

Part I. (100 points) Do all calculations in SAS. Use a word processor of your choice to write a report. Insert computer text output and graphics to support what you are saying, but you need to write something that looks like an academic paper — not a pile of computer output. Turn in a hard copy of your HW in class (i.e., don't email me your HW). Also:

1. Clearly specify parameters and hypotheses when appropriate.
2. Write a coherent conclusion.

(100^{pts})

1. Kangaroo skull regression analysis

The data to be analyzed here are selected skull measurements on 148 kangaroos of known sex and species. (See HW 5 for data.) There are 11 columns of data, corresponding to the following features. Columns, from left to right:

1. sex (1=M, 2=F)
2. species (0=M.\ giganteus, 1=M.f.\ melanops, 2=M.f.\ fuliginosus)
3. post orbit width
4. rostral width
5. supra-occipital - paroccipital depth
6. crest width
7. incisive foramina length
8. mandible length
9. mandible width
10. mandible depth
11. ascending ramus height (cols 3-11 are in mm times 10)

The first 4 observations in the data set are given below. Some of the observations in the data set are missing. These are represented by the SAS default missing value id of a period.

```
1 0 249 227 531 153 88 1086 131 179 591
1 0 233 248 632 141 100 1158 148 181 643
1 0 244 240 575 144 107 1131 116 169 610
1 0 224 242 568 116 79 1090 132 189 594
```

I am interested in you building a model, or models, that relates the (9) mandible width (i.e., the response) to the (8) mandible length and (10) mandible depth. Furthermore, I wish to understand how this relationship depends, if at all, on both the (2) species and the (1) sex of the kangaroo.

You can use any tools to answer this question. However, there is a lot you can do with this problem using simple plots and simple analyses.

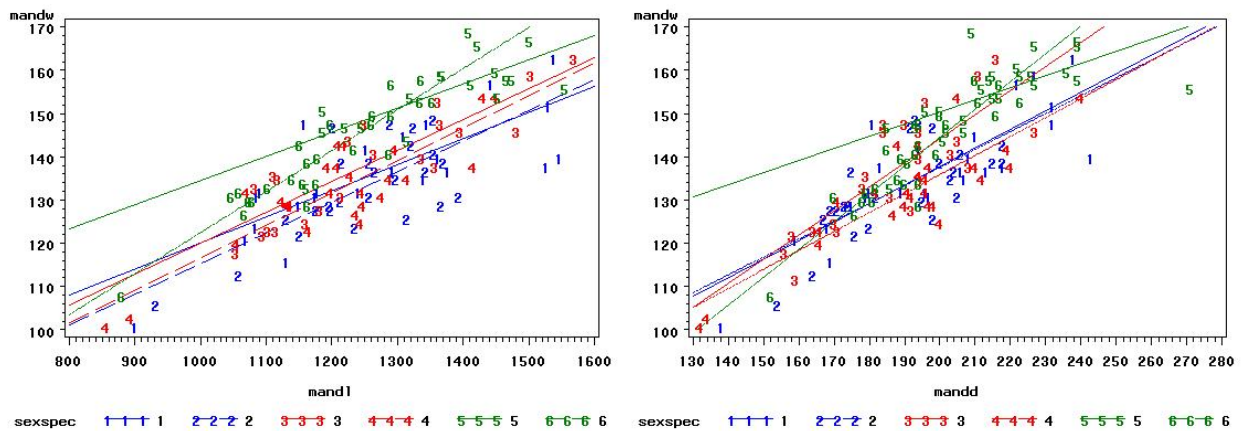
Write up a careful summary of your analysis and provide a careful summarization/interpretation of your findings. Below is a recommended strategy.

- (a) (20 pts) Plot mandible width against each of mandible length and mandible depth with colors/symbols indicating species and sex. Discuss the relationships you see, as well as any need for transformation.

Solution: In these plots, odd numbers are males and even are females, color indicates the three species, and line types are different for sexes.

100 pts

20 pts



Plots of the response variable mandw (mandible width) against each of mandl (mandible length) and mandd (mandible depth) show fairly strong linear relationships, with some differences across sexes and species. The differences across species seems to be larger than the differences between males and females. There are no obvious extreme points and no obvious need to transform the data at this point.

Program editor contents:

```
options ls=79 nodate nocenter;
** Kangaroo analysis;
data kang;
  infile "F:\Dropbox\UNM\teach\ADA2_stat528\assess\ADA2_HW_05_kang.dat";
  input sex spec postow rostw supocpd crestw incfl mandl mandw mandd ascrh;
  obs = _N_;
  sexspec = spec*2+sex; * odd is male, even is female for 3 species;
run;

proc sort data=kang;
  by sex spec;
run;

* define v=symbol, c=color, i=interpolation, l=line style;
symbol1 v="1" c=blue i=r l=1;
symbol2 v="2" c=blue i=r l=2;
symbol3 v="3" c=red i=r l=1;
symbol4 v="4" c=red i=r l=2;
symbol5 v="5" c=green i=r l=1;
symbol6 v="6" c=green i=r l=2;
proc gplot data=kang;
  plot mandw*(mandl mandd)=sexspec;
run;
```

- (b) (25 pts) Perform a model selection process to arrive at a parsimonious explanation of the data.

Solution: I began the model building process using a model with all main effects (mandl, mandd, sex, and spec), all 6 interactions between two effects, and all 4 interactions between three effects. SAS output from this model is given below.

Program editor contents:

```
proc glm data=kang;
  class sex spec;
  model mandw =
    sex spec mandl mandd
    sex*spec sex*mandl sex*mandd spec*mandl spec*mandd mandl*mandd
    sex*spec*mandl sex*spec*mandd sex*mandl*mandd spec*mandl*mandd;
run;
```

Output window contents:

The GLM Procedure

Class Level Information		
Class	Levels	Values
sex	2	1 2
spec	3	0 1 2

Number of Observations Read 148
 Number of Observations Used 136

Dependent Variable: mandw

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	21	20236.64696	963.64986	25.18	<.0001
Error	114	4363.47069	38.27606		
Corrected Total	135	24600.11765			

R-Square 0.822624 Coeff Var 4.479344 Root MSE 6.186765 mandw Mean 138.1176

Source	DF	Type III SS	Mean Square	F Value	Pr > F
sex	1	25.9113194	25.9113194	0.68	0.4124
spec	2	15.8683336	7.9341668	0.21	0.8131
mandl	1	659.4631278	659.4631278	17.23	<.0001
mandd	1	485.4758940	485.4758940	12.68	0.0005
sex*spec	2	30.9951935	15.4975967	0.40	0.6680
mandl*sex	1	30.0642085	30.0642085	0.79	0.3773
mandd*sex	1	23.7162571	23.7162571	0.62	0.4328
mandl*spec	2	30.2901450	15.1450725	0.40	0.6741
mandd*spec	2	9.0170839	4.5085419	0.12	0.8890
mandl*mandd	1	315.0923452	315.0923452	8.23	0.0049
mandl*sex*spec	2	48.8890906	24.4445453	0.64	0.5299
mandd*sex*spec	2	15.7903595	7.8951797	0.21	0.8139
mandl*mandd*sex	1	27.8808737	27.8808737	0.73	0.3952
mandl*mandd*spec	2	12.5345715	6.2672858	0.16	0.8492

Using the full model as a starting point, I performed a hierarchical backwards elimination of effects. The following effects were omitted, in the given order:

- mandl*mandd*spec
- mandd*sex*spec
- mandl*SEX*spec
- mandl*mandd*SEX
- mandd*sex
- sex*spec
- mandl*sex
- mandd*spec
- mandl*spec

Program editor contents:

```
* remove mandl*mandd*spec;
proc glm data=kang;
class sex spec;
model mandw =
sex spec mandl mandd
sex*spec sex*mandl sex*mandd spec*mandl spec*mandd mandl*mandd
sex*spec*mandl sex*spec*mandd sex*mandl*mandd;
run;

* remove mandd*sex*spec;
proc glm data=kang;
class sex spec;
model mandw =
sex spec mandl mandd
sex*spec sex*mandl sex*mandd spec*mandl spec*mandd mandl*mandd
sex*spec*mandl sex*mandl*mandd;
run;

* remove mandl*sex*spec;
proc glm data=kang;
class sex spec;
model mandw =
sex spec mandl mandd
sex*spec sex*mandl sex*mandd spec*mandl spec*mandd mandl*mandd
sex*mandl*mandd;
run;

* remove mandl*mandd*sex;
proc glm data=kang;
class sex spec;
model mandw =
sex spec mandl mandd
sex*spec sex*mandl sex*mandd spec*mandl spec*mandd mandl*mandd;
run;

* remove mandd*sex;
proc glm data=kang;
class sex spec;
model mandw =
sex spec mandl mandd
```

```

sex*spec sex*mandl spec*mandl spec*mandd mandl*mandd;
run;
* remove sex*spec;
proc glm data=kang;
class sex spec;
model mandw =
sex spec mandl mandd
sex*mandl spec*mandl spec*mandd mandl*mandd;
run;
* remove mandl*sex;
proc glm data=kang;
class sex spec;
model mandw =
sex spec mandl mandd
spec*mandl spec*mandd mandl*mandd;
run;
* remove mandd*spec;
proc glm data=kang;
class sex spec;
model mandw =
sex spec mandl mandd
spec*mandl mandl*mandd;
run;
* final reduced model, all effect significant;
proc glm data=kang;
class sex spec;
model mandw = sex spec mandl mandd mandl*mandd;
run;

```

The selected model has sex and spec effects, plus linear terms in mandl, mandd, and a mandl*mandd interaction.

The full model has 21 single degree of freedom regression effects and an $R^2 = 0.822$. The selected model has 6 single degree of freedom regression effects and an $R^2 = 0.806$. The observed decrease in R^2 is negligible relative to the simplicity of interpretation for the selected model.

Output window contents:

The GLM Procedure

Class Level Information		
Class	Levels	Values
sex	2	1 2
spec	3	0 1 2

Number of Observations Read 148
 Number of Observations Used 136

Dependent Variable: mandw

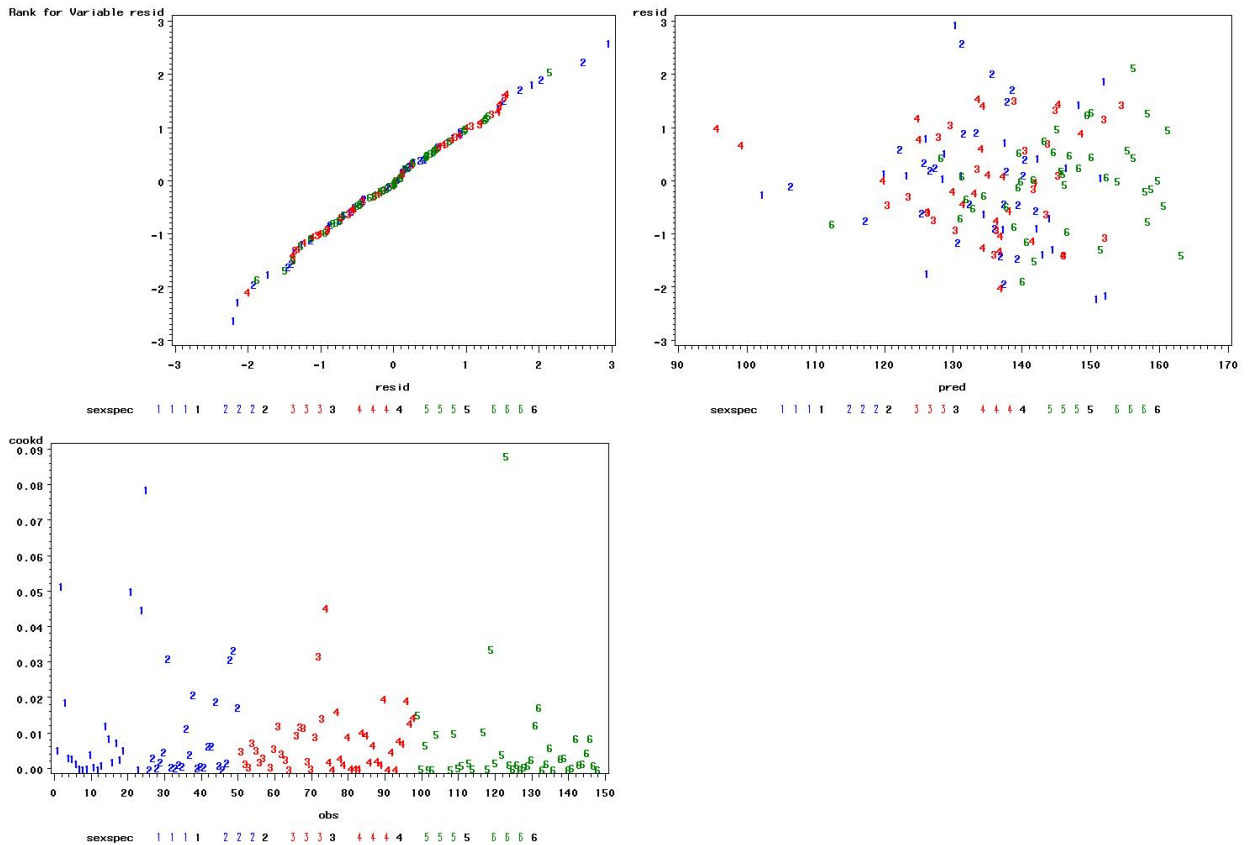
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	19827.21001	3304.53500	89.31	<.0001
Error	129	4772.90764	36.99928		
Corrected Total	135	24600.11765			

R-Square 0.805980 Coeff Var 4.404002 Root MSE 6.082704 mandw Mean 138.1176

Source	DF	Type III SS	Mean Square	F Value	Pr > F
sex	1	350.667318	350.667318	9.48	0.0025
spec	2	2208.861196	1104.430598	29.85	<.0001
mandl	1	1435.013178	1435.013178	38.78	<.0001
mandd	1	962.049561	962.049561	26.00	<.0001
mandl*mandd	1	674.598724	674.598724	18.23	<.0001

(c) (20 pts) Assess assumptions of your candidate final model visually (plots).

Solution: I made QQ (rankit) plots of the studentized residuals, and plots of the studentized residuals against the predicted values, using both sexespec as plotting symbols. The QQ-plot shows at most a modest deviation from normality. Two observations in species 0 (M. giganteus) that are not fit well by the model are seen here and in the plot of residuals against fitted values. The index plot of Cook's D (not provided) shows two potentially influential observations, but these are not very extreme. The overall picture does not suggest any gross deficiencies in the model. The typical residual appears to be about 5% of the mandw.



Program editor contents:

```
proc glm data=kang;
  class sex spec;
  model mandw = sex spec mandl mandd mandl*mandd/solution;
  lsmeans sex spec/pdiff;
  output out=outglm p=pred student=resid cookd=cookd;
run;

* diagnostics;
proc rank data=outglm out=outglm normal=blom;
  var resid;
  ranks rankres;
run;

* define v=symbol, c=color, i=interpolation, l=line style;
symbol1 v="1" c=blue i=none;
symbol2 v="2" c=blue i=none;
symbol3 v="3" c=red i=none;
symbol4 v="4" c=red i=none;
symbol5 v="5" c=green i=none;
symbol6 v="6" c=green i=none;
proc gplot data=outglm;
  plot rankres*resid=sexspec;
  plot resid*pred=sexspec;
  plot cookd*obs=sexspec;
run;
```

- (d) (35 pts) Interpret the model coefficients. Use GLM's LSMEANS to make any sex or species comparisons you feel are relevant.

Solution: The fitted model is fairly easy to understand. The sex baseline category is sex=2 (females). The species baseline category is spec=2 (M.f. fuliginosus). Thus the predicted mandw for females from M.f. fuliginosus is given by:

$$\widehat{\text{mandw}}_{\text{base}} = -64.321 + 0.140\text{mandl} + 0.779\text{mandd} - 0.0005\text{mandl} * \text{mandd}.$$

One could write down one fitted equation for each species and sex combination. However, sex and species do not interact so their effect on predicted values can be understood without resorting to this approach. For example, males (sex=1) are predicted to be 3.427 (mm times 10) larger than females, regardless of species, mandl, and mandd. Similarly, *M. giganteus* (spec=0) is predicted to be 10.190 (mm times 10) smaller than *M.f. fuliginosus* (spec=2), regardless of sex, mandl, and mandd, whereas *M.f. melanops* (spec=1) is predicted to be 7.957 (mm times 10) smaller than *M.f. fuliginosus* (spec=2), regardless of sex, mandl, and mandd.

This analysis provides a simple description of the effect of sex and species. One could take the analysis a bit further by noting that the LSMEANS for *M. giganteus* (spec=0) and *M.f. melanops* (spec=1) are not significantly different from each other. This is consistent with the adjustments for these species, relative to the baseline species, being very similar (-10.19 for *M. giganteus* and -7.957 for *M.f. melanops*). A sensible strategy would then be to combine species 0 and 1 together, so that there are only two species categories, and then to refit the same model. Output from this model leads to similar predictions as above, with the exception that *M. giganteus* and *M.f. melanops* have the same predicted mandw.

Output window contents:

Parameter		Estimate		Standard Error	t Value	Pr > t
Intercept		-64.32164579	B	27.42860286	-2.35	0.0206
sex	1	3.42715529	B	1.11322461	3.08	0.0025
sex	2	0.00000000	B			
spec	0	-10.19039325	B	1.35928231	-7.50	<.0001
spec	1	-7.95713591	B	1.36615525	-5.82	<.0001
spec	2	0.00000000	B			
mandl		0.14029736		0.02252777	6.23	<.0001
mandd		0.77914017		0.15279655	5.10	<.0001
mandl*mandd		-0.00048683		0.00011401	-4.27	<.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

		H0:LSMean1=LSMean2	
sex	mandw	LSMEAN	Pr > t
1		140.221240	0.0025
2		136.794085	

		LSMEAN	
spec	mandw	LSMEAN	Number
0		134.366445	1
1		136.599703	2
2		144.556839	3

Least Squares Means for effect spec
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: mandw			
i/j	1	2	3
1		0.0787	<.0001
2	0.0787		<.0001
3	<.0001	<.0001	

NOTE: To ensure overall protection level, only probabilities associated with pre-planned comparisons should be used.