variable met the 0.5000 significance level for entry into the model.

                        Summary of Forward Selection
      Variable   Number   Partial    Model
Step  Entered    Vars In  R-Square  R-Square    C(p)       F Value    Pr > F
  1   v3              1    0.2718    0.2718   12.6151      13.81    0.0007
  2   v11             2    0.2013    0.4731    1.4531      13.75    0.0007
  3   v5              3    0.0302    0.5033    1.4771       2.13    0.1534
  4   v4              4    0.0174    0.5207    2.3402       1.23    0.2745
```

In the summary table, the partial R^2 gives the increase in R^2 achieved by adding the given variables to the model, one at a time, in the specified order. The model R^2 column gives R^2 for the model that includes all variables added up to the specified point.

Backward selection output The output for the backward elimination method is below.

```
The REG Procedure
Model: MODEL2
Dependent Variable: v9 systolic bp

Number of Observations Read          39
Number of Observations Used          39

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.5259 and C(p) = 8.0000

                              Analysis of Variance
                                        Sum of           Mean
Source                     DF          Squares         Square    F Value    Pr > F
Model                       7       3434.98916      490.71274       4.91    0.0008
Error                      31       3096.44673       99.88538
Corrected Total            38       6531.43590

              Parameter     Standard
Variable       Estimate        Error   Type II SS  F Value  Pr > F
Intercept    106.45745     53.91303    389.46328     3.90  0.0573
v3             1.71094      0.38659   1956.48532    19.59  0.0001
v4            -0.04533      0.03945    131.87622     1.32  0.2593
v5            -1.15724      0.84612    186.84503     1.87  0.1812
v6            -0.70182      1.34986     27.00094     0.27  0.6068
v7             0.10357      0.61170      2.86340     0.03  0.8667 *
v8             0.07485      0.19570     14.61395     0.15  0.7047
v11          -29.31807      7.86839   1386.76004    13.88  0.0008

Bounds on condition number: 3.6493, 105.6
```

Remarks on Step 0: The full model has 7 predictors so REG df = 7. The F-test in the full model ANOVA table ($F = 4.91$ with p-value=0.0008) tests the hypothesis that the regression coefficient for each predictor variable is zero. This test is highly significant, indicating that one or more of the predictors is important in the model.

The underlined F-value column gives the square of the t-statistic for testing the significance of the individual predictors in the full model. The p-value is the same as the t-test p-value for testing the significance of individual regression effects.

The least important variable in the full model, as judged by the p-value, is $v7$ =calf skin fold. This variable, upon omission, reduces R^2 the least, or equivalently, increases the Residual SS the least. The p-value of 0.87 exceeds the default 0.10 cut-off, so $v7$ is omitted from the model.

```
Backward Elimination: Step 1

Variable v7 Removed: R-Square = 0.5255 and C(p) = 6.0287

                              Analysis of Variance
                                        Sum of           Mean
Source                     DF          Squares         Square    F Value    Pr > F
Model                       6       3432.12577      572.02096       5.91    0.0003
Error                      32       3099.31013       96.85344
Corrected Total            38       6531.43590

              Parameter     Standard
Variable       Estimate        Error   Type II SS  F Value  Pr > F
Intercept    106.13720     53.05580    387.60072     4.00  0.0540
v3             1.70900      0.38051   1953.76360    20.17  <.0001
v4            -0.04478      0.03871    129.55602     1.34  0.2560
v5            -1.14165      0.82823    184.02391     1.90  0.1776
v6            -0.56731      1.07462     26.99288     0.28  0.6012
v8             0.07103      0.19142     13.33727     0.14  0.7130 *
v11          -29.53996      7.63982   1448.00203    14.95  0.0005

Bounds on condition number: 2.8627, 68.302
```

Remarks on Step 1: After deleting $v7$, the six predictor model is fitted. At least one of the predictors left is important, as judged by the overall F-test p-value. The least important predictor left is $v8$ =pulse rate. This variable is omitted from the model because the p-value for including it exceeds the 0.10 threshold.

```
Backward Elimination: Step 2

Variable v8 Removed: R-Square = 0.5234 and C(p) = 4.1622

                          Analysis of Variance
                                   Sum of           Mean
Source                     DF       Squares         Square    F Value    Pr > F
Model                       5    3418.78850      683.75770       7.25    0.0001
Error                      33    3112.64740       94.32265
Corrected Total            38    6531.43590

             Parameter     Standard
Variable      Estimate        Error   Type II SS  F Value  Pr > F
Intercept    110.27853     51.18665    437.80889     4.64  0.0386
v3             1.71825      0.37470   1983.50473    21.03  <.0001
v4            -0.04504      0.03820    131.12231     1.39  0.2468
v5            -1.17715      0.81187    198.29475     2.10  0.1565
v6            -0.43384      0.99933     17.77746     0.19  0.6670 *
v11          -28.98166      7.39170   1450.01807    15.37  0.0004

Bounds on condition number: 2.8504, 48.657


Backward Elimination: Step 3

Variable v6 Removed: R-Square = 0.5207 and C(p) = 2.3402

                          Analysis of Variance
                                   Sum of           Mean
Source                     DF       Squares         Square    F Value    Pr > F
Model                       4    3401.01104      850.25276       9.23    <.0001
Error                      34    3130.42486       92.07132
Corrected Total            38    6531.43590

             Parameter     Standard
Variable      Estimate        Error   Type II SS  F Value  Pr > F
Intercept    104.52276     48.84626    421.58326     4.58  0.0396
v3             1.64631      0.33202   2263.63282    24.59  <.0001
v4            -0.03957      0.03563    113.56435     1.23  0.2745 *
v5            -1.31083      0.74220    287.19721     3.12  0.0863
v11          -28.32576      7.14878   1445.51712    15.70  0.0004

Bounds on condition number: 2.2929, 26.188


Backward Elimination: Step 4

Variable v4 Removed: R-Square = 0.5033 and C(p) = 1.4771

                          Analysis of Variance
                                   Sum of           Mean
Source                     DF       Squares         Square    F Value    Pr > F
Model                       3    3287.44669     1095.81556      11.82    <.0001
Error                      35    3243.98921       92.68541
Corrected Total            38    6531.43590

             Parameter     Standard
Variable      Estimate        Error   Type II SS  F Value  Pr > F
Intercept     52.90931     15.08952   1139.52747    12.29  0.0013
v3             1.44074      0.27656   2515.32927    27.14  <.0001
v5            -1.01353      0.69454    197.37104     2.13  0.1534 *
v11          -27.35223      7.11846   1368.43705    14.76  0.0005

Bounds on condition number: 1.5803, 12.43
```

```
Backward Elimination: Step 5

Variable v5 Removed: R-Square = 0.4731 and C(p) = 1.4531

                              Analysis of Variance
                                        Sum of           Mean
Source                     DF          Squares         Square    F Value    Pr > F
Model                       2       3090.07564     1545.03782      16.16    <.0001
Error                      36       3441.36025       95.59334
Corrected Total            38       6531.43590

              Parameter     Standard
Variable       Estimate        Error   Type II SS  F Value  Pr > F
Intercept      60.89595     14.28088   1738.18032    18.18  0.0001
v3              1.21686      0.23369   2592.00831    27.11  <.0001
v11           -26.76721      7.21779   1314.69602    13.75  0.0007

Bounds on condition number: 1.094, 4.3759


All variables left in the model are significant at the 0.1000 level.

                      Summary of Backward Elimination
      Variable   Number   Partial    Model
Step  Removed    Vars In  R-Square  R-Square  C(p)       F Value    Pr > F
  1   v7               6    0.0004    0.5255   6.0287       0.03    0.8667
  2   v8               5    0.0020    0.5234   4.1622       0.14    0.7130
  3   v6               4    0.0027    0.5207   2.3402       0.19    0.6670
  4   v4               3    0.0174    0.5033   1.4771       1.23    0.2745
  5   v5               2    0.0302    0.4731   1.4531       2.13    0.1534
```

In the Summary Table, the partial R^2 is the reduction in R^2 achieved by omitting variables sequentially.

Stepwise selection output The output for the stepwise selection is given below.

```
The REG Procedure
Model: MODEL3
Dependent Variable: v9 systolic bp

Number of Observations Read          39
Number of Observations Used          39

Stepwise Selection: Step 1

Variable v3 Entered: R-Square = 0.2718 and C(p) = 12.6151

                              Analysis of Variance
                                        Sum of           Mean
Source                     DF          Squares         Square    F Value    Pr > F
Model                       1     1775.37963     1775.37963      13.81    0.0007
Error                      37     4756.05627      128.54206
Corrected Total            38     6531.43590

              Parameter     Standard
Variable       Estimate        Error   Type II SS  F Value  Pr > F
Intercept      66.59687     16.46390   2103.23243    16.36  0.0003
v3              0.96286      0.25908   1775.37963    13.81  0.0007

Bounds on condition number: 1, 1
```

```
Stepwise Selection: Step 2

Variable v11 Entered: R-Square = 0.4731 and C(p) = 1.4531

                              Analysis of Variance
                                      Sum of           Mean
Source                     DF         Squares         Square    F Value    Pr > F
Model                       2     3090.07564     1545.03782      16.16    <.0001
Error                      36     3441.36025       95.59334
Corrected Total            38     6531.43590

              Parameter     Standard
Variable       Estimate        Error   Type II SS  F Value  Pr > F
Intercept      60.89595     14.28088   1738.18032    18.18  0.0001
v3              1.21686      0.23369   2592.00831    27.11  <.0001
v11           -26.76721      7.21779   1314.69602    13.75  0.0007

Bounds on condition number: 1.094, 4.3759

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

                          Summary of Stepwise Selection
      Variable  Variable  Number   Partial    Model
Step  Entered   Removed   Vars In  R-Square  R-Square   C(p)     F Value  Pr > F
  1   v3                      1     0.2718    0.2718   12.6151    13.81  0.0007
  2   v11                     2     0.2013    0.4731    1.4531    13.75  0.0007
```

Summary of three section methods

The forward selection method chooses four predictors for the model, in the following order: weight ($v3$), fraction ($v11$), chin skin fold ($v5$), and height ($v4$). As we progress from the simple linear regression model with weight as the predictor to the four predictor model, the value of R^2 increases as follows: $R^2 = 0.27$, 0.47, 0.50, and 0.52. The last two variables selected barely increase R^2.

The backward elimination procedure eliminates five variables from the full model, in the following order: calf skin fold ($v7$), pulse rate ($v8$), forearm skin fold ($v6$), height ($v4$), and chin skin fold ($v5$). The model selected by backward elimination includes two predictors: weight ($v3$) and fraction ($v11$). As we progress from the full model to the selected model, R^2 decreases as follows: $R^2 = 0.53$, 0.53, 0.52, 0.52, 0.50, and 0.47. The decrease is slight across this spectrum of models.

The stepwise selection method includes only two predictors: weight ($v3$) and fraction ($v11$).

10.4 Other Model Selection Procedures

10.4.1 R^2 Criterion

A substantial increase in R^2 is usually observed when an "important" effect is added to a regression model. With the R^2 criterion, variables are added to the model until further additions give inconsequential increases in R^2. The R^2 criterion is not well-defined in the sense of producing a single best model. All other things being equal, I prefer the simpler of two models with similar values of R^2. If several models with similar complexity have similar R^2s, then there may be no good reason, at this stage of the analysis, to prefer one over the rest.

In **SAS**, you can get the R^2 for all possible models when the number of candidate variables is 10 or less by using the **selection=rsquare** option on the **model** statement. Alternatively, you can get the m models (where m is a positive integer) with the highest R^2s among all models with

the same number of variables by choosing **selection=rsquare best=m**. The second approach is sufficient to highlight good models according to this criterion.

10.4.2 Adjusted-R^2 Criterion, maximize

The adjusted-R^2 criterion gives a way to compare R^2 across models with different numbers of variables. This eliminates some of the difficulty with calibrating R^2, which increases even when unimportant predictors are added to a model. For a model with p variables and an intercept, the adjusted-R^2 is defined by

$$\overline{R}^2 = 1 - \frac{n-1}{n-p-1}(1-R^2),$$

where n is the sample size.

There are four properties of $\overline{R}^2$ worth mentioning:

1. $\overline{R}^2 \leq R^2$,
2. if two models have the same number of variables, then the model with the larger R^2 has the larger $\overline{R}^2$,
3. if two models have the same R^2, then the model with fewer variables has the larger adjusted-R^2. Put another way, $\overline{R}^2$ *penalizes complex models with many variables.* And
4. $\overline{R}^2$ can be less than zero for models that explain little of the variation in Y.

The adjusted R^2 is easier to calibrate than R^2 because it tends to *decrease* when unimportant variables are added to a model. The model with the maximum $\overline{R}^2$ is judged best by this criterion. As I noted before, I do not take any of the criteria literally, and would also choose other models with $\overline{R}^2$ near the maximum value for further consideration.

The adjusted R^2 method is implemented in **reg** by the **selection=adjrsq** option. The $\overline{R}^2$ is given for each possible model when the number of candidate effects is 10 or less. If the **best=m** option is selected, then **SAS** provides $\overline{R}^2$ for the m models with the highest $\overline{R}^2$ **across all models**.

10.4.3 Mallows' C_p Criterion, minimize

Mallows' C_p statistic is defined for a given model with p variables by

$$C_p = \frac{\text{Residual SS}}{\hat{\sigma}^2_{\text{FULL}}} - \text{Residual df} + (p+1)$$

where $\hat{\sigma}^2_{\text{FULL}}$ is the Residual MS from the full model with k variables $X_1, X_2, \ldots, X_k$.

Mallows' C_p measures the adequacy of predictions from a model, relative to those obtained from the full model. If all the important effects from the candidate list are included in the model, then the difference between the first two terms of C_p should be approximately zero. Thus, if the model under consideration includes all the important variables from the candidate list, then C_p should be approximately $p+1$ (the number of variables in model plus one), or less. If important variables from the candidate list are excluded, C_p will tend to be much greater than $p+1$.

Two important properties of C_p are

1. the full model has $C_p = p + 1$, where $p = k$, and

2. if two models have the same number of variables, then the model with the larger R^2 has the smaller C_p.

Models with $C_p \approx p+1$, or less, merit further consideration. As with R^2 and $\overline{R}^2$, I prefer simpler models that satisfy this condition. The "best" model by this criterion has the minimum C_p.

The **selection=cp** option computes C_p in **reg**. If $k \leq 10$, you get C_p for all models. With the **best=m** option, you get C_p for the m models with the lowest C_ps among all models.

10.5 Illustration with Peru Indian data

Program editor contents:

```
* selection methods: rsquare, adjrsq, cp;
proc reg data=indian;
  model v9 = v3 v4 v5 v6 v7 v8 v11/ selection = rsquare best=5;
  model v9 = v3 v4 v5 v6 v7 v8 v11/ selection = adjrsq cp;
run;
```

selection=rsquare best=5 Output window contents:

```
The REG Procedure
Model: MODEL1
Dependent Variable: v9

R-Square Selection Method

Number of Observations Read          39
Number of Observations Used          39

Number in
  Model      R-Square    Variables in Model
        1       0.2718    v3                        *
        1       0.0763    v11
        1       0.0741    v6
        1       0.0629    v7
        1       0.0480    v4
-----------------------------------------------
        2       0.4731    v3 v11                    *
        2       0.2938    v3 v5
        2       0.2744    v3 v7
        2       0.2728    v3 v8
        2       0.2721    v3 v4
-----------------------------------------------
        3       0.5033    v3 v5 v11                 -
        3       0.4808    v3 v6 v11
        3       0.4777    v3 v7 v11
        3       0.4767    v3 v4 v11
        3       0.4740    v3 v8 v11
-----------------------------------------------
        4       0.5207    v3 v4 v5 v11
        4       0.5051    v3 v5 v8 v11
        4       0.5034    v3 v5 v6 v11
        4       0.5033    v3 v5 v7 v11
        4       0.4931    v3 v4 v6 v11
-----------------------------------------------
        5       0.5234    v3 v4 v5 v6 v11
        5       0.5213    v3 v4 v5 v8 v11
        5       0.5210    v3 v4 v5 v7 v11
        5       0.5056    v3 v5 v6 v8 v11
        5       0.5052    v3 v5 v7 v8 v11
-----------------------------------------------
        6       0.5255    v3 v4 v5 v6 v8 v11
        6       0.5237    v3 v4 v5 v6 v7 v11
        6       0.5218    v3 v4 v5 v7 v8 v11
        6       0.5057    v3 v5 v6 v7 v8 v11
        6       0.4973    v3 v4 v6 v7 v8 v11
-----------------------------------------------
        7       0.5259    v3 v4 v5 v6 v7 v8 v11
```

selection=adjrsq cp Output window contents:

```
The REG Procedure
Model: MODEL1
Dependent Variable: v9

Adjusted R-Square Selection Method

Number of Observations Read          39
Number of Observations Used          39

Number in    Adjusted
  Model      R-Square    R-Square        C(p)    Variables in Model
       4        0.4643      0.5207      2.3402    v3 v4 v5 v11
       3        0.4608      0.5033      1.4771    v3 v5 v11            -
       5        0.4512      0.5234      4.1622    v3 v4 v5 v6 v11
       5        0.4488      0.5213      4.2989    v3 v4 v5 v8 v11
       5        0.4485      0.5210      4.3189    v3 v4 v5 v7 v11
       4        0.4469      0.5051      3.3588    v3 v5 v8 v11
       4        0.4449      0.5034      3.4749    v3 v5 v6 v11
       4        0.4449      0.5033      3.4771    v3 v5 v7 v11
       2        0.4438      0.4731      1.4531    v3 v11               *
       6        0.4365      0.5255      6.0287    v3 v4 v5 v6 v8 v11

       3        0.4363      0.4808      2.9470    v3 v6 v11
       6        0.4344      0.5237      6.1463    v3 v4 v5 v6 v7 v11
       4        0.4334      0.4931      4.1474    v3 v4 v6 v11
       3        0.4330      0.4777      3.1506    v3 v7 v11
       6        0.4321      0.5218      6.2703    v3 v4 v5 v7 v8 v11
       3        0.4319      0.4767      3.2154    v3 v4 v11
       5        0.4307      0.5056      5.3257    v3 v5 v6 v8 v11
       5        0.4302      0.5052      5.3564    v3 v5 v7 v8 v11
       3        0.4289      0.4740      3.3942    v3 v8 v11
       5        0.4281      0.5034      5.4741    v3 v5 v6 v7 v11

       4        0.4244      0.4850      4.6743    v3 v6 v8 v11
       4        0.4235      0.4841      4.7315    v3 v4 v7 v11
       5        0.4211      0.4973      5.8710    v3 v4 v6 v8 v11
       4        0.4200      0.4810      4.9365    v3 v6 v7 v11
       7        0.4189      0.5259      8.0000    v3 v4 v5 v6 v7 v8 v11
       4        0.4181      0.4793      5.0455    v3 v7 v8 v11
       5        0.4163      0.4931      6.1452    v3 v4 v6 v7 v11
       4        0.4157      0.4772      5.1861    v3 v4 v8 v11
       6        0.4130      0.5057      7.3203    v3 v5 v6 v7 v8 v11
       5        0.4071      0.4851      6.6697    v3 v4 v7 v8 v11

       5        0.4070      0.4850      6.6730    v3 v6 v7 v8 v11
       6        0.4031      0.4973      7.8706    v3 v4 v6 v7 v8 v11
       2        0.2546      0.2938     13.1772    v3 v5
       1        0.2521      0.2718     12.6151    v3                   *
       3        0.2482      0.3076     14.2759    v3 v5 v7
       3        0.2393      0.2994     14.8119    v3 v4 v5
       3        0.2380      0.2981     14.8952    v3 v5 v6
       2        0.2341      0.2744     14.4482    v3 v7
       3        0.2338      0.2943     15.1477    v3 v5 v8
       2        0.2324      0.2728     14.5540    v3 v8

       2        0.2317      0.2721     14.5950    v3 v4
       2        0.2316      0.2720     14.6035    v3 v6
       4        0.2307      0.3117     16.0082    v3 v4 v5 v7
       4        0.2272      0.3086     16.2125    v3 v5 v7 v8
       4        0.2262      0.3076     16.2728    v3 v5 v6 v7
       4        0.2192      0.3014     16.6790    v3 v4 v5 v6
       4        0.2182      0.3005     16.7419    v3 v4 v5 v8
       4        0.2178      0.3002     16.7611    v3 v5 v6 v8
       3        0.2167      0.2785     16.1757    v3 v6 v7
       3        0.2136      0.2757     16.3641    v3 v7 v8

       3        0.2122      0.2744     16.4453    v3 v4 v7
       3        0.2110      0.2733     16.5201    v3 v4 v8
       3        0.2104      0.2728     16.5530    v3 v6 v8
       3        0.2103      0.2726     16.5610    v3 v4 v6
       5        0.2093      0.3133     17.9002    v3 v4 v5 v7 v8
       5        0.2083      0.3125     17.9557    v3 v4 v5 v6 v7
       5        0.2038      0.3086     18.2113    v3 v5 v6 v7 v8
       5        0.1983      0.3038     18.5253    v3 v4 v5 v6 v8
       4        0.1949      0.2797     18.1022    v3 v4 v6 v7
       4        0.1939      0.2788     18.1614    v3 v6 v7 v8

       4        0.1906      0.2758     18.3547    v3 v4 v7 v8
       4        0.1880      0.2735     18.5075    v3 v4 v6 v8
       6        0.1849      0.3136     19.8835    v3 v4 v5 v6 v7 v8
```

5	0.1708	0.2799	20.0868	v3 v4 v6 v7 v8
3	0.1523	0.2192	20.0553	v4 v6 v11
4	0.1342	0.2254	21.6522	v4 v6 v8 v11
4	0.1280	0.2198	22.0200	v4 v5 v6 v11
4	0.1274	0.2192	22.0550	v4 v6 v7 v11
5	0.1091	0.2263	23.5887	v4 v5 v6 v8 v11
3	0.1090	0.1793	22.6624	v4 v7 v11
5	0.1081	0.2254	23.6481	v4 v6 v7 v8 v11
2	0.1078	0.1547	22.2718	v6 v11
4	0.1076	0.2016	23.2092	v4 v7 v8 v11
4	0.1051	0.1993	23.3572	v4 v5 v8 v11
3	0.1034	0.1741	23.0025	v4 v5 v11
5	0.1016	0.2198	24.0184	v4 v5 v6 v7 v11
3	0.0988	0.1699	23.2776	v4 v8 v11
4	0.0955	0.1907	23.9203	v4 v5 v7 v11
5	0.0906	0.2103	24.6399	v4 v5 v7 v8 v11
3	0.0901	0.1619	23.8032	v6 v8 v11
3	0.0835	0.1558	24.1986	v5 v6 v11
2	0.0827	0.1310	23.8236	v4 v6
3	0.0827	0.1551	24.2483	v6 v7 v11
2	0.0825	0.1308	23.8360	v4 v11
6	0.0813	0.2264	25.5873	v4 v5 v6 v7 v8 v11
2	0.0765	0.1251	24.2094	v7 v11
3	0.0735	0.1467	24.7975	v7 v8 v11
3	0.0695	0.1430	25.0403	v5 v8 v11
2	0.0692	0.1182	24.6593	v5 v11
2	0.0654	0.1146	24.8937	v8 v11
4	0.0653	0.1637	25.6883	v5 v6 v8 v11
4	0.0643	0.1628	25.7454	v6 v7 v8 v11
2	0.0618	0.1112	25.1167	v4 v7
3	0.0614	0.1355	25.5269	v5 v7 v11
3	0.0601	0.1343	25.6049	v4 v6 v7
4	0.0568	0.1561	26.1838	v5 v6 v7 v11
3	0.0568	0.1312	25.8074	v4 v5 v6
3	0.0566	0.1311	25.8156	v4 v6 v8
4	0.0552	0.1546	26.2785	v5 v7 v8 v11
1	0.0513	0.0763	25.4029	v11
1	0.0491	0.0741	25.5416	v6
3	0.0421	0.1178	26.6885	v4 v7 v8
3	0.0377	0.1136	26.9580	v4 v5 v7
5	0.0377	0.1643	27.6463	v5 v6 v7 v8 v11
1	0.0376	0.0629	26.2766	v7
4	0.0330	0.1348	27.5762	v4 v6 v7 v8
4	0.0330	0.1348	27.5765	v4 v5 v6 v7
4	0.0291	0.1313	27.8008	v4 v5 v6 v8
2	0.0286	0.0797	27.1780	v6 v7
2	0.0263	0.0776	27.3170	v4 v5
2	0.0231	0.0745	27.5167	v6 v8
2	0.0227	0.0742	27.5402	v5 v6
1	0.0223	0.0480	27.2499	v4
2	0.0179	0.0696	27.8399	v7 v8
4	0.0156	0.1192	28.5917	v4 v5 v7 v8
2	0.0135	0.0654	28.1120	v4 v8
2	0.0132	0.0652	28.1284	v5 v7
3	0.0085	0.0868	28.7161	v4 v5 v8
5	0.0041	0.1352	29.5507	v4 v5 v6 v7 v8
1	0.0027	0.0290	28.4953	v5
3	0.0019	0.0807	29.1105	v6 v7 v8
3	0.0010	0.0798	29.1697	v5 v6 v7
3	-.0048	0.0745	29.5160	v5 v6 v8
3	-.0087	0.0709	29.7519	v5 v7 v8
1	-.0090	0.0176	29.2403	v8
2	-.0151	0.0384	29.8810	v5 v8
4	-.0273	0.0808	31.1045	v5 v6 v7 v8

10.5.1 R^2, $\overline{R}^2$, and C_p Summary for Peru Indian Data

Discussion of R^2 results:

1. The single predictor model with the highest value of R^2 has $v3$ = weight as a predictor: $R^2 = 0.272$. All the other single predictor models have $R^2 < 0.10$.

2. The two predictor model with the highest value of R^2 has weight and $v11$ = fraction as predictors: $R^2 = 0.473$. No other two predictor model has R^2 close to this.

3. All of the best three predictor models include weight and fraction as predictors. However, the increase in R^2 achieved by adding a third predictor is minimal.

4. None of the more complex models with four or more predictors provides a significant increase in R^2.

A good model using the R^2 criterion has two predictors, weight and fraction. The same conclusion is reached with $\overline{R}^2$, albeit the model with maximum $\overline{R}^2$ includes $v3$ (weight), $v4$ (height), $v5$ (chin skin fold), and $v11$ (fraction) as predictors.

Discussion of C_p results:

1. None of the single predictor models is adequate. Each has $C_p >> 1 + 1 = 2$, the target value.

2. The only adequate two predictor model has $v3$ = weight and $v11$ = fraction as predictors: $C_p = 1.45 < 2 + 1 = 3$. This is the minimum C_p model.

3. Every model with weight and fraction is adequate. Every model that excludes either weight or fraction is inadequate: $C_p >> p + 1$.

According to C_p, any reasonable model must include both weight and fraction as predictors. Based on simplicity, I would select the model with these two predictors as a starting point. I can always add predictors if subsequent analysis suggests this is necessary!

10.5.2 Peru Indian Data Summary

The model selection procedures suggest two models that warrant further consideration.

Predictors	Method suggesting model
v3, v11 weight, fraction	stepwise backward elimination R^2, adjusted R^2, C_p
v3, v11, v5, v4 weight, fraction, chin skin fold, and height	forward selection max adjusted R^2

I will give three reasons why I feel that the simpler model is preferable at this point:

1. It was suggested by 5 of the 6 methods.

2. Forward selection often chooses predictors that are not important, even when the significance level for inclusion is reduced from the default $\alpha = 0.50$ level.

3. The forward and backward elimination outputs suggest that neither chin skin fold nor height is significant at any of the standard levels of significance used in practice. Look at the third and fourth steps of forward selection to see this.

Using a mechanical approach, we are led to a model with weight and fraction as predictors of systolic blood pressure. At this point we should closely examine this model. We did this earlier this semester and found that observation 1 (the individual with the largest systolic blood pressure) was fitted poorly by the model and potentially influential.

As noted earlier this semester, model selection methods can be highly influenced by outliers and influential cases. Thus, we should hold out case 1, and re-evaluate the various procedures to see whether case 1 unduly influenced the models selected. I will just note (not shown) that the selection methods point to the same model when case 1 is held out. After deleting case 1, there are no large residuals, extremely influential points, or any gross abnormalities in plots.

Both analyses suggest that the "best model" for predicting systolic blood pressure is

$$\text{sys bp} = \beta_0 + \beta_1 \text{weight} + \beta_2 \text{fraction} + \varepsilon.$$

Should case 1 be deleted? I have not fully explored this issue, but I will note that eliminating this case does have a significant impact on the least squares estimates of the regression coefficients, and on predicted values. What do you think?

10.6 Example: Oxygen Uptake

An experiment was conducted to model oxygen uptake (o2up), in milligrams of oxygen per minute, from five chemical measurements: biological oxygen demand (bod), total Kjeldahl nitrogen (tkn), total solids (ts), total volatile solids (tvs), which is a component of ts, and chemical oxygen demand (cod), each measured in milligrams per liter. The data were collected on samples of dairy wastes kept in suspension in water in a laboratory for 220 days. All observations were on the same sample over time. We desire an equation relating o2up to the other variables. The goal is to find variables that should be further studied with the eventual goal of developing a prediction equation (day should not be considered as a predictor).

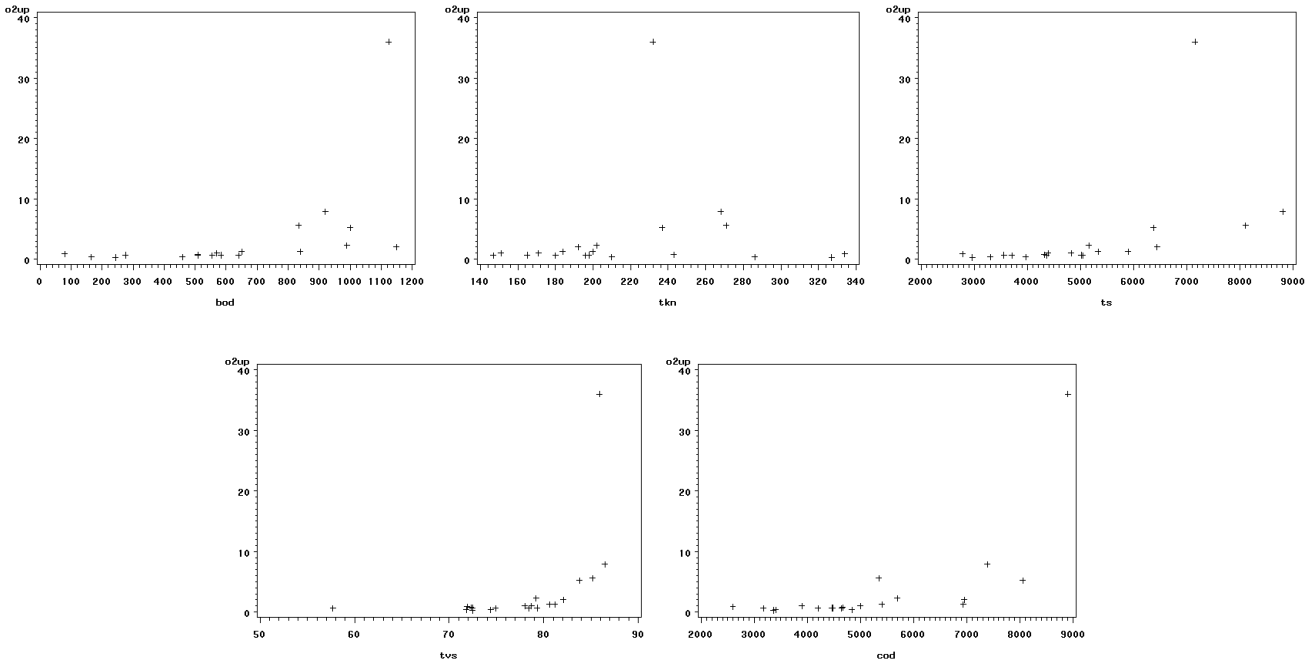
We are interested in developing a regression model with o2up, or some function of o2up, as a response. The researchers believe that the predictor variables are more likely to be linearly related to $\log_{10}$(o2up) rather than o2up, so $\log_{10}$(o2up) was included in the data set. As a first step, we should plot o2up against the different predictors, and see whether the relationship between o2up and the individual predictors is roughly linear. If not, we will consider appropriate transformations of the response and/or predictors.

A **SAS** program to read and plot the data is given below. The plots showed an exponential relationship between o2up and the predictors. To shorten the output, only one plot is given. An exponential relationship can often be approximately linearized by transforming o2up to the $\log_{10}$(o2up) scale suggested by the researchers. The extreme skewness in the marginal distribution of o2up also gives an indication that a transformation might be needed.

Program editor contents:

```
* Oxygen update example ;

options ls=79 nodate nocenter;

data oxygen;
  input day bod tkn ts tvs cod o2up logup;
datalines;
   0 1125 232 7160 85.9 8905 36.0 1.5563
   7  920 268 8804 86.5 7388  7.9  .8976
```

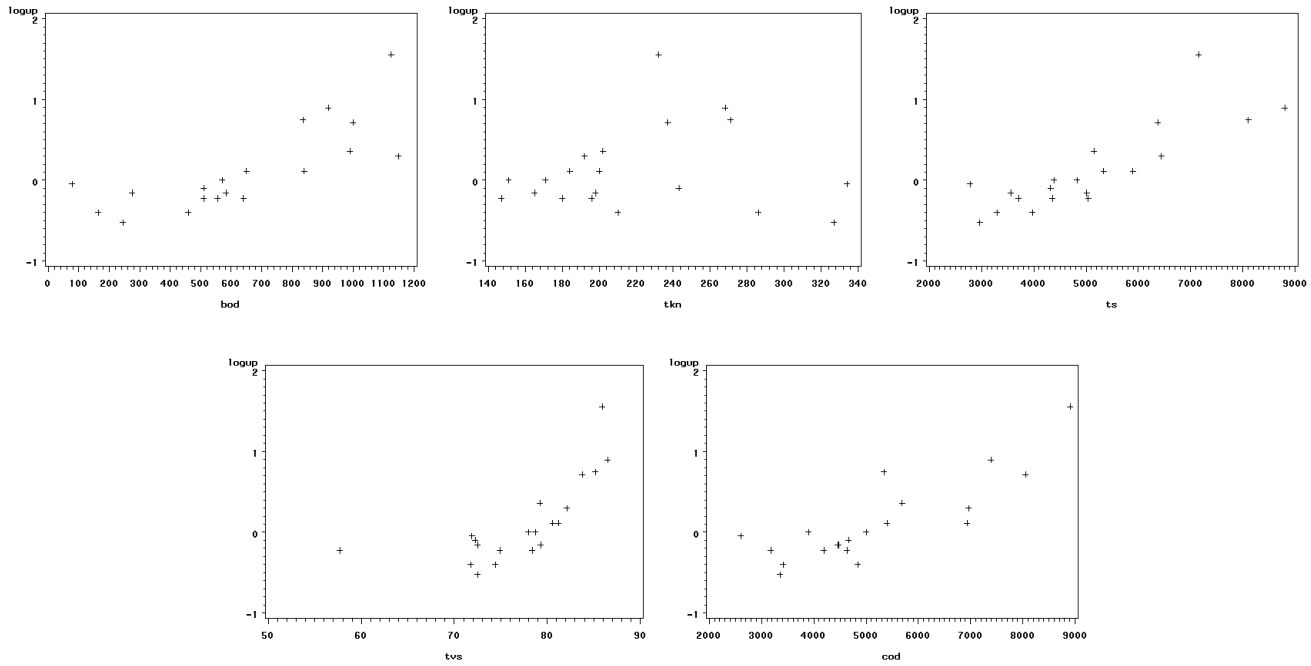
```
  15  835 271 8108 85.2 5348  5.6  .7482
  22 1000 237 6370 83.8 8056  5.2  .7160
  29 1150 192 6441 82.1 6960  2.0  .3010
  37  990 202 5154 79.2 5690  2.3  .3617
  44  840 184 5896 81.2 6932  1.3  .1139
  58  650 200 5336 80.6 5400  1.3  .1139
  65  640 180 5041 78.4 3177   .6 -.2218
  72  583 165 5012 79.3 4461   .7 -.1549
  80  570 151 4825 78.7 3901  1.0  .0000
  86  570 171 4391 78.0 5002  1.0  .0000
  93  510 243 4320 72.3 4665   .8 -.0969
 100  555 147 3709 74.9 4642   .6 -.2218
 107  460 286 3969 74.4 4840   .4 -.3979
 122  275 198 3558 72.5 4479   .7 -.1549
 129  510 196 4361 57.7 4200   .6 -.2218
 151  165 210 3301 71.8 3410   .4 -.3979
 171  244 327 2964 72.5 3360   .3 -.5229
 220   79 334 2777 71.9 2599   .9 -.0458
;
run;

proc gplot data=oxygen;
  plot o2up*( bod tkn ts tvs cod );
run;
```



After transformation, several plots show a roughly linear relationship. A sensible next step would be to build a regression model using log(o2up) as the response variable.

Program editor contents:

```
* log scale o2up;
proc gplot data=oxygen;
  plot logup*( bod tkn ts tvs cod );
run;
```



I used several of the model selection procedures to select out predictors. The summarized output was generated by the following **proc**.

Program editor contents:

```
* selection methods: backward, stepwise, rsquare, adjrsq, cp;
proc reg data=oxygen;
  model logup = bod tkn ts tvs cod /selection=backward;
  model logup = bod tkn ts tvs cod /selection=stepwise;
  model logup = bod tkn ts tvs cod /selection=rsquare adjrsq cp best=5;
run;
```

The model selection criteria below point to a more careful analysis of the model with ts and cod as predictors. This model has the minimum C_p and is selected by the backward and stepwise procedures. Furthermore, no other model has a substantially higher R^2 or $\overline{R}^2$. The fit of the model will not likely be improved substantially by adding any of the remaining three effects to this model.

These comments must be taken with a grain of salt because we have not critically assessed the underlying assumptions (linearity, normality, independence), nor have we considered whether the data contain influential points or outliers.

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	bod	4	0.0000	0.8094	4.0013	0.00	0.9718
2	tvs	3	0.0043	0.8050	2.3187	0.34	0.5685
3	tkn	2	0.0193	0.7857	1.7388	1.59	0.2258

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	ts		1	0.6965	0.6965	6.2942	41.30	<.0001
2	cod		2	0.0893	0.7857	1.7388	7.08	0.0165

```
R-Square Selection Method

Number of Observations Read          20
Number of Observations Used          20

Number in                 Adjusted
  Model      R-Square     R-Square        C(p)    Variables in Model
        1        0.6965       0.6796      6.2942    ts
        1        0.6926       0.6756      6.5744    cod
        1        0.5983       0.5760     13.5005    bod
        1        0.5054       0.4779     20.3267    tvs
        1        0.0082       -.0469     56.8431    tkn
-----------------------------------------------------------------
        2        0.7857       0.7605      1.7388    ts cod             *
        2        0.7376       0.7067      5.2735    tvs cod
        2        0.7158       0.6824      6.8701    tkn cod
        2        0.7157       0.6822      6.8829    bod ts
        2        0.7119       0.6780      7.1628    ts tvs
-----------------------------------------------------------------
        3        0.8050       0.7685      2.3187    tkn ts cod
        3        0.7900       0.7506      3.4241    ts tvs cod
        3        0.7898       0.7504      3.4386    bod ts cod
        3        0.7595       0.7144      5.6652    tkn tvs cod
        3        0.7515       0.7049      6.2510    bod tkn ts
-----------------------------------------------------------------
        4        0.8094       0.7585      4.0013    tkn ts tvs cod
        4        0.8050       0.7531      4.3184    bod tkn ts cod
        4        0.7948       0.7401      5.0685    bod ts tvs cod
        4        0.7716       0.7107      6.7757    bod tkn tvs cod
        4        0.7591       0.6948      7.6953    bod tkn ts tvs
-----------------------------------------------------------------
        5        0.8094       0.7413      6.0000    bod tkn ts tvs cod
```

Output for the model with ts and cod as predictors was obtained using the following **proc**.

Program editor contents:

```
* final model fit;
proc reg data=oxygen;
  model logup = ts cod /p r;
run;
```

The p-values for testing the importance of the individual predictors are small, indicating that both predictors are important. However, two observations (1 and 20) are poorly fitted by the model (both have $r_i > 2$) and are individually most influential (largest D_is). Recall that this experiment was conducted over 220 days, so these observations were the first and last data points collected. We have little information about the experiment, but it is reasonable to conjecture that the experiment may not have reached a steady state until the second time point, and that the experiment was ended when the experimental material dissipated. The end points of the experiment may not be typical of conditions under which we are interested in modelling oxygen uptake. A sensible strategy here is to delete these points and redo the entire analysis to see whether our model changes noticeably.

Output window contents:

```
The REG Procedure
Model: MODEL1
Dependent Variable: logup

Number of Observations Read          20
Number of Observations Used          20

                              Analysis of Variance
                                   Sum of           Mean
Source                     DF        Squares         Square    F Value    Pr > F
Model                       2        3.97836        1.98918      31.17    <.0001
Error                      17        1.08505        0.06383
Corrected Total            19        5.06340

Root MSE              0.25264    R-Square     0.7857
Dependent Mean        0.11860    Adj R-Sq     0.7605
Coeff Var           213.01742

                          Parameter Estimates
                       Parameter       Standard
Variable     DF         Estimate          Error    t Value    Pr > |t|
Intercept     1         -1.37004        0.19685      -6.96      <.0001
ts            1       0.00014916     0.00005489       2.72      0.0146
cod           1       0.00014150     0.00005318       2.66      0.0165

                            Output Statistics

    Dependent Predicted    Std Error                Std Error  Student                    Cook's
Obs  Variable     Value Mean Predict  Residual  Residual Residual Residual   -2-1 0 1 2          D
  1    1.5563    0.9580       0.1443    0.5983     0.207    2.885 |       |*****  |  1.342 *
  2    0.8976    0.9886       0.1481   -0.0910     0.205   -0.445 |       |       |  0.034
  3    0.7482    0.5961       0.1691    0.1521     0.188    0.810 |       |*      |  0.178
  4    0.7160    0.7201       0.1226 -0.004056     0.221  -0.0184 |       |       |  0.000
  5    0.3010    0.5756       0.0831   -0.2746     0.239   -1.151 |     **|       |  0.054
  6    0.3617    0.2039       0.0615    0.1578     0.245    0.644 |       |*      |  0.009
  7    0.1139    0.4903       0.0867   -0.3764     0.237   -1.586 |   ***|       |  0.112
  8    0.1139    0.1900       0.0572   -0.0761     0.246   -0.309 |       |       |  0.002
  9   -0.2218   -0.1686       0.1189   -0.0532     0.223   -0.239 |       |       |  0.005
 10   -0.1549  0.008794       0.0665   -0.1637     0.244   -0.672 |      *|       |  0.011
 11         0   -0.0983       0.0807    0.0983     0.239    0.411 |       |       |  0.006
 12         0 -0.007282       0.0645  0.007282     0.244   0.0298 |       |       |  0.000
 13   -0.0969   -0.0656       0.0626   -0.0313     0.245   -0.128 |       |       |  0.000
 14   -0.2218   -0.1600       0.0798   -0.0618     0.240   -0.258 |       |       |  0.002
 15   -0.3979   -0.0932       0.0745   -0.3047     0.241   -1.262 |     **|       |  0.051
 16   -0.1549   -0.2055       0.0823    0.0506     0.239    0.212 |       |       |  0.002
 17   -0.2218   -0.1252       0.0655   -0.0966     0.244   -0.396 |       |       |  0.004
 18   -0.3979   -0.3951       0.0862 -0.002761     0.237  -0.0116 |       |       |  0.000
 19   -0.5229   -0.4525       0.0936   -0.0704     0.235   -0.300 |       |       |  0.005
 20   -0.0458   -0.5881       0.1063    0.5423     0.229    2.366 |       |****   |  0.402 *

Sum of Residuals                           0
Sum of Squared Residuals             1.08505
Predicted Residual SS (PRESS)        1.81450
```

Here is a **SAS** program for reading in the data, excluding the end observations, and repeating the model selection steps. Summaries from the model selection are provided.

The model selection criteria again suggest ts and cod as predictors. After deleting observations 1 and 20 the R^2 for this two predictor model jumps from 0.786 to 0.892. Also note that the LS coefficients change noticeably after these observations are deleted.

Program editor contents:

```
* removing first and last observations (influential, outliers);
data oxygen2;
  set oxygen;
  if _n_ = 1 or _n_ = 20 then delete;
run;
```

```
proc reg data=oxygen2;
  model logup = bod tkn ts tvs cod /selection=backward;
  model logup = bod tkn ts tvs cod /selection=stepwise;
  model logup = bod tkn ts tvs cod /selection=rsquare adjrsq cp best=5;
run;
```

Output window contents:

```
                    Summary of Backward Elimination
        Variable    Number    Partial     Model
Step    Removed     Vars In   R-Square    R-Square    C(p)      F Value   Pr > F
  1     tkn             4      0.0000      0.8938     4.0036       0.00   0.9529
  2     bod             3      0.0001      0.8937     2.0202       0.02   0.8955
  3     tvs             2      0.0014      0.8923     0.1755       0.18   0.6771

                   Summary of Stepwise Selection
     Variable  Variable  Number  Partial   Model
Step Entered   Removed   Vars In R-Square R-Square  C(p)    F Value Pr > F
  1  ts                     1    0.8490   0.8490   3.0709   89.98 <.0001
  2  cod                    2    0.0433   0.8923   0.1755    6.03 0.0267

R-Square Selection Method

Number of Observations Read          18
Number of Observations Used          18

Number in                Adjusted
  Model      R-Square    R-Square        C(p)    Variables in Model
       1        0.8490      0.8396      3.0709    ts
       1        0.6608      0.6396     24.3563    bod
       1        0.6492      0.6273     25.6673    cod
       1        0.5139      0.4835     40.9716    tvs
       1        0.0135      -.0481     97.5464    tkn
-------------------------------------------------------------------
       2        0.8923      0.8780      0.1755    ts cod              *
       2        0.8669      0.8491      3.0508    bod ts
       2        0.8547      0.8354      4.4253    ts tvs
       2        0.8493      0.8292      5.0408    tkn ts
       2        0.7309      0.6950     18.4262    tvs cod
-------------------------------------------------------------------
       3        0.8937      0.8709      2.0202    ts tvs cod
       3        0.8926      0.8696      2.1452    bod ts cod
       3        0.8925      0.8695      2.1560    tkn ts cod
       3        0.8693      0.8413      4.7743    bod ts tvs
       3        0.8680      0.8397      4.9277    bod tkn ts
-------------------------------------------------------------------
       4        0.8938      0.8612      4.0036    bod ts tvs cod
       4        0.8938      0.8611      4.0084    tkn ts tvs cod
       4        0.8926      0.8596      4.1398    bod tkn ts cod
       4        0.8705      0.8306      6.6469    bod tkn ts tvs
       4        0.7752      0.7060     17.4181    bod tkn tvs cod
-------------------------------------------------------------------
       5        0.8939      0.8497      6.0000    bod tkn ts tvs cod
```

SAS output is given below for the model with ts and cod as predictors, after omitting the end observations. Both predictors are significant at the 0.05 level. Furthermore, there do not appear to be any extreme outliers. The rankit plot, and the plot of studentized residuals against predicted values do not show any extreme abnormalities.

Let us recall that the researcher's primary goal was to identify important predictors of o2up. Regardless of whether we are inclined to include the end observations in the analysis or not, it is reasonable to conclude that ts and cod are useful for explaining the variation in $\log_{10}$(o2up). If these data were the final experiment, I might be inclined to eliminate the end observations and use the following equation to predict oxygen uptake:

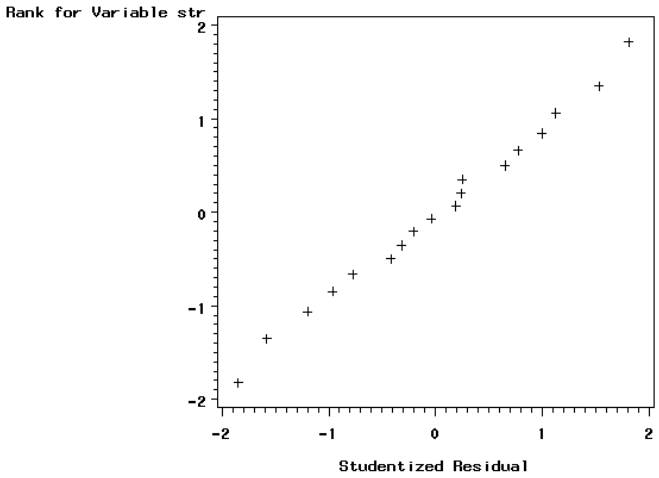
$$\log_{10}(\text{o2up}) = -1.335302 + 0.000185\text{ts} + 0.000086\text{cod}.$$

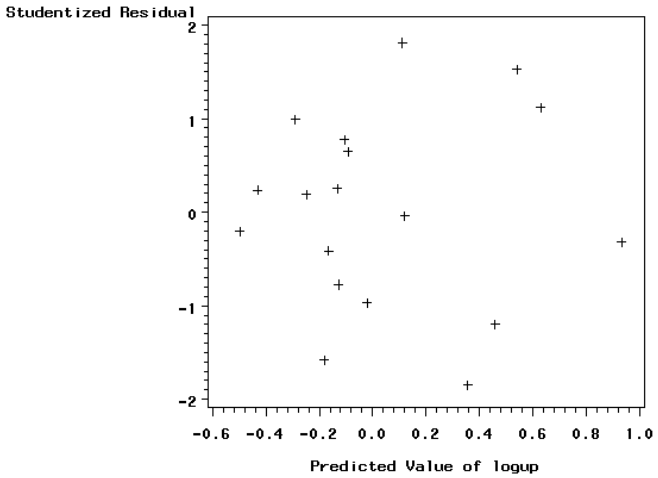
Program editor contents:

```
proc reg data=oxygen2;
  model logup = ts cod /p r;
  output out = outreg p = pv1 student = str;
run;

proc rank data=outreg out=outreg normal = blom;
  var str;
  ranks rankstr;
run;

proc gplot data=outreg;
  plot rankstr*str str*pv1;
run;
```

Output window contents:

```
The REG Procedure
Model: MODEL1
Dependent Variable: logup

Number of Observations Read          18
Number of Observations Used          18

                              Analysis of Variance
                                   Sum of           Mean
Source                     DF      Squares         Square    F Value    Pr > F
Model                       2      2.56929        1.28465      62.15    <.0001
Error                      15      0.31003        0.02067
Corrected Total            17      2.87932

Root MSE              0.14377    R-Square     0.8923
Dependent Mean        0.04786    Adj R-Sq     0.8780
Coeff Var           300.38326

                          Parameter Estimates
                       Parameter       Standard
Variable     DF         Estimate          Error    t Value    Pr > |t|
Intercept     1         -1.33530        0.13385      -9.98      <.0001
ts            1       0.00018521     0.00003182       5.82      <.0001
cod           1       0.00008638     0.00003517       2.46      0.0267

                                                Output Statistics
    Dependent Predicted    Std Error                Std Error  Student                          Cook's
Obs  Variable     Value Mean Predict  Residual       Residual Residual    -2-1 0 1 2              D
  1    0.8976    0.9335       0.0893   -0.0359          0.113   -0.318  |       |       |  0.021
  2    0.7482    0.6283       0.0964    0.1199          0.107    1.124  |       |**     |  0.344
  3    0.7160    0.5403       0.0866    0.1757          0.115    1.531  |       |***    |  0.445
  4    0.3010    0.4588       0.0568   -0.1578          0.132   -1.195  |    **|       |  0.088
  5    0.3617    0.1108       0.0389    0.2509          0.138    1.813  |       |***    |  0.086
  6    0.1139    0.3555       0.0597   -0.2416          0.131   -1.847  |   ***|       |  0.237
  7    0.1139    0.1194       0.0346 -0.005521          0.140  -0.0396  |       |       |  0.000
  8   -0.2218   -0.1272       0.0749   -0.0946          0.123   -0.771  |     *|       |  0.074
  9   -0.1549   -0.0217       0.0399   -0.1332          0.138   -0.964  |     *|       |  0.026
 10         0   -0.1047       0.0501    0.1047          0.135    0.777  |       |*      |  0.028
 11         0   -0.0900       0.0392    0.0900          0.138    0.651  |       |*      |  0.011
 12   -0.0969   -0.1322       0.0380    0.0353          0.139    0.255  |       |       |  0.002
 13   -0.2218   -0.2474       0.0481    0.0256          0.135    0.189  |       |       |  0.001
 14   -0.3979   -0.1821       0.0450   -0.2158          0.137   -1.580  |   ***|       |  0.090
 15   -0.1549   -0.2894       0.0496    0.1345          0.135    0.997  |       |*      |  0.045
 16   -0.2218   -0.1648       0.0405   -0.0570          0.138   -0.413  |       |       |  0.005
 17   -0.3979   -0.4294       0.0555    0.0315          0.133    0.237  |       |       |  0.003
 18   -0.5229   -0.4961       0.0594   -0.0268          0.131   -0.205  |       |       |  0.003

Sum of Residuals                           0
Sum of Squared Residuals             0.31003
Predicted Residual SS (PRESS)        0.47079
```


Rank for Variable str
2
1
0
-1
-2
-2
-1
0
1
2
Studentized Residual


Studentized Residual
2
1
0
-1
-2
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
Predicted Value of logup

Chapter 11

Logistic Regression

11.1 Example: Age of Menarche in Warsaw

The data below are from a study conducted by Milicer and Szczotka on pre-teen and teenage girls in Warsaw, Poland. The subjects were classified into 25 age categories. The number of girls in each group (sample size) and the number that reached menarche (# RM) at the time of the study were recorded. The age for a group corresponds to the midpoint for the age interval.

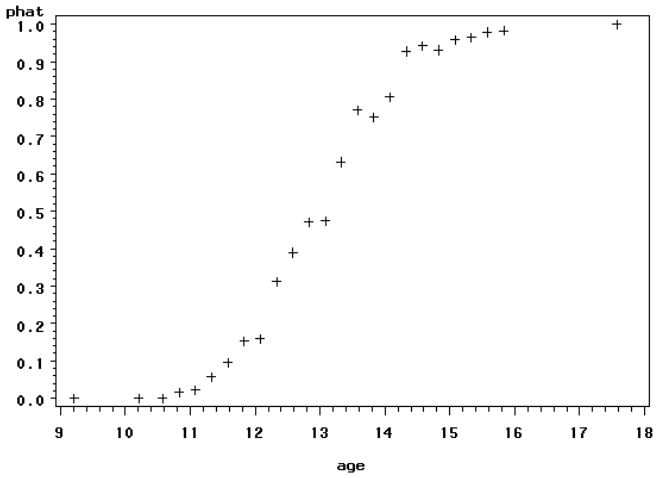
Sample size	# RM	$\hat{p}$	Age	Sample size	# RM	$\hat{p}$	Age
376	0	0.00	9.21	106	67	0.63	13.33
200	0	0.00	10.21	105	81	0.77	13.58
93	0	0.00	10.58	117	88	0.75	13.83
120	2	0.02	10.83	98	79	0.81	14.08
90	2	0.02	11.08	97	90	0.93	14.33
88	5	0.06	11.33	120	113	0.94	14.58
105	10	0.10	11.58	102	95	0.93	14.83
111	17	0.15	11.83	122	117	0.96	15.08
100	16	0.16	12.08	111	107	0.96	15.33
93	29	0.31	12.33	94	92	0.98	15.58
100	39	0.39	12.58	114	112	0.98	15.83
108	51	0.47	12.83	1049	1049	1.00	17.58
99	47	0.47	13.08				

The researchers were curious about how the proportion of girls that reached menarche ($\hat{p}$ = # RM/sample size) varied with age. One could perform a test of homogeneity by arranging the data as a 2-by-25 contingency table with columns indexed by age and two rows: ROW1 = # RM and ROW2 = # that have not RM = sample size − # RM. A more powerful approach treats these as regression data, using the proportion of girls reaching menarche as the response and age as a predictor.

A plot of the observed proportion $\hat{p}$ of girls that have reached menarche (**phat** ($\hat{p}$) in the plot) shows that the proportion increases as age increases, but that the relationship is nonlinear. The observed proportions, which are bounded between zero and one, have a lazy S-shape (a **sigmoidal function**) when plotted against age. The change in the observed proportions for a given change in

age is much smaller when the proportion is near 0 or 1 than when the proportion is near 1/2. This phenomenon is common with regression data where the response is a proportion.
Program editor contents:

```
* Example: Age of Menarche in Warsaw ;
options ls=79 nodate nocenter;
data menarche;
  infile "F:\Dropbox\UNM\teach\ADA2_stat528\sas\ADA2_11_menarche.dat";
  input ntot nres age;
  phat = nres/ntot;
run;

proc gplot data=menarche;
  plot phat*age;
run;
```

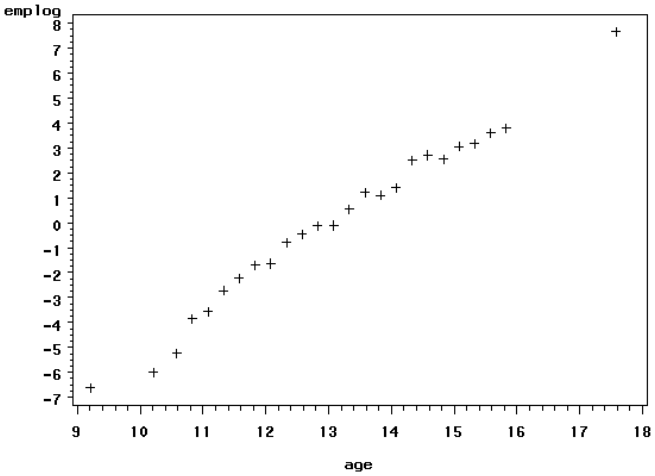


The trend is nonlinear so linear regression is inappropriate. A sensible alternative might be to transform the response or the predictor to achieve near linearity. A common transformation of response proportions following a sigmoidal curve is to the **logit scale** $\log_e\{\hat{p}/(1-\hat{p})\}$. This transformation is the basis for the **logistic regression model**. The natural logarithm (base e) is traditionally used in logistic regression.

The logit transformation is undefined when $\hat{p} = 0$ or $\hat{p} = 1$. To overcome this problem, researchers use the **empirical logits**, defined by $\log\{(\hat{p} + 0.5/n)/(1 - \hat{p} + 0.5/n)\}$, where n is the sample size or the number of observations on which $\hat{p}$ is based.

A plot of the empirical logits against age is roughly linear, which supports a logistic transformation for the response.

```
* emperical logits ;
data menarche;
  set menarche;
  emplog = log( (phat + 0.5/ntot) / (1 - phat + 0.5/ntot) );
run;

proc gplot data=menarche;
  plot emplog*age;
run;
```



11.2 The Simple Logistic Regression Model

The simple logistic regression model expresses the population proportion p of individuals with a given attribute (called the probability of success) as a function of a single predictor variable X. The model assumes that p is related to X through

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

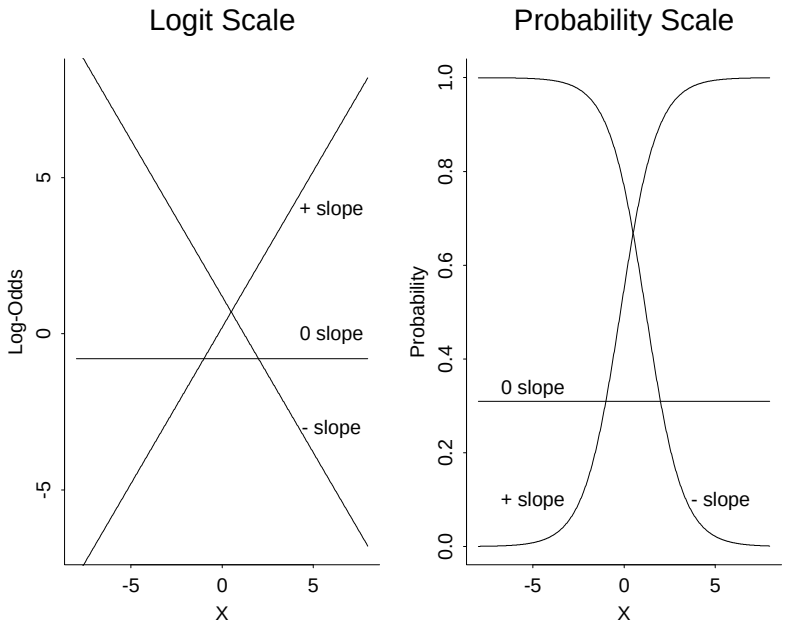
or, equivalently, as

$$p = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}.$$

The logistic regression model is a **binary response model**, where the response for each case falls into one of 2 exclusive and exhaustive categories, success (cases with the attribute of interest) and failure (cases without the attribute of interest).

The odds of success are $p/(1-p)$. For example, when $p = 1/2$ the odds of success are 1 (or 1 to 1). When $p = 0.9$ the odds of success are 9 (or 9 to 1). The logistic model assumes that the log-odds of success is linearly related to X. Graphs of the logistic model relating p to X are given below. The sign of the slope refers to the sign of β_1.

I should write $p = p(X)$ to emphasize that p is the proportion of all individuals with score X that have the attribute of interest. In the menarche data, $p = p(X)$ is the population proportion of girls at age X that have reached menarche.



The data in a logistic regression problem are often given in summarized or **aggregate** form:

X	n	y
X_1	n_1	y_1
X_2	n_2	y_2
.	.	.
.	.	.
X_m	n_m	y_m

where y_i is the number of individuals with the attribute of interest among n_i randomly selected or representative individuals with predictor variable value X_i. For **raw data** on individual cases, $y_i = 1$ or 0, depending on whether the case at X_i is a success or failure. The sample size column n is omitted with raw data.

For logistic regression, a plot of the sample proportions $\hat{p}_i = y_i/n_i$ against X_i should be roughly sigmoidal, and a plot of the empirical logits against X_i should be roughly linear. If not, then some other model is probably appropriate. I find the second plot easier to calibrate, but neither plot is very informative when the sample sizes are small, say 1 or 2. (Why?).

There are a variety of other binary response models that are used in practice. The **probit** regression model or the **complementary log-log** regression model might be appropriate when the logistic model does fit the data.

11.2.1 Estimating Regression Parameters

There are two unknown population parameters in the logistic regression model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X.$$

A simple way to estimate β_0 and β_1 is by least squares (LS), using the empirical logits as responses and the X_is as the predictor values. (We will discuss a better way in the next section.)

A **SAS** program for generating the plots we looked at earlier and the output for the LS fit is given below. The LS estimates for the menarche data are $b_0 = -22.03$ and $b_1 = 1.68$, which gives the fitted relationship

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -22.03 + 1.68\text{AGE}$$

or

$$\hat{p} = \frac{\exp(-22.03 + 1.68\text{AGE})}{1 + \exp(-22.03 + 1.68\text{AGE})},$$

where $\hat{p}$ is the predicted proportion (under the model) of girls having reached menarche at the given age. I used $\hat{p}$ to identify a predicted probability, in contrast to $\hat{p}$ which is the observed proportion at a given age.

The power of the logistic model versus the contingency table analysis discussed earlier is that the model gives estimates for the population proportion reaching menarche at all ages within the observed age range. The observed proportions allow you to estimate only the population proportions at the observed ages.

Program editor contents:

```
* estimating regression parameters by LS;
proc reg data=menarche;
  model emplog = age;
run;
```

Output window contents:

```
The REG Procedure
Model: MODEL1
Dependent Variable: emplog

Number of Observations Read          25
Number of Observations Used          25

                            Analysis of Variance
                                    Sum of           Mean
Source                     DF      Squares         Square    F Value    Pr > F
Model                       1    278.61565      278.61565    1250.44    <.0001
Error                      23      5.12472        0.22281
Corrected Total            24    283.74037

Root MSE              0.47203    R-Square     0.9819
Dependent Mean       -0.06649    Adj R-Sq     0.9812
Coeff Var          -709.93966

                       Parameter Estimates
                    Parameter       Standard
Variable     DF      Estimate          Error    t Value    Pr > |t|
Intercept     1     -22.02793        0.62819     -35.07      <.0001
age           1       1.67639        0.04741      35.36      <.0001
```

11.2.2 Maximum Likelihood Estimation for Logistic Regression Model

There are better ways to the fit the logistic regression model than LS which assumes that the responses are normally distributed with constant variance. A deficiency of the LS fit to the logistic model is that the observed counts y_i have a **Binomial distribution** under random sampling. The Binomial distribution is a discrete probability model associated with counting the number of successes in a fixed size sample, and other equivalent experiments such as counting the number of heads in repeated flips of a coin. The distribution of the empirical logits depend on the y_is so they are not normal (but are approximately normal in large samples), and are extremely skewed when

the sample sizes n_i are small. The response variability depends on the population proportions, and is not roughly constant when the observed proportions or the sample sizes vary appreciably across groups.

The differences in variability among the empirical logits can be accounted for using **weighted least squares** (WLS) when the sample sizes are large. An alternative approach called maximum likelihood uses the exact Binomial distribution of the responses y_i to generate optimal estimates of the regression coefficients. Software for maximum likelihood estimation is widely available, so LS and WLS methods are not really needed.

In **maximum likelihood estimation** (MLE), the regression coefficients are estimated iteratively by minimizing the **deviance** function (also called the likelihood ratio chi-squared statistic)

$$D = 2\sum_{i=1}^{m}\left\{y_i \log\left(\frac{y_i}{n_i p_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i p_i}\right)\right\}$$

over all possible values of β_0 and β_1, where the p_is satisfy the logistic model

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_i.$$

The ML method also gives standard errors and significance tests for the regression estimates.

The deviance is an analog of the residual sums of squares in linear regression. The choices for β_0 and β_1 that minimize the deviance are the parameter values that make the observed and fitted proportions as close together as possible in a "likelihood sense".

Suppose that b_0 and b_1 are the MLEs of β_0 and β_1. The deviance evaluated at the MLEs:

$$D = 2\sum_{i=1}^{m}\left\{y_i \log\left(\frac{y_i}{n_i \hat{p}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i \hat{p}_i}\right)\right\},$$

where the fitted probabilities $\hat{p}_i$ satisfy

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = b_0 + b_1 X_i,$$

is used to test the adequacy of the model. The deviance is small when the data fits the model, that is, when the observed and fitted proportions are close together. Large values of D occur when one or more of the observed and fitted proportions are far apart, which suggests that the model is inappropriate.

If the logistic model holds, then D has a chi-squared distribution with $m-r$ degrees of freedom, where m is the the number of groups and r (here 2) is the number of estimated regression parameters. A p-value for the deviance is given by the area under the chi-squared curve to the right of D. A small p-value indicates that the data does not fit the model.

Alternatively, the fit of the model can be evaluated using the chi-squared approximation to the Pearson X^2 statistic:

$$X^2 = \sum_{i=1}^{m}\left\{\frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i} + \frac{((n_i - y_i) - n_i(1-\hat{p}_i))^2}{n_i(1-\hat{p}_i)}\right\} = \sum_{i=1}^{m}\frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1-\hat{p}_i)}.$$

11.2.3 Fitting the Logistic Model by Maximum Likelihood

The `logistic` procedure in **SAS** fits logistic regression models by maximum likelihood. The logistic procedure has model selection methods and diagnostic tools. Cook's distances and residuals can be obtained by using the `influence` option on the `model` statement. Index plots of diagnostics are obtained with the `iplots` option. Unlike the `reg` procedure, the `logistic` procedure does not generate high resolution plots. I will not consider diagnostics here. Comments on the **SAS** program and output for the logistic fit to the menarche data are given below.

The `model` statement specifies the response in the form of a ratio of two variables from the data set **menarche**. The numerator **nres** is the number of successes at a given AGE. The denominator **ntot** is the group sample size. Do not use observed proportions computed in the `data` step as the "response" in the model statement. The `scale=none aggregate` option gives output for D and X^2.

Program editor contents:

```
* estimating regression parameters by ML;
proc logistic data=menarche;
  model nres/ntot = age  /scale=none aggregate;
  output out=outlog p=prob u=u95 l=l95;
run;

proc print data=outlog;
run;
```

The `output` statement creates a data set (**outlog**) that contains the input data set **menarche**, plus the fitted probabilities and 95% CIs for the population proportions at the ages in **menarche**. The fitted probabilities and the limits are stored in columns labeled `prob`, `l95`, and `u95`, respectively. This printed summary information is easily interpreted. For example, the estimated population proportion of girls aged 15.08 (more precisely, among girls in the age interval with midpoint 15.08) that have reached menarche is 0.967. You are 95% confident that the population proportion is between 0.958 and 0.975. A variety of other summaries and diagnostics can be saved here. See the online help for more details.

Obs	ntot	nres	age	phat	emplog	prob	l95	u95	
1	376	0	9.21	0.00000	-6.62407	0.00203	0.00129	0.00321	
2	200	0	10.21	0.00000	-5.99396	0.01031	0.00730	0.01455	
3	93	0	10.58	0.00000	-5.23111	0.01871	0.01380	0.02530	
4	120	2	10.83	0.01667	-3.85862	0.02787	0.02115	0.03663	
5	90	2	11.08	0.02222	-3.56671	0.04132	0.03226	0.05279	
6	88	5	11.33	0.05682	-2.72010	0.06088	0.04889	0.07557	
7	105	10	11.58	0.09524	-2.20775	0.08882	0.07335	0.10717	
8	111	17	11.83	0.15315	-1.68640	0.12784	0.10853	0.15002	
9	100	16	12.08	0.16000	-1.63339	0.18062	0.15744	0.20638	
10	93	29	12.33	0.31183	-0.78227	0.24895	0.22237	0.27758	
11	100	39	12.58	0.39000	-0.44274	0.33265	0.30358	0.36306	
12	108	51	12.83	0.47222	-0.11020	0.42844	0.39795	0.45947	
13	99	47	13.08	0.47475	-0.10008	0.52990	0.49899	0.56058	
14	106	67	13.33	0.63208	0.53583	0.62895	0.59851	0.65840	
15	105	81	13.58	0.77143	1.20193	0.71823	0.68925	0.74551	
16	117	88	13.83	0.75214	1.09861	0.79310	0.76660	0.81730	
17	98	79	14.08	0.80612	1.40534	0.85216	0.82897	0.87269	
18	97	90	14.33	0.92784	2.49045	0.89657	0.87707	0.91327	
19	120	113	14.58	0.94167	2.71690	0.92875	0.91293	0.94188	
20	102	95	14.83	0.93137	2.54422	0.95146	0.93898	0.96149	
21	122	117	15.08	0.95902	3.06169	0.96719	0.95756	0.97469	<---
22	111	107	15.33	0.96396	3.17341	0.97794	0.97064	0.98345	
23	94	92	15.58	0.97872	3.61092	0.98522	0.97977	0.98922	
24	114	112	15.83	0.98246	3.80666	0.99012	0.98609	0.99299	
25	1049	1049	17.58	1.00000	7.64922	0.99943	0.99902	0.99967	

The deviance and Pearson statistics are $D = 26.70$ and $X^2 = 21.87$, on $25-2 = 23$ df. The large p-values for D and X^2 suggest no gross deficiencies with the logistic model. The observed and fitted proportions (`phat` and `prob` in the output from **outlog**) are reasonably close at each observed

age. This is consistent with D and X^2 being fairly small. The data fits the logistic regression model reasonably well.

```
The LOGISTIC Procedure

               Model Information
Data Set                          WORK.MENARCHE
Response Variable (Events)        nres
Response Variable (Trials)        ntot
Model                             binary logit
Optimization Technique            Fisher's scoring

Number of Observations Read          25
Number of Observations Used          25
Sum of Frequencies Read            3918
Sum of Frequencies Used            3918

          Response Profile
 Ordered      Binary           Total
   Value      Outcome      Frequency
       1      Event             2308
       2      Nonevent          1610

                  Model Convergence Status
        Convergence criterion (GCONV=1E-8) satisfied.

       Deviance and Pearson Goodness-of-Fit Statistics
Criterion          Value       DF     Value/DF     Pr > ChiSq
Deviance         26.7035       23       1.1610         0.2688
Pearson          21.8684       23       0.9508         0.5282
Number of unique profiles: 25
```

The **Analysis of Maximum Likelihood Estimates** table gives MLEs and standard errors for the regression parameters. The (Wald) chi-square column is the square of the ratio of the parameter estimate divided by its standard error. The Wald statistics are used to test whether the corresponding parameters of the logistic model are zero. A Wald test p-value, based on a chi-squared approximation, is given in the column directly to the right of the Wald statistic.

The MLEs $b_0 = -21.23$ and $b_1 = 1.63$ for the intercept and slope are close to the LS estimates of $b_0 = -22.03$ and $b_1 = 1.68$, respectively from page 219. The two estimation methods give similar predicted probabilities here. The MLE of the predicted probabilities satisfy:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -21.23 + 1.63\text{AGE}$$

or

$$\hat{p} = \frac{\exp(-21.23 + 1.63\text{AGE})}{1 + \exp(-21.23 + 1.63\text{AGE})}.$$

```
        Testing Global Null Hypothesis: BETA=0
Test                 Chi-Square       DF     Pr > ChiSq
Likelihood Ratio      3667.1801        1         <.0001
Score                 2546.1146        1         <.0001
Wald                   766.3344        1         <.0001

             Analysis of Maximum Likelihood Estimates
                                Standard          Wald
Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq
Intercept     1    -21.2258      0.7707      758.5894        <.0001
age           1      1.6319      0.0590      766.3344        <.0001
```

If the model holds, then a slope of $\beta_1 = 0$ implies that p does not depend on AGE, i.e., the proportion of girls that have reached menarche is identical across age groups. The Wald p-value

for the slope is 0.0001, which leads to rejecting $H_0 : \beta_1 = 0$ at any of the usual test levels. The proportion of girls that have reached menarche is not constant across age groups. Again, the power of the model is that it gives you a simple way to quantify the effect of age on the proportion reaching menarche. This is more appealing than testing homogeneity across age groups followed by multiple comparisons.

The **Testing Global Null Hypothesis** table gives several statistics and p-values for testing the hypothesis that the regression coefficients are zero for each predictor in the model. This is the logistic regression analog of the overall model F-test in ANOVA and regression. The only predictor is AGE, so the implied test is that the slope of the regression line is zero. The Wald test statistic and p-value reported here are identical to the Wald test and p-value for the AGE effect given in the parameter estimates table.

For now I will ignore the few tables below that are routinely printed by the **logistic** procedure: the **Model Fit Statistics** table, the **Odds Ratio** table and the **Association** table.

```
         Model Fit Statistics
                                  Intercept
                Intercept                and
Criterion            Only     Covariates
AIC               5308.485       1643.305
SC                5314.758       1655.851
-2 Log L          5306.485       1639.305

            Odds Ratio Estimates
              Point           95% Wald
Effect     Estimate      Confidence Limits
age           5.114       4.556       5.740

Association of Predicted Probabilities and Observed Responses
Percent Concordant       96.9    Somers' D    0.944
Percent Discordant        2.4    Gamma        0.951
Percent Tied              0.7    Tau-a        0.457
Pairs                 3715880    c            0.972
```

11.2.4 Example: Leukemia white blood cell types

Feigl and Zelen reported the survival time in weeks and the white cell blood count (WBC) at time of diagnosis for 33 patients who eventually died of acute leukemia. Each person was classified as AG+ or AG−, indicating the presence or absence of a certain morphological characteristic in the white cells. Four variables are given in the **SAS** data set: WBC, a binary factor or **indicator variable** IAG (1 for AG+, 0 for AG−), NTOTAL (the number of patients with the given combination of IAG and WBC), and NRES (the number of NTOTAL that survived at least one year from the time of diagnosis).

The researchers are interested in modelling the probability p of surviving at least one year as a function of WBC and IAG. They believe that WBC should be transformed to a log scale, given the skewness in the WBC values.

Program editor contents:

```
* Leukemia white blood cell types example;

options ls=79 nodate nocenter;
data blood;
  input ntotal nres Iag wbc @@;
  lwbc  = log(wbc);
  phat = nres/ntotal;
  label ntotal = 'number of patients with IAG and WBC combination'
        nres   = 'number surviving at least one year'
```

```
        Iag    = '1 for AG+, 0 for AG-'
        wbc    = 'white cell blood count'
        lwbc   = 'log white cell blood count'
        phat   = 'Emperical Probability';
  datalines;
1 1 1    75     1 1 1   230     1 1 1   430
1 1 1   260     1 0 1   600     1 1 1  1050
1 1 1  1000     1 0 1  1700     1 0 1   540
1 1 1   700     1 1 1   940     1 0 1  3200
1 0 1  3500     1 0 1  5200     3 1 1 10000
1 1 0   440     1 1 0   300     1 0 0   400
1 0 0   150     1 0 0   900     1 0 0   530
1 0 0  1000     1 0 0  1900     1 0 0  2700
1 0 0  2800     1 0 0  3100     1 0 0  2600
1 0 0  2100     1 0 0  7900     2 0 0 10000
;
run;
```

As an initial step in the analysis, consider the following model:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{LWBC} + \beta_2 \text{IAG},$$

where LWBC = log(WBC). The model is best understood by separating the AG+ and AG− cases. For AG− individuals, IAG=0 so the model reduces to

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{LWBC} + \beta_2 * 0 = \beta_0 + \beta_1 \text{LWBC}.$$

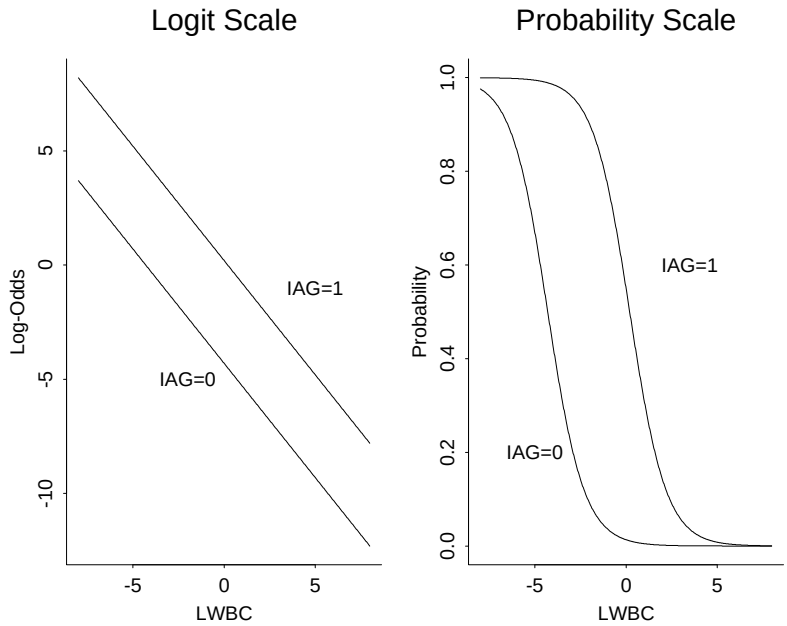
For AG+ individuals, IAG=1 and the model implies

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{LWBC} + \beta_2 * 1 = (\beta_0 + \beta_2) + \beta_1 \text{LWBC}.$$

The model without IAG (i.e., $\beta_2 = 0$) is a simple logistic model where the log-odds of surviving one year is linearly related to LWBC, and is independent of AG. The reduced model with $\beta_2 = 0$ implies that there is no effect of the AG level on the survival probability once LWBC has been taken into account.

Including the **binary predictor** IAG in the model implies that there is a linear relationship between the log-odds of surviving one year and LWBC, with a constant slope for the two AG levels. This model includes an effect for the AG morphological factor, but more general models are possible. A natural extension would be to include a product or interaction effect, a point that I will return to momentarily.

The parameters are easily interpreted: β_0 and $\beta_0 + \beta_2$ are intercepts for the population logistic regression lines for AG− and AG+, respectively. The lines have a common slope, β_1. The β_2 coefficient for the IAG indicator is the difference between intercepts for the AG+ and AG− regression lines. A picture of the assumed relationship is given below for $\beta_1 < 0$. The population regression lines are parallel on the logit scale only, but the order between IAG groups is preserved on the probability scale.




```
proc logistic data=blood;
  model nres/ntotal = Iag lwbc  /scale=none aggregate;
  output out=outlog p=prob;
run;

symbol1 v="0" c=blue  i=none;
symbol2 v="1" c=red   i=none;
proc gplot data=outlog;
  plot phat*lwbc=Iag;
  plot prob*lwbc=Iag;
run;
```

Before looking at output for the equal slopes model, note that the data set has 30 distinct IAG and LWBC combinations, or 30 "groups" or samples. Only two samples have more than 1 observation. The majority of the observed proportions surviving at least one year (number surviving ≥ 1 year/ group sample size) are 0 (i.e., 0/1) or 1 (i.e., 1/1). This sparseness of the data makes it difficult to graphically assess the suitability of the logistic model (because the estimated proportions are almost all 0 or 1). Although significance tests on the regression coefficients do not require large group sizes, the chi-squared approximation to the deviance and Pearson statistics is suspect in sparse data settings. With small group sizes as we have here, most researchers would not interpret the p-values for D and X^2 literally. Instead, they would use the p-values to informally check the fit of the model. Diagnostics would be used to highlight problems with the model.

Output window contents:

```
The LOGISTIC Procedure

               Model Information

Data Set                          WORK.BLOOD
Response Variable (Events)        nres
Response Variable (Trials)        ntotal
Model                             binary logit
Optimization Technique            Fisher's scoring

               Model Information

number surviving at least one year
number of patients with IAG and WBC combination

Number of Observations Read           30
Number of Observations Used           30
Sum of Frequencies Read               33
Sum of Frequencies Used               33

          Response Profile
 Ordered      Binary           Total
   Value      Outcome      Frequency
       1      Event               11
       2      Nonevent            22

                  Model Convergence Status
         Convergence criterion (GCONV=1E-8) satisfied.
```

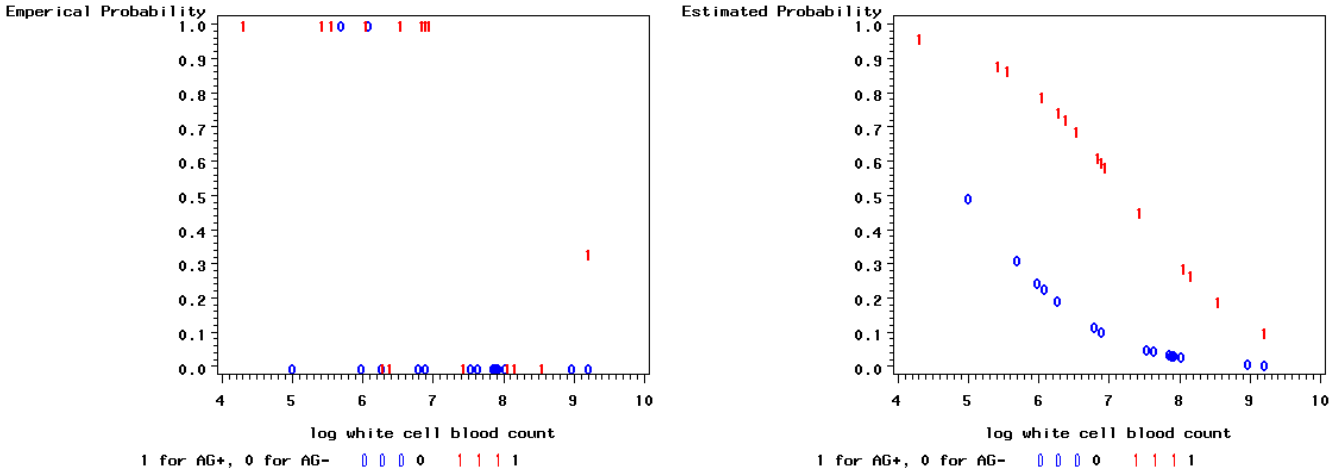
The large p-values for D and X^2 indicate that there are no gross deficiencies with the model. Recall that the **Testing Global Null Hypothesis** table gives p-values for testing the hypothesis that the regression coefficients are zero for each predictor in the model. The two predictors are LWBC and IAG, so the small p-values indicate that LWBC or IAG, or both, are important predictors of survival. The Wald test p-values in the estimates table suggest that LWBC and IAG are both important. If either predictor was insignificant, I would consider refitting the model omitting the least significant effect, as in regression.

```
       Deviance and Pearson Goodness-of-Fit Statistics
Criterion          Value       DF     Value/DF     Pr > ChiSq
Deviance         23.0136       27       0.8524         0.6843
Pearson          19.8083       27       0.7336         0.8387

Number of unique profiles: 30

        Testing Global Null Hypothesis: BETA=0
Test                 Chi-Square       DF     Pr > ChiSq
Likelihood Ratio        15.1772        2         0.0005
Score                   12.6236        2         0.0018
Wald                     8.1956        2         0.0166
```

Given that the model fits reasonably well, a test of $H_0 : \beta_2 = 0$ might be a primary interest here. This checks whether the regression lines are identical for the two AG levels, which is a test for whether AG affects the survival probability, after taking LWBC into account. This test is rejected at any of the usual significance levels, suggesting that the AG level affects the survival probability (assuming a very specific model).

```
             Analysis of Maximum Likelihood Estimates
                                Standard          Wald
Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq
Intercept     1      5.5428      3.0222        3.3636        0.0667
Iag           1      2.5193      1.0906        5.3362        0.0209
lwbc          1     -1.1086      0.4609        5.7856        0.0162
```

A plot of the predicted survival probabilities as a function of LWBC is given with the **SAS** output, using IAG as the plotting symbol. The model indicates that the probability of surviving at least one year from the time of diagnosis is a decreasing function of LWBC. For a given LWBC the survival probability is greater for AG+ patients than for AG− patients. This tendency is consistent with the observed proportions, which show little information about the exact form of the trend.



The estimated survival probabilities satisfy

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 5.54 - 1.11\text{LWBC} + 2.52\text{IAG}.$$

For AG− individuals with IAG=0, this reduces to

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 5.54 - 1.11\text{LWBC},$$

or equivalently,

$$\hat{p} = \frac{\exp(5.54 - 1.11\text{LWBC})}{1 + \exp(5.54 - 1.11\text{LWBC})}.$$

For AG+ individuals with IAG=1,

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 5.54 - 1.11\text{LWBC} + 2.52(1) = 8.06 - 1.11\text{LWBC},$$

or

$$\hat{p} = \frac{\exp(8.06 - 1.11\text{LWBC})}{1 + \exp(8.06 - 1.11\text{LWBC})}.$$

Using the **logit scale**, the difference between AG+ and AG− individuals in the estimated log-odds of surviving at least one year, at a fixed but arbitrary LWBC, is the estimated IAG regression coefficient:

$$(8.06 - 1.11\text{LWBC}) - (5.54 - 1.11\text{LWBC}) = 2.52.$$

Using properties of exponential functions, the odds that an AG+ patient lives at least one year is $\exp(2.52) = 12.42$ times larger than the odds that an AG− patient lives at least one year, regardless of LWBC.

This summary, and a CI for the IAG odds ratio, is given in the **Odds Ratio** table. Similarly, the estimated odds ratio of 0.33 for LWBC implies that the odds of surviving at least one year is reduced by a factor of 3 for each unit increase of LWBC.

```
            Odds Ratio Estimates
              Point          95% Wald
Effect     Estimate      Confidence Limits
Iag          12.420       1.465     105.310
lwbc          0.330       0.134       0.814
```

Although the equal slopes model appears to fit well, a more general model might fit better. A natural generalization here would be to add an **interaction**, or product term, IAG $*$ LWBC to the model. The logistic model with an IAG effect and the IAG $*$ LWBC interaction is equivalent to fitting separate logistic regression lines to the two AG groups. This interaction model provides an easy way to test whether the slopes are equal across AG levels. I will note that the interaction term is not needed here.

11.3 Example: The UNM Trauma Data

The data to be analyzed here were collected on 3132 patients admitted to The University of New Mexico Trauma Center between the years 1991 and 1994. For each patient, the attending physician recorded their age, their revised trauma score (RTS), their injury severity score (ISS), whether their injuries were blunt (i.e., the result of a car crash: BP=0) or penetrating (i.e., gunshot/knife wounds: BP=1), and whether they eventually survived their injuries (SURV=0 if not, SURV=1 if survived). Approximately 10% of patients admitted to the UNM Trauma Center eventually die from their injuries.

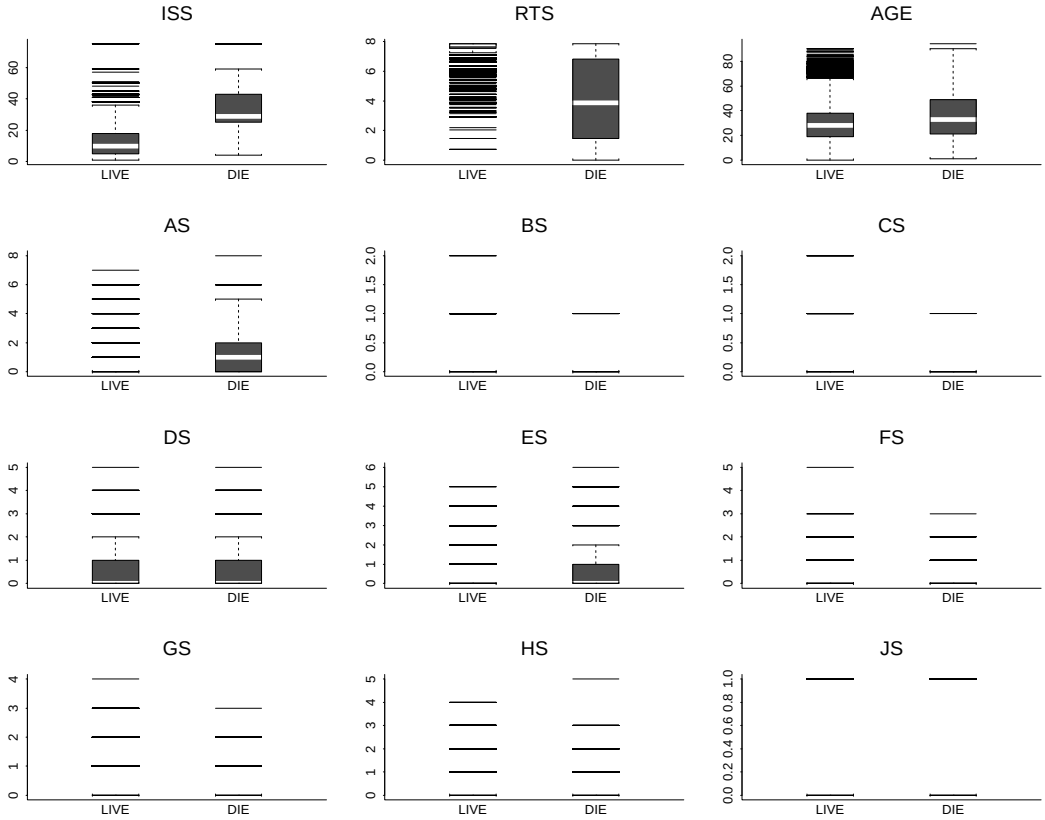
The ISS is an overall index of a patient's injuries, based on the approximately 1300 injuries cataloged in the Abbreviated Injury Scale. The ISS can take on values from 0 for a patient with no injuries to 75 for a patient with 3 or more life threatening injuries. The ISS is the standard injury index used by trauma centers throughout the U.S. The RTS is an index of physiologic injury, and is constructed as a weighted average of an incoming patient's systolic blood pressure, respiratory rate, and Glasgow Coma Scale. The RTS can take on values from 0 for a patient with no vital signs to 7.84 for a patient with normal vital signs.

Champion et al. (1981) proposed a logistic regression model to estimate the probability of a patient's survival as a function of RTS, the injury severity score ISS, and the patient's age, which is used as a surrogate for physiologic reserve. Subsequent survival models included the binary effect BP as a means to differentiate between blunt and penetrating injuries.

We will develop a logistic model for predicting survival from ISS, AGE, BP, and RTS. Data on the number of severe injuries in each of the nine body regions is also included in the database, so we will also assess whether these features have any predictive power. The following labels were used to identify the number of severe injuries in the nine regions: AS = HEAD, BS = FACE, CS = NECK, DS = THORAX, ES = ABDOMEN, FS = SPINE, GS = UPPER EXTREMITIES, HS = LOWER EXTREMITIES, and JS = SKIN.

I made side-by-side boxplots of the distributions of ISS, AGE, RTS, and AS through JS for the survivors and non-survivors. In several body regions the number of injuries is limited, so these

boxplots are not very enlightening. Survivors tend to have lower ISS scores, tend to be slightly younger, tend to have higher RTS scores, and tend to have fewer severe head (AS) and abdomen injuries (ES) than non-survivors. The importance of the effects individually towards predicting survival is directly related to the separation between the survivors and non-survivors scores.



11.3.1 Selecting Predictors in the Trauma Data

The logistic procedure has automated methods for model building, including **backward elimination**, forward selection, and stepwise methods, among others. A **SAS** proc is given below for performing a backward elimination, starting with a full model having 13 effects: ISS, BP, RTS, AGE, and AS–JS. The elimination process begins by identifying the least significant effect in the full model, based on Wald test p-values. If the least significant effect is not important, the effect is omitted from the model and the resulting model is fitted. The process is repeated until the least significant effect in the model is too important to exclude. This gives the final model. The default significance level for testing effects in backward elimination is 0.05. The test level was changed to 0.10 using the `slstay=0.10` option on the `model` statement.

In our previous logistic regression analyses, the cases in the data set were pre-aggregated into groups of observations having identical levels of the predictor variables. The numbers of cases in the success category and the group sample sizes were specified in the model statement, along with the names of the predictors. The trauma data set, which is not reproduced here, is **raw data** consisting of one record per patient (i.e., 3132 lines). The logistic model is fitted to data on individual cases by specifying the binary response variable (SURV) and the predictors on the

model statement. Some care is needed because **SAS** will define the logistic model in such a way that the **success** category corresponds to the **lower** of the two levels of response variable. Here SURV=0 corresponds to patients that died whereas SURV=1 corresponds to patients that lived, so the default model is for the log-odds of **dying** as a function of the predictor variables. To reverse the role of the response levels and model the log-odds of **surviving**, I used the descending option on the model statement (see the printed note from the log window contained with the **SAS** output). This switch has no impact on the model building process. In particular, you can convert a model for the log-odds of surviving to a model for the log-odds of dying by simply changing the signs of each regression coefficient in the model.

Program editor contents:

```
* Example: The UNM Trauma Data ;

options ls=79 nodate nocenter;
data trauma;
  infile "F:\Dropbox\UNM\teach\ADA2_stat528\sas\ADA2_11_trauma.dat";
  *infile 'x:\turner\albuq.bed';

  input id surv a1 a2 a3 a4 a5 a6  b1 b2 b3 b4 b5 b6  c1 c2 c3 c4 c5 c6
                d1 d2 d3 d4 d5 d6  e1 e2 e3 e4 e5 e6  f1 f2 f3 f4 f5 f6
                g1 g2 g3 g4 g5 g6  h1 h2 h3 h4 h5 h6  j1 j2 j3 j4 j5 j6
                iss iciss bp rts age prob;

  * add the severe injuries 3--6;
  as = a3+a4+a5+a6;  * as = head                ;
  bs = b3+b4+b5+b6;  * bs = face                ;
  cs = c3+c4+c5+c6;  * cs = neck                ;
  ds = d3+d4+d5+d6;  * ds = thorax              ;
  es = e3+e4+e5+e6;  * es = abdomen             ;
  fs = f3+f4+f5+f6;  * fs = spine               ;
  gs = g3+g4+g5+g6;  * gs = upper extremities   ;
  hs = h3+h4+h5+h6;  * hs = lower extremities   ;
  js = j3+j4+j5+j6;  * js = skin                ;
run;

proc logistic data=trauma descending;
  model surv = as bs cs ds es fs gs hs js
               iss rts age bp
        / selection=backward scale=none aggregate slstay=0.10
          ctable pprob=(0.10 to 1 by 0.10);
run;
```

I only included the summary table from the backward elimination, and information on the fit of the selected model. The final model includes effects for ES (number of severe abdominal injuries), ISS, RTS, AGE, and BP. All of the effects in the selected model are significant at the 5% level.

Output window contents:

```
FROM LOG FILE:  NOTE: PROC LOGISTIC is modeling the probability that surv=1.

The LOGISTIC Procedure

NOTE: No (additional) effects met the 0.1 significance level for removal from
      the model.

                Summary of Backward Elimination
          Effect                Number          Wald
    Step  Removed      DF          In    Chi-Square    Pr > ChiSq
       1  as            1          12        0.0489        0.8250
       2  bs            1          11        0.1196        0.7294
       3  fs            1          10        0.1468        0.7016
       4  cs            1           9        0.6680        0.4137
       5  ds            1           8        0.8078        0.3688
       6  gs            1           7        1.1229        0.2893
       7  hs            1           6        1.1977        0.2738
       8  js            1           5        1.7047        0.1917

      Deviance and Pearson Goodness-of-Fit Statistics
Criterion          Value       DF     Value/DF     Pr > ChiSq
Deviance        866.0687     2573       0.3366         1.0000
Pearson        2534.2677     2573       0.9849         0.7031

Number of unique profiles: 2579

             Analysis of Maximum Likelihood Estimates
                              Standard          Wald
Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq
Intercept     1      0.3558      0.4429        0.6454        0.4218
es            1     -0.4613      0.1101       17.5567        <.0001
iss           1     -0.0569     0.00741       58.9887        <.0001
rts           1      0.8431      0.0553      232.1313        <.0001
age           1     -0.0497     0.00529       88.2479        <.0001
bp            1     -0.6351      0.2496        6.4751        0.0109

           Odds Ratio Estimates
             Point          95% Wald
Effect    Estimate      Confidence Limits
es           0.630       0.508       0.782
iss          0.945       0.931       0.958
rts          2.324       2.085       2.590
age          0.952       0.942       0.961
bp           0.530       0.325       0.864
```

Letting p be the probability of survival, the estimated survival probability is given by

$$\log\left(\frac{\tilde{p}}{1-\tilde{p}}\right) = 0.3558 - 0.4613\text{ES} - 0.6351\text{BP} - 0.0569\text{ISS} + 0.8431\text{RTS} - 0.0497\text{AGE}.$$

The p-values for D and X^2 are large, indicating no gross deficiencies with the selected model.

Let us interpret the sign of the coefficients, and the odds ratios, in terms of the impact that individual predictors have on the survival probability.

11.3.2 Checking Predictions for the Trauma Model

To assess the ability of the selected model to accurately predict survival, assume that the two types of errors (a false positive prediction of survival, and a false negative prediction) have equal costs. Under this assumption, the optimal classification rule is to predict survival for an individual with an estimated survival probability of 0.50 or larger.

In the **SAS** program above, the `ctable pprob=` option gives classifications for cutoff probabilities $0.1, 0.2, \ldots, 1.0$ based on the selected model. This summary uses the **jackknife** method

for assessing classification accuracy. To implement the jackknife method, each observation is temporarily held out and the selected model is fitted to the remaining (3131) cases. This leads to an estimated survival probability for the case, which is compared to the actual result. The results are summarized in terms of total number of correct and incorrect classifications for each possible outcome. In the table below, the columns labeled **Event** (a survival) and **Non-Event** (a death) refer to the classification of observations, and not to the actual outcomes. The columns labeled **Correct** and **Incorrect** identify whether the classifications are accurate.

The data set has 2865 survivors and 267 people that died. Using a 0.50 cutoff, 2825 of the survivors would be correctly identified, and 40 misclassified. Similarly, 138 of the patients that died would be correctly classified and 129 would not. The overall percentage of cases correctly classified is $100 * (2825 + 138)/3132 = 94.6$. The **false positive rate**, which is the % of those predicted to survive that did not, is 4.4% or $129/(129 + 2825)$. The **false negative rate**, which is the % of those predicted to die that did not is $40/(40 + 138) = 22.5\%$. The **sensitivity** is the percentage of survivors that are correctly classified, $2825/(2825 + 40) = 98.6\%$. The **specificity** is the percentage of patients that died that are correctly classified, $138/(138 + 129) = 51.7\%$.

The misclassification rate seems small, but you should remember that approximately 10% of patients admitted to UNM eventually die from their injuries. Given this historical information only, you could achieve a 10% misclassification rate by completely ignoring the data and classifying each admitted patient as a survivor. Using the data reduces the misclassification rate by about 50% (from 10% down to 4.4%), which is an important reduction in this problem.

Classification Table

	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG	
0.100	2860	77	190	5	93.8	99.8	28.8	6.2	6.1	
0.200	2853	105	162	12	94.4	99.6	39.3	5.4	10.3	
0.300	2848	117	150	17	94.7	99.4	43.8	5.0	12.7	
0.400	2836	125	142	29	94.5	99.0	46.8	4.8	18.8	
0.500	2825	138	129	40	94.6	98.6	51.7	4.4	22.5	<---
0.600	2801	156	111	64	94.4	97.8	58.4	3.8	29.1	
0.700	2773	171	96	92	94.0	96.8	64.0	3.3	35.0	
0.800	2726	195	72	139	93.3	95.1	73.0	2.6	41.6	
0.900	2626	227	40	239	91.1	91.7	85.0	1.5	51.3	
1.000	0	267	0	2865	8.5	0.0	100.0	.	91.5	

11.4 Historical Example: O-Ring Data

The table below presents data from Dalal, Fowlkes, and Hoadley (1989) on field O-ring failures in the 23 pre-*Challenger* space shuttle launches. Temperature at lift-off and O-ring joint pressure are predictors. The binary version of the data views each flight as an independent trial. The result of a trial (y) is a 1 if at least one field O-ring failure occurred on the flight and a 0 if all six O-rings functioned properly. A regression for the binary data would model the probability that **any** O-ring failed as a function of temperature and pressure. The binomial version of these data is also presented. It views the six field O-rings on each flight as independent trials. A regression for the binomial data would model the probability that **a particular** O-ring on a given flight failed as a function of temperature and pressure. The two regressions model different probabilities. The *Challenger* explosion occurred during a takeoff at 31 degrees Fahrenheit.

Program editor contents:

```
* Shuttle O-ring data;
```

```
options ls=79 nodate nocenter;
data shuttle;
  input case flight y six temp pressure @@;
  datalines;
           1     14 1   2   53       50         13  2 1 1 70 200
           2      9 1   1   57       50         14 11 1 1 70 200
           3     23 1   1   58      200         15  6 0 0 72 200
           4     10 1   1   63       50         16  7 0 0 73 200
           5      1 0   0   66      200         17 16 0 0 75 100
           6      5 0   0   67       50         18 21 1 2 75 200
           7     13 0   0   67      200         19 19 0 0 76 200
           8     15 0   0   67       50         20 22 0 0 76 200
           9      4 0   0   68      200         21 12 0 0 78 200
          10      3 0   0   69      200         22 20 0 0 79 200
          11      8 0   0   70       50         23 18 0 0 81 200
          12     17 0   0   70      200
;
run;
```

The regression model for the binomial data (i.e., six trials per launch) is suspect on substantive grounds. The binomial model makes no allowance for an effect due to having the six O-rings on the same flight. If these data were measurements rather than counts, they would almost certainly be treated as dependent repeated measures. If interest is focused on whether one or more O-rings failed, the simplest, most direct data are the binary data.

Consider fitting a logistic regression model using temperature and pressure as predictors. Let p_i be the probability that any O-ring fails in case i and model this as

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \text{Temp}_i + \beta_2 \text{Pressure}_i.$$

A **SAS** program for fitting the logistic model is given below. I read in the binary data with $Y = 1$ if at least one O-ring failed, and 0 otherwise. I used the **descending** option, so I am modelling the chance of one or more O-ring failures as a function of temperature and pressure.

Program editor contents:

```
proc logistic data=shuttle descending;
  model y = temp pressure   /scale=none aggregate;
run;
```

The D and X^2 goodness-of-fit statistics suggest no gross deviations from the model. Furthermore, the test of $H_0 : \beta_1 = \beta_2 = 0$ (no regression effects) based on the likelihood ratio has a p-value of 0.016, which suggests that either temperature or pressure, or both, are useful predictors of the probability of O-ring failure. The Wald test p-value is slightly higher, 0.10. The Wald test p-values for testing $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$ individually are 0.038 and 0.576, respectively, which indicates pressure is not important (when added last to the model), but that temperature is important. This conclusion might be anticipated by looking at data plots (not provided).

Output window contents:

```
The LOGISTIC Procedure

              Model Information
Data Set                          WORK.SHUTTLE
Response Variable                 y
Number of Response Levels         2
Model                             binary logit
Optimization Technique            Fisher's scoring

Number of Observations Read          23
Number of Observations Used          23

          Response Profile
 Ordered                          Total
   Value             y     Frequency
       1             1             7
       2             0            16

Probability modeled is y=1.

                  Model Convergence Status
         Convergence criterion (GCONV=1E-8) satisfied.


       Deviance and Pearson Goodness-of-Fit Statistics
Criterion          Value       DF     Value/DF     Pr > ChiSq
Deviance         16.1647       16       1.0103         0.4415
Pearson          18.6293       16       1.1643         0.2884

Number of unique profiles: 19

        Testing Global Null Hypothesis: BETA=0
Test                 Chi-Square       DF     Pr > ChiSq
Likelihood Ratio         8.2833        2         0.0159
Score                    7.4657        2         0.0239
Wald                     4.6020        2         0.1002

             Analysis of Maximum Likelihood Estimates
                                Standard          Wald
Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq
Intercept     1     16.3852      8.0274        4.1663        0.0412
temp          1     -0.2634      0.1264        4.3446        0.0371
pressure      1     0.00518     0.00926        0.3128        0.5760
```

A reasonable next step would be to refit the model, after omitting pressure as a predictor. The target model is now

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \text{Temp}_i.$$

Our conclusions on the overall fit of the model and the significance of the effect of temperature on the probability of O-ring failure are consistent with the results for two predictor model. The model estimates the log-odds of (at least one) O-ring failure to be

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 15.043 - 0.2322\text{Temp}.$$

The estimated probability of (at least one) O-ring failure is

$$\hat{p} = \frac{\exp(15.043 - 0.2322\text{Temp})}{1 + \exp(15.043 - 0.2322\text{Temp})}.$$

This is an decreasing function of temperature.
Program editor contents:

```
proc logistic data=shuttle descending;
  model y = temp  /scale=none aggregate;
run;
```

Output window contents:

```
       Deviance and Pearson Goodness-of-Fit Statistics
Criterion          Value       DF     Value/DF     Pr > ChiSq
Deviance         11.9974       14       0.8570         0.6065
Pearson          11.1302       14       0.7950         0.6758

Number of unique profiles: 16

        Testing Global Null Hypothesis: BETA=0
Test                 Chi-Square       DF     Pr > ChiSq
Likelihood Ratio         7.9520        1         0.0048
Score                    7.2312        1         0.0072
Wald                     4.6008        1         0.0320

             Analysis of Maximum Likelihood Estimates
                                Standard          Wald
Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq
Intercept     1     15.0429      7.3786        4.1563        0.0415
temp          1     -0.2322      0.1082        4.6008        0.0320
```

The *Challenger* was set to launch on a morning where the predicted temperature at lift-off was 31 degrees. Given that temperature appears to affect the probability of O-ring failure (a point that NASA missed), what can/should we say about the potential for O-ring failure?

Clearly, the launch temperature is outside the region for which data were available. Thus, we really have no prior information about what is likely to occur. If we assume the logistic model holds, and we can extrapolate back to 31 degrees, what is the estimated probability of O-ring failure?

The following **SAS** program gives output to answer this question. I augmented the original data set to obtain predicted probabilities for temperatures not observed in the data set. I included a snipit of the summary information for the fitted model to convince you that the data augmentation does not influence the fit of the model.

Program editor contents:

```
* add missing values to predict 31 F temp ;
data shuttle;
  input case flight y six temp pressure @@;
  datalines;
          1     14 1   2   53       50         13  2 1 1 70 200
          2      9 1   1   57       50         14 11 1 1 70 200
          3     23 1   1   58      200         15  6 0 0 72 200
          4     10 1   1   63       50         16  7 0 0 73 200
          5      1 0   0   66      200         17 16 0 0 75 100
          6      5 0   0   67       50         18 21 1 2 75 200
          7     13 0   0   67      200         19 19 0 0 76 200
          8     15 0   0   67       50         20 22 0 0 76 200
          9      4 0   0   68      200         21 12 0 0 78 200
         10      3 0   0   69      200         22 20 0 0 79 200
         11      8 0   0   70       50         23 18 0 0 81 200
         12     17 0   0   70      200

         24      . .   .   31        .         28  . . . 43   .
         25      . .   .   34        .         29  . . . 46   .
         26      . .   .   37        .         30  . . . 49   .
         27      . .   .   40        .         31  . . . 52   .
;
run;

proc logistic data=shuttle descending;
  model y = temp /scale=none aggregate;
  output out=outlog p=fitprob l=l95 u=u95/alpha=0.05;
```

```
run;

proc print data=outlog;
run;

symbol1 v=dot c=black  i=none;
proc gplot data=outlog;
  plot fitprob*temp;
run;
```

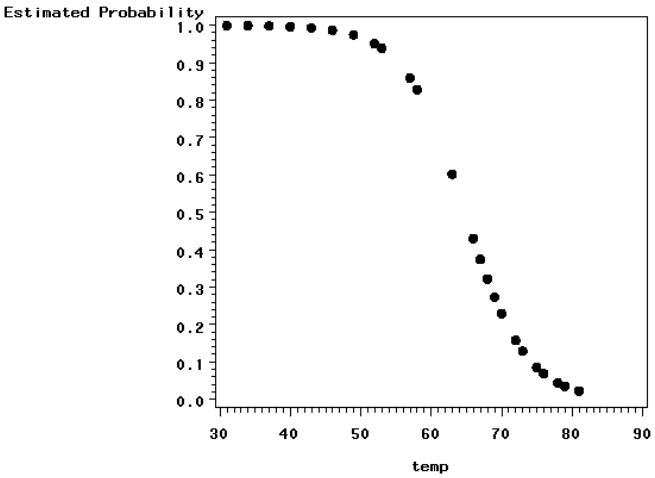
Output window contents:

```
NOTE: 8 observations were deleted due to missing values for the response or
      explanatory variables.
```

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	15.0429	7.3786	4.1563	0.0415
temp	1	-0.2322	0.1082	4.6008	0.0320

Obs	case	flight	y	six	temp	pressure	_LEVEL_	fitprob	l95	u95	
1	1	14	1	2	53	50	1	0.93925	0.34988	0.99775	
2	13	2	1	1	70	200	1	0.22997	0.08509	0.48955	
3	2	9	1	1	57	50	1	0.85932	0.31868	0.98762	
4	14	11	1	1	70	200	1	0.22997	0.08509	0.48955	
5	3	23	1	1	58	200	1	0.82884	0.30972	0.98123	
6	15	6	0	0	72	200	1	0.15805	0.04509	0.42736	
7	4	10	1	1	63	50	1	0.60268	0.25018	0.87335	
8	16	7	0	0	73	200	1	0.12955	0.03151	0.40504	
9	5	1	0	0	66	200	1	0.43049	0.19103	0.70757	
10	17	16	0	0	75	100	1	0.08554	0.01460	0.37131	
11	6	5	0	0	67	50	1	0.37472	0.16565	0.64400	
12	18	21	1	2	75	200	1	0.08554	0.01460	0.37131	
13	7	13	0	0	67	200	1	0.37472	0.16565	0.64400	
14	19	19	0	0	76	200	1	0.06904	0.00976	0.35815	
15	8	15	0	0	67	50	1	0.37472	0.16565	0.64400	
16	20	22	0	0	76	200	1	0.06904	0.00976	0.35815	
17	9	4	0	0	68	200	1	0.32209	0.13833	0.58441	
18	21	12	0	0	78	200	1	0.04454	0.00427	0.33647	
19	10	3	0	0	69	200	1	0.27362	0.11078	0.53249	
20	22	20	0	0	79	200	1	0.03564	0.00280	0.32729	
21	11	8	0	0	70	50	1	0.22997	0.08509	0.48955	
22	23	18	0	0	81	200	1	0.02270	0.00119	0.31119	
23	12	17	0	0	70	200	1	0.22997	0.08509	0.48955	
24	24	.	.	.	31	.	1	0.99961	0.48161	1.00000	**
25	28	.	.	.	43	.	1	0.99369	0.41359	0.99997	
26	25	.	.	.	34	.	1	0.99922	0.46501	1.00000	
27	29	.	.	.	46	.	1	0.98743	0.39560	0.99989	
28	26	.	.	.	37	.	1	0.99843	0.44818	1.00000	
29	30	.	.	.	49	.	1	0.97508	0.37684	0.99961	
30	27	.	.	.	40	.	1	0.99685	0.43107	0.99999	
31	31	.	.	.	52	.	1	0.95122	0.35690	0.99854	



11.5 Using genmod for generalized linear models

The **genmod** procedure fits **generalized linear models**, a class of models that includes normal linear regression and logistic regression as special cases. The structure of **genmod** mimics **glm** in that factors and interactions can be directly entered into the model. The following example illustrates some basic **genmod** features. The online help gives more details.

The purpose of this experiment was to assess the resistance of tobacco budworm larvae to cypermethrin. Batches of 20 pyrethroid-resistant moths (i.e., moths that have grown resistant to the pesticide pyrethroid) of each sex were exposed to a range of doses of cypermethrin two days after emergence from pupation. The number of moths which were either knocked down (uncoordinated movements) or dead was recorded 72 hours after treatment. We will model the probability of uncoordinated movement, or death, as a function of sex and dose.

The **SAS** program below inputs the number of responders for each combination of sex and dose. The sample size of 20 for each sex-dose combination is defined directly in the `data` step, rather than entered with the data. A plot of the observed proportion of responders against the log(dose) shows that increased doses are more effective, and that males tend to be more affected than females.

The first **genmod** proc fits a logistic model with main effects for **sex** (a factor) and **log(dose)** (a predictor) and a sex-by-log(dose) interaction. The model implies that the log-odds of a response is linearly related to log(dose), but that males and females have distinct lines.

The `class` statement identifies sex as a factor. The `model` statement specifies the response in the number of events divided by sample size format used in the **logistic** procedure, but the **logistic** format for binary response data is also allowed here. The `dist=bin link=logit` option on the `model` statement implies that a binomial response model with a logit link (i.e., logistic regression) is fitted to the data. The output should be easy to interpret given our experience with **glm**.

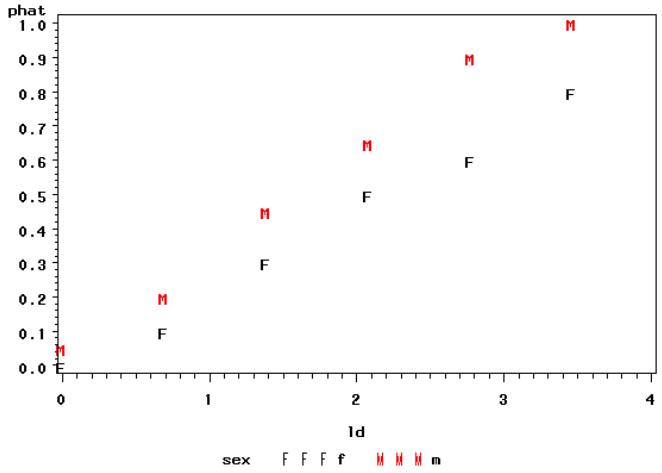
Looking at the output, we see that the sex-by-log(dose) interaction is not significant (p-value = 0.18) so an equal slopes model (i.e., the original model with no interaction) was then considered. Let us interpret the output.

Program editor contents:

```
* Moth data using genmod;
options ls=79 nodate nocenter;
```

```
data moths;
  input sex $ dose nresp @@;     *$;
  ntot  = 20;
  phat  = nresp/ntot;
  ld    = log(dose);
  cards;
m  1  1   m  2  4   m  4  9   m  8  13
m 16 18   m 32 20
f  1  0   f  2  2   f  4  6   f  8  10
f 16 12   f 32 16
;
run;

symbol1 v="f" c=black  i=none;
symbol2 v="m" c=red    i=none;
proc gplot data=moths;
  plot phat*ld=sex;
run;
```

The logistic model on log(dose) looks reasonable (more so than on dose?).



Model with unequal slopes

Program editor contents:

```
* unequal slopes model;
proc genmod data=moths;
  class sex;
  model nresp/ntot = sex ld sex*ld
        / dist=bin link=logit noscale type3;
run;
```

Output window contents:

```
The GENMOD Procedure

           Model Information
Data Set                      WORK.MOTHS
Distribution                    Binomial
Link Function                      Logit
Response Variable (Events)         nresp
Response Variable (Trials)          ntot

Number of Observations Read           12
Number of Observations Used           12
Number of Events                     111
Number of Trials                     240

  Class Level Information
Class      Levels    Values
sex             2    f m



           Criteria For Assessing Goodness Of Fit
Criterion                 DF           Value          Value/DF
Deviance                   8          4.9937            0.6242
Scaled Deviance            8          4.9937            0.6242
Pearson Chi-Square         8          3.5047            0.4381
Scaled Pearson X2          8          3.5047            0.4381
Log Likelihood                     -105.7388

Algorithm converged.

                        Analysis Of Parameter Estimates
                                  Standard       Wald 95%          Chi-
Parameter     DF  Estimate     Error   Confidence Limits   Square  Pr > ChiSq
Intercept      1   -2.8186    0.5480   -3.8926   -1.7445    26.46      <.0001
sex        f   1   -0.1750    0.7783   -1.7004    1.3505     0.05      0.8221
sex        m   0    0.0000    0.0000    0.0000    0.0000        .           .
ld             1    1.8163    0.3059    1.2166    2.4159    35.24      <.0001
ld*sex     f   1   -0.5091    0.3895   -1.2726    0.2543     1.71      0.1912
ld*sex     m   0    0.0000    0.0000    0.0000    0.0000        .           .
Scale          0    1.0000    0.0000    1.0000    1.0000

NOTE: The scale parameter was held fixed.

     LR Statistics For Type 3 Analysis
                            Chi-
Source            DF      Square    Pr > ChiSq
sex                1        0.05        0.8221
ld                 1      112.73        <.0001
ld*sex             1        1.76        0.1842
```

Model with equal slopes

Program editor contents:

```
* equal slopes model;
proc genmod data=moths;
  class sex;
  model nresp/ntot = sex ld
        / dist=bin link=logit noscale type3;
run;
```

Output window contents:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	9	6.7571	0.7508
Scaled Deviance	9	6.7571	0.7508
Pearson Chi-Square	9	5.3060	0.5896
Scaled Pearson X2	9	5.3060	0.5896
Log Likelihood		-106.6204	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-2.3724	0.3855	-3.1280	-1.6168	37.87	<.0001
sex	f	1	-1.1007	0.3558	-1.7982	-0.4033	9.57	0.0020
sex	m	0	0.0000	0.0000	0.0000	0.0000	.	.
ld		1	1.5353	0.1891	1.1647	1.9060	65.92	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
sex	1	10.23	0.0014
ld	1	112.04	<.0001

Chapter 12

An Introduction to Multivariate Methods

Multivariate statistical methods are used to display, analyze, and describe data on two or more features or variables simultaneously. I will discuss multivariate methods for measurement data. Methods for multi-dimensional count data, or mixtures of counts and measurements are available, but are beyond the scope of what I can do here. I will give a brief overview of the type of problems where multivariate methods are appropriate.

Example: Turtle shells Jolicouer and Mosimann provided data on the height, length, and width of the carapace (shell) for a sample of female painted turtles. **Cluster analysis** is used to identify which shells are similar on the three features. **Principal component analysis** is used to identify the linear combinations of the measurements that account for most of the variation in size and shape of the shells.

Cluster analysis and principal component analysis are primarily descriptive techniques.

Example: Fisher's Iris data Random samples of 50 flowers were selected from three iris species: Setosa, Virginica, and Versicolor. Four measurements were made on each flower: sepal length, sepal width, petal length, and petal width. Suppose the sample means on each feature are computed within the three species. Are the means on the four traits significantly different across species? This question can be answered using four separate one-way ANOVAs. A more powerful **MANOVA** (multivariate analysis of variance) method compares species on the four features simultaneously.

Discriminant analysis is a technique for comparing groups on multi-dimensional data. Discriminant analysis can be used with Fisher's Iris data to find the linear combinations of the flower features that best distinguish species. The linear combinations are optimally selected, so insignificant differences on one or all features may be significant (or better yet, important) when the features are considered simultaneously! Furthermore, the discriminant analysis could be used to classify flowers into one of these three species when their species is unknown.

MANOVA, discriminant analysis and **classification** are primarily inferential techniques.

12.1 Linear Combinations

Suppose data are collected on p measurements or features $X_1, X_2, \ldots, X_p$. Most multivariate methods use **linear combinations** of the features as the basis for analysis. A linear combination has the form

$$Y = a_1X_1 + a_2X_2 + \cdots + a_pX_p,$$

where the coefficients $a_1, a_2, \ldots, a_p$ are known constants. Y is evaluated for each observation in the data set, keeping the coefficients constant.

For example, three linear combinations of $X_1, X_2, \ldots, X_p$ are:

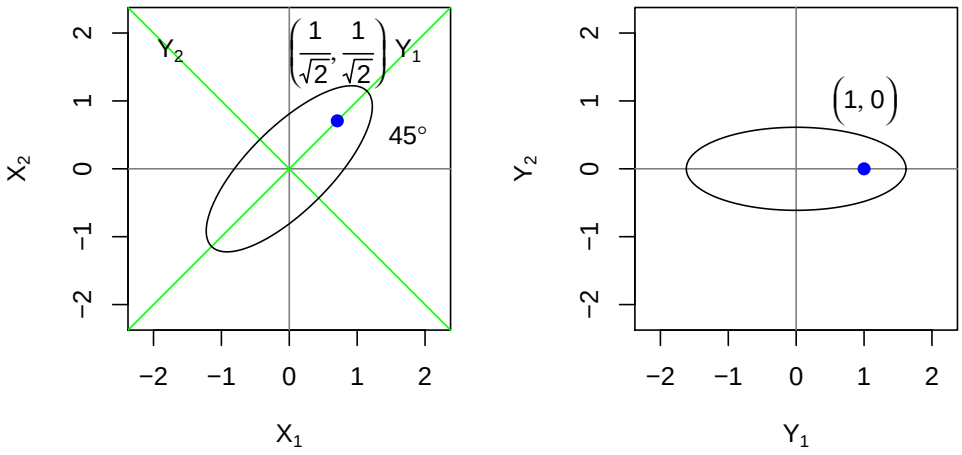
$$\begin{aligned} Y &= 1X_1 + 0X_2 + 0X_3 + \cdots + 0X_p = X_1, \\ Y &= \frac{1}{p}(X_1 + X_2 + \cdots + X_p), \text{ and} \\ Y &= 2X_1 - 4X_2 + 55X_3 - 1954X_4 + \cdots + 44X_p. \end{aligned}$$

Vector and **matrix** notation are useful for representing and summarizing multivariate data. Before introducing this notation, let us try to understand linear combinations geometrically when $p = 2$.

Example: $-45°$ rotation A plot of data on two features X_1 and X_2 is given below. Also included is a plot for the two linear combinations

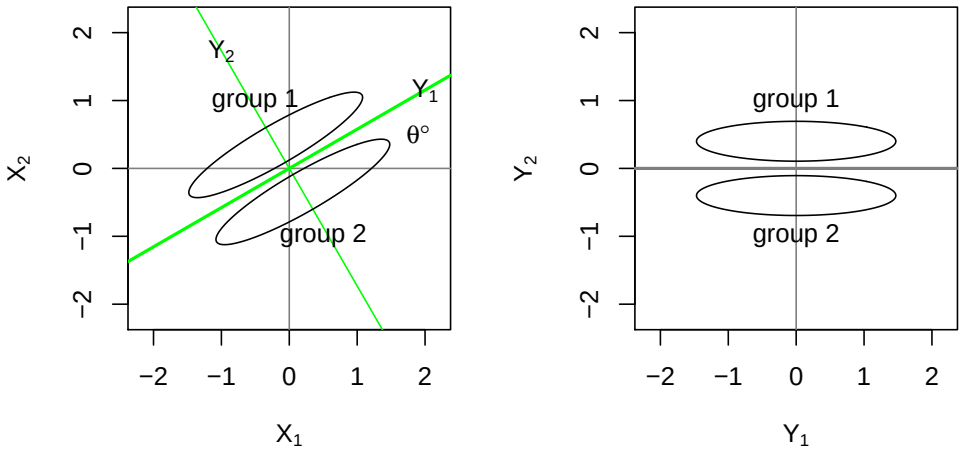
$$\begin{aligned} Y_1 &= \frac{1}{\sqrt{2}}(X_1 + X_2) \qquad \text{and} \\ Y_2 &= \frac{1}{\sqrt{2}}(X_2 - X_1). \end{aligned}$$

This transformation creates two (roughly) uncorrelated linear combinations Y_1 and Y_2 from two highly correlated features X_1 and X_2. The transformation corresponds to a **rotation** of the original coordinate axes by -45 degrees. Each data point is then expressed relative to the new axes. The new features are uncorrelated!



The $\sqrt{2}$ divisor in Y_1 and Y_2 does not alter the interpretation of these linear combinations: Y_1 is essentially the sum of X_1 and X_2, whereas Y_2 is essentially the difference between X_2 and X_1.

Example: Two groups The plot below shows data on two features X_1 and X_2 from two distinct groups.



If you compare the groups on X_1 and X_2 separately, you may find no significant differences because the groups overlap substantially on each feature. The plot on the right was obtained by rotating the coordinate axes $-\theta$ degrees, and then plotting the data relative to the new coordinate axes. The rotation corresponds to creating two linear combinations:

$$\begin{aligned} Y_1 &= \cos(\theta)X_1 + \sin(\theta)X_2 \\ Y_2 &= -\sin(\theta)X_1 + \cos(\theta)X_2. \end{aligned}$$

The two groups differ substantially on Y_2. This linear combination is used with discriminant analysis and MANOVA to distinguish between the groups.

The linear combinations used in certain multivariate methods do not correspond to a rotation of the original coordinate axes. However, the pictures given above should provide some insight into the motivation for the creating linear combinations of two features. The ideas extend to three or more features, but are more difficult to represent visually.

12.2 Vector and Matrix Notation

A **vector** is a string of numbers or variables that is stored in either a row or in a column. For example, the collection $X_1, X_2, \ldots, X_p$ of features can be represented as a **column-vector** with p rows, using the notation

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}.$$

The entry in the j^{th} row is X_j. The transpose of X, represented by X', is a **row-vector** with p columns: $X' = (X_1, X_2, \ldots, X_p)$. The j^{th} column of X' contains X_j.

Suppose you collect data on p features $X_1, X_2, \ldots, X_p$ for a sample of n individuals. The data for the i^{th} individual can be represented as the column-vector:

$$X_i = \begin{bmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ip} \end{bmatrix}.$$

or as the row-vector $X_i' = (X_{i1}, X_{i2}, \cdots, X_{ip})$. Here X_{ij} is the value on the j^{th} variable. Two subscripts are needed for the data values. One subscript identifies the individual and the other subscript identifies the feature.

A **matrix** is a rectangular array of numbers or variables. A **SAS** data set can be viewed as a matrix with n rows and p columns, where n is the sample size. Each row contains data for a given individual:

$$\begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}.$$

Vector and matrix notation are used for summarizing multivariate data. For example, the **sample mean vector** is

$$\overline{X} = \begin{bmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \vdots \\ \overline{X}_p \end{bmatrix},$$

where $\overline{X}_j$ is the sample average on the j^{th} feature. Using matrix algebra, $\overline{X}$ is defined using a familiar formula:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

This mathematical operation is well-defined because vectors are added elementwise.

The sample variances and covariances on the p variables can be grouped together in a $p \times p$ **sample variance-covariance matrix** S (i.e., p rows and p columns)

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix},$$

where

$$s_{ii} = \frac{1}{n-1}\sum_{k=1}^{n}(X_{ki} - \overline{X}_i)^2$$

is the sample variance for the i^{th} feature, and

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(X_{ki} - \overline{X}_i)(X_{kj} - \overline{X}_j)$$

is the sample covariance between the i^{th} and j^{th} features. The subscripts on the elements in S identify where the element is found in the matrix: s_{ij} is stored in the i^{th} row and the j^{th} column. The variances are found on the **main diagonal** of the matrix. The covariances are **off-diagonal** elements. S is symmetric, meaning that the elements above the main diagonal are a reflection of the entries below the main diagonal. More formally, $s_{ij} = s_{ji}$.

Matrix algebra allows you to express S using a formula analogous to the sample variance for a single feature:

$$S = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X})(X_k - \overline{X})'.$$

Here $(X_k - \overline{X})(X_k - \overline{X})'$ is the matrix product of a column vector with p entries times a row vector with p entries. This matrix product is a $p \times p$ matrix with $(X_{ki} - \overline{X}_i)(X_{kj} - \overline{X}_j)$ in the i^{th} row and j^{th} column. The matrix products are added up over all n observations and then divided by $n - 1$.

The interpretation of covariances is enhanced by standardizing them to give correlations. The **sample correlation matrix** is denoted by the $p \times p$ symmetric matrix

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}.$$

The i^{th} row and j^{th} column element of R is the correlation between the i^{th} and j^{th} features. The diagonal elements are one: $r_{ii} = 1$. The off-diagonal elements satisfy

$$r_{ij} = r_{ji} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}.$$

In many applications the data are standardized to have mean 0 and variance 1 on each feature. The data are standardized through the so-called **Z-score transformation**: $(X_{ki}-\overline{X}_i)/s_{ii}$ which, on each feature, subtracts the mean from each observation and divides by the corresponding standard deviation. The sample variance-covariance matrix for the standardized data is the correlation matrix R for the raw data.

Example: Let X_1, X_2 and X_3 be the reaction times for three visual stimuli named A, B and C, respectively. Suppose you are given the following summaries based on a sample of 30 individuals:

$$\overline{X} = \begin{bmatrix} 4 \\ 5 \\ 4.7 \end{bmatrix},$$

$$S = \begin{bmatrix} 2.26 & 2.18 & 1.63 \\ 2.18 & 2.66 & 1.82 \\ 1.63 & 1.82 & 2.47 \end{bmatrix},$$

$$R = \begin{bmatrix} 1.00 & 0.89 & 0.69 \\ 0.89 & 1.00 & 0.71 \\ 0.69 & 0.71 & 1.00 \end{bmatrix}.$$

The average response time on B is 5. The sample variance of response times on A is 2.26. The sample covariance between response times on A and C is 1.63. The sample correlation between response times on B and C is 0.71.

12.3 Matrix Notation to Summarize Linear Combinations

Matrix algebra is useful for computing sample summaries for linear combinations of the features $X' = (X_1, X_2, \ldots, X_p)$ from the sample summaries on these features. For example, suppose you define the linear combination

$$Y_1 = a_1X_1 + a_2X_2 + \cdots + a_pX_p.$$

Using matrix algebra, Y_1 is the matrix product $Y_1 = a'X$, where $a' = (a_1, a_2, \ldots, a_p)$. The sample mean and variance of Y_1 are

$$\overline{Y} = a_1\overline{X}_1 + a_2\overline{X}_2 + \cdots + a_p\overline{X}_p = a'\overline{X}$$

and

$$s_Y^2 = \sum_{ij} a_i a_j s_{ij} = a'Sa,$$

where $\overline{X}$ and S are the sample mean vector and sample variance-covariance matrix for $X' = (X_1, X_2, \ldots, X_p)$.

Similarly, the sample covariance between Y_1 and

$$Y_2 = b'X = b_1X_1 + b_2X_2 + \cdots + b_pX_p$$

is

$$s_{Y_1,Y_2} = \sum_{ij} a_i b_j s_{ij} = a'Sb = b'Sa.$$

Example: In the stimuli example, the total reaction time per individual is

$$Y \;=\; [1\ 1\ 1]\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \;=\; X_1 + X_2 + X_3.$$

The mean total reaction time is

$$\begin{aligned} \overline{Y} &= [1\ 1\ 1] \\ \overline{X} &= [1\ 1\ 1]\begin{bmatrix} 4 \\ 5 \\ 4.7 \end{bmatrix} \;=\; 4 + 5 + 4.7 \;=\; 13.7. \end{aligned}$$

The variance of Y is the sum of the elements in the variance-covariance matrix:

$$s_Y^2 \;=\; [1\ 1\ 1]\, S \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \;=\; \sum_{ij} s_{ij} \;=\; 2.26 + 2.18 + \cdots + 2.47 \;=\; 18.65.$$

Chapter 13

Principal Component Analysis: PCA

Principal component analysis is a multivariate technique for understanding variation, and for summarizing measurement data possibly through **variable reduction**. Principal components (the variables created in PCA) are sometimes used in addition to, or in place of, the original variables in certain analyses. I will illustrate the use and misuse of principal components in a series of examples.

Given data on p variables or features $X_1, X_2, \ldots, X_p$, PCA uses a rotation of the original coordinate axes to produce a **new** set of p **uncorrelated** variables, called principal components, that are **unit-length linear combinations** of the original variables. A unit-length linear combination $a_1X_1 + a_2X_2 + \cdots + a_pX_p$ has $a_1^2 + a_2^2 + \cdots + a_p^2 = 1$.

The principal components have the following properties. The **first principal component**

$$\text{PRIN1} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p$$

has the **largest variability** among all unit-length linear combinations of the original variables. The **second principal component**

$$\text{PRIN2} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p$$

has the largest variability among all unit-length linear combinations of $X_1, X_2, \ldots, X_p$ that are uncorrelated with PRIN1. In general, the j^{th} principal component PRINj for $j = 1, 2, ..., p$, has the largest variability among all unit-length linear combinations of the features that are uncorrelated with $\text{PRIN1}, \text{PRIN2}, \ldots, \text{PRIN}(j-1)$. The **last** or p^{th} **principal component** PRINp has the **smallest variability** among all unit-length linear combinations of the features.

I have described PCA on the raw or unstandardized data. This method is often called PCA on the sample covariance matrix, because the principal components are computed numerically using a **singular value decomposition** of the sample covariance matrix for the X_is. The variances of the PCs are **eigenvalues** of the sample covariance matrix. The coefficients in the PCs are **eigenvectors** of the sample covariance matrix. The sum of the variances of the principal components is equal to the sum of the variances in the original features. An alternative method for PCA uses standardized data, which is often called PCA on the correlation matrix.

The ordered principal components are uncorrelated variables with progressively less variation. Principal components are often viewed as separate dimensions corresponding to the collection of features. The variability of each component divided by the total variability of the components is the proportion of the total variation in the data captured by each component. If data reduction

is your goal, then you might need only the first few principal components to capture most of the variability in the data. This issue will be returned to later.

The unit-length constraint on the coefficients in PCA is needed to make the maximization well-defined. Without this constraint, there does not exist a linear combination with maximum variation. For example, the variability of an arbitrary linear combination $a_1X_1+a_2X_2+\cdots+a_pX_p$ is increased by 100 when each coefficient is multiplied by 10!

The principal components are unique only up to a change of the sign for each coefficient. For example,

$$\text{PRIN1} = 0.2X_1 - 0.4X_2 + 0.4X_3 + 0.8X_4$$

and

$$\text{PRIN1} = -0.2X_1 + 0.4X_2 - 0.4X_3 - 0.8X_4$$

have the same variability, so either could play the role of the first principal component. This non-uniqueness does not have an important impact on the analysis.

13.1 Example: Temperature Data

The `princomp` procedure is used in **SAS** for PCA. The `cov` option computes principal components based on the covariance matrix. The default method uses the correlation matrix. The `out=prin` option creates a data set PRIN containing the input data plus the principal component variables or scores, called PRIN1, PRIN2, ...,PRINp by default. The **principal component scores** are the values of the principal components across cases. **SAS** standardizes the principal component scores PRIN1, PRIN2, ..., PRINp to have mean zero. A `var` statement specifies the variables used in the PCA.

The following **SAS** program contains data on the mean monthly temperatures in January and July for 64 U.S. cities. Output from a PCA on the covariance matrix is given. Two principal components are created because $p = 2$.

Some comments on the output:

1. You can visualize PCA when $p = 2$. In the temperature plot, the direction of maximal variability corresponds to the first PC axis. The PRIN1 score for each city is obtained by projecting the temperature pairs perpendicularly onto this axis. The direction of minimum variation corresponds to the second PC axis, which is perpendicular to the first PC axis. The PRIN2 score for each city is obtained by projecting the temperature pairs onto this axis.

2. The **total variance** is the sum of variances for the monthly temperatures: 163.38 = 26.20 + 138.18.

3. The **eigenvalues of the covariance matrix** are variances for the PCs. The variability of PRIN1 = 0.343 JULY + 0.939 JAN is 154.236. The variability of PRIN2 = 0.939 JULY − 0.343 JAN is 9.146. The proportion of the total variability due to PRIN1 is 0.944 = 154.23/163.38. The proportion of the total variability due to PRIN2 is 0.056 = 9.146/163.08.

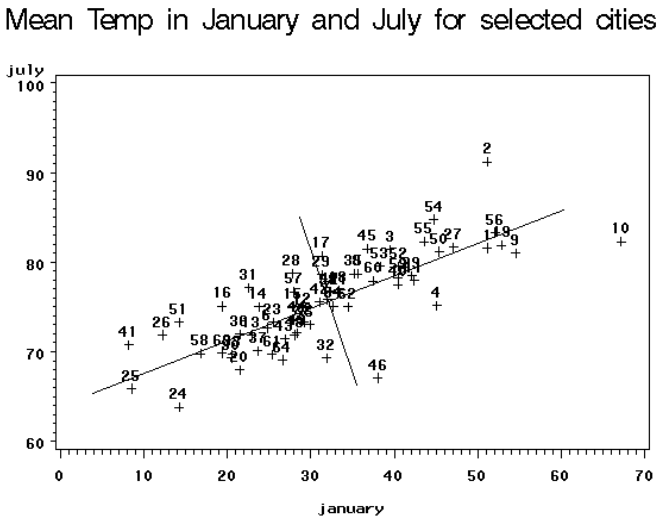
4. Almost all of the variability (94.4%) in the original temperatures is captured by the first PC. The second PC accounts for the remaining 5.6% of the total variability.

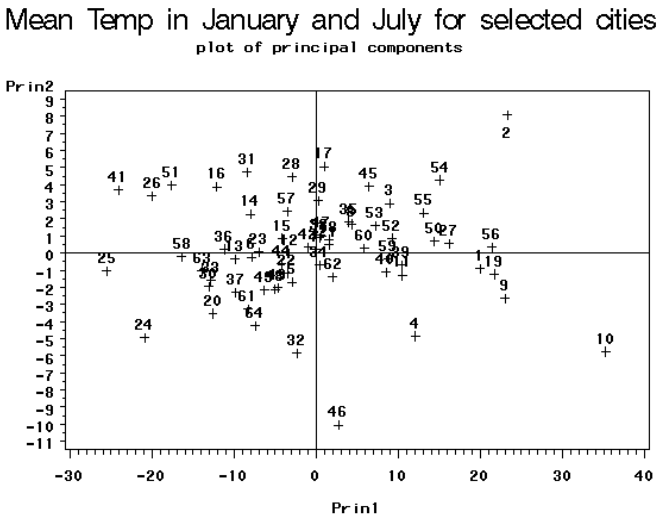
5. PRIN1 weights the January temperature about three times the July temperature. This is sensible because PRIN1 maximizes variation among linear combinations of the January and July temperatures. January temperatures are more variable, so they are weighted heavier in this linear combination.

6. **SAS** standardizes PRIN1 and PRIN2 to have mean zero. This explains why some PRIN1 scores are negative, even though PRIN1 is a weighted average of the January and July temperatures, each of which is non-negative.

Program editor contents:

```
* temperature PCA example;
options ls=79 nodate nocenter;

title 'Mean Temp in January and July for selected cities';
data temp;
  infile "F:\Dropbox\UNM\teach\ADA2_stat528\sas\ADA2_13_temperature.dat";
  input city $1-14 january july;        *$;
  obs = _N_;    * for plotting observation number on plot;
run;

* Create an annotate data set that draws the PC reference lines;
data anno;
  * PC1;
  function='move';
  xsys='2'; ysys='2';
  x=32.09531250-30*0.939390;
  y=75.61093750-30*0.342849;
  output;

  function='draw';
  xsys='2'; ysys='2';
  x=32.09531250+30*0.939390;
  y=75.61093750+30*0.342849;
  output;

  * PC2;
  function='move';
  xsys='2'; ysys='2';
  x=32.09531250-10*0.342849;
  y=75.61093750+10*0.939390;
  output;

  function='draw';
  xsys='2'; ysys='2';
  x=32.09531250+10*0.342849;
  y=75.61093750-10*0.939390;
  output;
run;

  * Use the POINTLABEL option to label the plot points;
  symbol1 pointlabel=('#obs');
proc gplot data=temp;
  plot july*january /anno=anno;   * plot raw data;
run;

proc princomp data=temp cov out=prin;
  var july january;
run;

title2 'plot of principal components';
proc gplot data=prin;
  plot prin2*prin1 /vref=0 href=0;    * plot PC scores;
run;
```





Output window contents:

```
The PRINCOMP Procedure

Observations          64
Variables              2

           Simple Statistics
                   july           january
Mean        75.61093750       32.09531250
StD          5.11862820       11.71243309

             Covariance Matrix
                     july           january
july           26.2003547        46.7290997
january        46.7290997       137.1810888

Total Variance    163.38144345

           Eigenvalues of the Covariance Matrix
        Eigenvalue    Difference    Proportion    Cumulative
   1    154.235808    145.090173        0.9440        0.9440
   2      9.145635                      0.0560        1.0000

           Eigenvectors
                Prin1         Prin2
july          0.342849      0.939390
january       0.939390      -.342849
```

13.2 PCA on Correlation Matrix

The coefficients in the first principal component reflect the relative sizes of the feature variances. The features with large variances have larger coefficients or loadings. This might be considered a problem, because variability is **scale dependent** and the principal component analysis on the raw data does not take scale into account. For example, if you measure height in meters but then change height to centimeters, the variability increases by a factor of 100*100 = 10,000. This might have a dramatic effect on the PCA.

You might prefer to standardize the features when they are measured on different scales, or when the features have wildly different variances. The features are standardized to have mean zero and variance one by using the Z-score transformation: (Obs − Mean)/Std Dev. The PCA is then performed on the standardized data.

The covariance matrix computed from the standardized data is the correlation matrix. Thus, principal components based on the standardized data are computed from the correlation matrix. This is the default method for PCA in **SAS**, and is implemented by removing the `cov` option on the `princomp` procedure statement.

The standardized features are dimensionless, so the PCs are not influenced by the original units of measure, nor are they affected by the variability in the features. The only important factor is the correlation between the features, which is not changed by standardization.

PCA is an exploratory tool, so neither a PCA on the covariance matrix nor a PCA on the correlation matrix is always the "right" method. I often do both and see which analysis is more informative.

13.3 Interpreting Principal Components

You should try to interpret the linear combinations created by multivariate techniques. The interpretations are usually non-trivial, and some degree of creativity is involved.

The **coefficients** or **loadings** in a principal component reflect the **relative** contribution of the features to the linear combination. Most researchers focus more on the **signs** of the coefficents than on the **magnitude** of the coefficients. The principal components are then interpreted as weighted averages or comparisons of weighted averages.

A **weighted average** is a linear combination of features with non-negative loadings. The simplest case of a weighted average is an arithmetic average, where each feature has the same coefficient.

The difference $Z = X - Y$ is a **comparison** of X and Y. The sign and magnitude of Z indicates which of X and Y is larger, and by how much. To see this, note that $Z = 0$ if and only if $X = Y$, whereas $Z < 0$ when $X < Y$ and $Z > 0$ when $X > Y$.

In the temperature data, PRIN1 is a weighted average of January and July temperatures (signs of the loadings: JAN is + and JULY is +):

$$\text{PRIN1} = 0.34\ \text{JULY} + 0.94\ \text{JAN}.$$

PRIN2 is a comparison of January and July temperatures (signs of the loadings: JAN is − and JULY is +):

$$\text{PRIN2} = 0.94\ \text{JULY} - 0.34\ \text{JAN}.$$

Principal components often have positive and negative loadings when $p \geq 3$. To interpret the components, group the features with + and − signs together and then interpret the linear combination as a comparison of weighted averages.

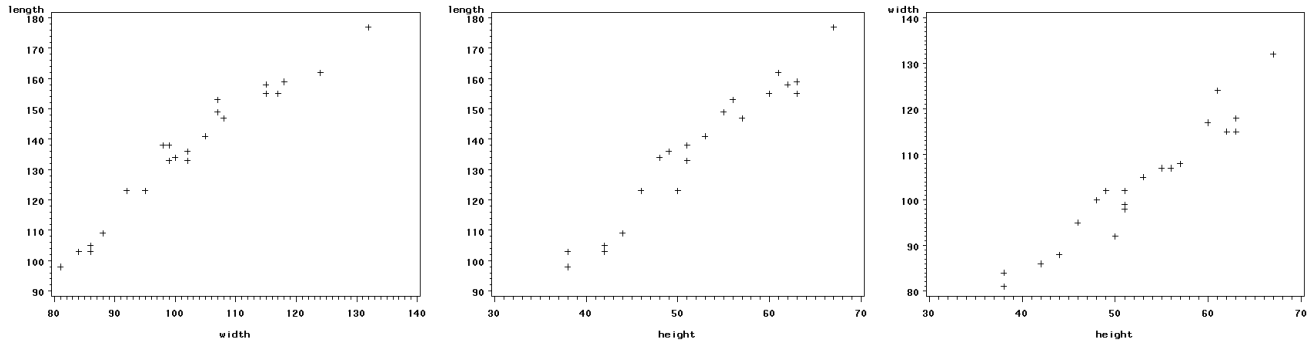
You can often simplify the interpretation of principal components by **mentally eliminating** features from the linear combination that have relatively small (in magnitude) loadings or coefficients. This strategy does not carry over to all multivariate analyses, so I will be careful about this issue when necessary.

13.4 Example: Painted turtle shells

Jolicouer and Mosimann gave the length, width, and height in mm of the carapace (shell) for a sample of 24 female painted turtles. I used **SAS** for a PCA on the original data and on the

standardized data.
Program editor contents:

```
* Painted turtle shells PCA example;
options ls=79 nodate nocenter;

* carapace measurements for female painted turtles ;
data shells;
  input length width height @@;
  datalines;
   98  81 38   103  84 38   103  86 42   105  86 42   109 88  44
  123  92 50   123  95 46   133  99 51   133 102 51   133 102 51
  134 100 48   136 102 49   138  98 51   138  99 51   141 105 53
  147 108 57   149 107 55   153 107 56   155 115 63   155 117 60
  158 115 62   159 118 63   162 124 61   177 132 67
;
run;

proc gplot data=shells;
  plot length*(width height) width*height;
run;
```



13.4.1 PCA on covariance matrix

Let us look at the programs and interpret the output, beginning with the analysis on the covariance matrix. The plots show that the shell measurements are strongly positively correlated, which is not surprising.

The three principal components from the raw data are given below. Length and width are grouped in PRIN3 because they have negative loadings.

$$\begin{aligned} \text{PRIN1} &= 0.81\ \text{Length} + 0.50\ \text{Width} + 0.30\ \text{Height} \\ \text{PRIN2} &= -0.55\ \text{Length} + (0.82\ \text{Width} + 0.15\ \text{Height}) \\ \text{PRIN3} &= -(0.17\ \text{Length} + 0.29\ \text{Width}) + 0.94\ \text{Height}. \end{aligned}$$

PRIN1 is a weighted average of the carapace measurements, and can be viewed as an overall measure of shell size. Jolicouer and Mosimann interpreted the second and third principal components as measures of **shape**, for they appear to be a comparison of length with an average of width and height, and a comparison of height with length and width, respectively.
Program editor contents:

```
proc princomp data=shells cov;
  var length width height;
run;
```

Output window contents:

```
The PRINCOMP Procedure

Observations          24
Variables              3

                  Simple Statistics
              length             width            height
Mean      136.0416667       102.5833333       52.04166667
StD        21.2489983        13.1046512        8.04595046

                   Covariance Matrix
                length             width            height
length      451.5199275       270.9746377       165.9547101
width       270.9746377       171.7318841       101.8442029
height      165.9547101       101.8442029        64.7373188

Total Variance    687.98913043

           Eigenvalues of the Covariance Matrix
        Eigenvalue    Difference    Proportion    Cumulative
   1    678.365651    671.595953        0.9860        0.9860
   2      6.769697      3.915915        0.0098        0.9959
   3      2.853783                      0.0041        1.0000

                  Eigenvectors
                Prin1         Prin2         Prin3
length       0.813881      -.554896      -.172302
width        0.496106      0.818027      -.291052
height       0.302452      0.151401      0.941064
```

Jolicouer and Mosimann argue that the size of female turtle shells can be characterized by PRIN1 with little loss of information because this linear combination accounts for 98.6% of the total variability in the measurements. The form of PRIN1 makes sense conceptually. The carapace measurements are positively correlated with each other, so larger lengths tend to occur with larger widths and heights. The primary way the shells vary is with regards to their overall size, as measured by a weighted average of length, width and height.

Question: Can PRIN2 and PRIN3, which have relatively little variation, be used in any meaningful way? To think about this, suppose the variability in PRIN2 and PRIN3 was zero.

13.4.2 PCA on correlation matrix

For the analysis on the correlation matrix, replace the procedure statement with:

```
proc princomp data=shells;  * no cov option for analysis on corr matrix ;
  var length width height;
run;
```

The three principal components for the standardized data are

$$\begin{aligned}
\text{PRIN1} &= 0.58\ \text{Length} + 0.58\ \text{Width} + 0.58\ \text{Height} \\
\text{PRIN2} &= -(0.14\ \text{Length} + 0.63\ \text{Width}) + 0.77\ \text{Height} \\
\text{PRIN3} &= -0.80\ \text{Length} + (0.52\ \text{Width} + 0.28\ \text{Height}).
\end{aligned}$$

The first principal component accounts for 98% of the total variability in the standardized data. **SAS** does not report the total variability, which is always the number p of features. (Why? — because it is the sum of the variances.). Here, $p = 3$. Little loss of information is obtained by

summarizing the standardized data using PRIN1, which is essentially an average of length, width, and height. PRIN2 and PRIN3 are measures of shape.

The loadings in the first principal component are approximately equal because the correlations between pairs of features are almost identical. The standardized features are essentially interchangeable with regards to the construction of the first principal component, so they must be weighted similarly. True, but not obvious.

Output window contents:

The PRINCOMP Procedure

Observations 24
Variables 3

Simple Statistics

	length	width	height
Mean	136.0416667	102.5833333	52.04166667
StD	21.2489983	13.1046512	8.04595046

Correlation Matrix

	length	width	height
length	1.0000	0.9731	0.9707
width	0.9731	1.0000	0.9659
height	0.9707	0.9659	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.93979909	2.90546140	0.9799	0.9799
2	0.03433769	0.00847446	0.0114	0.9914
3	0.02586323		0.0086	1.0000

Eigenvectors

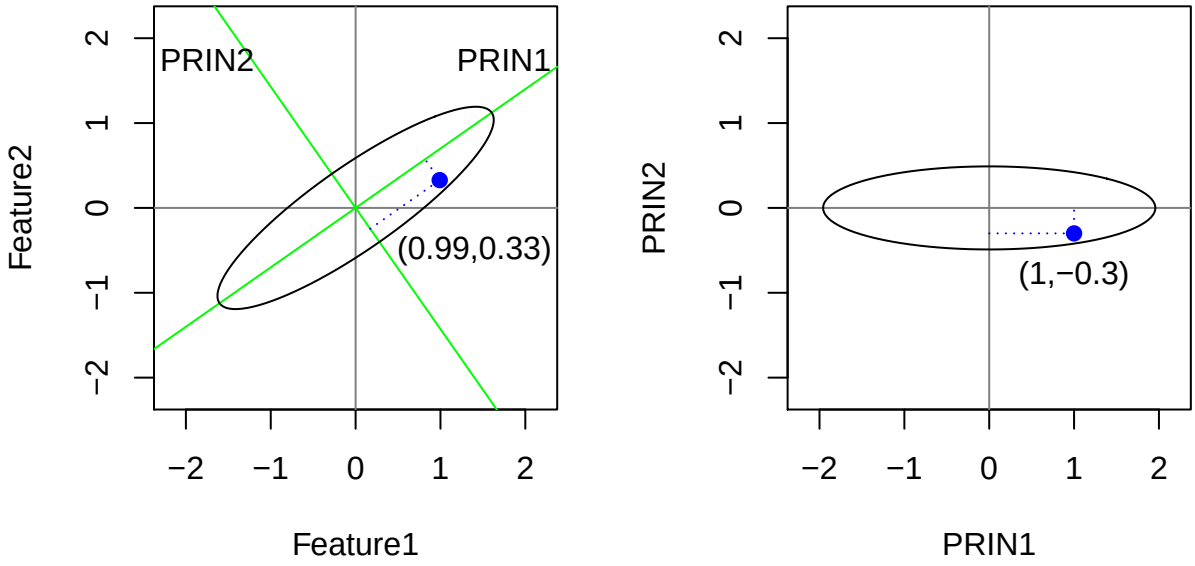
	Prin1	Prin2	Prin3
length	0.578142	-.137395	-.804285
width	0.577197	-.627848	0.522159
height	0.576711	0.766113	0.283681

13.5 Why is PCA a Sensible Variable Reduction Technique?

My description of PCA provides little insight into why principal components are reasonable summaries for **variable reduction**. Specifically, why does it make sense for researchers to consider the linear combination of the original features with the **largest variance** as the best single variable summary of the data?

There is an alternative description of PCA that provides more insight into this issue. For simplicity, consider the following data plot of two features, and the implied principal components. The PC scores for an observation are obtained by projecting the feature scores onto the axes of maximal and minimal variation, and then rotating the axes appropriately.

One can show mathematically that PRIN1 is the best (in some sense) linear combination of the two features to predict the original two features simultaneously. Intuitively, this is plausible. In a PCA, you know the direction for the axis of maximal variation. Given the value of PRIN1, you get a good prediction of the original feature scores by moving PRIN1 units along the axis of maximal variation in the feature space.



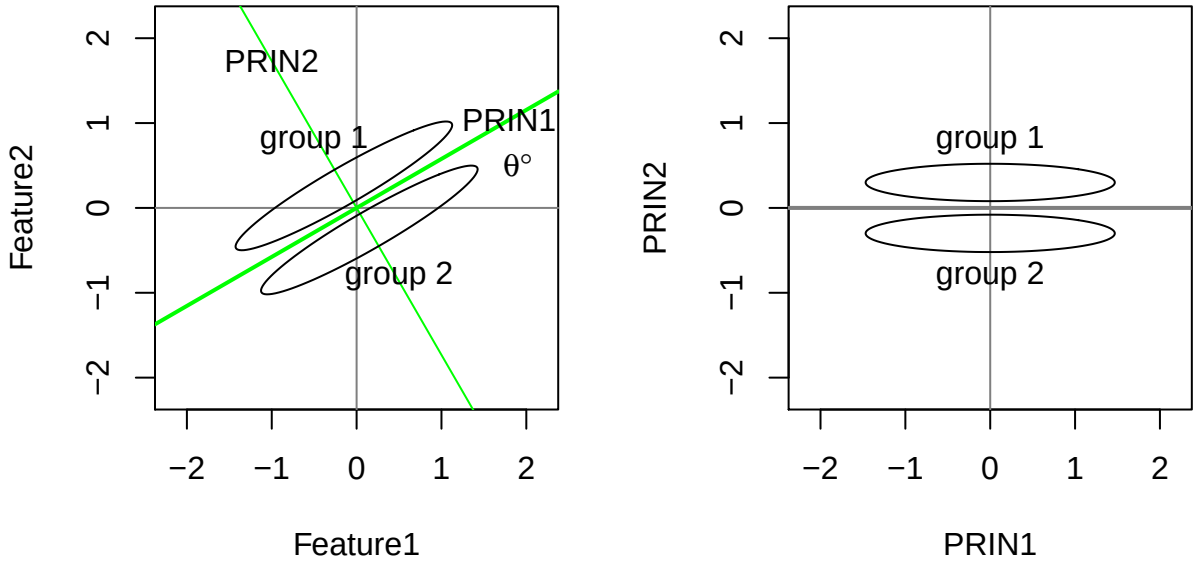
The LS line from regressing feature 2 on feature 1 gives the best prediction for feature 2 scores when the feature 1 score is known. Similarly, the LS line from regressing feature 1 on feature 2 gives the best prediction for feature 1 scores when the feature 2 score is known. PRIN1 is the best linear combination of features 1 and 2 to predict both features *simultaneously*. Note that feature 1 and feature 2 are linear combinations as well!

This idea generalizes. The first k principal components give the best simultaneous prediction of the original p features, among all possible choices of k uncorrelated unit-length linear combinations of the features. Prediction of the original features improves as additional components are added, but the improvement is slight when the added principal components have little variability. Thus, summarizing the data using the principal components with maximum variation is a sensible strategy for data reduction.

13.5.1 A Warning on Using PCA as a Variable Reduction Technique

Some researchers view PCA as a "catchall" technique for reducing the number of variables that need to be considered in an analysis. They will replace the original variables with a small number of principal components that explain most of the variation in the original data and proceed with an analysis on the principal components.

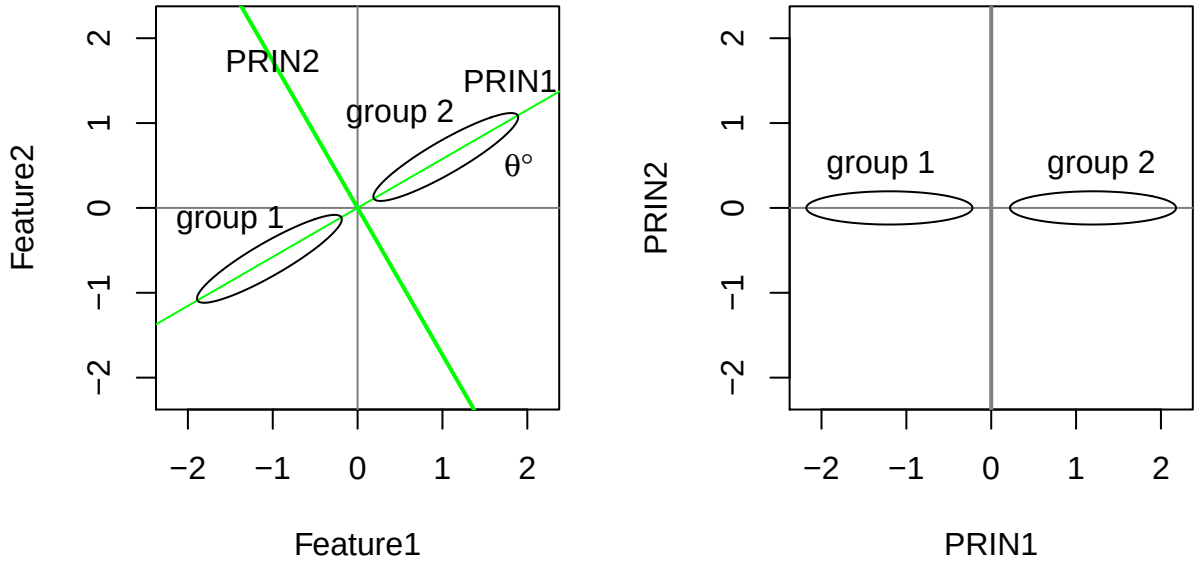
This strategy is not always sensible, especially if a primary interest in the analysis is a comparison of heterogeneous groups. If the group structure was ignored in the PCA analysis, then the linear combinations retained by the researcher may contain little information for distinguishing among groups. For example, consider the following data plot on two features and two groups, and the implied principal components.



Although PRIN1 explains most of the variability in the two features (ignoring the groups), little of the total variation is due to group differences. If the researcher reduced the two features to the first principal component, he would be throwing away most of the information for distinguishing between the groups. PRIN2 accounts for little of the total variation in the features, but most of the variation in PRIN2 is due to group differences.

If a comparison of the two groups was the primary interest, then the researcher should use discriminant analysis instead. Although there is little gained by reducing two variables to one, this principle always applies in multivariate problems. In discriminant analysis, a stepwise selection of variables can be implemented to eliminate features that have no information for distinguishing among groups. This is data reduction as it should be practiced — with a final goal in mind.

Variable reduction using PCA followed by group comparisons might be fruitful if you are fortunate enough to have the directions with large variation correspond to the directions of group differences. For example, in the plot below, the first principal component is a linear combination of the features that distinguishes between the groups. A comparison of the groups based on the first principal component will lead to similar conclusions as a discriminant analysis.

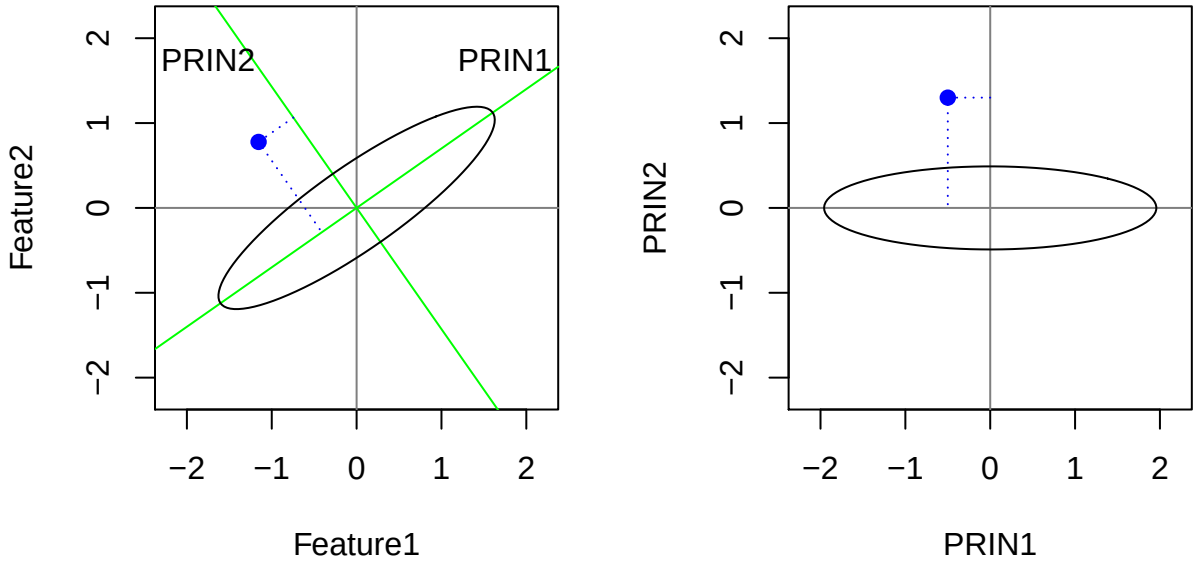


There is nothing unreasonable about using principal components in a comparison of groups, provided you recognize that the principal component scores with the largest variability need not be informative for group comparisons!

In summary, PCA should be used to summarize the variation **within a homogeneous group**, and should not, in general, be used as a data reduction tool prior to a comparison across groups. The same concern applies to using PCA for identifying groups of similar objects (use cluster analysis instead), or when **factor analysis** is used prior to a group comparison.

13.5.2 PCA is Used for Multivariate Outlier Detection

An outlier in a multidimensional data set has atypical readings on one or on several features *simultaneously*. These observations can often be found in univariate plots of the lead principal component scores, even when outliers are not extreme in any individual feature.



13.6 Example: Sparrows, for Class Discussion

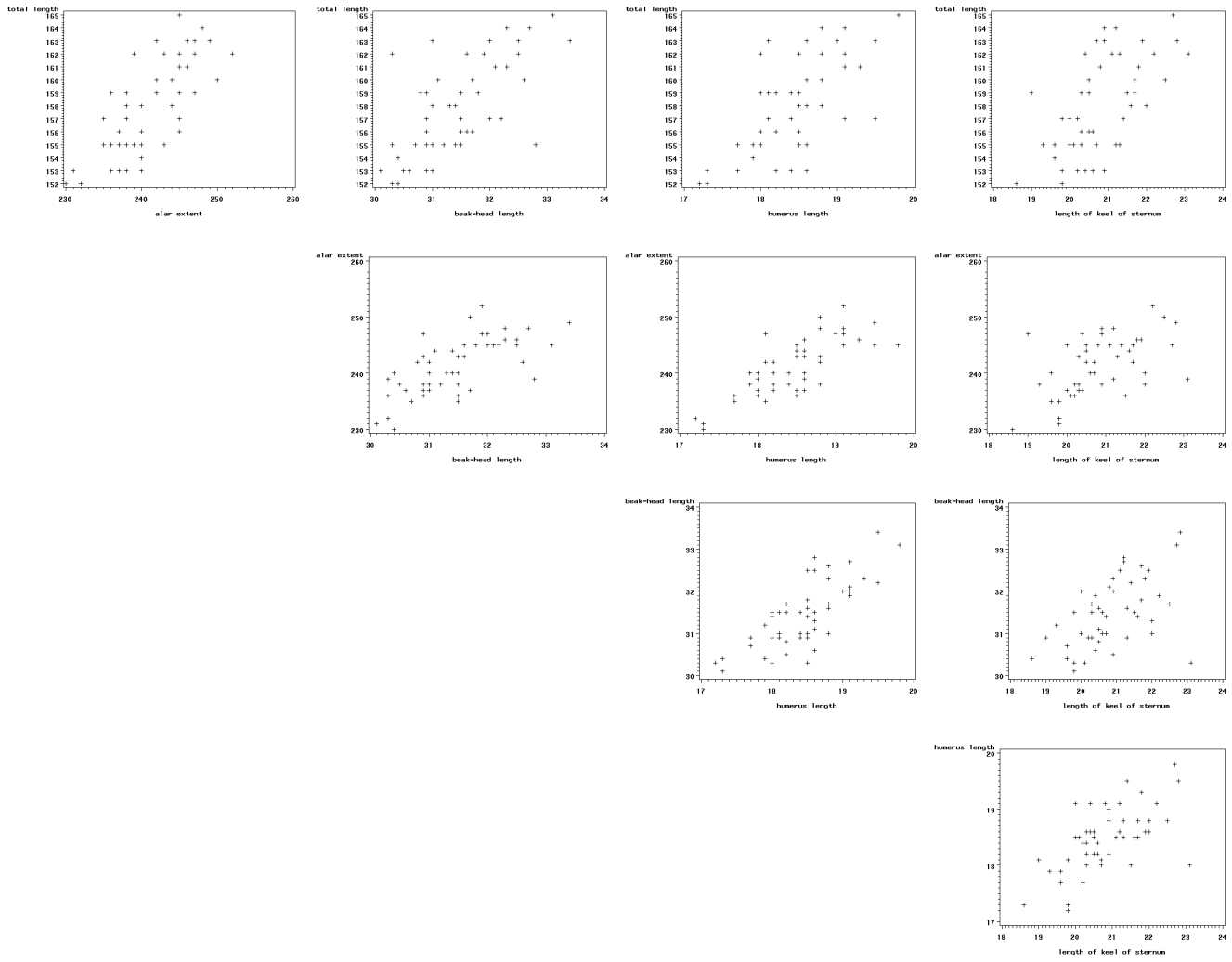
After a severe storm in 1898, a number of sparrows were taken to the biological laboratory at the University of Rhode Island. H. Bumbus measured several morphological characteristics on each bird. The data here correspond to five measurements on a sample of 49 females.

Let us look at the **SAS** output, paying careful attention to the interpretations of the principal components (zeroing out small loadings). How many components seem sufficient to capture the total variation in the morphological measurements?

Program editor contents:

```
* example: body measurements of female sparrows;
options ls=79 nodate nocenter;

data sparrows;
  input x1-x5 @@;
  label x1='total length'
        x2='alar extent'
        x3='beak-head length'
        x4='humerus length'
        x5='length of keel of sternum';
  datalines;                                    * 2 birds per line;
  156 245 31.6 18.5 20.5   154 240 30.4 17.9 19.6
  153 240 31.0 18.4 20.6   153 236 30.9 17.7 20.2
  155 243 31.5 18.6 20.3   163 247 32.0 19.0 20.9
  157 238 30.9 18.4 20.2   155 239 32.8 18.6 21.2
  164 248 32.7 19.1 21.2   158 238 31.0 18.8 22.0
  158 240 31.3 18.6 22.0   160 244 31.1 18.6 20.5
  161 246 32.3 19.3 21.8   157 245 32.0 19.1 20.0
  157 235 31.5 18.1 19.8   156 237 30.9 18.0 20.3
  158 244 31.4 18.5 21.6   153 238 30.5 18.2 20.9
  155 236 30.3 18.5 20.1   163 246 32.5 18.6 21.9
```

```
  159 236 31.5 18.0 21.5   155 240 31.4 18.0 20.7
  156 240 31.5 18.2 20.6   160 242 32.6 18.8 21.7
  152 232 30.3 17.2 19.8   160 250 31.7 18.8 22.5
  155 237 31.0 18.5 20.0   157 245 32.2 19.5 21.4
  165 245 33.1 19.8 22.7   153 231 30.1 17.3 19.8
  162 239 30.3 18.0 23.1   162 243 31.6 18.8 21.3
  159 245 31.8 18.5 21.7   159 247 30.9 18.1 19.0
  155 243 30.9 18.5 21.3   162 252 31.9 19.1 22.2
  152 230 30.4 17.3 18.6   159 242 30.8 18.2 20.5
  155 238 31.2 17.9 19.3   163 249 33.4 19.5 22.8
  163 242 31.0 18.1 20.7   156 237 31.7 18.2 20.3
  159 238 31.5 18.4 20.3   161 245 32.1 19.1 20.8
  155 235 30.7 17.7 19.6   162 247 31.9 19.1 20.4
  153 237 30.6 18.6 20.4   162 245 32.5 18.5 21.1
  164 248 32.3 18.8 20.9
  ;
run;

* create all pair-wise scatterplots;
proc gplot data=sparrows;
  plot x1*(x2-x5) x2*(x3-x5) x3*(x4 x5) x4*x5;
run;

  ** Two ways to get scatterplot matrices in SAS;
  * use the "output delivery system" in SAS v9.1+ ;
  *    cite: http://www.ats.ucla.edu/stat/sas/faq/sas9_stat_plots.htm ;
  * scatterplot matrix;
  ods html style=journal;           * turn on html output;
  ods graphics on;                  * turn on ods graphics;
  proc corr data=sparrows plots;
    var x1-x5;
  run;
  ods graphics off;
  ods html close;

* perform PCA on correlation matrix
*   output observations with eigenvectors coordinates;
proc princomp data=sparrows out=prin;
  var x1-x5;
run;
```



Output window contents:

```
The PRINCOMP Procedure

Observations          49
Variables              5

                                Simple Statistics
                  x1           x2           x3           x4           x5
Mean     157.9795918  241.3265306  31.45918367  18.46938776  20.82857143
StD        3.6542772    5.0678223   0.79475320   0.56428571   0.99205175

                             Correlation Matrix
                                     x1       x2       x3       x4       x5
x1   total length                1.0000   0.7350   0.6618   0.6453   0.6082
x2   alar extent                 0.7350   1.0000   0.6737   0.7685   0.5314
x3   beak-head length            0.6618   0.6737   1.0000   0.7632   0.5292
x4   humerus length              0.6453   0.7685   0.7632   1.0000   0.6086
x5   length of keel of sternum   0.6082   0.5314   0.5292   0.6086   1.0000

           Eigenvalues of the Correlation Matrix
        Eigenvalue    Difference    Proportion    Cumulative
   1    3.61972961    3.09149508        0.7239        0.7239
   2    0.52823453    0.14210281        0.1056        0.8296
   3    0.38613172    0.08470633        0.0772        0.9068
   4    0.30142539    0.13694664        0.0603        0.9671
   5    0.16447875                      0.0329        1.0000

                                  Eigenvectors
                                    Prin1     Prin2     Prin3     Prin4     Prin5
x1  total length                 0.451659  0.057569  -.688674  0.421547  -.375121
x2  alar extent                  0.461418  -.300644  -.344761  -.544832  0.530099
x3  beak-head length             0.450356  -.326162  0.453466  0.606942  0.341742
x4  humerus length               0.470433  -.189042  0.408937  -.389564  -.651127
x5  length of keel of sternum    0.398710  0.874174  0.184409  -.072927  0.193713
```

13.7 PCA for Variable Reduction in Regression

I will outline an analysis where PCA was used to create predictors in a regression model. The data were selected from the Berkeley Guidance Study, a longitudinal monitoring of children born in Berkeley, California, between 1928 and 1929. The variables selected from the study are an identification number (ID), weight at age 2 in kg (WT2), height at age 2 in cm (HT2), weight at age 9 (WT9), height at age 9 (HT9), leg circumference at age 9 in cm (LG9), a composite measure of strength at age 9 (ST9; higher is stronger), weight at age 18 (WT18), height at age 18 (HT18), leg circumference at age 18 (LG18), a composite measure of strength at age 18 (ST18), and somatotype on a 7-point scale, as a measure of fatness (SOMA; 1=slender to 7=fat) determined using a photo taken at age 18. Data on 26 boys are given in the table below.

We are interested in building a regression model to predict somatotype from the other variables. Program editor contents:

```
* example: Berkeley Guidance Study;
options ls=79 nodate nocenter;

data bgs;
  input  ID  WT2  HT2  WT9   HT9  LG9 ST9  WT18  HT18 LG18 ST18 SOMA;
  datalines;
        201 13.6 90.2 41.5 139.4 31.6  74 110.2 179.0 44.1  226  7.0
        202 12.7 91.4 31.0 144.3 26.0  73  79.4 195.1 36.1  252  4.0
        203 12.6 86.4 30.1 136.5 26.6  64  76.3 183.7 36.9  216  6.0
        204 14.8 87.6 34.1 135.4 28.2  75  74.5 178.7 37.3  220  3.0
        205 12.7 86.7 24.5 128.9 24.2  63  55.7 171.5 31.0  200  1.5
        206 11.9 88.1 29.8 136.0 26.7  77  68.2 181.8 37.0  215  3.0
        207 11.5 82.2 26.0 128.5 26.5  45  78.2 172.5 39.1  152  6.0
        209 13.2 83.8 30.1 133.2 27.6  70  66.5 174.6 37.3  189  4.0
```

```
        210 16.9 91.0 37.9 145.6 29.0  61  70.5 190.4 33.9  183  3.0
        211 12.7 87.4 27.0 132.4 26.0  74  57.3 173.8 33.3  193  3.0
        212 11.4 84.2 25.9 133.7 25.8  68  50.3 172.6 31.6  202  3.0
        213 14.2 88.4 31.1 138.3 27.3  59  70.8 185.2 36.6  208  4.0
        214 17.2 87.7 34.6 134.6 30.6  87  73.7 178.4 39.2  227  3.0
        215 13.7 89.6 34.6 139.0 28.9  71  75.2 177.6 36.8  204  2.5
        216 14.2 91.4 43.1 146.0 32.4  98  83.1 183.5 38.0  226  4.0
        217 15.9 90.0 33.2 133.2 28.5  82  74.3 178.1 37.8  233  2.5
        218 14.3 86.4 30.7 133.3 27.3  73  72.2 177.0 36.5  237  2.0
        219 13.3 90.0 31.6 130.3 27.5  68  88.6 172.9 40.4  230  7.0
        221 13.8 91.4 33.4 144.5 27.0  92  75.9 188.4 36.5  250  1.0
        222 11.3 81.3 29.4 125.4 27.7  70  64.9 169.4 35.7  236  3.0
        223 14.3 90.6 30.2 135.8 26.7  70  65.6 180.2 35.4  177  4.0
        224 13.4 92.2 31.1 139.9 27.2  63  66.4 189.0 35.3  186  4.0
        225 12.2 87.1 27.6 136.8 25.8  73  59.0 182.4 33.5  199  3.0
        226 15.9 91.4 32.3 140.6 27.9  69  68.1 185.8 34.2  227  1.0
        227 11.5 89.7 29.0 138.6 24.6  61  67.7 180.7 34.3  164  4.0
        228 14.2 92.2 31.4 140.0 28.2  74  68.5 178.7 37.0  219  2.0
;
run;
```

It is reasonable to expect that the characteristics measured over time, for example HT2, HT9, and HT18 are strongly correlated. Evidence supporting this hypothesis is given in the following output, which summarizes correlations within subsets of the predictors. Two of the three subsets include measures over time on the same characteristic.

Program editor contents:

```
proc corr data=bgs noprob;
  var WT2  WT9 WT18;
run;
proc corr data=bgs noprob;
  var HT2  HT9 HT18;
run;
proc corr data=bgs noprob;
  var LG9  ST9 LG18 ST18;
run;
```

Output window contents:

```
   Pearson Correlation Coefficients, N = 26
              WT2           WT9          WT18
WT2        1.00000       0.57922      0.21587
WT9        0.57922       1.00000      0.70890
WT18       0.21587       0.70890      1.00000

              HT2           HT9          HT18
HT2        1.00000       0.77583      0.68471
HT9        0.77583       1.00000      0.86446
HT18       0.68471       0.86446      1.00000

              LG9           ST9          LG18          ST18
LG9        1.00000       0.52395      0.66448       0.34408
ST9        0.52395       1.00000      0.20853       0.65959
LG18       0.66448       0.20853      1.00000       0.28454
ST18       0.34408       0.65959      0.28454       1.00000
```

Strong correlation among predictors can cause collinearity problems in regression. The presence of collinearity makes the interpretation of regression effects more difficult, and can reek havoc with the numerical stability of certain algorithms used to compute least squares summaries. A natural way to avoid collinearity and improve interpretation of regression effects is to use uncorrelated linear combinations of the original predictors, such as principal components, to build a model.

The interpretation of regression effects changes when linear combinations of predictors are used in place of the original predictors. However, two models will be equivalent in the sense of giving identical fitted values when p linearly independent combinations of p predictors are used. The

fitted model can be expressed in terms of the original predictors, or in terms of the linear combinations. Similarly, standardizing the original predictors does not change the significance of individual regression effects nor does it change the interpretation of the model.

A reasonable strategy with the Berkeley data might be to find principal components for the three height variables separately, the three weights separately, and so on. It may not make sense to combine the four different types of measures (HT, WT, ST, and LG) together because the resulting linear combinations would likely be uninterpretable.

Output from a PCA on the four sets of standardized measures is given below. Two sets (ST9, ST18) and (LG9, LG18) have two predictors. The PCs on the strength and leg measures are essentially the sum and difference between the two standardized features. The loadings have magnitude $0.707 = (1/2)^{1/2}$ to satisfy the unit-length restriction.

Here are some comments on how I might use the PCA to build a regression model to predict somatotype. First, I would not necessarily use the given PCS, but would instead use interpretable linear combinations that were nearly identical to the PCs. For example, the first principal component of the heights is roughly an unweighted average of the weights at ages 2, 9, and 18. I would use (WT2+WT9+WT18)/3 instead. The overall sum of squared loadings is not important in the regression analysis. Only the relative sizes are important. Following this idea, what linear combinations of the heights are reasonable?

Feel free to explore these ideas at your leisure. You can get a copy of the data file from me via e-mail.

Program editor contents:

```
* perform PCA on correlation matrix
*   output observations with eigenvectors coordinates;
proc princomp data=bgs out=prinwt;
  var WT2 WT9 WT18;
run;
proc princomp data=bgs out=prinht;
  var HT2 HT9 HT18;
run;
proc princomp data=bgs out=prinlg;
  var LG9 LG18;
run;
proc princomp data=bgs out=prinst;
  var ST9 ST18;
run;
```

Output window contents:
PCA on WT2, WT9, and WT18:

```
The PRINCOMP Procedure
Observations          26
Variables              3

           Correlation Matrix
             WT2         WT9        WT18
WT2       1.0000      0.5792      0.2159
WT9       0.5792      1.0000      0.7089
WT18      0.2159      0.7089      1.0000

          Eigenvalues of the Correlation Matrix
        Eigenvalue    Difference    Proportion    Cumulative
   1    2.02813956    1.23915935        0.6760        0.6760
   2    0.78898020    0.60609996        0.2630        0.9390
   3    0.18288024                      0.0610        1.0000

                Eigenvectors
              Prin1         Prin2         Prin3
WT2       0.492458      0.780915      0.384262
WT9       0.664767      -.052525      -.745202
WT18      0.561756      -.622425      0.544992
```

PCA on HT2, HT9, and HT18:

```
           Correlation Matrix
             HT2         HT9        HT18
HT2       1.0000      0.7758      0.6847
HT9       0.7758      1.0000      0.8645
HT18      0.6847      0.8645      1.0000

          Eigenvalues of the Correlation Matrix
        Eigenvalue    Difference    Proportion    Cumulative
   1    2.55232277    2.22472072        0.8508        0.8508
   2    0.32760205    0.20752687        0.1092        0.9600
   3    0.12007518                      0.0400        1.0000

                Eigenvectors
              Prin1         Prin2         Prin3
HT2       0.554278      0.799587      0.231163
HT9       0.598903      -.190273      -.777889
HT18      0.578006      -.569611      0.584339
```

PCA on LG9 and LG18:

```
     Correlation Matrix
             LG9        LG18
LG9       1.0000      0.6645
LG18      0.6645      1.0000

          Eigenvalues of the Correlation Matrix
        Eigenvalue    Difference    Proportion    Cumulative
   1    1.66447530    1.32895060        0.8322        0.8322
   2    0.33552470                      0.1678        1.0000

           Eigenvectors
              Prin1         Prin2
LG9       0.707107      0.707107
LG18      0.707107      -.707107
```

PCA on ST9 and ST18:

Correlation Matrix

	ST9	ST18
ST9	1.0000	0.6596
ST18	0.6596	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	1.65959471	1.31918941	0.8298	0.8298
2	0.34040529		0.1702	1.0000

Eigenvectors

	Prin1	Prin2
ST9	0.707107	0.707107
ST18	0.707107	-.707107

Chapter 14

Cluster Analysis

Cluster analysis is an exploratory tool for locating and grouping observations that are similar to each other across features. Cluster analysis can also be used to group variables that are similar across observations.

Clustering or grouping is distinct from discriminant analysis and classification. In discrimination problems there are a given number of known groups to compare or distinguish. The aim in cluster analysis is to define groups based on similarities. The clusters are then examined for underlying characteristics that might help explain the grouping.

There are a variety of clustering algorithms. I will discuss a simple **(agglomerative) hierarchical clustering method** for grouping observations. The method begins with each observation as an individual cluster or group. The two most similar observations are then grouped, giving one cluster with two observations. The remaining clusters have one observation. The clusters are then joined sequentially until one cluster is left.

To illustrate the steps, suppose eight observations are collected on two features X_1 and X_2. A plot of the data is given below.

Step 1. Each observation is a cluster.

Step 2. Form a new cluster by grouping the two clusters that are most similar, or closest to each other. This leaves seven clusters.

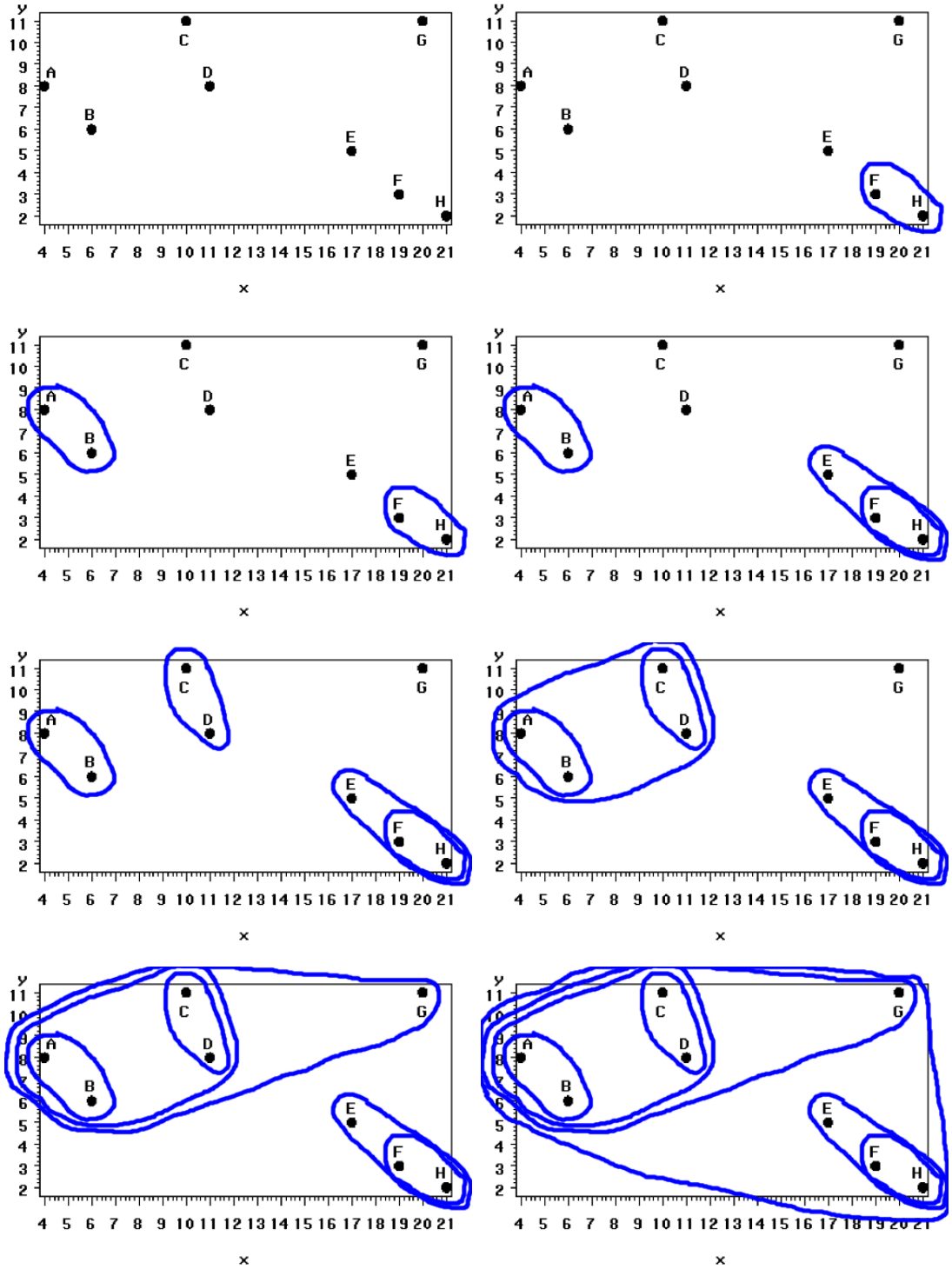
Step 3. Form a new cluster by grouping the two clusters that are most similar, or closest to each other. This leaves six clusters.

Step 4–7. Continue the process of merging clusters one at a time.

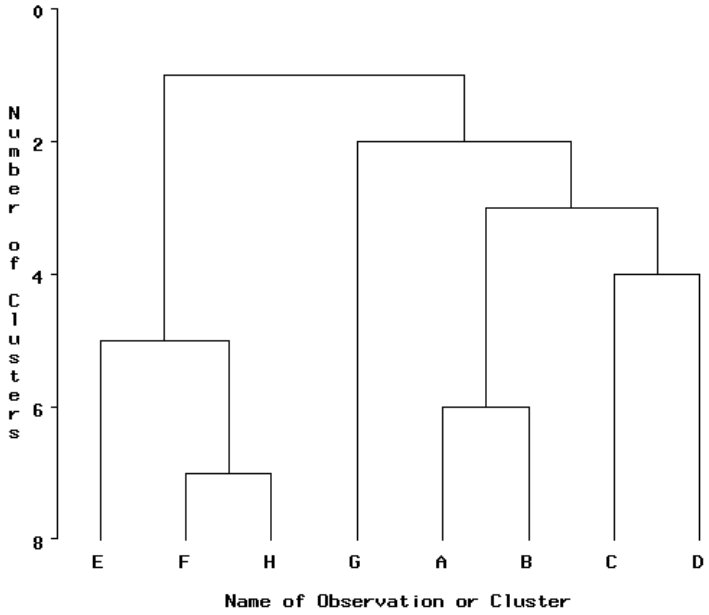
Step 8. Merge (fuse or combine) the remaining two clusters.

Finally Use a tree or dendrogram to summarize the steps in the cluster formation.

Clustering steps 1–8

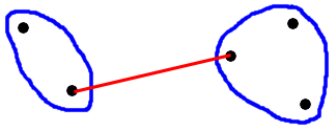


This order of clustering is summarized in this dendrogram.

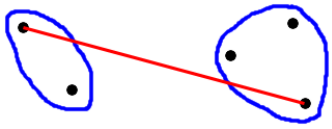


There are several accepted measures of distance between clusters. The **single linkage** distance is the minimum distance between points across two clusters. The **complete linkage** distance is the maximum distance between points across two clusters. The **average linkage** distance is the average distance between points across two clusters. In these three cases the distance between points is the Euclidean or "ruler" distance. The pictures below illustrate the measures.

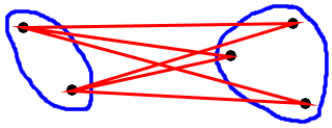
Given a distance measure, the distance between each pair of clusters is evaluated at each step. The two clusters that are closest to each other are merged. The observations are usually standardized prior to clustering to eliminate the effect of different variability on the distance measure.



Single uses the length of the **shortest** line between points in clusters



Complete uses the length of the **longest** line between points in clusters



Average uses the **average** length of all line between points in clusters

Different distance measures can produce different **shape** clusters. Average linkage tends to produce clusters with similar variability. Complete linkage is biased towards producing clusters with roughly equal diameters. Single linkage has the ability to produce and detect elongated and irregular clusters. You should try different distances to decide the most sensible measure for your problem.

14.1 Example: Mammal teeth

The table below gives the numbers of different types of teeth for 32 mammals. The columns, from left to right, give the numbers of (v1) top incisors, (v2) bottom incisors, (v3) top canines, (v4) bottom canines, (v5) top premolars, (v6) bottom premolars, (v7) top molars, (v8) bottom molars, respectively. A cluster analysis will be used to identify the mammals that have similar counts across the eight types of teeth.

Mammal	Numbers of Teeth
Brown Bat	2 3 1 1 3 3 3 3
Mole	3 2 1 0 3 3 3 3
Silver Hair Bat	2 3 1 1 2 3 3 3
Pigmy Bat	2 3 1 1 2 2 3 3
House Bat	2 3 1 1 1 2 3 3
Red Bat	1 3 1 1 2 2 3 3
Pika	2 1 0 0 2 2 3 3
Rabbit	2 1 0 0 3 2 3 3
Beaver	1 1 0 0 2 1 3 3
Groundhog	1 1 0 0 2 1 3 3
Gray Squirrel	1 1 0 0 1 1 3 3
House Mouse	1 1 0 0 0 0 3 3
Porcupine	1 1 0 0 1 1 3 3
Wolf	3 3 1 1 4 4 2 3
Bear	3 3 1 1 4 4 2 3
Raccoon	3 3 1 1 4 4 3 2
Marten	3 3 1 1 4 4 1 2
Weasel	3 3 1 1 3 3 1 2
Wolverine	3 3 1 1 4 4 1 2
Badger	3 3 1 1 3 3 1 2
River Otter	3 3 1 1 4 3 1 2
Sea Otter	3 2 1 1 3 3 1 2
Jaguar	3 3 1 1 3 2 1 1
Cougar	3 3 1 1 3 2 1 1
Fur Seal	3 2 1 1 4 4 1 1
Sea Lion	3 2 1 1 4 4 1 1
Grey Seal	3 2 1 1 3 3 2 2
Elephant Seal	2 1 1 1 4 4 1 1
Reindeer	0 4 1 0 3 3 3 3
Elk	0 4 1 0 3 3 3 3
Deer	0 4 0 0 3 3 3 3
Moose	0 4 0 0 3 3 3 3

The **SAS** program given below produces cluster analysis summaries for the mammal teeth data. The `method=average` option on the `cluster` procedure statement specifies average linkage. Other popular choices are `method=` complete (for complete linkage), single (for single linkage), centroid, density (a method that uses non-parametric density estimation), and ward (for Ward's two-stage density linkage). The `std` option standardizes the data prior to clustering. The `pseudo` option prints pseudo F and t-statistics for selecting the number of clusters. The `ccc` option prints the cubic clustering criterion and related information for selecting the number of clusters. The `noeigen` option suppresses default output on eigenvalues. The `id` statement labels each case by the value of **mammal**.

The `cluster` procedure **always** produces an output data set to be used as input to the `tree` procedure, which creates dendrograms. You should name this data set to keep straight that it is the default input for later procedures until another data set is created. The `outtree=treedat` option names the data set `treedat`. The `sort` and `height=h` options on `tree` produce a dendrogram that is plotted by the RMS distance (root mean squared distance) at which the clusters were joined. With the `height=ncl` option, the dendrogram is plotted using the number of clusters present at the time of merger.

The first output table gives a chronology of the cluster formation. The first two clusters joined are Beaver and Groundhog. Many sets of mammals have identical responses on all eight teeth. There is no unique method to determine the order that the clusters will merge here in the first few steps.

Program editor contents:

```
* Mammal teeth example;
options ls=179 nodate nocenter;

data teeth;
  infile "F:\Dropbox\UNM\teach\ADA2_stat528\sas\ADA2_14_teeth.dat";
  input mammal $ v1-v8;                  *$;
  label v1='top incisors'  v2='bottom incisors'
        v3='top canines'   v4='bottom canines'
        v5='top premolars' v6='bottom premolars'
        v7='top molars'    v8='bottom molars';
run;

* cluster analysis;
*   method=average, =complete, =single, centroid, density, and ward are options;
*   std standardizes the data prior to clustering                          ;
*   pseudo prints pseudo F and t-stats to select number of clusters        ;
*   ccc prints cubic clustering criterion to select number of clusters     ;
*   noeigen suppresses default output on eigenvalues                       ;
*   id labels each case by the value of mammal                             ;
*   outtree=treedat outputs dataset to plot with proc tree                 ;
proc cluster data=teeth method=average std pseudo ccc noeigen outtree=treedat;
  id mammal;
  var v1-v8;
run;

* plot cluster analysis dendrogram or tree by average distance between clusters;
proc tree data=treedat sort height=h;
run;

* plot cluster analysis dendrogram or tree by number of clusters;
proc tree data=treedat sort height=ncl;
run;
```

Output window contents:

The CLUSTER Procedure
Average Linkage Cluster Analysis

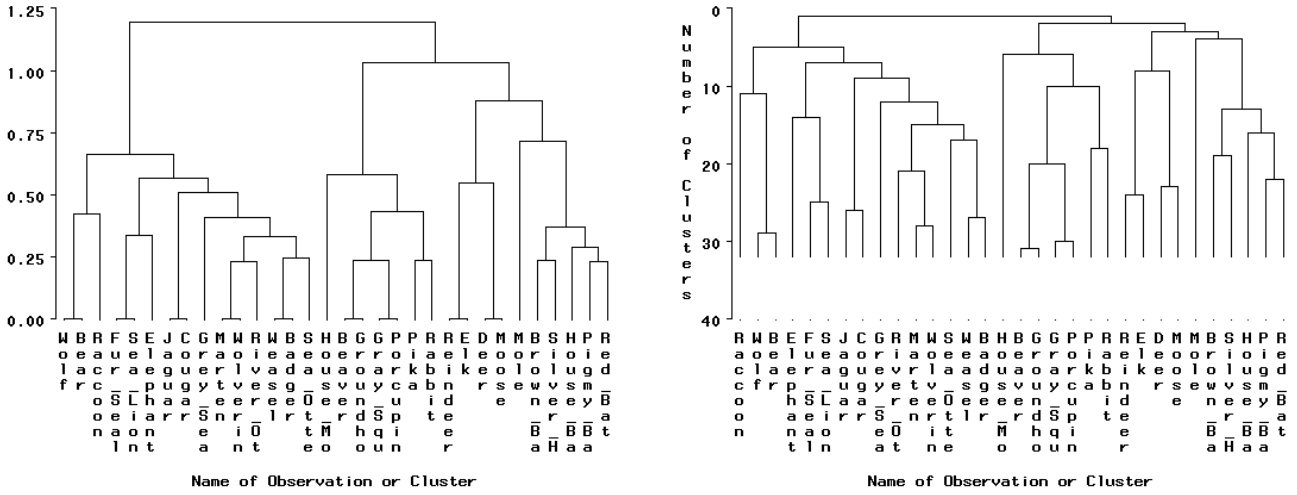
The data have been standardized to mean 0 and variance 1
Root-Mean-Square Total-Sample Standard Deviation = 1
Root-Mean-Square Distance Between Observations = 4

Cluster History

NCL	--Clusters Joined---		FREQ	SPRSQ	RSQ	ERSQ	CCC	PSF	PST2	Norm RMS Dist	Tie	
31	Beaver	Groundho	2	0.0000	1.00	.	.	.	.	0	T	
30	Gray_Squ	Porcupin	2	0.0000	1.00	.	.	.	.	0	T	
29	Wolf	Bear	2	0.0000	1.00	.	.	.	.	0	T	
28	Marten	Wolverin	2	0.0000	1.00	.	.	.	.	0	T	
27	Weasel	Badger	2	0.0000	1.00	.	.	.	.	0	T	
26	Jaguar	Cougar	2	0.0000	1.00	.	.	.	.	0	T	
25	Fur_Seal	Sea_Lion	2	0.0000	1.00	.	.	.	.	0	T	
24	Reindeer	Elk	2	0.0000	1.00	.	.	.	.	0	T	
23	Deer	Moose	2	0.0000	1.00	.	.	.	.	0		
22	Pigmy_Ba	Red_Bat	2	0.0017	.998	.	.	281	.	0.2289		
21	CL28	River_Ot	3	0.0023	.996	.	.	139	.	0.2292		
20	CL31	CL30	4	0.0036	.992	.	.	83.2	.	0.2357	T	
19	Brown_Ba	Silver_H	2	0.0018	.991	.	.	76.7	.	0.2357	T	
18	Pika	Rabbit	2	0.0018	.989	.	.	73.2	.	0.2357		
17	CL27	Sea_Otte	3	0.0026	.986	.	.	67.4	.	0.2462		
16	CL22	House_Ba	3	0.0030	.983	.	.	62.9	1.7	0.2859		
15	CL21	CL17	6	0.0083	.975	.	.	47.4	6.8	0.3328		
14	CL25	Elephant	3	0.0049	.970	.	.	45.0	.	0.3362		
13	CL19	CL16	5	0.0075	.963	.	.	40.8	3.5	0.3672		
12	CL15	Grey_Sea	7	0.0073	.955	.	.	38.9	2.8	0.4078		
11	CL29	Raccoon	3	0.0077	.948	.	.	38.0	.	0.423		
10	CL18	CL20	6	0.0138	.934	.	.	34.5	10.3	0.4339		
9	CL12	CL26	9	0.0213	.913	.	.	30.0	7.3	0.5071		
8	CL24	CL23	4	0.0193	.893	.	.	28.7	.	0.5473		
7	CL9	CL14	12	0.0326	.861	.	.	25.7	7.0	0.5668		
6	CL10	House_Mo	7	0.0158	.845	.516	22.4	28.3	4.1	0.5792		*
5	CL11	CL7	15	0.0459	.799	.446	21.3	26.8	6.9	0.6621		
4	CL13	Mole	6	0.0252	.774	.365	23.9	31.9	7.2	0.7156		*
3	CL4	CL8	10	0.0926	.681	.270	22.5	31.0	12.7	0.8799		
2	CL3	CL6	17	0.1999	.481	.157	17.8	27.8	16.1	1.0316		
1	CL2	CL5	32	0.4813	.000	.000	0.00	.	27.8	1.1938		

COMMENT: CCC is cubic clustering criterion,
PSF is pseudo-F, and
PST2 is pseudo-t (squared).

Dendrograms (trees) of **average distance between clusters** and **number of clusters**.



14.2 Identifying the Number of Clusters

Cluster analysis can be used to produce an "optimal" splitting of the data into a prespecified number of groups or clusters, with different algorithms usually giving different clusters. However, the important issue in many analyses revolves around identifying the number of clusters in the data. A simple empirical method is to continue grouping until the clusters being fused are relatively dissimilar, as measured by the normalized RMS between clusters. Experience with your data is needed to provide a reasonable stopping rule.

The **SAS** manual discusses several statistical methods for selecting the number of clusters. No method is best. They suggest using the cubic clustering criteria (`ccc`), a pseudo-F statistic, and a pseudo-t statistic. At a given step, the pseudo-t statistic is the distance between the center of the two clusters to be merged, relative to the variability within these clusters. A large pseudo-t statistic implies that the clusters to be joined are relatively dissimilar (i.e., much more variability between the clusters to be merged than within these clusters). The pseudo-F statistic at a given step measures the variability among the centers of the current clusters relative to the variability within the clusters. A large pseudo-F value implies that the clusters merged consist of fairly similar observations. As clusters are joined, the pseudo-t statistic tends to increase, and the pseudo-F statistic tends to decrease. The ccc is more difficult to describe.

The RSQ summary is also useful for determining the number of clusters. RSQ is a pseudo-R^2 statistic that measures the proportion of the total variation explained by the differences among the existing clusters at a given step. RSQ will typically decrease as the pseudo-F statistic decreases.

A common recommendation on cluster selection is to choose a cluster size where the values of ccc and the pseudo-F statistic are relatively high (compared to what you observe with other numbers of clusters), and where the pseudo-t statistic is relatively low and increases substantially at the next proposed merger. For the mammal teeth data this corresponds to four clusters. Six clusters is a sensible second choice.

The following **SAS** procedures were added to the data step to summarize the composition of the four clusters. The `noprint` option on the `tree` procedure suppresses the printing of output. The `out=treepart` option plus the `copy` command create an output data set that includes the

original data set plus a label that identifies the clusters. Four clusters are requested using the `nclusters=4` option. The final two procedures are self-explanatory.

Program editor contents:

```
proc tree data=treedat noprint out=treepart nclusters=4;
  copy mammal v1-v8;
run;

proc sort data=treepart;
  by cluster;
run;

proc print;
  id mammal;
  var v1-v8;
  by cluster;
run;
```

Output window contents:

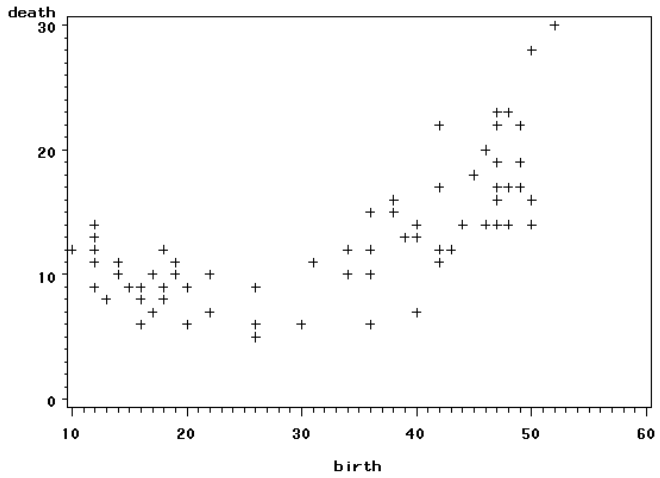
```
CLUSTER=1

mammal       v1     v2     v3     v4     v5     v6     v7     v8
Beaver        1      1      0      0      2      1      3      3
Groundho      1      1      0      0      2      1      3      3
Gray_Squ      1      1      0      0      1      1      3      3
Porcupin      1      1      0      0      1      1      3      3
Pika          2      1      0      0      2      2      3      3
Rabbit        2      1      0      0      3      2      3      3
House_Mo      1      1      0      0      0      0      3      3

CLUSTER=2

mammal       v1     v2     v3     v4     v5     v6     v7     v8
Wolf          3      3      1      1      4      4      2      3
Bear          3      3      1      1      4      4      2      3
Marten        3      3      1      1      4      4      1      2
Wolverin      3      3      1      1      4      4      1      2
Weasel        3      3      1      1      3      3      1      2
Badger        3      3      1      1      3      3      1      2
Jaguar        3      3      1      1      3      2      1      1
Cougar        3      3      1      1      3      2      1      1
Fur_Seal      3      2      1      1      4      4      1      1
Sea_Lion      3      2      1      1      4      4      1      1
River_Ot      3      3      1      1      4      3      1      2
Sea_Otte      3      2      1      1      3      3      1      2
Elephant      2      1      1      1      4      4      1      1
Grey_Sea      3      2      1      1      3      3      2      2
Raccoon       3      3      1      1      4      4      3      2


CLUSTER=3

mammal       v1     v2     v3     v4     v5     v6     v7     v8
Reindeer      0      4      1      0      3      3      3      3
Elk           0      4      1      0      3      3      3      3
Deer          0      4      0      0      3      3      3      3
Moose         0      4      0      0      3      3      3      3

CLUSTER=4

mammal       v1     v2     v3     v4     v5     v6     v7     v8
Pigmy_Ba      2      3      1      1      2      2      3      3
Red_Bat       1      3      1      1      2      2      3      3
Brown_Ba      2      3      1      1      3      3      3      3
Silver_H      2      3      1      1      2      3      3      3
House_Ba      2      3      1      1      1      2      3      3
Mole          3      2      1      0      3      3      3      3
```

14.3 Example: 1976 birth and death rates

The **SAS** program below contains data on the 1976 crude birth and death rates in 74 countries. A data plot and output from a complete and single linkage cluster analyses are given. The `print=16` option prints summary information on the last 16 clusters formed.

Program editor contents:

```
* Example: 1976 birth and death rates;

options ls=79 nodate nocenter;
data birthdeath;
  input country $ birth death @@;   *$;
  datalines;
  afghan    52 30   algeria    50 16   angola    47 23
  argentina 22 10   australia  16  8   austria   12 13
  banglades 47 19   belguim    12 12   brazil    36 10
  bulgaria  17 10   burma      38 15   cameroon  42 22
  canada    17  7   chile      22  7   china     31 11
  taiwan    26  5   columbia   34 10   cuba      20  6
  czechosla 19 11   ecuador    42 11   egypt     39 13
  ethiopia  48 23   france     14 11   german_dr 12 14
  german_fr 10 12   ghana      46 14   greece    16  9
  guatamala 40 14   hungary    18 12   india     36 15
  indonesia 38 16   iran       42 12   iraq      48 14
  italy     14 10   ivory_cst  48 23   japan     16  6
  kenya     50 14   nkorea     43 12   skorea    26  6
  madagasca 47 22   malaysia   30  6   mexico    40  7
  morocco   47 16   mozambique 45 18   nepal     46 20
  netherlan 13  8   nigeria    49 22   pakistan  44 14
  peru      40 13   phillip    34 10   poland    20  9
  portugal  19 10   rhodesia   48 14   romania   19 10
  saudi_ar  49 19   sth_africa 36 12   spain     18  8
  sri_lanka 26  9   sudan      49 17   sweden    12 11
  switzer   12  9   syria      47 14   tanzania  47 17
  thailand  34 10   turkey     34 12   ussr      18  9
  uganda    48 17   uk         12 12   usa       15  9
  upp_volta 50 28   venez      36  6   vietnam   42 17
  yugoslav  18  8   zaire      45 18
;
run;

proc gplot data=birthdeath;
  plot death*birth;
run;

* complete linkage;
proc cluster data=birthdeath method=complete std pseudo ccc noeigen print=16;
  id country;
  var birth death;
run;

* single linkage;
proc cluster data=birthdeath method=single   std pseudo ccc noeigen print=16;
  id country;
  var birth death;
run;
```



Output window contents:

```
The CLUSTER Procedure
Complete Linkage Cluster Analysis

The data have been standardized to mean 0 and variance 1
Root-Mean-Square Total-Sample Standard Deviation =         1
Mean Distance Between Observations                = 1.713213

                                Cluster History

                                                                                  Norm   T
                                                                                   Max   i
NCL  --Clusters Joined---  FREQ   SPRSQ   RSQ  ERSQ   CCC   PSF  PST2    Dist   e
 16  CL36         CL41         6  0.0018  .983  .      .     218  15.8  0.3406   T
 15  CL31         CL33         8  0.0017  .981  .      .     216  11.2  0.3554
 14  CL25         CL19        10  0.0029  .978  .953  6.93   205   8.6  0.4479   T
 13  CL18         CL30         9  0.0049  .973  .947  6.14   184  19.0  0.4479
 12  CL43         CL22         9  0.0043  .969  .941  5.87   175  32.4  0.4723
 11  CL17         CL20        16  0.0047  .964  .934  5.75   169  12.8  0.4762
 10  CL28         CL27         5  0.0048  .959  .924  5.85   168  15.0  0.5752
  9  CL16         CL12        15  0.0075  .952  .913  5.64   161  13.1  0.6015
  8  CL23         CL29         9  0.0075  .944  .899  5.80   160  22.6  0.6652
  7  CL15         CL9         23  0.0169  .927  .881  4.95   143  20.2  0.7801   T
  6  CL10         CL11        21  0.0116  .916  .856  5.51   148  14.2  0.8288
  5  CL6          CL14        31  0.0315  .884  .822  4.67   132  28.0  1.1628
  4  CL24         CL13        11  0.0236  .861  .770  5.86   144  29.4   1.228
  3  CL7          CL8         32  0.0592  .801  .684  6.29   143  40.0  1.3733
  2  CL3          CL5         63  0.3934  .408  .393  0.47  49.6   143  1.8456
  1  CL4          CL2         74  0.4080  .000  .000  0.00     .  49.6   3.046
```

```
The CLUSTER Procedure
Single Linkage Cluster Analysis

The data have been standardized to mean 0 and variance 1
Root-Mean-Square Total-Sample Standard Deviation =        1
Mean Distance Between Observations                = 1.713213

                                Cluster History
                                                                                    Norm  T
                                                                                     Min  i
NCL  --Clusters Joined---  FREQ   SPRSQ   RSQ  ERSQ   CCC   PSF  PST2     Dist   e
 16  CL57        china        5  0.0005  .948  .      .     70.9  15.2  0.1678  T
 15  CL16        CL48         7  0.0012  .947  .      .     75.5   8.3  0.1678  T
 14  CL18        vietnam     11  0.0010  .946  .953  -1.2  81.0   1.9  0.1678  T
 13  CL15        CL20        16  0.0116  .934  .947  -2.0  72.5  18.4  0.1678
 12  CL13        CL53        22  0.0184  .916  .941  -3.3  61.5  17.4   0.169  T
 11  CL35        malaysia     3  0.0004  .916  .934  -2.2  68.3   3.5   0.169
 10  CL17        CL11        30  0.0229  .893  .924  -3.3  59.2  16.6  0.2016  T
  9  mexico      venez        2  0.0004  .892  .913  -2.1  67.3    .    0.2016  T
  8  CL10        sri_lank    31  0.0027  .890  .899  -.87  76.0   1.3  0.2016  T
  7  CL14        CL12        33  0.0552  .834  .881  -3.3  56.3  37.6  0.2016
  6  CL34        cameroon     6  0.0011  .833  .856  -1.5  68.0  11.0  0.2112
  5  CL7         CL6         39  0.0855  .748  .822  -3.8  51.2  31.0   0.224
  4  afghan      upp_volt     2  0.0006  .747  .770  -1.1  69.0    .    0.2356
  3  CL8         CL9         33  0.0319  .715  .684  1.42  89.2  15.3  0.2534
  2  CL5         CL3         72  0.5629  .152  .393  -6.2  13.0   139  0.3049
  1  CL4         CL2         74  0.1525  .000  .000  0.00    .   13.0  0.5563
```

The two methods suggest three clusters. Complete linkage also suggests 14 clusters, but the clusters were unappealing so this analysis will not be presented here.

The following **SAS** statements provide output on the three clusters selected by complete linkage. To get the three cluster output from single linkage replace `method=complete` with `method=single`. The three clusters generated by the two methods are very different. The same tendency was observed using average linkage and Ward's method.

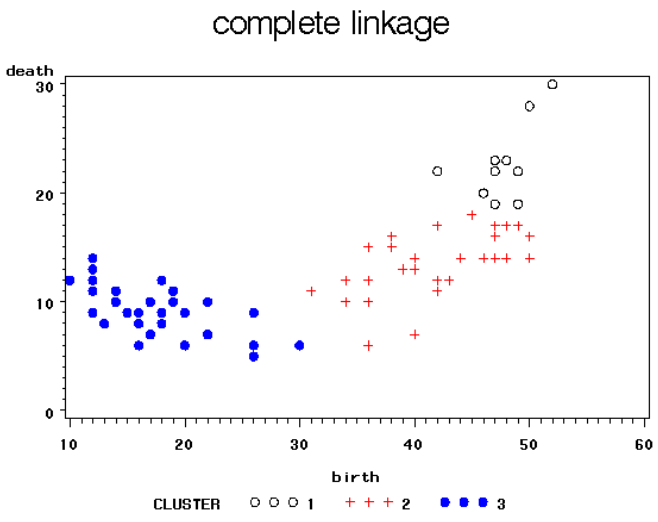
An important point to recognize is that different clustering algorithms may agree on the number of clusters, but they may not agree on the composition of the clusters.

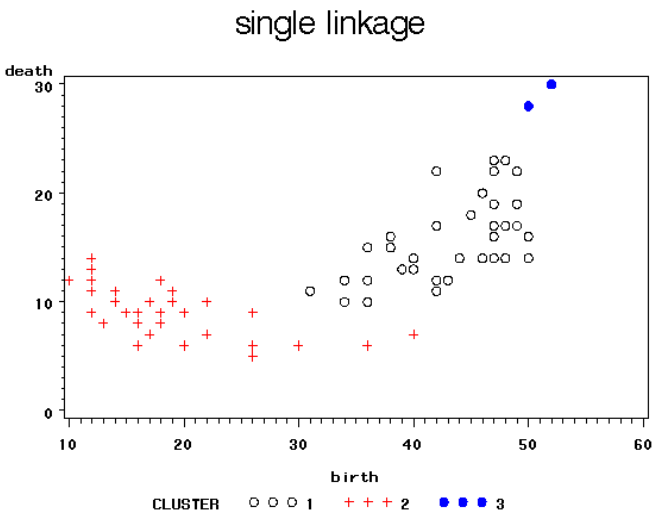
Program editor contents:

```
title 'complete linkage';
proc cluster data=birthdeath method=complete std noeigen out=tree print=0;
  id country;
  var birth death;
run;

proc tree data=tree out=out nclusters=3 noprint;
  id country;
  copy birth death;
run;

proc sort data=out;
  by cluster;
run;

proc print data=out;
  id country;
  var birth death;
  by cluster;
run;

symbol1 v=circle c=black i=none;
symbol2 v=plus   c=red   i=none;
symbol3 v=dot    c=blue  i=none;
proc gplot data=out;
  plot death*birth=cluster;
run;
```

Program editor contents:

```
title 'single linkage';
proc cluster data=birthdeath method=single std noeigen out=tree print=0;
  id country;
  var birth death;
run;

proc tree data=tree out=out nclusters=3 noprint;
  id country;
  copy birth death;
run;

proc sort data=out;
  by cluster;
run;

proc print data=out;
  id country;
  var birth death;
  by cluster;
run;

symbol1 v=circle c=black i=none;
symbol2 v=plus   c=red   i=none;
symbol3 v=dot    c=blue  i=none;
proc gplot data=out;
  plot death*birth=cluster;
run;
```





Output window contents:

```
The CLUSTER Procedure
Complete Linkage Cluster Analysis

The data have been standardized to mean 0 and variance 1
Root-Mean-Square Total-Sample Standard Deviation =        1
Mean Distance Between Observations                = 1.713213

CLUSTER=1

country     birth    death
ethiopia      48       23
ivory_cs      48       23
angola        47       23
madagasc      47       22
nigeria       49       22
banglade      47       19
saudi_ar      49       19
nepal         46       20
afghan        52       30
upp_volt      50       28
cameroon      42       22

CLUSTER=2

country     birth    death
columbia      34       10
phillip       34       10
iraq          48       14
rhodesia      48       14
thailand      34       10
mozambiq      45       18
zaire         45       18
iran          42       12
nkorea        43       12
syria         47       14
sudan         49       17
uganda        48       17
egypt         39       13
peru          40       13
brazil        36       10
ghana         46       14
pakistan      44       14
tanzania      47       17
sth_afri      36       12
turkey        34       12
burma         38       15
india         36       15
ecuador       42       11
guatamal      40       14
kenya         50       14
algeria       50       16
morocco       47       16
indonesi      38       16
vietnam       42       17
mexico        40        7
venez         36        6
china         31       11
```

```
CLUSTER=3

country     birth    death
portugal      19       10
romania       19       10
belguim       12       12
uk            12       12
spain         18        8
yugoslav      18        8
greece        16        9
usa           15        9
german_f      10       12
bulgaria      17       10
australi      16        8
france        14       11
sweden        12       11
poland        20        9
ussr          18        9
austria       12       13
german_d      12       14
taiwan        26        5
skorea        26        6
canada        17        7
japan         16        6
czechosl      19       11
hungary       18       12
netherla      13        8
switzer       12        9
chile         22        7
cuba          20        6
italy         14       10
argentin      22       10
sri_lank      26        9
malaysia      30        6
```

Output window contents:

```
The CLUSTER Procedure
Single Linkage Cluster Analysis

The data have been standardized to mean 0 and variance 1
Root-Mean-Square Total-Sample Standard Deviation =        1
Mean Distance Between Observations                = 1.713213

CLUSTER=1

country     birth    death
ethiopia      48       23
ivory_cs      48       23
columbia      34       10
phillip       34       10
iraq          48       14
rhodesia      48       14
thailand      34       10
mozambiq      45       18
zaire         45       18
angola        47       23
iran          42       12
nkorea        43       12
syria         47       14
sudan         49       17
uganda        48       17
tanzania      47       17
egypt         39       13
peru          40       13
ghana         46       14
brazil        36       10
kenya         50       14
madagasc      47       22
nigeria       49       22
pakistan      44       14
banglade      47       19
saudi_ar      49       19
sth_afri      36       12
turkey        34       12
burma         38       15
india         36       15
indonesi      38       16
ecuador       42       11
guatamal      40       14
morocco       47       16
algeria       50       16
nepal         46       20
china         31       11
vietnam       42       17
cameroon      42       22
```

CLUSTER=2

country	birth	death
portugal	19	10
romania	19	10
belguim	12	12
uk	12	12
spain	18	8
yugoslav	18	8
greece	16	9
usa	15	9
german_f	10	12
bulgaria	17	10
australi	16	8
france	14	11
sweden	12	11
ussr	18	9
poland	20	9
czechosl	19	11
italy	14	10
austria	12	13
german_d	12	14
taiwan	26	5
skorea	26	6
canada	17	7
japan	16	6
hungary	18	12
netherla	13	8
switzer	12	9
argentin	22	10
chile	22	7
cuba	20	6
malaysia	30	6
mexico	40	7
venez	36	6
sri_lank	26	9

CLUSTER=3

country	birth	death
afghan	52	30
upp_volt	50	28

Chapter 15

Multivariate Analysis of Variance

Jolicouer and Mosimann studied the relationship between the size and shape of painted turtles. The table below gives the length, width, and height (all in mm) for 24 males and 24 females.

Sex	Length	Width	Height	Sex	Length	Width	Height
f	98	81	38	m	93	74	37
f	103	84	38	m	94	78	35
f	103	86	42	m	96	80	35
f	105	86	42	m	101	84	39
f	109	88	44	m	102	85	38
f	123	92	50	m	103	81	37
f	123	95	46	m	104	83	39
f	133	99	51	m	106	83	39
f	133	102	51	m	107	82	38
f	133	102	51	m	112	89	40
f	134	100	48	m	113	88	40
f	136	102	49	m	114	86	40
f	138	98	51	m	116	90	43
f	138	99	51	m	117	90	41
f	141	105	53	m	117	91	41
f	147	108	57	m	119	93	41
f	149	107	55	m	120	89	40
f	153	107	56	m	121	93	44
f	155	115	63	m	121	95	42
f	155	117	60	m	125	93	45
f	158	115	62	m	127	96	45
f	159	118	63	m	128	95	45
f	162	124	61	m	131	95	46
f	177	132	67	m	135	106	47

MANOVA considers the following two questions. Are the population mean length, width, and height the same for males and females? If not, then what combination of features is most responsible for the differences?

To describe MANOVA, suppose you measure p features on independent random samples from k strata, groups, or populations. Let

$$\mu_i' = \begin{bmatrix} \mu_{i1} & \mu_{i2} & \cdots & \mu_{ip} \end{bmatrix}'$$

be the vector of population means for the i^{th} population, where μ_{ij} is the i^{th} population mean on the j^{th} feature. For the turtles, $p = 3$ and $k = 2$.

A one-way MANOVA tests the hypothesis that the population mean vectors are identical: $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ against H_A : not H_0. For the carapace data, you are simultaneously testing that the sexes have equal population mean lengths, equal population mean widths, and equal population mean heights.

Assume that the sample sizes from the different groups are $n_1, n_2, \ldots, n_k$. The total sample size is $n = n_1 + n_2 + \cdots + n_k$. Let

$$X_{ij}' = \begin{bmatrix} X_{ij1} & X_{ij2} & \cdots & X_{ijp} \end{bmatrix}'$$

be the vector of responses for the j^{th} individual from the i^{th} sample. Let

$$\overline{X}_i' = \begin{bmatrix} \overline{X}_{i1} & \overline{X}_{i2} & \cdots & \overline{X}_{ip} \end{bmatrix}'$$

and S_i be the mean vector and variance-covariance matrix for the i^{th} sample. Finally, let

$$\overline{X}' = \begin{bmatrix} \overline{X}_1 & \overline{X}_2 & \cdots & \overline{X}_p \end{bmatrix}'$$

be the vector of means ignoring samples (combine all the data across samples and compute the average on each feature), and let

$$S = \frac{\sum_i (n_i - 1) S_i}{n - k}$$

be the pooled variance-covariance matrix. The pooled variance-covariance matrix is a weighted average of the variance-covariance matrices from each group.

To test H_0, construct the following MANOVA table, which is the multivariate analog of the ANOVA table:

Source	*df*	SS	MS
Between	$k-1$	$\sum_i n_i(\overline{X}_i - \overline{X})(\overline{X}_i - \overline{X})'$	
Within	$n-k$	$\sum_i (n_i - 1) S_i$	
Total	$n-1$	$\sum_{ij}(X_{ij} - \overline{X})(X_{ij} - \overline{X})'$,	

where all the MSs are SS/df.

The expressions for the SS have the same form as SS in univariate analysis of variance, except that each SS is a $p \times p$ symmetric matrix. The diagonal elements of the **SS matrices** are the SS for one-way ANOVAs on the individual features. The off-diagonal elements are SS between features. The Error MS matrix is the pooled variance-covariance matrix S.

The standard MANOVA assumes that you have independent samples from multivariate normal populations with identical variance-covariance matrices. This implies that each feature is normally distributed in each population, that a feature has the same variability across populations, and that the correlation (or covariance) between two features is identical across populations. The Error MS

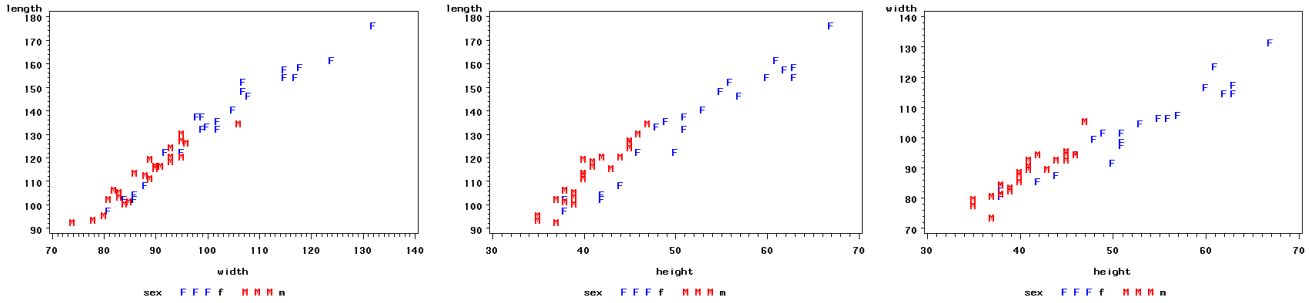
matrix estimates the common population variance-covariance matrix when the population variance-covariance matrices are identical.

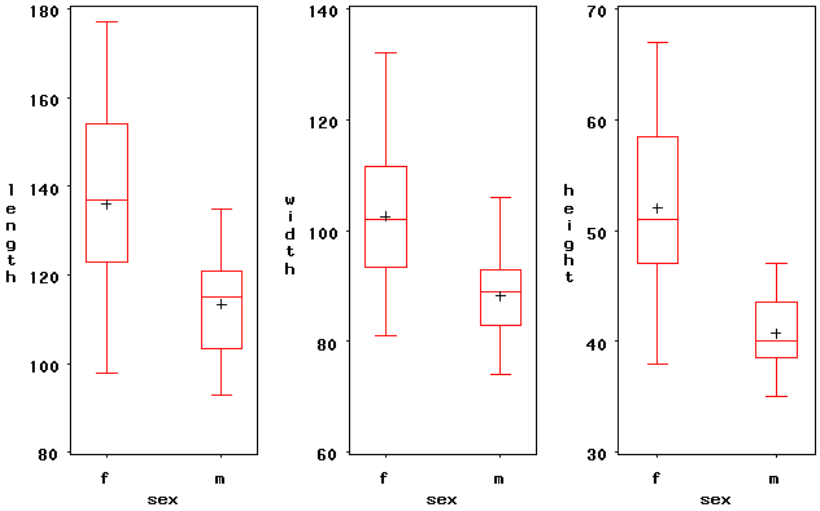
The H_0 of equal population mean vectors should be rejected when the difference among mean vectors, as measured by the Between MS matrix, is large relative to the variability within groups, as measured by the Error MS matrix. Equivalently, H_0 is implausible if a significant portion of the total variation in the data, as measured by the Total SS matrix, is due to differences among the groups. The same idea is used in a one-way ANOVA to motivate the F-test of no differences in population means. However, some care is needed to quantify these ideas in a MANOVA because there are several natural matrix definitions for comparing the Between MS matrix to the Error MS matrix. As a result, several MANOVA tests of H_0 have been proposed.

Graphical and numerical summaries for the carapace data are given below. The features are positively correlated within each sex. The correlations between pairs of features are similar for males and females. Females tend to be larger on each feature. The distributions for length, width and height are fairly symmetric within sexes. No outliers are present. Although females are more variable on each feature than males, the MANOVA assumptions do not appear to be grossly violated here.

Program editor contents:

```
options ls=79 nodate nocenter;

* example: carapace measurements for painted turtles;
data turtle;
  infile "F:\Dropbox\UNM\teach\ADA2_stat528\sas\ADA2_15_turtle.dat";
  input sex $ length width height @@;  *$;
run;

proc sort data=turtle;
  by sex;
run;

* define v=symbol, c=color, i=interpolation, l=line style;
symbol1 v="f" c=blue  i=none; * i=r l=1;
symbol2 v="m" c=red   i=none; * i=r l=2;
proc gplot data=turtle;
  plot length*(width height)=sex width*height=sex;
run;

symbol1;
symbol2;
proc boxplot data=turtle;
  plot length*sex /boxstyle=schematic;
  plot width *sex /boxstyle=schematic;
  plot height*sex /boxstyle=schematic;
run;

* generate corr and cov matrices by sex ;
proc corr data=turtle cov;
  var length width height;
  by sex;
run;

proc univariate data=turtle plot;
  var length width height;
  by sex;
run;
```





Output window contents:

```
sex=f

The CORR Procedure

   3  Variables:    length   width    height

                Covariance Matrix, DF = 23
                  length              width             height
length       451.5199275       270.9746377       165.9547101
width        270.9746377       171.7318841       101.8442029
height       165.9547101       101.8442029        64.7373188

                              Simple Statistics
Variable          N        Mean     Std Dev         Sum     Minimum     Maximum
length           24    136.04167    21.24900        3265    98.00000   177.00000
width            24    102.58333    13.10465        2462    81.00000   132.00000
height           24     52.04167     8.04595        1249    38.00000    67.00000

   Pearson Correlation Coefficients, N = 24
          Prob > |r| under H0: Rho=0

                 length            width           height
length        1.00000           0.97312          0.97067
                                 <.0001           <.0001
width         0.97312           1.00000          0.96590
               <.0001                             <.0001
height        0.97067           0.96590          1.00000
               <.0001            <.0001
```

```
sex=m

The CORR Procedure

   3  Variables:    length   width    height

                Covariance Matrix, DF = 23
                  length           width          height
length       139.3840580       79.3514493       37.5181159
width         79.3514493       50.0416667       21.6539855
height        37.5181159       21.6539855       11.2590580

                            Simple Statistics
Variable          N        Mean     Std Dev         Sum     Minimum     Maximum
length           24   113.41667    11.80610        2722    93.00000   135.00000
width            24    88.29167     7.07401        2119    74.00000   106.00000
height           24    40.70833     3.35545   977.00000    35.00000    47.00000

   Pearson Correlation Coefficients, N = 24
           Prob > |r| under H0: Rho=0

              length          width         height
length       1.00000        0.95013        0.94707
                             <.0001         <.0001
width        0.95013        1.00000        0.91226
              <.0001                        <.0001
height       0.94707        0.91226        1.00000
              <.0001         <.0001
```

The **glm** procedure in **SAS** can be used for one-way MANOVA. The `class` and `model` statements are similar to those used for a one-way ANOVA, except that all three features are identified as responses on the `model` statement. The `model` statement produces one-way ANOVAs for each feature separately. The `lsmeans` statement gives averages for each feature by sex, and p-values for multiple comparison on the features. The `manova h=sex` statement gives p-values for four MANOVA test statistics.

For the carapace data, the univariate ANOVAs and the `lsmeans` output indicate significant differences between sexes on length, width, and height. Females are larger on average than males on each feature. The p-values on the `lsmeans` output are not needed when comparing two groups because they are identical to the one-way ANOVA p-values.

The four MANOVA tests of no differences between sexes are significant at the 0.0001 level. These tests reinforce the univariate analyses. Of the four tests, I prefer Roy's test because it has an intuitive interpretation. I will mostly ignore the other three tests.

Roy's test locates the linear combination of the features that produces the most significant one-way ANOVA test for no differences among groups. If the groups are not significantly different on the linear combination that best separates the groups in a one-way ANOVA sense, then there is no evidence that the population mean vectors are different. The critical value for Roy's test accounts for the linear combination being suggested by the data. That is, the critical value for Roy's test is not the same critical value that is used in a one-way ANOVA. The idea is similar to a Bonferroni-type correction with multiple comparisons.

Roy's method has the ability to locate linear combinations of the features on which the groups differ, even when the differences across groups are not significant on any feature. This is a reason for treating multivariate problems using multivariate methods rather than through individual univariate analyses on each feature.

Program editor contents:

```
* MANOVA model has multiple y variables       ;
*   on the left-hand side of the equation =   ;
```

```
proc glm data=turtle;
  class sex;
  model length width height = sex;
  lsmeans sex / pdiff;
  manova h=sex /printe printh;
run;
```

Output window contents:

```
The GLM Procedure

   Class Level Information
Class         Levels    Values
sex                2    f m

Number of Observations Read          48
Number of Observations Used          48

Dependent Variable: length
                                     Sum of
Source                       DF        Squares    Mean Square   F Value   Pr > F
Model                         1     6142.68750     6142.68750     20.79   <.0001
Error                        46    13590.79167      295.45199
Corrected Total              47    19733.47917

R-Square     Coeff Var      Root MSE    length Mean
0.311283      13.78083      17.18872       124.7292

Source                       DF    Type III SS    Mean Square   F Value   Pr > F
sex                           1    6142.687500    6142.687500     20.79   <.0001


Dependent Variable: width
                                     Sum of
Source                       DF        Squares    Mean Square   F Value   Pr > F
Model                         1    2451.020833    2451.020833     22.10   <.0001
Error                        46    5100.791667     110.886775
Corrected Total              47    7551.812500

R-Square     Coeff Var      Root MSE    width Mean
0.324561      11.03369      10.53028      95.43750

Source                       DF    Type III SS    Mean Square   F Value   Pr > F
sex                           1    2451.020833    2451.020833     22.10   <.0001


Dependent Variable: height
                                     Sum of
Source                       DF        Squares    Mean Square   F Value   Pr > F
Model                         1    1541.333333    1541.333333     40.56   <.0001
Error                        46    1747.916667      37.998188
Corrected Total              47    3289.250000

R-Square     Coeff Var      Root MSE    height Mean
0.468597      13.29222      6.164267       46.37500

Source                       DF    Type III SS    Mean Square   F Value   Pr > F
sex                           1    1541.333333    1541.333333     40.56   <.0001
```

```
Least Squares Means
                                 H0:LSMean1=
                 length            LSMean2
sex              LSMEAN             Pr > |t|
f            136.041667               <.0001
m            113.416667

                                 H0:LSMean1=
                                   LSMean2
sex     width LSMEAN                Pr > |t|
f         102.583333                <.0001
m          88.291667

                                 H0:LSMean1=
                 height            LSMean2
sex              LSMEAN             Pr > |t|
f          52.0416667               <.0001
m          40.7083333
```

The `printe` and `printh` options on the `manova` statement give the Error SS and Between or Hypothesis SS matrices for the MANOVA table. A MANOVA table is not explicitly given.

The **partial correlation coefficients** are correlations between features constructed from the pooled covariance matrix. Recall how to compute a correlation matrix from a covariance matrix. The partial correlations provide the best estimate of the correlation between features, which is assumed to be constant across groups. The partial correlations are approximately the average of the within sex correlation matrices (compare with earlier output). Partial correlations should not be confused with correlations between features ignoring groups, which do not account for differences in the locations of the groups.

The **characteristic vector** (eigenvector) output gives one linear combination of the features for each variable in the data set. By construction, the linear combinations are uncorrelated (adjusting for groups). In general, the first $a = \text{minimum}(p, k-1)$ linear combinations contain information in decreasing amounts for distinguishing among the groups. The first linear combination is used by Roy's test.

The three linear combinations for the carapace data are (reading across the rows in the matrix)

$$\begin{aligned} \text{D1} &= -0.017\text{Length} - 0.008\text{Width} + 0.075\text{Height} \\ \text{D2} &= -0.016\text{Length} - 0.013\text{Width} + 0.049\text{Height} \\ \text{D3} &= -0.032\text{Length} + 0.056\text{Width} - 0.007\text{Height}. \end{aligned}$$

Here $p = 3$ and $k = 2$ gives $a = \text{minimum}(p, k-1) = \text{minimum}(3, 2-1) = 1$ so only D1 from D1, D2, and D3 contains information for distinguishing between male and female painted turtles.

As in PCA, the linear combinations should be interpreted. However, do not discount the contribution of a feature with a small loading. In particular, the most important feature for distinguishing among the groups might have a small loading because of the measurement scale.

To interpret D1, I multiplied each loading by 100, giving the following comparison of height with a weighted average of length and width:

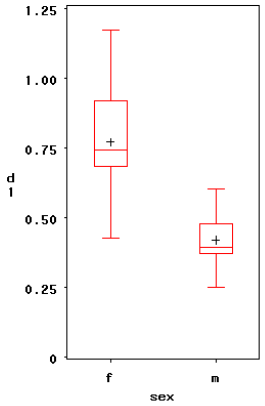
$$\text{D1} = -1.6\text{Length} - 0.8\text{Width} + 7.5\text{Height}.$$

```
Multivariate Analysis of Variance

                  E = Error SSCP Matrix
                  length             width            height
length      13590.791667            8057.5          4679.875
width             8057.5      5100.7916667      2840.4583333
height          4679.875      2840.4583333      1747.9166667

Partial Correlation Coefficients from the Error SSCP Matrix / Prob > |r|
DF = 46          length            width           height
length         1.000000         0.967741         0.960177
                                  <.0001           <.0001
width          0.967741         1.000000         0.951282
                 <.0001                            <.0001
height         0.960177         0.951282         1.000000
                 <.0001           <.0001


Multivariate Analysis of Variance

              H = Type III SSCP Matrix for sex
                  length             width            height
length         6142.6875         3880.1875              3077
width          3880.1875      2451.0208333      1943.6666667
height              3077      1943.6666667      1541.3333333

        Characteristic Roots and Vectors of: E Inverse * H, where
                      H = Type III SSCP Matrix for sex
                           E = Error SSCP Matrix
Characteristic                 Characteristic Vector  V'EV=1
          Root    Percent            length           width          height
    1.58431733     100.00      -0.01667253      -0.00785883       0.07525466
    0.00000000       0.00      -0.03211615       0.05636945      -0.00696931
    0.00000000       0.00      -0.01600765      -0.01364539       0.04916368

                  MANOVA Test Criteria and Exact F Statistics
                  for the Hypothesis of No Overall sex Effect
                       H = Type III SSCP Matrix for sex
                             E = Error SSCP Matrix
                             S=1    M=0.5    N=21
Statistic                         Value    F Value    Num DF    Den DF    Pr > F
Wilks' Lambda                  0.38694938      23.24         3        44    <.0001
Pillai's Trace                 0.61305062      23.24         3        44    <.0001
Hotelling-Lawley Trace         1.58431733      23.24         3        44    <.0001
Roy's Greatest Root            1.58431733      23.24         3        44    <.0001
```

The following output shows the differences between sexes on D1. The separation between sexes on D1 is greater than on any single feature. Females typically have much larger D1 scores than males.

Program editor contents:

```
* characteristic vector, plotting univariate boxplot ;
data turtle;
  set turtle;
  d1 = - 0.017*length - 0.008*width + 0.075*height;
run;

proc sort data=turtle;
  by sex;
run;

proc boxplot data=turtle;
  plot d1*sex /boxstyle=schematic;
run;
```



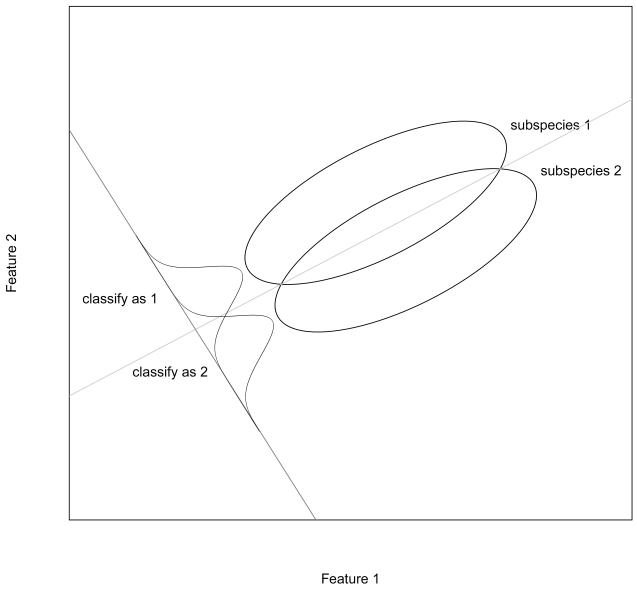
Question: What are reasonable methods for multiple comparisons among three or more groups, on two or more features?

Chapter 16

Discriminant Analysis

A researcher collected data on two external features for two (known) sub-species of an insect. She can use **discriminant analysis** to find linear combinations of the features that best distinguish the sub-species. The analysis can then be used to classify insects with unknown sub-species origin into one of the two sub-species based on their external features.

To see how this might be done, consider the following data plot. CAN1 is the linear combination of the two features that best distinguishes or discriminates the two sub-species. The value of CAN1 could be used to classify insects into one of the two groups, as illustrated.



The method generalizes to more than two features and sub-species.

There are three **SAS** procedures for discriminant analysis and classification. **Candisc** evaluates the linear combinations of features that best distinguish the groups. **Stepdisc** performs a stepwise selection of features. **Discrim** is used for classification. I will not discuss the non-parametric **npar** method available in **discrim**.

16.1 Canonical Discriminant Analysis Using CANDISC

The **SAS** manual highlights a connection between **canonical discriminant analysis** and **canonical correlation**. I prefer to emphasize the connection between canonical discriminant analysis and MANOVA because these techniques are essentially identical.

Assume that you have representative samples from k groups, strata, or sub-populations. Each selected individual is measured on p features (measurements) $X_1, X_2, \ldots, X_p$. As in MANOVA, canonical discriminant analysis assumes you have independent samples from multivariate normal populations with identical variance-covariance matrices.

Candisc computes $r = \text{minimum}(p, k-1)$ linear combinations of the features with the following properties. The first linear combination, called the **first linear discriminant function**

$$\text{CAN1} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p$$

gives the most significant F-test for no group differences in a one-way ANOVA, among all linear combinations of the features. The second linear combination or the **second linear discriminant function**:

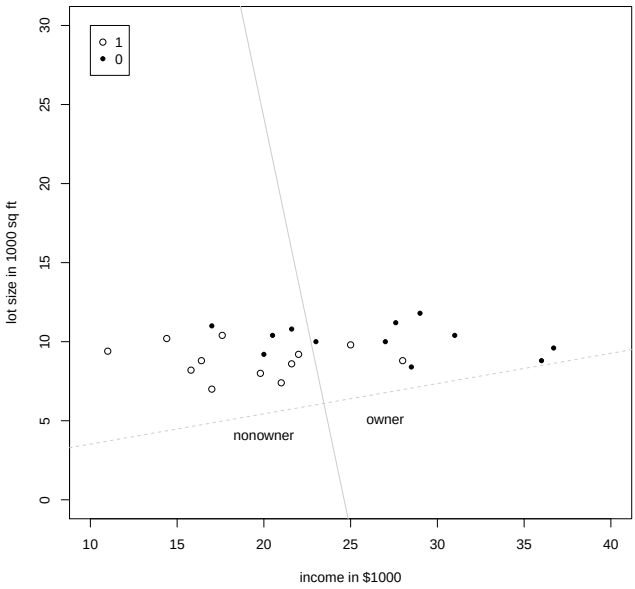
$$\text{CAN2} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p$$

gives the most significant F-test for no group differences in a one-way ANOVA, among all linear combinations of the features that are uncorrelated (adjusting for groups) with CAN1. In general, the j^{th} linear combination CANj ($j = 1, 2, \ldots, r$) gives the most significant F-test for no group differences in a one-way ANOVA, among all linear combinations of the features that are uncorrelated with $CAN1, CAN2, \ldots, CAN(j-1)$. **SAS** also provides significance tests on the differences among groups on the canonical variables.

The coefficients in the canonical discriminant functions can be multiplied by a constant, or all the signs can be changed, without changing their properties or interpretations.

16.2 Example: Owners of riding mowers

The manufacturer of a riding lawn mower wishes to identify the best prospects for buying their product using data on the incomes (X_1) and lot sizes (X_2) of homeowners (Johnson and Wichern, 1988). The data plot below gives incomes and lot sizes from a random sample of 12 current owners of the mowers and from an independent sample of 12 non-owners.



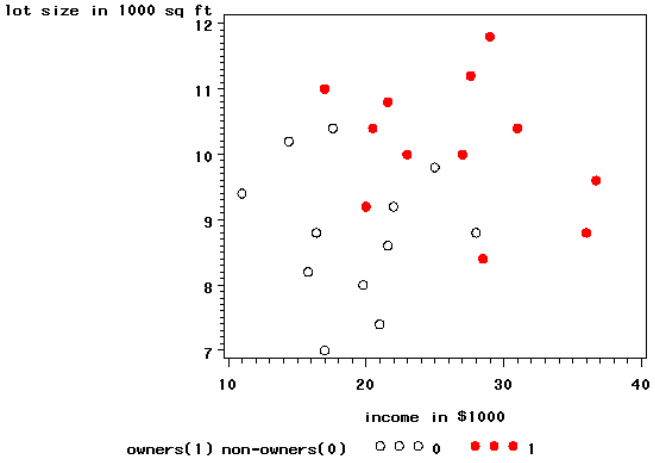
Although the two groups overlap, the owners tend to have higher incomes and larger lots than the non-owners. Income seems to distinguish owners and non-owners better than lot size, but both variables seem to be useful for discriminating between groups. Qualitatively, one might classify prospects based on their location relative to the solid line on the plot. A discriminant analysis gives similar results to this heuristic approach because the CAN1 scores will roughly correspond to the projection of the two features onto the dashed line perpendicular to the solid line. **Candisc** computes one discriminant function here because $p = 2$ and $k = 2$ gives $r = \text{minimum}(p, k-1) = \text{minimum}(2, 1) = 1$.

A **SAS** program for the discriminant analysis on the mower data is given below. The `anova` option on the `candisc` statement prints univariate summary statistics. The `out=outdisc` option creates an output data set `outdisc` that contains the input data set `mower` plus the values on the discriminant functions CAN1, CAN2, ..., CANr. Only CAN1 is generated here. Other popular options are `distance` for the Mahalanobis distance between groups, `wcorr` for within group correlation matrices, and `wcov` for within group covariance matrices. The `all` option prints everything possible, and `short` gives abbreviated output.

Program editor contents:

```
options ls=79 nodate nocenter;

* canonical discriminant analysis on rider mower ownership;
data mower;
  input income lotsize owner $ @@;
  label income='income in $1000'
        lotsize='lot size in 1000 sq ft'
        owner='non-owners(0) owners(1)';
  datalines;
  20.0  9.2 1   28.5  8.4 1   21.6 10.8 1   20.5 10.4 1   29.0 11.8 1
  36.7  9.6 1   36.0  8.8 1   27.6 11.2 1   23.0 10.0 1   31.0 10.4 1
  17.0 11.0 1   27.0 10.0 1
  25.0  9.8 0   17.6 10.4 0   21.6  8.6 0   14.4 10.2 0   28.0  8.8 0
  19.8  8.0 0   22.0  9.2 0   15.8  8.2 0   11.0  9.4 0   17.0  7.0 0
  16.4  8.8 0   21.0  7.4 0
;
```

```
run;

* plot data by ownership;
symbol1 v=circle c=black i=none;  * define v=symbol, c=color, ;
symbol2 v=dot    c=red   i=none;  *   and i=interpolation     ;
proc gplot data=mower;
  plot lotsize*income=owner;   * plot given earlier;
run;

* perform discriminant analysis with (anova) univariate summary statistics;
proc candisc anova out=outdisc;
  class owner;
  var income lotsize;
run;

* boxplot of two groups by CAN1 direction;
proc boxplot data=outdisc;
  plot can1*owner /boxstyle=schematic;
run;
```



Output window contents:

```
The CANDISC Procedure

Observations      24          DF Total                 23
Variables          2          DF Within Classes        22
Classes            2          DF Between Classes        1

                 Class Level Information
          Variable
owner     Name         Frequency       Weight    Proportion
0         _0                  12      12.0000      0.500000   <--- proportions
1         _1                  12      12.0000      0.500000      in each group
```

The univariate summaries given below are related to one-way ANOVAs on each variable. The Pr>F p-values are for one-way ANOVA comparing owners to non-owners. These p-values suggest that income and lot size are important individually for distinguishing between the groups.

```
                          Univariate Test Statistics
                     F Statistics,    Num DF=1,   Den DF=22
              Total    Pooled   Between
           Standard  Standard  Standard           R-Square
Variable  Deviation Deviation Deviation R-Square / (1-RSq)  F Value    Pr > F
income       6.5977    5.5445    5.2031   0.3245    0.4804    10.57    0.0037
lotsize      1.2141    1.0336    0.9310   0.3068    0.4426     9.74    0.0050

        Average R-Square
Unweighted              0.315637
Weighted by Variance    0.3239058
```

The MANOVA p-values given below are for testing differences between the two groups using income and lot size simultaneously.

Multivariate Statistics and Exact F Statistics

S=1 M=0 N=9.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.46139563	12.26	2	21	0.0003
Pillai's Trace	0.53860437	12.26	2	21	0.0003
Hotelling-Lawley Trace	1.16733737	12.26	2	21	0.0003
Roy's Greatest Root	1.16733737	12.26	2	21	0.0003

I ignore the canonical correlation summary that follows the MANOVA tests.

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation
1	0.733897	0.727591	0.096208	0.538604

Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)

	Eigenvalue	Difference	Proportion	Cumulative
1	1.1673		1.0000	1.0000

The likelihood ratio p-values below correspond to tests of no differences between groups on the canonical discriminant functions. There is only one canonical discriminant function here. The tests of no differences based on the first canonical discriminant function is Roy's MANOVA test.

Test of H0: The canonical correlations in the current row and all that follow are zero

	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F	
1	0.46139563	12.26	2	21	0.0003	<---

NOTE: The F statistic is exact.

I ignore the total, between, and within canonical structure output given below. Candisc gives two standardized canonical coefficients (total sample and pooled within class), and the raw canonical coefficients.

Total Canonical Structure

Variable	Label	Can1
income	income in $1000	0.776180
lotsize	lot size in 1000 sq ft	0.754718

Between Canonical Structure

Variable	Label	Can1
income	income in $1000	1.000000
lotsize	lot size in 1000 sq ft	1.000000

Pooled Within Canonical Structure

Variable	Label	Can1
income	income in $1000	0.641479
lotsize	lot size in 1000 sq ft	0.615727

Total-Sample Standardized Canonical Coefficients

Variable	Label	Can1
income	income in $1000	0.9589145576
lotsize	lot size in 1000 sq ft	0.9215860160

Pooled Within-Class Standardized Canonical Coefficients

Variable	Label	Can1
income	income in $1000	0.8058418698
lotsize	lot size in 1000 sq ft	0.7845512333

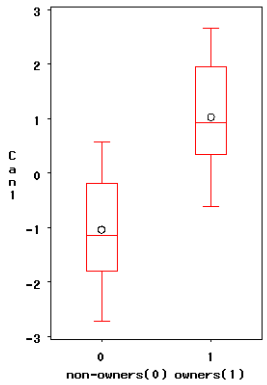
The raw canonical coefficients define the canonical discriminant variables and are identical to the feature loadings in a one-way MANOVA, except for an unimportant multiplicative factor. The `class means` output gives the mean score on the canonical discriminant variables by group, after centering the scores to have mean zero over all groups.

```
              Raw Canonical Coefficients
Variable      Label                              Can1
income        income in $1000              0.1453404136                <---
lotsize       lot size in 1000 sq ft       0.7590456922                <---

Class Means on Canonical Variables
owner                Can1
0           -1.034436686
1            1.034436686
```

The linear combination of income and lot size that best distinguishes owners from non-owners

$$\text{CAN1} = 0.145\text{INCOME} + 0.759\text{LOT SIZE}$$

is a weighted average of income and lot size. This is the direction indicated by the dashed line in the first scatterplot. The standardized coefficients (use the pooled within-class coefficients) indicate the relative contributions of the features to the discrimination. The standardized coefficients are roughly equal, which suggests that income and lot size contribute similarly to distinguishing the owners from non-owners.



The p-value of 0.0003 on the likelihood ratio test indicates that CAN1 strongly distinguishes between owners and non-owners. This is consistent with the separation between owners and non-owners in the boxplot of CAN1 scores.

I noted above that CAN1 is essentially the same linear combination given in a MANOVA comparison of owners to non-owners. Here is some `glm` output to support this claim. The MANOVA test p-values agree with the `candisc` output. The first canonical discriminant function is obtained by multiplying the first characteristic vector given in MANOVA by 4.69:

$$\begin{aligned} \text{CAN1} &= 0.145\text{INCOME} + 0.759\text{LOT SIZE} \\ &= 4.69(0.03099\text{INCOME} + 0.16183\text{LOT SIZE}). \end{aligned}$$

Program editor contents:

```
* MANOVA result matches discriminant analysis for two dimensions;
proc glm data=mower;
  class owner;
  model lotsize income = owner;
  manova h=owner /printe printh;
run;
```

Output window contents:

```
Characteristic Roots and Vectors of: E Inverse * H, where
            H = Type III SSCP Matrix for owner
                  E = Error SSCP Matrix
Characteristic              Characteristic Vector  V'EV=1
          Root    Percent          lotsize         income
    1.16733737     100.00       0.16182909      0.03098668         <---
    0.00000000       0.00      -0.13532047      0.02421363

                MANOVA Test Criteria and Exact F Statistics for
                   the Hypothesis of No Overall owner Effect
                      H = Type III SSCP Matrix for owner
                             E = Error SSCP Matrix
                              S=1    M=0    N=9.5
Statistic                        Value    F Value    Num DF    Den DF    Pr > F
Wilks' Lambda                0.46139563      12.26         2        21    0.0003
Pillai's Trace               0.53860437      12.26         2        21    0.0003
Hotelling-Lawley Trace       1.16733737      12.26         2        21    0.0003
Roy's Greatest Root          1.16733737      12.26         2        21    0.0003
```

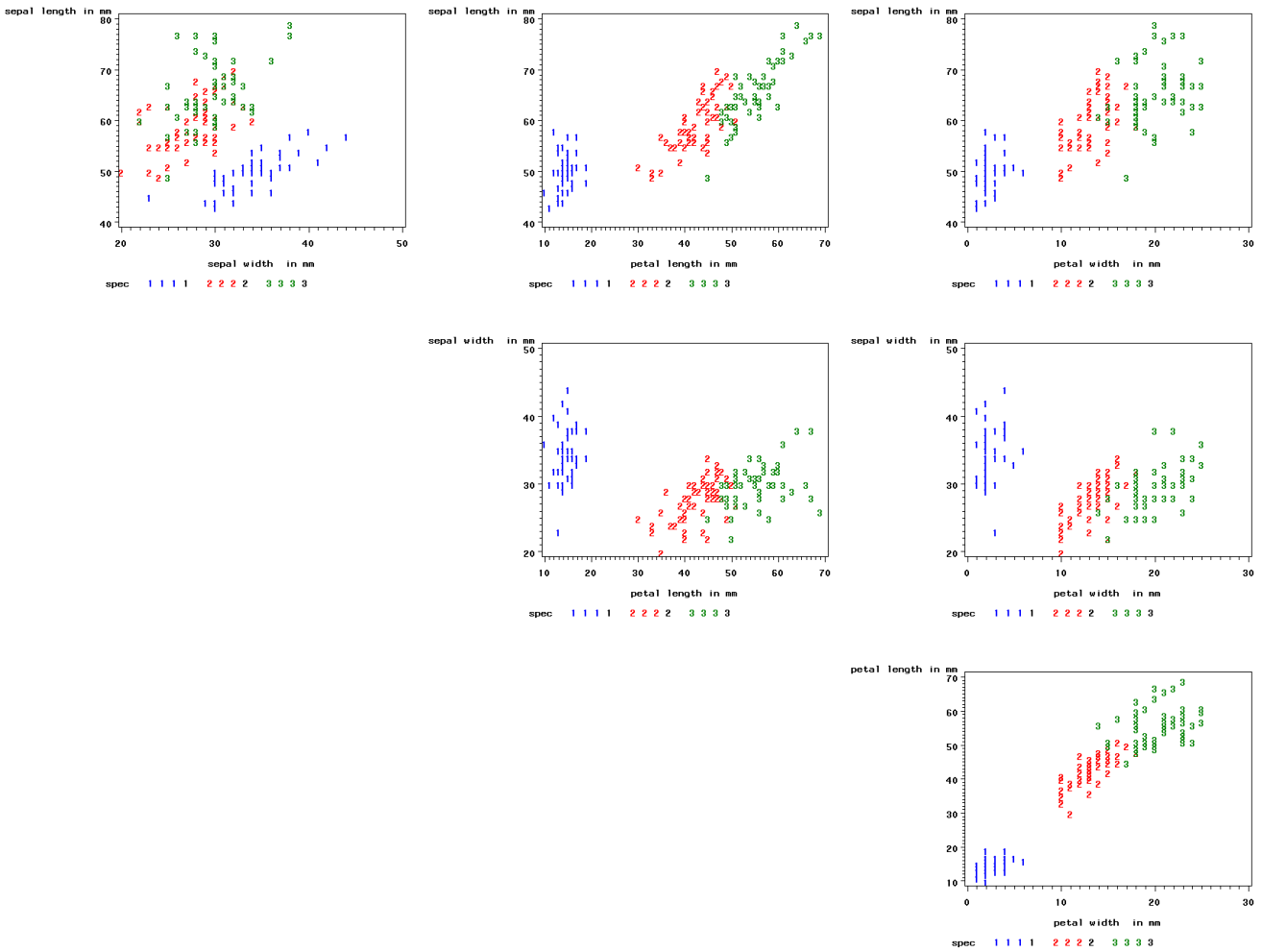
16.3 Discriminant Analysis on Fisher's Iris Data

Fisher's iris data consists of samples of 50 flowers from each of three species of iris: Virginica, Versicolor, and Setosa. Four measurements (in mm) were taken on each flower: sepal length, sepal width, petal length, and petal width. The following **SAS** output gives all possible plots of two variables in the iris data, using species ID as a plotting symbol (1=Setosa, 2=Versicolor, and 3=Virginica). The plots show big differences between Setosa and the other two species. The differences between Versicolor and Virginica are smaller, and appear to be mostly due to differences in the petal widths and lengths.

Program editor contents:

```
options ls=79 nodate nocenter;

* fisher iris data;
data iris;
  infile "F:\Dropbox\UNM\teach\ADA2_stat528\sas\ADA2_16_iris.dat";
  input sepalL sepalW petalL petalW spec @@;
  if      spec=1 then species='setosa';
  else if spec=2 then species='versicolor';
  else if spec=3 then species='virginica';

  label sepalL='sepal length in mm'
        sepalW='sepal width  in mm'
        petalL='petal length in mm'
        petalW='petal width  in mm';
run;

* scatterplot;
* define v=symbol, c=color, and i=interpolation;
symbol1 v="1" c=blue  i=none;
symbol2 v="2" c=red   i=none;
symbol3 v="3" c=green i=none;
proc gplot data=iris;
  plot sepalL*sepalW=spec sepalL*petalL=spec sepalL*petalW=spec
                          sepalW*petalL=spec sepalW*petalW=spec
                                             petalL*petalW=spec;
run;

* perform canonical discrimination;
proc candisc wcorr wcov out=outcandisc;
  class species;
  var sepalL sepalW petalL petalW;
run;
```

```
* scatterplot of canonical discriminant functions;
* define v=symbol, c=color, and i=interpolation;
symbol1 v="1" c=blue  i=none;
symbol2 v="2" c=red   i=none;
symbol3 v="3" c=green i=none;
proc gplot data=outcandisc;
  plot can2*can1=spec;
run;
```



Candisc was used to discriminate among species. There are $k = 3$ species and $p = 4$ features, so the number of discriminant functions is 2 (the minimum of 4 and $3 - 1$). CAN1 is a comparison of petal and sepal measurements (from Raw Canonical Coefficients):

$$\text{CAN1} = -0.083\ \text{sepalL} - 0.154\ \text{sepalW} + 0.220\ \text{petalL} + 0.283\ \text{petalW}.$$

CAN2 is not easily interpreted:

$$\text{CAN2} = +0.001\ \text{sepalL} + 0.218\ \text{sepalW} - 0.091\ \text{petalL} + 0.281\ \text{petalW}.$$

The CANDISC Procedure

Observations	150	DF Total	149
Variables	4	DF Within Classes	147
Classes	3	DF Between Classes	2

Class Level Information

species	Variable Name	Frequency	Weight	Proportion
setosa	setosa	50	50.0000	0.333333
versic	versic	50	50.0000	0.333333
virgin	virgin	50	50.0000	0.333333

Within-Class Covariance Matrices

species = setosa, DF = 49

Variable	Label	sepalL	sepalW	petalL	petalW
sepalL	sepal length in mm	12.42489796	9.92163265	1.63551020	1.15673469
sepalW	sepal width in mm	9.92163265	14.36897959	1.16979592	1.01714286
petalL	petal length in mm	1.63551020	1.16979592	3.01591837	0.64000000
petalW	petal width in mm	1.15673469	1.01714286	0.64000000	1.14938776

species = versic, DF = 49

Variable	Label	sepalL	sepalW	petalL	petalW
sepalL	sepal length in mm	26.64326531	8.51836735	18.28979592	5.57795918
sepalW	sepal width in mm	8.51836735	9.84693878	8.26530612	4.12040816
petalL	petal length in mm	18.28979592	8.26530612	22.08163265	7.31020408
petalW	petal width in mm	5.57795918	4.12040816	7.31020408	3.91061224

species = virgin, DF = 49

Variable	Label	sepalL	sepalW	petalL	petalW
sepalL	sepal length in mm	40.43428571	9.37632653	30.32897959	4.90938776
sepalW	sepal width in mm	9.37632653	10.40040816	7.13795918	4.76285714
petalL	petal length in mm	30.32897959	7.13795918	30.45877551	4.88244898
petalW	petal width in mm	4.90938776	4.76285714	4.88244898	7.54326531

Within-Class Correlation Coefficients / Pr > |r|

species = setosa

Variable	sepalL	sepalW	petalL	petalW
sepalL	1.00000	0.74255	0.26718	0.30609
sepal length in mm		<.0001	0.0607	0.0306
sepalW	0.74255	1.00000	0.17770	0.25029
sepal width in mm	<.0001		0.2170	0.0796
petalL	0.26718	0.17770	1.00000	0.34375
petal length in mm	0.0607	0.2170		0.0145
petalW	0.30609	0.25029	0.34375	1.00000
petal width in mm	0.0306	0.0796	0.0145	

species = versic

Variable	sepalL	sepalW	petalL	petalW
sepalL	1.00000	0.52591	0.75405	0.54646
sepal length in mm		<.0001	<.0001	<.0001
sepalW	0.52591	1.00000	0.56052	0.66400
sepal width in mm	<.0001		<.0001	<.0001
petalL	0.75405	0.56052	1.00000	0.78667
petal length in mm	<.0001	<.0001		<.0001
petalW	0.54646	0.66400	0.78667	1.00000
petal width in mm	<.0001	<.0001	<.0001	

species = virgin

Variable	sepalL	sepalW	petalL	petalW
sepalL	1.00000	0.45723	0.86422	0.28111
sepal length in mm		0.0008	<.0001	0.0480
sepalW	0.45723	1.00000	0.40104	0.53773
sepal width in mm	0.0008		0.0039	<.0001
petalL	0.86422	0.40104	1.00000	0.32211
petal length in mm	<.0001	0.0039		0.0225
petalW	0.28111	0.53773	0.32211	1.00000
petal width in mm	0.0480	<.0001	0.0225	

Multivariate Statistics and F Approximations
S=2 M=0.5 N=71

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.02340737	199.30	8	288	<.0001
Pillai's Trace	1.19078426	53.34	8	290	<.0001
Hotelling-Lawley Trace	32.57098618	583.88	8	203.4	<.0001
Roy's Greatest Root	32.28757663	1170.42	4	145	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.
NOTE: F Statistic for Wilks' Lambda is exact.

There are significant differences among species on both discriminant functions; see the p-values under the likelihood ratio tests. Of course, CAN1 produces the largest differences — the overlap among species on CAN1 is small. Setosa has the lowest CAN1 scores because this species has the smallest petal measurements relative to its sepal measurements. Virginica has the highest CAN1 scores.

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation
1	0.984865	0.984553	0.002461	0.969959
2	0.469921	0.460120	0.063832	0.220825

Eigenvalues of Inv(E)*H
= CanRsq/(1-CanRsq)

	Eigenvalue	Difference	Proportion	Cumulative
1	32.2876	32.0042	0.9913	0.9913
2	0.2834		0.0087	1.0000

Test of H0: The canonical correlations in the current row and all that follow are zero

	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F	
1	0.02340737	199.30	8	288	<.0001	<---
2	0.77917450	13.70	3	145	<.0001	<---

Again, I ignore the total, between, and pooled canonical structures.

Total Canonical Structure

Variable	Label	Can1	Can2
sepalL	sepal length in mm	0.791833	0.218497
sepalW	sepal width in mm	-0.530804	0.759838
petalL	petal length in mm	0.984903	0.046551
petalW	petal width in mm	0.972777	0.222863

Between Canonical Structure

Variable	Label	Can1	Can2
sepalL	sepal length in mm	0.991444	0.130535
sepalW	sepal width in mm	-0.825764	0.564016
petalL	petal length in mm	0.999746	0.022546
petalW	petal width in mm	0.994078	0.108666

Pooled Within Canonical Structure

Variable	Label	Can1	Can2
sepalL	sepal length in mm	0.222260	0.312344
sepalW	sepal width in mm	-0.118850	0.866456
petalL	petal length in mm	0.705016	0.169705
petalW	petal width in mm	0.632036	0.737437

Total-Sample Standardized Canonical Coefficients

Variable	Label	Can1	Can2
sepalL	sepal length in mm	-0.689261976	0.010193102
sepalW	sepal width in mm	-0.671925794	0.948163297
petalL	petal length in mm	3.881120486	-1.614362126
petalW	petal width in mm	2.155968978	2.143560936

Pooled Within-Class Standardized Canonical Coefficients

Variable	Label	Can1	Can2
sepalL	sepal length in mm	-.4284981258	0.0063368140
sepalW	sepal width in mm	-.5236581881	0.7389409337
petalL	petal length in mm	0.9461177106	-.3935401140
petalW	petal width in mm	0.5790384430	0.5757059586

The raw values give the coefficients in the original scale (centered at 0).

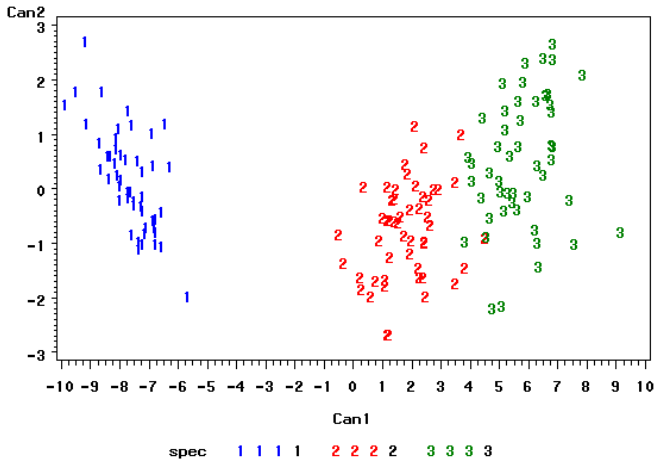
Raw Canonical Coefficients

Variable	Label	Can1	Can2	
sepalL	sepal length in mm	-.0832375523	0.0012309526	<---
sepalW	sepal width in mm	-.1541586991	0.2175353610	<---
petalL	petal length in mm	0.2198563627	-.0914498239	<---
petalW	petal width in mm	0.2825052260	0.2808793507	<---

Class Means on Canonical Variables

species	Can1	Can2
setosa	-7.619324451	0.214250317
versic	1.829218125	-0.725335997
virgin	5.790106326	0.511085680

The canonical directions provide a maximal separation the species. Two vertical lines along CAN1 will provide a classification rule.



Two questions: What is the most striking feature of the plot of the CAN1 scores? Does the assumption of equal population covariance matrices across species seem plausible?

16.3.1 Variable Selection Using STEPDISC

Stepwise methods can be used to select features for a discriminant analysis. Three algorithms are available in **SAS**: **backward elimination**, **forward selection**, and a **pure stepwise** approach. The motivation behind the algorithms is similar to variable selection in regression in the sense that some features may be unnecessary for distinguishing groups.

I prefer stepwise selection to forward selection. The backward elimination is useful when the number of features is small. I will describe forward selection, given samples from k groups, strata or sub-populations on p features $X_1, X_2, \ldots, X_p$.

Step 1. Identify the feature that gives the smallest p-value for testing no differences among groups in a one-way ANOVA. Suppose this feature is X_1.

Step 2. Identify the remaining feature from $X_2, X_3, \ldots, X_p$ that gives the most significant F-test for no group differences **after adjusting** for X_1.

The second step requires fitting all possible regression models:

$$X_i = \beta_1 I_1 + \beta_2 I_2 + \cdots + \beta_k I_k + \beta_{k+1} X_1 + \varepsilon,$$

where X_i is one of $X_2, X_3, \ldots, X_p$ and $I_1, I_2, \ldots, I_k$ are binary $(0, 1)$ indicator variables for observations in groups $1, 2, \ldots, k$. (Recall the discussion of response models with factors and predictors.)

This model assumes that X_i is linearly related to X_1 with a constant slope across the groups. The coefficients $\beta_1, \beta_2, \ldots, \beta_k$ for the indicator variables are intercepts for the individual regression lines. An F-test of $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k$ (similar to the ANOVA F-test in multiple regression) checks whether the groups have identical means on X_i at each level of X_1 (i.e., adjusting for X_1). The largest F-statistic, or smallest p-value, across the remaining features gives the variable on which the groups differ most after adjusting for X_1. Suppose this is X_2.

Step 3. From $X_3, \ldots, X_p$ identify the feature that gives the most significant F-test for no group differences after adjusting for X_1 and X_2.

The process is continued until the most important remaining feature gives a p-value larger than the specified cutoff. The default level is 0.15 for all the selection methods.

Stepdisc identifies the features that are useful for discriminating among groups but does not compute the canonical discriminant functions. After selecting variables, you need to use `candisc` to complete the analysis.

The **SAS** program below illustrates stepwise variable selection on Fisher's iris data. The `simple` option on the `stepdisc` statement gives simple descriptive statistics. The `wcov` and `wcorr` options give the within sample (species) covariance and correlation matrices. The `short` option prints a summary of the variable selection. If short is omitted, more detailed information is given for each step in the selection. The default method is stepwise selection. `Forward` or `backward` selection are obtained by specifying these keywords prior to the semi-colon at the end of the `stepdisc` statement. Other useful options are `slstay=p` and `slentry=p` for setting the test levels on the backward and forward steps, respectively. Most of the printed output below was omitted.

Program editor contents:

```
* stepwise discrimination analysis;
proc stepdisc data=iris simple wcov wcorr short;
  class species;
  var sepalL sepalW petalL petalW;
run;
```

Output window contents:

```
The STEPDISC Procedure

           The Method for Selecting Variables is STEPWISE
Observations       150          Variable(s) in the Analysis        4
Class Levels         3          Variable(s) will be Included       0
                                Significance Level to Enter     0.15  <--- default
                                Significance Level to Stay      0.15  <--- levels

                 Class Level Information
           Variable
species    Name        Frequency       Weight    Proportion
setosa     setosa             50      50.0000      0.333333
versic     versic             50      50.0000      0.333333
virgin     virgin             50      50.0000      0.333333
```

I'm ignoring the other summaries.

The stepwise discriminant analysis selects all four features, in the following order: petal length, sepal width, petal width, and sepal length. This ordering is not surprising given the data plots. At each step, the F-test p-value checks the importance of the selected feature for distinguishing among groups, after adjusting for the effects of features already selected. The F-test p-value is compared to the default test level in the stopping rule. The output for Wilk's lambda gives a MANOVA p-value for testing no difference in means for the features selected up to and including the given step. For example, in the second step, the sepal width is identified as the most important remaining variable. The F-test p-value of 0.0001 indicates that the mean sepal widths are significantly different across species, after accounting for the differences in petal lengths across species. Using Wilk's test, the p-value for testing no differences in the mean petal lengths and sepal widths across species is 0.0001.

Stepwise Selection Summary

Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	petalL		0.9414	1180.16	<.0001	0.05862828	<.0001	0.47068586	<.0001
2	2	sepalW		0.3709	43.04	<.0001	0.03688411	<.0001	0.55995394	<.0001
3	3	petalW		0.3234	34.65	<.0001	0.02495618	<.0001	0.59439702	<.0001
4	4	sepalL		0.0621	4.76	0.0099	0.02340737	<.0001	0.59539213	<.0001

The four petal measurements are important for distinguishing among species, so we would use our earlier analysis to summarize the discrimination of species.

16.3.2 Stepwise Selection of Features and MANOVA

A stepwise selection of features can be used with MANOVA to eliminate the features that have little information for distinguishing groups. This helps interpret the MANOVA results, especially when the groups are different on all features individually. For example, a strong correlation between all pairs of features within groups may lead to significant differences among groups on each feature, yet only one feature may be needed to distinguish the groups.

Here is a **SAS** program and output for a `backward` and `stepwise` selection of carapace measurements for comparing male and female painted turtles. The `short` option was used to reduce the amount of output.

The carapace measurements are strongly correlated. We saw earlier that the differences between males and females are significant on each feature, and on all features collectively. The stepwise and backward selection indicates that the differences in widths between males and females are insignificant once the differences on height and length are taken into account.

Program editor contents:

```
options ls=79 nodate nocenter;

* example: carapace measurements for painted turtles;
data turtle;
  infile "F:\Dropbox\UNM\teach\ADA2_stat528\sas\ADA2_15_turtle.dat";
  input sex $ length width height @@;  *$;
run;

* stepwise ;
proc stepdisc data=turtle short;
  class sex;
  var length width height;
run;

* forward ;
```

```
proc stepdisc data=turtle forward short;
  class sex;
  var length width height;
run;

* backwards ;
proc stepdisc data=turtle backwards short;
  class sex;
  var length width height;
run;
```

Stepwise:

The STEPDISC Procedure

The Method for Selecting Variables is STEPWISE

Observations	48	Variable(s) in the Analysis	3
Class Levels	2	Variable(s) will be Included	0
		Significance Level to Enter	0.15
		Significance Level to Stay	0.15

Class Level Information

sex	Variable Name	Frequency	Weight	Proportion
f	f	24	24.0000	0.500000
m	m	24	24.0000	0.500000

Stepwise Selection Summary

Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	height		0.4686	40.56	<.0001	0.53140280	<.0001	0.46859720	<.0001
2	2	length		0.2637	16.12	0.0002	0.39127425	<.0001	0.60872575	<.0001

Forward:

The Method for Selecting Variables is FORWARD

Observations	48	Variable(s) in the Analysis	3
Class Levels	2	Variable(s) will be Included	0
		Significance Level to Enter	0.15

Class Level Information

sex	Variable Name	Frequency	Weight	Proportion
f	f	24	24.0000	0.500000
m	m	24	24.0000	0.500000

Forward Selection Summary

Step	Number In	Entered	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	height	0.4686	40.56	<.0001	0.53140280	<.0001	0.46859720	<.0001
2	2	length	0.2637	16.12	0.0002	0.39127425	<.0001	0.60872575	<.0001

Backward:

The Method for Selecting Variables is BACKWARD

Observations	48	Variable(s) in the Analysis	3
Class Levels	2	Variable(s) will be Included	0
		Significance Level to Stay	0.15

Class Level Information

sex	Variable Name	Frequency	Weight	Proportion
f	f	24	24.0000	0.500000
m	m	24	24.0000	0.500000

Backward Elimination Summary

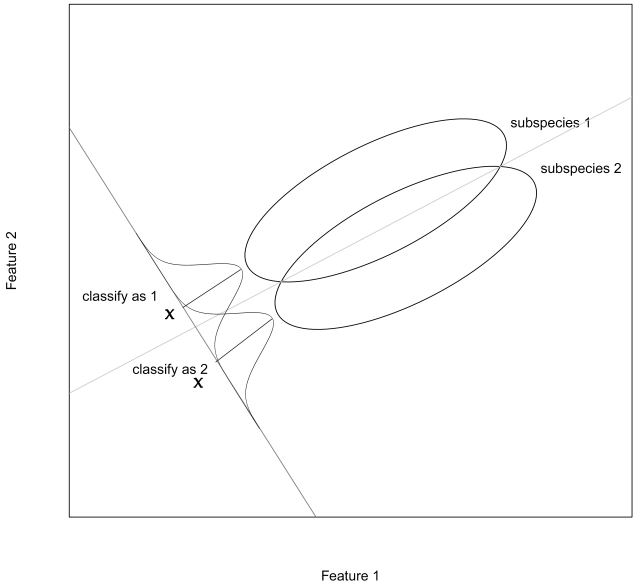
Step	Number In	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
0	3		.	.	.	0.38694938	<.0001	0.61305062	<.0001
1	2	width	0.0111	0.49	0.4868	0.39127425	<.0001	0.60872575	<.0001

Chapter 17

Classification

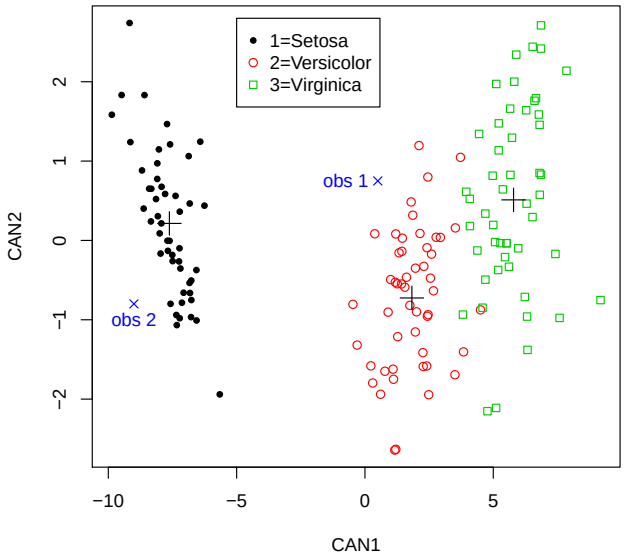
A goal with discriminant analysis might be to classify individuals of unknown origin into one of several known groups. How should this be done? Recall our example where two subspecies of an insect are compared on two external features. The discriminant analysis gives one discriminant function (CAN1) for distinguishing the subspecies. It makes sense to use CAN1 to classify insects because CAN1 is the best (linear) combination of the features to distinguish between subspecies.

Given the score on CAN1 for each insect to be classified, assign insects to the sub-species that they most resemble. Similarity is measured by the distance on CAN1 to the average CAN1 scores for the two subspecies, identified by X's on the plot.



To classify using r discriminant functions, compute the average response on CAN1, ..., CANr in each sample. Then compute the canonical discriminant function scores for each individual to be classified. Each observation is classified into the group it is closest to, as measured by the distance from the observation in r-space to the sample mean vector on the canonical variables. How do you measure distance in r-space?

The plot below illustrates the idea with the $r = 2$ discriminant functions in Fisher's iris data. Observation 1 is classified as Versicolor. Observation 2 is classified as Setosa.



17.1 Classification in SAS

Classification using **discrim** is based on the **original features** and not the **canonical discriminant function scores**. Although **SAS**'s classification rule is identical to the method I just outlined, the equivalence is not obvious. I will discuss **SAS**'s method without justifying the equivalence.

Suppose p features $X = (X_1, X_2, \ldots, X_p)'$ are used to discriminate among the k groups. Let $\overline{X}_i = (\overline{X}_{i1}, \overline{X}_{i2}, \ldots, \overline{X}_{ip})'$ be the vector of mean responses for the i^{th} sample, and let S_i be the p-by-p variance-covariance matrix for the i^{th} sample. The pooled variance-covariance matrix is given by

$$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2 + \cdots + (n_k - 1)S_k}{n - k},$$

where the n_is are the group sample sizes and $n = n_1 + n_2 + \cdots + n_k$ is the total sample size.

SAS's classification rule can be defined using either the **Mahalanobis generalized squared distance** or in terms of a probability model. I will describe each, starting with the Mahalanobis or M-distance.

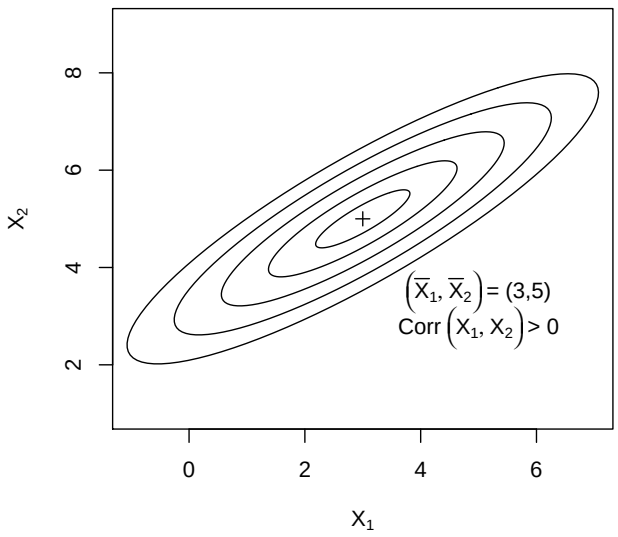
The M-distance from an observation X to (the center of) the i^{th} sample is

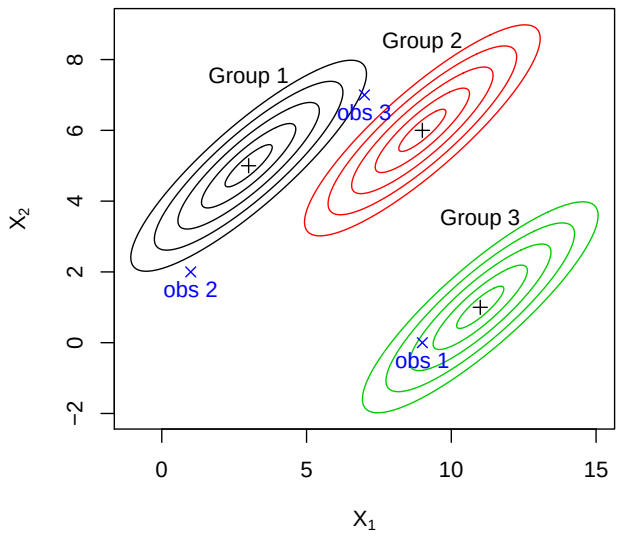
$$D_i^2(X) = (X - \overline{X}_i)'S^{-1}(X - \overline{X}_i),$$

where $(X - \overline{X}_i)'$ is the transpose of the column vector $(X - \overline{X}_i)$, and S^{-1} is the matrix inverse of S. Note that if S is the identity matrix (a matrix with 1s on the diagonal and 0s on the off-diagonals), then this is the Euclidean distance. Given the M-distance from X to each sample, classify X into the group which has the minimum M-distance.

The M-distance is an elliptical distance measure that accounts for correlation between features, and adjusts for different scales by standardizing the features to have unit variance. The picture

below (left) highlights the idea when $p = 2$. All of the points on a given ellipse are the same M-distance to the center $(\overline{X}_1, \overline{X}_2)'$. As the ellipse expands, the M-distance to the center increases.





To see how classification works, suppose you have three groups and two features, as in the plot above (right). Observations 1 is closest in M-distance to the center of group 3. Observation 2 is closest to group 1. Thus, classify observations 1 and 2 into groups 3 and 1, respectively. Observation 3 is closest to the center of group 2 in terms of the standard Euclidean (walking) distance. However, observation 3 is more similar to data in group 1 than it is to either of the other groups. The M-distance from observation 3 to group 1 is substantially smaller than the M-distances to either group 2 or 3. The M-distance accounts for the elliptical cloud of data within each group, which reflects the correlation between the two features. Thus, you would classify observation 3 into group 1.

The M-distance from the i^{th} group to the j^{th} group is the M-distance between the centers of the groups:

$$D^2(i,j) = D^2(j,i) = (\overline{X}_i - \overline{X}_j)' S^{-1} (\overline{X}_i - \overline{X}_j).$$

Larger values suggest relatively better potential for discrimination between groups. In the plot above, $D^2(1,2) < D^2(1,3)$ which implies that it should be easier to distinguish between groups 1 and 3 than groups 1 and 2.

M-distance classification is equivalent to classification based on a probability model that assumes the samples are independently selected from multivariate normal populations with identical covariance matrices. This assumption is consistent with the plot above where the data points form elliptical clouds with similar orientations and spreads across samples. Suppose you can assume **a priori** (without looking at the data for the individual that you wish to classify) that a randomly selected individual from the **combined population** (i.e., merge all sub-populations) is equally likely to be from any group:

$$\text{PRIOR}_j \equiv \Pr(\text{observation is from group } j) = \frac{1}{k},$$

where k is the number of groups. Then, given the observed features X for an individual

$$\begin{aligned} \Pr(j|X) &\equiv \Pr(\text{observation is from group } j \text{given } X) \\ &= \frac{\exp\{-0.5D_j^2(X)\}}{\sum_k \exp\{-0.5D_k^2(X)\}}. \end{aligned}$$

To be precise, I will note that $\Pr(j|X)$ is unknown, and the expression for $\Pr(j|X)$ is an estimate based on the data.

The group with the largest **posterior** probability $\Pr(j|X)$ is the group into which X is classified. Maximizing $\Pr(j|X)$ across groups is equivalent to minimizing the M-distance $D_j^2(X)$ across groups, so the two classification rules used by **SAS** are equivalent.

17.2 Evaluating the Accuracy of a Classification Rule

The **misclassification rate**, or the expected proportion of misclassified observations, is a good yardstick to gauge a classification rule. Better rules have smaller misclassification rates, but there is no universal cutoff for what is considered good in a given problem. You should judge a classification rule relative to the current standards in your field for "good classification".

By default, **discrim** evaluates the misclassification rate using the data from which the classification rule is constructed. This method is called **resubstitution**. The resubstitution estimate of the error rate is optimistic (too small). A greater percentage of misclassifications is expected when the rule is used on new data, or on data from which the rule is not constructed.

A better way to estimate the misclassification rate is through some form of **cross-validation**. In many statistical packages, you can implement cross-validation by randomly splitting the data into a **training** or **calibration set** from which the classification rule is constructed. The remaining data, called the **test data set**, is used with the classification rule to estimate the error rate. In particular, the proportion of test cases misclassified estimates the misclassification rate. This process is often repeated, say 10 times, and the error rate estimated to be the average of the error rates from the individual splits. With repeated random splitting, it is common to use 10% of each split as the test data set (a 10-fold cross-validation).

Repeated random splitting can not be directly implemented in **SAS**. As an alternative, you might consider using one random 50-50 split (a 2-fold) to estimate the misclassification rate, provided you have a reasonably large data base.

Another form of cross-validation uses a **jackknife** method where single cases are held out of the data (an n-fold), then classified after constructing the classification rule from the remaining data. The process is repeated for each case, giving an estimated misclassification rate as the proportion of cases misclassified.

Discrim allows for jackknife cross-validation and cross-validation using a single test data set. The jackknife method is necessary with small sized data sets. You can also classify observations with unknown group membership, by treating the observations to be classified as a test data set.

17.3 Example: Error Estimation on the Carapace Data

The **SAS** program below illustrates the **discrim** procedure on the carapace data. As suggested in an earlier analysis, classification is based on the height and length of the carapace. The default

analysis assumes equal prior probabilities for males and females. I will highlight the basic commands and output assuming that the misclassification rate is to be estimated using the data from which the classification rule is constructed, and then using the jackknife method.

The `list` option on the `discrim` statement gives the posterior probability of being male or female for each observation in the data set (treating sex as unknown). The `listerr` option, which was not used here, gives identical output but only for the misclassified observations.

The `crossvalidate` option performs jackknife error estimation and classification. The `crosslist` option gives the jackknife based classification and posterior probabilities of being male or female for each observation in the data set. The `crosslisterr` option, which was not used here, is analogous to the `listerr` option.

Program editor contents:

```
options ls=79 nodate nocenter;

* example: carapace measurements for painted turtles;
data turtle;
  infile "F:\Dropbox\UNM\teach\ADA2_stat528\sas\ADA2_15_turtle.dat";
  input sex $ length width height @@;  *$;
run;

proc discrim data=turtle list crossvalidate crosslist;
  class sex;
  var height length;
run;
```

Output window contents:

```
The DISCRIM Procedure

Observations      48          DF Total                47
Variables          2          DF Within Classes       46
Classes            2          DF Between Classes       1

                        Class Level Information
       Variable                                                  Prior
sex    Name         Frequency       Weight    Proportion    Probability
f      f                   24      24.0000      0.500000       0.500000
m      m                   24      24.0000      0.500000       0.500000

Pooled Covariance Matrix Information
               Natural Log of the
 Covariance    Determinant of the
Matrix Rank     Covariance Matrix
          2               6.77576
```

The **pairwise generalized squared distances** between groups is the M-distance between the centers of the male and female samples. Larger values suggest relatively better potential for discrimination. The linear discriminant function summary after the M-distance table can be ignored. This information can be used for hand-classification.

```
Pairwise Generalized Squared Distances Between Groups
 2                                -1
D (i|j) = (X - X )' COV    (X - X )
            i   j           i   j

Generalized Squared Distance to sex
From
sex                 f              m
f                   0        5.96372
m            5.96372              0

Linear Discriminant Function
                                -1                                    -1
Constant = -.5 X' COV    X      Coefficient Vector = COV    X
                j         j                                          j

Linear Discriminant Function for sex
Variable              f              m
Constant     -35.87304      -22.23022
height         1.75208        0.55772
length        -0.14286        0.19183
```

The **classification summary table** is constructed from the canonical discriminant functions by comparing the predicted group membership for each observation to its actual group label. To be precise, if you assume that the sex for the 24 males are unknown then you would classify each of them correctly. Similarly, 20 of the 24 females are classified correctly, with the other four classified as males. The **Total Error** of 0.0833 is the estimated miscassification rate, computed as the sum of Rates*Prior over sexes: $0.0833 = 0.1667 * 0.5 + 0 * 0.5$. Are the misclassification results sensible, given the data plots that you saw earlier?

The listing of the posterior probabilities for each sex, by case, gives you an idea of the **clarity of classification**, with larger differences between the male and female posteriors corresponding to more definitive (but not necessarily correct!) classifications.

Classification Results for Calibration Data: WORK.TURTLE
Resubstitution Results using Linear Discriminant Function

Generalized Squared Distance Function

$$D_j^2(X) = (X-\overline{X}_j)' \; COV^{-1} \; (X-\overline{X}_j)$$

Posterior Probability of Membership in Each sex

$$Pr(j|X) = \exp(-.5\, D_j^2(X)) \; / \; SUM_k \exp(-.5\, D_k^2(X))$$

Posterior Probability of Membership in sex

Obs	From sex	Classified into sex		f	m
1	f	m	*	0.2579	0.7421
2	m	m		0.3595	0.6405
3	f	m	*	0.0612	0.9388
4	m	m		0.0355	0.9645
5	f	f		0.8857	0.1143
6	m	m		0.0185	0.9815
7	f	f		0.7986	0.2014
8	m	m		0.2960	0.7040
9	f	f		0.9189	0.0811
10	m	m		0.0835	0.9165
11	f	f		0.9927	0.0073
12	m	m		0.0194	0.9806
13	f	f		0.5327	0.4673
14	m	m		0.1335	0.8665
15	f	f		0.9402	0.0598
16	m	m		0.0731	0.9269
17	f	f		0.9402	0.0598
18	m	m		0.0168	0.9832
19	f	f		0.9402	0.0598
20	m	m		0.0338	0.9662
21	f	m	*	0.2383	0.7617
22	m	m		0.0244	0.9756
23	f	m	*	0.3459	0.6541
24	m	m		0.0176	0.9824
25	f	f		0.7469	0.2531
26	m	m		0.2480	0.7520
27	f	f		0.7469	0.2531
28	m	m		0.0212	0.9788
29	f	f		0.9218	0.0782
30	m	m		0.0212	0.9788
31	f	f		0.9947	0.0053
32	m	m		0.0110	0.9890
33	f	f		0.8983	0.1017
34	m	m		0.0024	0.9976
35	f	f		0.8843	0.1157
36	m	m		0.1696	0.8304
37	f	f		0.9999	0.0001
38	m	m		0.0184	0.9816
39	f	f		0.9979	0.0021
40	m	m		0.1502	0.8498
41	f	f		0.9995	0.0005
42	m	m		0.0830	0.9170
43	f	f		0.9998	0.0002
44	m	m		0.0608	0.9392
45	f	f		0.9933	0.0067
46	m	m		0.0727	0.9273
47	f	f		0.9992	0.0008
48	m	m		0.0635	0.9365

* Misclassified observation

```
Number of Observations and Percent Classified into sex
From
sex                f             m          Total
f                 20             4             24
                83.33         16.67         100.00
m                  0            24             24
                0.00        100.00         100.00
Total             20            28             48
                41.67         58.33         100.00
Priors           0.5           0.5

          Error Count Estimates for sex
                        f           m         Total
Rate               0.1667      0.0000      0.0833
Priors             0.5000      0.5000
```

The cross-validation results follow. The posterior probabilities of membership in each group have changed, but the same 4 observations are misclassified.

```
Classification Results for Calibration Data: WORK.TURTLE
Cross-validation Results using Linear Discriminant Function

  Posterior Probability of Membership in sex
       From      Classified
Obs    sex       into sex           f          m
  1    f         m          *    0.1662     0.8338
  2    m         m               0.4811     0.5189
  3    f         m          *    0.0307     0.9693
  4    m         m               0.0403     0.9597
  5    f         f               0.8466     0.1534
  6    m         m               0.0204     0.9796
  7    f         f               0.7479     0.2521
  8    m         m               0.3461     0.6539
  9    f         f               0.9001     0.0999
 10    m         m               0.0917     0.9083
 11    f         f               0.9922     0.0078
 12    m         m               0.0210     0.9790
 13    f         f               0.5171     0.4829
 14    m         m               0.1458     0.8542
 15    f         f               0.9367     0.0633
 16    m         m               0.0784     0.9216
 17    f         f               0.9367     0.0633
 18    m         m               0.0182     0.9818
 19    f         f               0.9367     0.0633
 20    m         m               0.0362     0.9638
 21    f         m          *    0.1844     0.8156
 22    m         m               0.0263     0.9737
 23    f         m          *    0.2936     0.7064
 24    m         m               0.0190     0.9810
 25    f         f               0.7333     0.2667
 26    m         m               0.2562     0.7438
 27    f         f               0.7333     0.2667
 28    m         m               0.0230     0.9770
 29    f         f               0.9169     0.0831
 30    m         m               0.0230     0.9770
 31    f         f               0.9944     0.0056
 32    m         m               0.0119     0.9881
 33    f         f               0.8864     0.1136
 34    m         m               0.0023     0.9977
 35    f         f               0.8645     0.1355
 36    m         m               0.1752     0.8248
 37    f         f               1.0000     0.0000
 38    m         m               0.0201     0.9799
 39    f         f               0.9979     0.0021
 40    m         m               0.1574     0.8426
 41    f         f               0.9996     0.0004
 42    m         m               0.0899     0.9101
 43    f         f               0.9999     0.0001
 44    m         m               0.0670     0.9330
 45    f         f               0.9929     0.0071
 46    m         m               0.0812     0.9188
 47    f         f               0.9995     0.0005
 48    m         m               0.0739     0.9261

* Misclassified observation
```

```
Number of Observations and Percent Classified into sex
From
sex                f             m         Total
f                 20             4            24
                83.33         16.67        100.00
m                  0            24            24
                 0.00        100.00        100.00
Total             20            28            48
                41.67         58.33        100.00
Priors             0.5           0.5

          Error Count Estimates for sex
                        f             m         Total
Rate               0.1667        0.0000        0.0833
Priors             0.5000        0.5000
```

17.4 Example: Cross-Validation Using Fisher's Iris Data

I will illustrate cross-validation on Fisher's iris data first using a test data set, and then using the jackknife method. The 150 observations were randomly rearranged and separated into two batches of 75. The 75 observations in the calibration set were used to develop a classification rule. This rule was applied to the remaining 75 flowers, which form the test data set. There is no general rule about the relative sizes of the test data and the training data. Many researchers use a 50-50 split. Regardless of the split, you should combine the two data sets at the end of the cross-validation to create the actual rule for classifying future data.

In the **SAS** program, the `iris` data set was randomly rearranged, giving the `raniris` data set. A `set` statement was used in two new data steps to create the training and the test data sets. Observations 76 to 150 from `raniris` are stored in `train`, whereas cases 1 to 75 are stored in `test`.

The `data=train` option on the `discrim` statement develops a classification rule using the `train` data set. The `testdata = test` option uses this classification rule to create a classification table for the `test` data set. The `pool=yes` option specifies equal population covariance matrices, which is the default method. The `listerr` and `testlisterr` options list the observations misclassified in the training data and test data, respectively. The `list` and `testlist` options could be used instead to give the classifications for each observation. The `class` and `testclass` statements specify the group labeling variables in the training data and test data, respectively.

As an aside, note that the M-distances between centers in the calibration set indicate that the Setosa data cloud is further from Versicolor and Virginica than Versicolor and Virginica are from each other. This is what we expect, given the data plots.

The classification rule based on the training set works well with the training data. One of the 75 observations (1.33%) is misclassified. When applied to the test data, two of the 75 observations (2.66%) are misclassified. The misclassification rate is higher with the test data, which is expected. Regardless of which rate you consider, the classification rule is very accurate. Do not expect such nice results on all classification problems!

It is important to recognize that significant differences among groups on linear discriminant function scores do not necessarily translate into accurate classification rules! (WHY?)
Program editor contents:

```
options ls=79 nodate nocenter;

* fisher iris data;
```

```
data iris;
  infile "F:\Dropbox\UNM\teach\ADA2_stat528\sas\ADA2_16_iris.dat";
  input sepalL sepalW petalL petalW spec @@;
  if      spec=1 then species='setosa';
  else if spec=2 then species='versicolor';
  else if spec=3 then species='virginica';

  xrn   = ranuni(1998); * xrn is a set of random numbers, 1998 is the "seed";

  label sepalL='sepal length in mm'
        sepalW='sepal width  in mm'
        petalL='petal length in mm'
        petalW='petal width  in mm';
run;

* sort by the random number column;
proc sort data = iris out=raniris;
  by xrn;
run;

* take the last 75 for the training set;
data train;
  set raniris;
  if _n_ > 75;
run;

* take the first 75 for the test set;
data test;
  set raniris;
  if _n_ <=  75;
run;

* use proc discrim to create a classification rule using the training set;
*   then test the quality of the rule on the test set;
* pool=yes uses the linear discriminant function with equal covariance;
proc discrim data = train testdata = test pool=yes listerr testlisterr;
  class species;
  testclass species;
  var sepalL sepalW petalL petalW;
run;
```

Training:

```
...

     Generalized Squared Distance to species

From
species        setosa        versic        virgin
setosa              0     100.80048     206.47721
versic      100.80048             0      21.17140
virgin      206.47721      21.17140             0

...

Classification Results for Calibration Data: WORK.TRAIN
Resubstitution Results using Linear Discriminant Function

...

      Posterior Probability of Membership in species
                 Classified
       From      into
Obs    species   species       setosa    versic    virgin
 23    versic    virgin   *    0.0000    0.3236    0.6764

* Misclassified observation
```

Testing:

```
Classification Results for Test Data: WORK.TEST
Classification Results using Linear Discriminant Function

...

      Posterior Probability of Membership in species
                  Classified
       From       into
Obs    species    species       setosa    versic    virgin
 53    virgin     versic   *    0.0000    0.8993    0.1007
 68    versic     virgin   *    0.0000    0.4838    0.5162

* Misclassified observation
```

A jackknife estimate of the misclassification rate (using all data) was obtained using the **SAS** code given below. Regardless of the analysis, the error rate is very low.
Program editor contents:

```
* crossvalidate performs jackknife error estimation;
* crosslisterr gives the posterior probability for the misclassified observations;
proc discrim data=iris listerr crossvalidate crosslisterr;
  class species;
  var sepalL sepalW petalL petalW;
run;
```

Output window contents:

```
...

Classification Results for Calibration Data: WORK.IRIS
Resubstitution Results using Linear Discriminant Function

...

      Posterior Probability of Membership in species
                  Classified
       From       into
Obs    species    species       setosa    versic    virgin
  5    virgin     versic   *    0.0000    0.7289    0.2711
  9    versic     virgin   *    0.0000    0.2531    0.7469
 12    versic     virgin   *    0.0000    0.1415    0.8585

* Misclassified observation

...

Number of Observations and Percent Classified into species
From
species       setosa       versic       virgin        Total
setosa            50            0            0           50
                 100.00         0.00         0.00       100.00
versic             0           48            2           50
                   0.00        96.00         4.00       100.00
virgin             0            1           49           50
                   0.00         2.00        98.00       100.00
Total             50           49           51          150
                  33.33        32.67        34.00       100.00
Priors         0.33333      0.33333      0.33333

              Error Count Estimates for species
                   setosa       versic       virgin        Total
Rate               0.0000       0.0400       0.0200       0.0200
Priors             0.3333       0.3333       0.3333
```

Question: Which of the two cross-validation methods is preferred?

In future analyses, I will usually report only the cross-validation results, and not the resubstitution results.

17.5 Example: Analysis of Admissions Data

The admissions officer of a business school has used an index of undergraduate GPA and management aptitude test scores (GMAT) to help decide which applicants should be admitted to graduate school. The data below gives the GPA and GMAT scores for recent applicants who are classified as admit (A), borderline (B), or not admit (N). An equal number of A, B, and N's (roughly) were selected from their corresponding populations (Johnson and Wichern, 1988).

The officer wishes to use these data to develop a more quantitative (i.e., less subjective) approach to classify prospective students. Historically, about 20% of all applicants have been admitted initially, 10% are classified as borderline, and the remaining 70% are not admitted. The officer would like to keep these percentages roughly the same in the future.

This is a natural place to use discriminant analysis. Let us do a more careful analysis here, paying attention to underlying assumptions of normality and equal covariance matrices. I will begin with a simple program for summarizing the data.

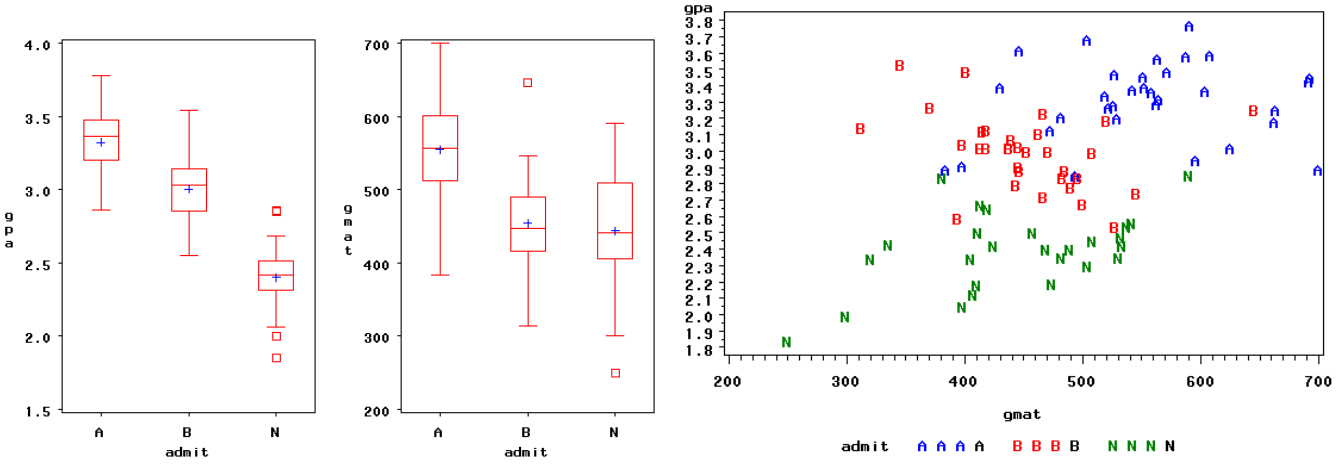
Program editor contents:

```
options ls=79 nodate nocenter;

data business;
  input admit $ gpa gmat @@;  *$;
  datalines;
  A 2.96 596   A 3.14 473   A 3.22 482   A 3.29 527
  A 3.69 505   A 3.46 693   A 3.03 626   A 3.19 663
  A 3.63 447   A 3.59 588   A 3.30 563   A 3.40 553
  A 3.50 572   A 3.78 591   A 3.44 692   A 3.48 528
  A 3.47 552   A 3.35 520   A 3.39 543   A 3.28 523
  A 3.21 530   A 3.58 564   A 3.33 565   A 3.40 431
  A 3.38 605   A 3.26 664   A 3.60 609   A 3.37 559
  A 2.90 384   A 2.86 494   A 2.92 399   A 2.90 700
  N 2.00 300   N 1.85 250   N 2.86 590   N 2.85 381
  N 2.43 425   N 2.20 474   N 2.36 531   N 2.57 542
  N 2.35 406   N 2.51 412   N 2.51 458   N 2.06 399
  N 2.36 482   N 2.66 420   N 2.68 414   N 2.48 533
  N 2.46 509   N 2.43 534   N 2.44 336   N 2.13 408
  N 2.41 469   N 2.55 538   N 2.31 505   N 2.41 489
  N 2.19 411   N 2.35 321   B 2.60 394   B 2.55 528
  B 2.85 496   B 3.14 419   B 3.28 371   B 2.89 447
  B 3.15 313   B 3.50 402   B 2.89 485   B 2.80 444
  B 3.13 416   B 3.01 471   B 2.79 490   B 2.69 501
  B 2.91 446   B 2.75 546   B 2.73 467   B 3.12 463
  B 3.08 440   B 3.03 419   B 3.00 509   B 3.03 438
  B 3.05 399   B 2.85 483   B 3.01 453   B 3.03 414
  B 3.20 521   B 3.26 646   B 3.24 467   B 3.54 346
  B 3.04 446
;

proc sort data=business;
   by admit;
run;

symbol1 v=plus c=blue  i=none;
proc boxplot data=business;
  plot (gpa gmat)*admit /boxstyle=schematic;
run;

* scatterplot;
* define v=symbol, c=color, and i=interpolation;
symbol1 v="A" c=blue  i=none;
symbol2 v="B" c=red   i=none;
symbol3 v="N" c=green i=none;
proc gplot data=business;
```

```
   plot gpa*gmat=admit;
run;
```

The GPA and GMAT distributions are reasonably symmetric. Although a few outliers are present, it does not appear that any transformation will eliminate the outliers and preserve the symmetry. Given that the outliers are not very extreme, I would analyze the data on this scale. Except for the outliers, the spreads (IQRs) are roughly equal across groups within GPA and GMAT. I will look carefully at the variance-covariance matrices later.

There is a fair amount of overlap between the borderline and other groups, but this should not be too surprising. Otherwise these applicants would not be borderline!



17.5.1 Further Analysis of the Admissions Data

Summary information from `stepdisc` is given below. The assumption of constant variance-covariance matrices is suspect. The GPA and GMAT variances are roughly constant across groups, but the correlation between GPA and GMAT varies greatly over groups.

The `candisc` and `stepdisc` procedures do not allow you to relax the assumption of constant variance-covariance matrices. You must assume equal variance-covariance matrices when selecting variables and testing hypotheses. A formal test of equal variance-covariance matrices is possible in `discrim`.

Assuming equal variance-covariance matrices, both GPA and GMAT are important for discriminating among entrance groups. This is consistent with the original data plots.

Program editor contents:

```
proc stepdisc data=business simple wcorr;
  class admit;
  var gpa gmat;
run;
```

Output window contents:

```
...
Within-Class Correlation Coefficients  /  Pr > |r|

             admit = A
Variable            gpa             gmat
gpa             1.00000          0.14656
                                  0.4235
gmat            0.14656          1.00000
                 0.4235
-------------------------------------------------------------------------------
             admit = B
Variable            gpa             gmat
gpa             1.00000         -0.33107
                                  0.0689
gmat           -0.33107          1.00000
                 0.0689
-------------------------------------------------------------------------------
             admit = N
Variable            gpa             gmat
gpa             1.00000          0.50651
                                  0.0083
gmat            0.50651          1.00000
                 0.0083
-------------------------------------------------------------------------------

Simple Statistics
                                Total-Sample
                                                                  Standard
Variable          N           Sum          Mean      Variance    Deviation
gpa              89     261.85000       2.94213       0.19559       0.4423
gmat             89         43358     487.16854          8213      90.6245
-------------------------------------------------------------------------------
                                admit = A
                                                                  Standard
Variable          N           Sum          Mean      Variance    Deviation
gpa              32     106.30000       3.32188       0.05867       0.2422
gmat             32         17741     554.40625          6480      80.4984
-------------------------------------------------------------------------------
                                admit = B
                                                                  Standard
Variable          N           Sum          Mean      Variance    Deviation
gpa              31      93.14000       3.00452       0.05423       0.2329
gmat             31         14080     454.19355          4003      63.2674
-------------------------------------------------------------------------------
                                admit = N
                                                                  Standard
Variable          N           Sum          Mean      Variance    Deviation
gpa              26      62.41000       2.40038       0.05603       0.2367
gmat             26         11537     443.73077          6973      83.5055
-------------------------------------------------------------------------------

...
                          Stepwise Selection Summary
                                                                        Average
                                                                        Squared
     Number                    Partial                        Wilks'   Pr <   Canonical   Pr >
Step     In Entered Removed R-Square F Value Pr > F     Lambda Lambda Correlation   ASCC
   1      1 gpa               0.7184  109.72 <.0001 0.28155501 <.0001  0.35922250 <.0001
   2      2 gmat              0.1618    8.20 0.0006 0.23600494 <.0001  0.42937837 <.0001
```

The two linear combinations of GPA and GMAT that best distinguish the groups are

$$\text{CAN1} = 3.978\ \text{GPA} + 0.003\ \text{GMAT},$$

which is a weighted average of GPA and GMAT, and

$$\text{CAN2} = -1.483\ \text{GPA} + 0.013\ \text{GMAT},$$

which is a comparison of GPA and GMAT.

The groups are significantly different on each canonical variable (p-values are starred), so CAN1 and CAN2 each provide information for distinguishing among groups. The plot of CAN1*CAN2 shows that CAN1 is more informative for distinguishing the groups, as expected. The loadings on CAN1 are positive, so it is not surprising that the CAN1 scores for the A group are generally

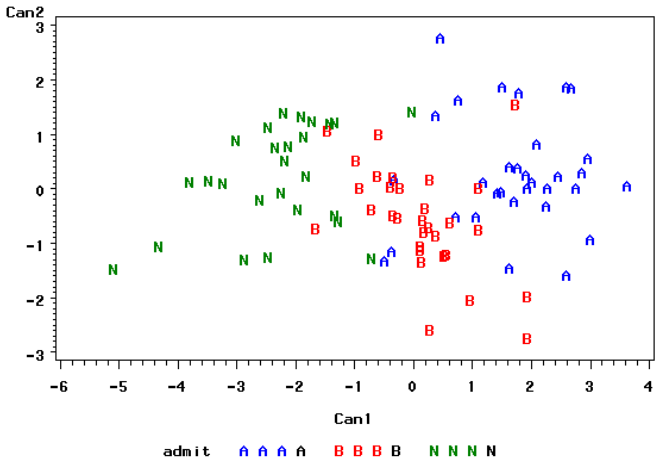
highest, whereas the CAN1 scores for the N group are typically lowest. A **SAS** program and output follow.

Program editor contents:

```
proc candisc data=business out=outcandisc;
  class admit;
  var gpa gmat;
run;

* scatterplot;
* define v=symbol, c=color, and i=interpolation;
symbol1 v="A" c=blue  i=none;
symbol2 v="B" c=red   i=none;
symbol3 v="N" c=green i=none;
proc gplot data=outcandisc;
  plot can2*can1=admit;
run;
```

Output window contents:

```
...
                Multivariate Statistics and F Approximations
                              S=2    M=-0.5    N=41.5
Statistic                        Value    F Value    Num DF    Den DF    Pr > F
Wilks' Lambda                0.23600494      44.98         4       170    <.0001
Pillai's Trace               0.85875673      32.36         4       172    <.0001
Hotelling-Lawley Trace       2.83567542      60.03         4    100.97    <.0001
Roy's Greatest Root          2.68619873     115.51         2        86    <.0001

         NOTE: F Statistic for Roy's Greatest Root is an upper bound.
                NOTE: F Statistic for Wilks' Lambda is exact.

...

                Test of H0: The canonical correlations in the
                   current row and all that follow are zero
            Likelihood    Approximate
                 Ratio        F Value    Num DF    Den DF    Pr > F
       1    0.23600494          44.98         4       170    <.0001
       2    0.86996110          12.85         1        86    0.0006

...

         Raw Canonical Coefficients
Variable                Can1                Can2
gpa            3.977912929      -1.483464562
gmat           0.003057846       0.012923188
```



17.5.2 Classification Using Unequal Prior Probabilities

A new twist to this problem is that we have prior information on the relative sizes of the populations. Historically 20% of applicants are admitted, 10% are borderline, and 70% are not admitted. This prior information is incorporated into the classification rule by using a `priors` statement with `discrim`. The prior probabilities are assumed to be equal when this statement is omitted. The classification rule for unequal prior probabilities uses both the M-distance and the prior probability for making decisions, as outlined below.

When the prior probabilities are unequal, classification is based on the **generalized distance** to group j:

$$D_j^2(X) = (X - \overline{X}_j)' S^{-1} (X - \overline{X}_j) - 2\log(\text{PRIOR}_j),$$

or on the estimated posterior probability of membership in group j:

$$\Pr(j|X) = \frac{\exp\{-0.5 D_j^2(X)\}}{\sum_k \exp\{-0.5 D_k^2(X)\}}.$$

Here S is the pooled covariance matrix, and $\log(\text{PRIOR}_j)$ is the (natural) log of the prior probability of being in group j. As before, you classify observation X into the group that it is closest to in terms of generalized distance, or equivalently, into the group with the maximum posterior probability.

Note that $-2\log(\text{PRIOR}_j)$ exceeds zero, and is extremely large when PRIOR_j is near zero. The generalized distance is the M-distance plus a penalty term that is large when the prior probability for a given group is small. If the prior probabilities are equal, the penalty terms are equal so the classification rule depends only on the M-distance.

The penalty makes it harder (relative to equal probabilities) to classify into a low probability group, and easier to classify into high probability groups. In the admissions data, an observation has to be very close to the B or A groups to not be classified as N.

A **SAS** program for classification with unequal priors is given below. The analysis makes the tenuous assumption that the population covariance matrices are equal. We also have 6 new observations that we wish to classify. These observations are entered as a test data set with missing class levels. Some comments follow the output.

Program editor contents:

```
* new observations to classify;
data test;
  input admit $ gpa gmat;     *$;
  datalines;
  .  2.7  630
  .  3.3  450
  .  3.4  540
  .  2.8  420
  .  3.5  340
  .  3.0  500
;
run;

proc discrim data=business testdata=test pool=yes
             crossvalidate crosslisterr testlisterr;
   class admit;
   testclass admit;
   var gpa gmat;
   priors A=0.2 B=0.1 N=0.7;
run;
```

Output window contents:

...

Class Level Information

admit	Variable Name	Frequency	Weight	Proportion	Prior Probability
A	A	32	32.0000	0.359551	0.200000
B	B	31	31.0000	0.348315	0.100000
N	N	26	26.0000	0.292135	0.700000

...

Pairwise Generalized Squared Distances Between Groups

$$D^2(i|j) = (\bar{X}_i - \bar{X}_j)' \, COV^{-1} \, (\bar{X}_i - \bar{X}_j) - 2 \ln PRIOR_j$$

Generalized Squared Distance to admit

From admit	A	B	N
A	3.21888	7.74593	16.74967
B	6.35963	4.60517	7.22255
N	19.25519	11.11437	0.71335

...

Classification Results for Calibration Data: WORK.BUSINESS
Cross-validation Results using Linear Discriminant Function <- cross-validation

Posterior Probability of Membership in admit

Obs	From admit	Classified into admit		A	B	N
29	A	B	*	0.0273	0.5570	0.4157
30	A	N	*	0.1144	0.3828	0.5027
31	A	B	*	0.0454	0.5949	0.3597
33	B	N	*	0.0009	0.0437	0.9553
34	B	N	*	0.0024	0.0160	0.9815
35	B	N	*	0.1258	0.3287	0.5455
40	B	A	*	0.7471	0.2527	0.0002
41	B	N	*	0.1605	0.4117	0.4277
42	B	N	*	0.0366	0.2770	0.6863
45	B	N	*	0.0591	0.2271	0.7138
46	B	N	*	0.0170	0.0887	0.8942
48	B	N	*	0.0701	0.1291	0.8008
49	B	N	*	0.0197	0.1463	0.8340
56	B	N	*	0.1072	0.3443	0.5485
59	B	A	*	0.7581	0.2326	0.0094
60	B	A	*	0.9788	0.0206	0.0006
61	B	A	*	0.6273	0.3645	0.0082
62	B	A	*	0.6326	0.3673	0.0001
66	N	A	*	0.4420	0.2729	0.2851
67	N	B	*	0.0327	0.5164	0.4508

* Misclassified observation

```
...
Cross-validation Summary using Linear Discriminant Function

  Number of Observations and Percent Classified into admit
From
admit              A             B             N         Total
A                 29             2             1            32
               90.63          6.25          3.13        100.00
B                  5            16            10            31
               16.13         51.61         32.26        100.00
N                  1             1            24            26
                3.85          3.85         92.31        100.00
Total             35            19            35            89
               39.33         21.35         39.33        100.00
Priors           0.2           0.1           0.7

               Error Count Estimates for admit
                      A             B             N         Total
Rate             0.0938        0.4839        0.0769        0.1210
Priors           0.2000        0.1000        0.7000

Classification Results for Test Data: WORK.TEST                    <--- test data

       Posterior Probability of Membership in admit
       From        Classified
Obs    admit       into admit             A             B             N
  1                N             *     0.1015        0.0744        0.8241
  2                A             *     0.6294        0.3665        0.0040
  3                A             *     0.9190        0.0807        0.0004
  4                N             *     0.0261        0.2974        0.6764
  5                B             *     0.4611        0.5384        0.0005
  6                B             *     0.3853        0.4669        0.1478
```

About 1/2 of the borderlines in the calibration set are misclassified. This is due to the overlap of the B group with the other 2 groups, but also reflects the low prior probability for the borderline group. The classification rule requires strong evidence that an observation is borderline before it can be classified as such.

The test data cases were entered with missing group IDs. The **SAS** classification table compares the group IDs, which are unknown, to the ID for the group into which an observation is classified. These two labels differ, so all the test data cases are identified as misclassified. Do not be confused by this! Just focus on the classification for each case, and ignore the other summaries.

Except for observation 5, the posterior probabilities for the test cases give overwhelming evidence in favor of classification into a specific group.

17.5.3 Classification With Unequal Covariance Matrices

The assumption of multivariate normal populations with equal covariance matrices is often unrealistic. Although there are no widely available procedures for MANOVA or stepwise variable selection in discriminant analysis that relax these assumptions, there are a variety of classification rules that weaken one or both of these assumptions.

The **discrim** procedure includes a **quadratic discriminant classification rule** that assumes normality but **allows unequal covariance matrices**. This rule is invoked by specifying `pool=no` in the `discrim` statement. Alternatively, you can test equality of the population covariance matrices by specifying `pool=test`. The quadratic discriminant rule is used when the test is significant at the default 0.10 level. If the test is not significant, then the classification is based on the standard rule using the pooled covariance matrix (called **linear discrimination**.

The quadratic discriminant classification rule is based on the generalized distance to group j:

$$D_j^2(X) = (X - \overline{X}_j)' S_j^{-1} (X - \overline{X}_j) - 2\log(\text{PRIOR}_j) + \log|S_j|,$$

or equivalently, the posterior probability of membership in group j:

$$\Pr(j|X) = \frac{\exp\{-0.5 D_j^2(X)\}}{\sum_k \exp\{-0.5 D_k^2(X)\}}.$$

Here S_j is the sample covariance matrix from group j and $\log|S_j|$ is the log of the determinant of this covariance matrix. The determinant penalty term is large for groups having large variability. The rule is not directly tied to linear discriminant function variables, so interpretation and insight into this method is less straightforward.

There is evidence that quadratic discrimination does not improve misclassification rates in many problems with small to modest sample sizes, in part, because the quadratic rule requires an estimate of the covariance matrix for each population. A modest to large number of observations is needed to accurately estimate variances and correlations. I often compute the linear and quadratic rules, but use the linear discriminant analysis unless the quadratic rule noticeably reduces the misclassification rate.

The following **SAS** `discrim` procedure was used with the data step given earlier to test for equal covariance matrices in the admissions data. The within group covariance and correlation matrices were printed using `wcov` and `wcorr`. The pooled covariance and correlation matrices were printed using the `pcov` and `pcorr` options.

The GPA and GMAT sample variances are roughly constant across admission groups, but the correlation between GPA and GMAT varies widely across groups. The test for equal population covariance matrices is significant, so the `pool=test` analysis uses quadratic discrimination. As an aside, note that the pooled covariance matrix is roughly the average of the within group covariance matrices, and the pooled correlation matrix is roughly an average of the within group correlation matrices.

The quadratic rule does not classify the training data noticeably better than the linear discriminant analysis. The individuals in the test data have the same classifications under both approaches. Assuming that the optimistic error rates for the two rules were "equally optimistic", I would be satisfied with the standard linear discriminant analysis, and would summarize my analysis based on this approach. Additional data is needed to decide whether the quadratic rule might help reduce the misclassification rates.

Program editor contents:

```
proc discrim data=business testdata=test wcov wcorr pcov pcorr
             pool=test crossvalidate crosslisterr testlisterr;
  class admit;
  testclass admit;
  var gpa gmat;
  priors A=0.2 B=0.1 N=0.7;
run;
```

Output window contents:

```
                    Class Level Information
          Variable                                              Prior
admit     Name         Frequency       Weight    Proportion    Probability
A         A                   32      32.0000      0.359551       0.200000
B         B                   31      31.0000      0.348315       0.100000
N         N                   26      26.0000      0.292135       0.700000
```

Within-Class Covariance Matrices

admit = A, DF = 31

Variable	gpa	gmat
gpa	0.058667	2.857601
gmat	2.857601	6479.990927

admit = B, DF = 30

Variable	gpa	gmat
gpa	0.054226	-4.877570
gmat	-4.877570	4002.761290

admit = N, DF = 25

Variable	gpa	gmat
gpa	0.056028	10.011708
gmat	10.011708	6973.164615

Pooled Within-Class Covariance Matrix, DF = 86

Variable	gpa	gmat
gpa	0.056351	2.238968
gmat	2.238968	5759.205498

Within-Class Correlation Coefficients / Pr > |r|

admit = A

Variable	gpa	gmat
gpa	1.00000	0.14656 0.4235
gmat	0.14656 0.4235	1.00000

admit = B

Variable	gpa	gmat
gpa	1.00000	-0.33107 0.0689
gmat	-0.33107 0.0689	1.00000

admit = N

Variable	gpa	gmat
gpa	1.00000	0.50651 0.0083
gmat	0.50651 0.0083	1.00000

Pooled Within-Class Correlation Coefficients / Pr > |r|

Variable	gpa	gmat
gpa	1.00000	0.12428 0.2514
gmat	0.12428 0.2514	1.00000

```
Test of Homogeneity of Within Covariance Matrices

Notation: K    = Number of Groups
          P    = Number of Variables
          N    = Total Number of Observations - Number of Groups
          N(i) = Number of Observations in the i'th Group - 1
                   __                                N(i)/2
                   ||  |Within SS Matrix(i)|
          V    = ----------------------------------
                                  N/2
                       |Pooled SS Matrix|

                       _                          _       2
                      |       1         1          |  2P + 3P - 1
          RHO  = 1.0 - | SUM -----  -  ---   | -------------
                      |_     N(i)       N   _|  6(P+1)(K-1)

          DF   = .5(K-1)P(P+1)

                                                 _                  _
                                                |    PN/2             |
                                                |   N        V        |
Under the null hypothesis:      -2 RHO ln       | ------------------ |
                                                |   __       PN(i)/2  |
                                                |_  ||  N(i)         _|

is distributed approximately as Chi-Square(DF).

  Chi-Square        DF    Pr > ChiSq
   12.320968         6        0.0552

Since the Chi-Square value is significant at the 0.1 level, the within
covariance matrices will be used in the discriminant function.
Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.
```

Because the covariance matrices are statistically different, quadratic discrimination will be used.

```
Pairwise Generalized Squared Distances Between Groups
 2                                _    _'  -1  _    _                          _
D (i|j) = (X - X )' COV    (X - X ) + ln |COV | - 2 ln PRIOR
           i   j      j    i   j           j             j

     Generalized Squared Distance to admit
From
admit                A             B             N
A              9.13776      16.37831      22.10278
B             11.98738       9.86921      14.73233
N             24.29429      17.77821       6.38481
```

Classification Summary for Calibration Data: WORK.BUSINESS
Resubstitution Summary using Quadratic Discriminant Function <--- Quad discrim

Generalized Squared Distance Function

$$D_j^2(X) = (X-\bar{X}_j)' \, COV_j^{-1} \, (X-\bar{X}_j) + \ln |COV_j|$$

Posterior Probability of Membership in Each admit

$$Pr(j|X) = \exp(-.5\, D_j^2(X)) \,/\, SUM_k \exp(-.5\, D_k^2(X))$$

...

Posterior Probability of Membership in admit

Obs	From admit	Classified into admit		A	B	N
2	A	B	*	0.4716	0.5175	0.0109
29	A	B	*	0.0443	0.8001	0.1556
30	A	N	*	0.0903	0.4429	0.4668
31	A	B	*	0.0639	0.8115	0.1246
32	A	N	*	0.1962	0.0068	0.7971
33	B	N	*	0.0024	0.0145	0.9831
34	B	N	*	0.0025	0.0252	0.9723
35	B	N	*	0.1114	0.3929	0.4957
40	B	A	*	0.6853	0.3147	0.0000
42	B	N	*	0.0534	0.3844	0.5622
45	B	N	*	0.0545	0.2840	0.6616
46	B	N	*	0.0149	0.1122	0.8729
48	B	N	*	0.0436	0.1228	0.8336
49	B	N	*	0.0229	0.1880	0.7892
56	B	N	*	0.1040	0.4335	0.4625
59	B	A	*	0.8385	0.1527	0.0088
60	B	A	*	0.9911	0.0000	0.0089
61	B	A	*	0.6211	0.3771	0.0018
62	B	A	*	0.5598	0.4402	0.0000
67	N	B	*	0.1084	0.8269	0.0647

* Misclassified observation

Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into admit

From admit	A	B	N	Total
A	27	3	2	32
	84.38	9.38	6.25	100.00
B	5	17	9	31
	16.13	54.84	29.03	100.00
N	0	1	25	26
	0.00	3.85	96.15	100.00
Total	32	21	36	89
	35.96	23.60	40.45	100.00
Priors	0.2	0.1	0.7	

Error Count Estimates for admit

	A	B	N	Total
Rate	0.1563	0.4516	0.0385	0.1033
Priors	0.2000	0.1000	0.7000	

Classification Results for Test Data: WORK.TEST
Classification Results using Quadratic Discriminant Function

	Posterior Probability of Membership in admit				
Obs	From admit	Classified into admit		A	B	N
1	N		*	0.0432	0.0379	0.9188
2	A		*	0.5972	0.4024	0.0004
3	A		*	0.9778	0.0219	0.0003
4	N		*	0.0509	0.4231	0.5260
5	B		*	0.2921	0.7079	0.0000
6	B		*	0.3632	0.5135	0.1233

* Misclassified observation