Advanced Data Analysis 2, Stat 428/528
Spring 2012

Homework 5

Name: Answer Key
Prof. Erik B. Erhardt

Part I. (140 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.

(70$^{\text{pts}}$) **1. Kangaroos skull measurements: mandible length**

70 pts

The data to be analyzed here are selected skull measurements on 299 kangaroos of known sex and species. 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. rostal 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 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 whether there is any effect of sex or species on the mandible length.

(a) (10 pts) Provide side-by-side boxplots of the data, comparing the mandible lengths across the 6 combinations of sex and species. Comment on the distributional shapes and compare the typical mandible lengths across groups.

Solution: 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_;
run;

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

*********************;
** mandible length **;
*********************;

* part (a) *****************************;
proc univariate data=kang normal plot;
  var mandl;
  by sex spec;
  output out=meanmandl mean=meanmandl; * output mean;
run;

* part (b) *****************************;
* print means;
proc print data=meanmandl;
run;

* plot means in two profile plots;
```

10 pts

```
* define v=symbol, c=color, i=interpolation, l=line style;
symbol1 v="0" c=blue  i=join l=1;
symbol2 v="1" c=red   i=join l=2;
symbol3 v="2" c=green i=join l=3;
proc gplot data=meanmandl;
  plot meanmandl*sex=spec;
run;

* 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;
proc gplot data=meanmandl;
  plot meanmandl*spec=sex;
run;

* part (c,d) ****************************;
proc glm data=kang;
  class sex spec;
  model mandl = sex spec sex*spec;
* part (e) ****************************;
  lsmeans sex spec sex*spec/pdiff;
  *means sex spec sex*spec/bon;

* part (f) ****************************;
  * 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;
```

The side-by-side boxplots given below show that the distributions of mandible lengths are fairly symmetric and reasonably normal within groups. Within each species, males tend to be larger than females on average. In addition, the difference in mean lengths between males and females is different across species, with larger sex differences for M.f. fuliginosus (species 2) than for the other two species. This is also clearly seen in the interaction plot, discussed below.
Output window contents:

0 pts

```
The UNIVARIATE Procedure
Variable:  mandl

Schematic Plots

     |
1600 +
     |                |
     |    |           |           |
     |    |           |           |
1500 +    |           |           |
     |    |           |           |
     |    |           |        +-----+                    |
     |    |           |        |     |                    |
1400 +    |           |        *-----*        |           |
     | +-----+        |        |     |        |           |
     | |     |     +-----+     |  +  |        |           |           |
     | |     |     |     |     |     |     +-----+        |           |
1300 + |     |     |     |     |     |     |     |        |           |
     | |  +  |     |     |     |     |     |     |     +-----+     +-----+
     | *-----*     |  +  |     +-----+     *--+--*     |     |     |     |
     | |     |     *-----*        |        |     |     *-----*     |     |
1200 + |     |     |     |        |        +-----+     |  +  |     |     |
     | |     |     |     |        |           |        |     |     *--+--*
     | |     |     |     |        |           |        |     |     |     |
     | +-----+     +-----+                    |        +-----+     |     |
1100 +    |           |                       |           |        +-----+
     |    |           |                       |           |           |
     |    |           |                       |           |           |
     |    |                                                           |
1000 +    |                                                           |
     |    |                                                           |
     |    |                                                           |
     |    |                                   0                       |
 900 +    |                                               0           |
     |                                                                |
     |                                                    0
     |
 800 +
      ----+-----------+-----------+-----------+-----------+-----------+------
spec      0           1           2           0           1           2
 sex      1           1           1           2           2           2
```

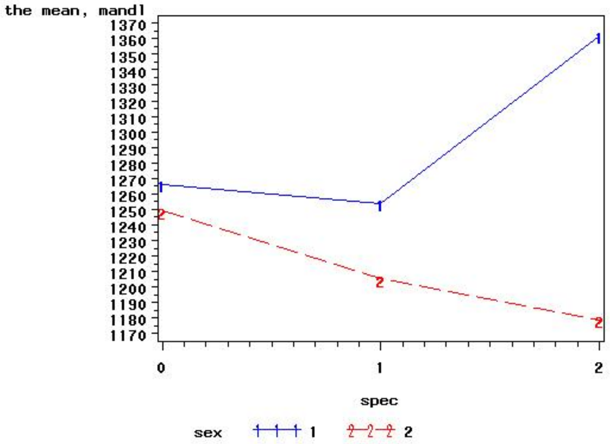
(b) (10 pts) Produce an interaction plot (i.e., profile plot of cell means) for these data. Discuss what you see in the plot.

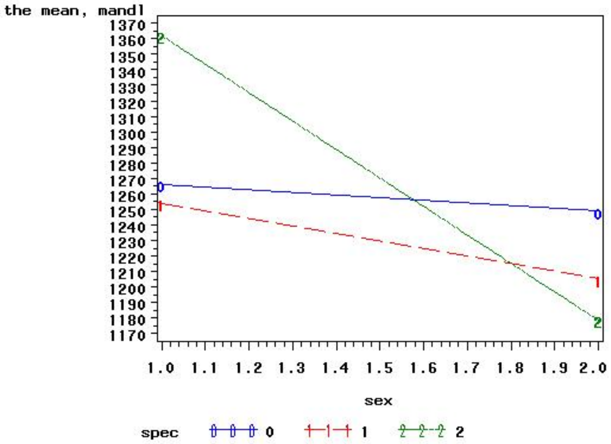
Solution:

The profile plot of sample means (i.e., the interaction plot) shows a clear interaction between sex and species. As noted above, the difference in the mean lengths for males and females depends on the species, with larger sex differences for M.f. fuliginosus than for the other two species, a point clearly seen in the plot below. Another interpretation of the interaction is that differences among species depend on the sex being considered. For example, the largest average mandible length for males is found in M.f. fuliginosus, the same species that produces the lowest average mandible lengths for females.

```
Obs    sex    spec    meanmandl
 1      1       0      1266.26
 2      1       1      1254.05
 3      1       2      1361.83
 4      2       0      1248.84
 5      2       1      1205.20
 6      2       2      1178.46
```

10 pts





(c) (10 pts) Provide an ANOVA table for this two-factor setting.

Solution:

An overall ANOVA table is given below, with Type III SS and F-tests for the sex and species main effect, and the sex by species interaction.

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: mandl
                                            Sum of
Source                      DF           Squares     Mean Square    F Value    Pr > F
Model                        5        403451.353       80690.271       4.30    0.0012
Error                      130       2441207.205       18778.517
Corrected Total            135       2844658.559

Source                      DF       Type III SS     Mean Square    F Value    Pr > F
sex                          1       232267.3688     232267.3688      12.37    0.0006
spec                         2        37880.5099      18940.2549       1.01    0.3676
sex*spec                     2       167853.2326      83926.6163       4.47    0.0133
```

(d) (10 pts) Test for the presence of interaction between sex and species. Also test for the presence of main effects, effects due to the sex and species.

Solution:

There is a statistically significant interaction between sex and species (p-value=0.0133), and a significant main effect for sex (p-value=0.0006), but the species main effect is not significant at any of the standard levels (p-value=0.368).

Note that the analysis does not say that there is no effect of species, only that the average mandible lengths for the species, when averaged over sexes, do not differ significantly. The significant interaction between sex and species reinforces that species do differ in that the size of the differences in mean mandible lengths between sexes depends on species.

(e) (10 pts) Summarize differences, if any, in sexes and species using relevant multiple comparisons. Give clear interpretations of any significant effects.

Solution:

Looking at the LSMEANS output, we see that the sample mean mandible length for males, averaged over species, is 1294. This exceeds the average for females by approximately 8.4mm. Note that the

30 pts

unit is 0.1 mm so a mean difference of 1294-1210 = 84 is 8.4mm. This difference is statistically significant, with a p-value of 0.0006. There is no need for multiple comparisons here because there are only 2 sexes.

```
Least Squares Means
                                H0:LSMean1=
                                  LSMean2
sex    mandl LSMEAN               Pr > |t|
1        1294.04727                 0.0006
2        1210.83278
```

The main effect for species is far from significant, as is each comparison between pairs of species.

```
                              LSMEAN
spec    mandl LSMEAN          Number
0         1257.55043               1
1         1229.62381               2
2         1270.14583               3

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

          Dependent Variable: mandl
i/j              1              2              3
   1                       0.3263         0.6661
   2        0.3263                        0.1713
   3        0.6661         0.1713
```

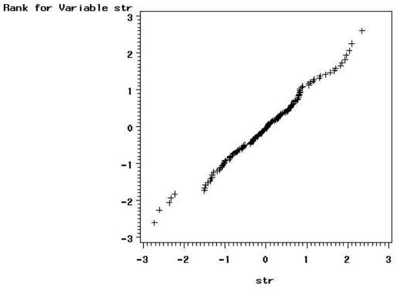
Looking at the individual cell means we see several pairwise comparisons with extremely small p-values, for example the comparison of male and females from M.f. fuliginosus. In this species, male mandible lengths exceed those for females, on average, by about 18.3mm (i.e., 0.1mm*183). The p-value for this comparison (p-value=0.0001) would be significant even after a Bonferroni adjustment.

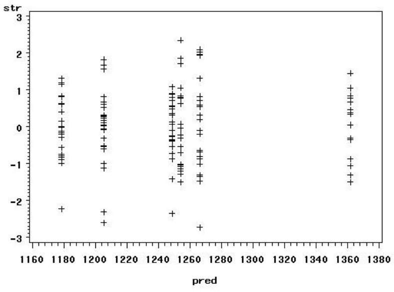
```
                                       LSMEAN
sex     spec    mandl LSMEAN           Number
1       0         1266.26087                1
1       1         1254.04762                2
1       2         1361.83333                3
2       0         1248.84000                4
2       1         1205.20000                5
2       2         1178.45833                6

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

                         Dependent Variable: mandl
i/j            1            2            3            4            5            6
   1                   0.7682       0.0284       0.6607       0.1255       0.0299
   2      0.7682                    0.0157       0.8980       0.2307       0.0672
   3      0.0284       0.0157                    0.0086       0.0003       <.0001
   4      0.6607       0.8980       0.0086                    0.2623       0.0746
   5      0.1255       0.2307       0.0003       0.2623                    0.4959
   6      0.0299       0.0672       <.0001       0.0746       0.4959
```

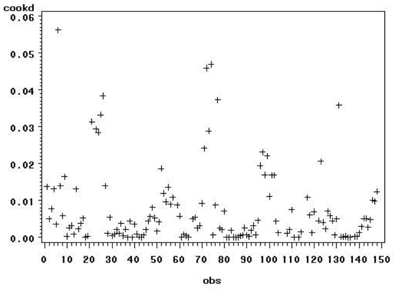
(f) (10 pts) Do the assumptions for a standard two-factor analysis appear reasonable? If not, suggest and carry out an alternative analysis.

Solution:

As noted in the part (a), the side-by-side boxplots show that the distributions of mandible lengths are fairly symmetric and reasonably normal. Although the variability of male mandible lengths appear to be somewhat larger than the variability for females, the differences in spreads across the six sex-species combinations would not appear to be sufficient enough to cause concern with doing a standard normal theory-based analysis on the mandible lengths.







10 pts

(g) (10 pts) Summarize your findings on the effect of sex and species on the mandible lengths.

Solution:

Our analysis shows significant main effects for sex, with males on average being larger than females by approximately 8.4mm, averaging over species. There is no significant species main effect, but a significant interaction between sex and species, which appears to be mostly due to a much larger sex difference in M.f. fuliginosus than in the other two species. In particular, the sample mean mandible lengths of males exceed that for females by 1.8mm, 4.9mm, and 18.3mm for M. giganteus, M.f. melanops, and M.f. fuliginosus, respectively.

(70^{pts}) **2. Kangaroos skull measurements: crest width**

70 pts

Repeat the analysis, using the crest width.

(a) (10 pts) Provide side-by-side boxplots of the data, comparing the crest widths across the 6 combinations of sex and species. Comment on the distributional shapes and compare the typical crest widths across groups.

Solution: Program editor contents:

```
*********************;
**   crest width   **;
*********************;

* part (a) ****************************;
proc univariate data=kang normal plot;
  var crestw;
  by sex spec;
  output out=meancrestw mean=meancrestw; * output mean;
run;

* part (b) ****************************;
* print means;
proc print data=meancrestw;
run;

* plot means in two profile plots;

* define v=symbol, c=color, i=interpolation, l=line style;
symbol1 v="0" c=blue  i=join l=1;
symbol2 v="1" c=red   i=join l=2;
symbol3 v="2" c=green i=join l=3;
proc gplot data=meancrestw;
  plot meancrestw*sex=spec;
run;

* 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;
proc gplot data=meancrestw;
  plot meancrestw*spec=sex;
run;


* part (c,d) ****************************;
proc glm data=kang;
  class sex spec;
  model crestw = sex spec sex*spec;
* part (e) ****************************;
  lsmeans sex spec sex*spec/pdiff;
  *means sex spec sex*spec/bon;

* part (f) ****************************;
  * 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;
```

20 pts

```
run;
```

Except for the heavy tailed distribution of crest widths for males from M. giganteus, the side-by-side boxplots show that the distributions of crest widths are fairly symmetric and reasonably normal. Within each species, females tend to be larger than males on average. Within each sex, M.f. fuliginosus (species 2) tends to be larger, on average, than the other two species, which are similar on this feature.

Output window contents:

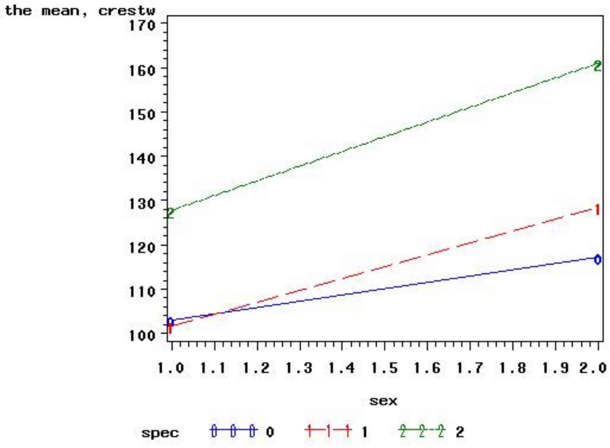
```
The UNIVARIATE Procedure
Variable:  crestw

Schematic Plots

     |
 220 +
     |                                                    |           |
     |                                                    |           |
     |                                        0           |           |
 200 +                                                    |           |
     |                                        0           |           |
     |    |                       |                       |           |
     |    |                       |                       |           |
 180 +    |                       |                       |        +-----+
     |    |                       |                       |        |     |
     |    |                       |                       |        |     |
     |    |                       |           |           |        *-----*
 160 +    |                       |           |           |        |  +  |
     |    |           |           |           |           |        |     |
     |    |           |        +-----+        |        +-----+     |     |
     |    |           |        |     |        |        |     |     +-----+
 140 +    |           |        |     |        |        |     |        |
     |    |        +-----+     |     |        |        |     |        |
     | +-----+     |     |     |  +  |     +-----+     *--+--*        |
     | |     |     |     |     *-----*     |     |     |     |        |
 120 + |     |     |     |     |     |     *-----*     |     |        |
     | |     |     *-----*     |     |     |  +  |     |     |
     | |     |     |     |     +-----+     |     |     |     |
     | *--+--*     |     |        |        |     |     |     |
 100 + |     |     |  +  |        |        +-----+     +-----+
     | |     |     |     |        |           |           |
     | |     |     |     |        |           |           |
     | |     |     |     |        |           |           |
  80 + +-----+     |     |        |           |           |           0
     |    |        |     |        |           |           |
     |    |        +-----+        |           |           |
     |    |           |           |                       |
  60 +    |           |           |
     |    |           |
     |    |           |
     |    |           |
  40 +    |           |
     |    |           |
     |    |           |                       0
     |    |           |
  20 +    |           |
     |                |
     |
     |
   0 +
      ----+-----------+-----------+-----------+-----------+-----------+-----------
spec      0           1           2           0           1           2
 sex      1           1           1           2           2           2
```

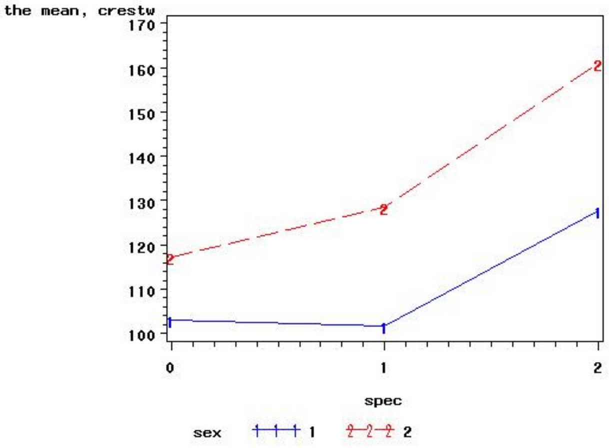
(b) (10 pts) Produce an interaction plot (i.e., profile plot of cell means) for these data. Discuss what you see in the plot.

10 pts

Solution:
The profile plot of the sample means shows that the mean difference between sexes is approximately the same for each species. That is, the profile plot is fairly parallel. I would not be surprised to find that interaction between sex and species is not significant.

```
Obs     sex     spec     meancrestw
 1       1       0        103.080
 2       1       1        101.652
 3       1       2        127.800
 4       2       0        117.160
 5       2       1        128.480
 6       2       2        161.000
```





(c) (10 pts) Provide an ANOVA table for this two-factor setting.

Solution: An ANOVA for the two-factor analysis is given below.
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         148

Dependent Variable: crestw
                                              Sum of
Source                        DF             Squares     Mean Square    F Value    Pr > F
Model                          5          58646.3156      11729.2631       8.59    <.0001
Error                        142         193852.6574       1365.1596
Corrected Total              147         252498.9730

Source                        DF         Type III SS     Mean Square    F Value    Pr > F
sex                            1         22556.30489     22556.30489      16.52    <.0001
spec                           2         34194.72740     17097.36370      12.52    <.0001
sex*spec                       2          2367.13979      1183.56990       0.87    0.4224
```

(d) (10 pts) Test for the presence of interaction between sex and species. Also test for the presence of main effects, effects due to the sex and species.

Solution:
Using the Type III SS and F-tests, we see that the main effects for sex and species are highly significant (p-values = 0.0001), but that the interaction between sex and species is not significant (p-value = 0.422).

(e) (10 pts) Summarize differences, if any, in sexes and species using relevant multiple comparisons. Give clear interpretations of any significant effects.

Solution:

30 pts

Looking at the LSMEANS output below, we see that the sample mean crest widths for females, averaged over species, is about 25.3mm larger than that for males. This difference is statistically significant, with a p-value of 0.0006. Because the interaction is far from significant, this difference in mean crest widths applies to each of the 3 species.

```
The GLM Procedure
Least Squares Means

                       H0:LSMean1=
             crestw      LSMean2
sex          LSMEAN       Pr > |t|
1        110.844058         <.0001
2        135.546667
```

I will do a Bonferroni comparison of all possible pairs of species. For an overall or family error rate of 0.05, each comparison should be done at the 0.05/3 = 0.0167 level. Any comparison with a p-value of 0.0167 or less is significant. The Bonferroni comparison gives 2 groups, one of which includes M. f. fuliginosus (species 2), the other of which includes M. giganteus (0) and M.f. melanops (1). The summaries show that the mean crest width for M.f. fuliginosus, averaged over sexes, exceeds the mean crest widths for the the other 2 species by approximately 3mm. Because the interaction is far from significant, this mean difference applies to each sex.

```
              crestw      LSMEAN
spec          LSMEAN      Number
0         110.120000           1
1         115.066087           2
2         144.400000           3

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

          Dependent Variable: crestw
i/j              1             2             3
   1                      0.5089        <.0001
   2        0.5089                      0.0001
   3        <.0001        0.0001
```

I provided the cell means and p-values for comparing cell means. This might be useful for cross species comparisons, but almost all of the other relevant summaries can be derived from the main effects, given the lack of interaction.

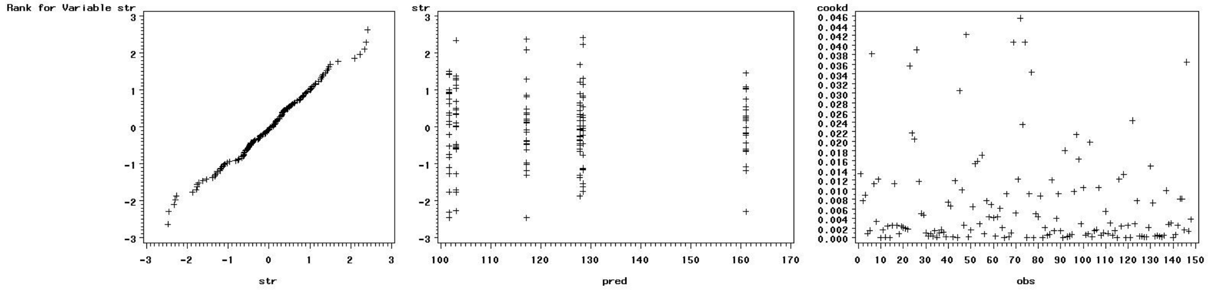
```
                     crestw      LSMEAN
sex    spec          LSMEAN      Number
1      0         103.080000           1
1      1         101.652174           2
1      2         127.800000           3
2      0         117.160000           4
2      1         128.480000           5
2      2         161.000000           6

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

                         Dependent Variable: crestw
i/j            1           2           3           4           5           6
   1                  0.8938      0.0194      0.1800      0.0163      <.0001
   2      0.8938                  0.0155      0.1485      0.0131      <.0001
   3      0.0194      0.0155                  0.3103      0.9482      0.0018
   4      0.1800      0.1485      0.3103                  0.2806      <.0001
   5      0.0163      0.0131      0.9482      0.2806                  0.0022
   6      <.0001      <.0001      0.0018      <.0001      0.0022
```

(f) (10 pts) Do the assumptions for a standard two-factor analysis appear reasonable? If not, suggest and carry out an alternative analysis.

Solution:

As noted in the part (a), the side-by-side boxplots show that the distributions of mandible lengths are fairly symmetric and reasonably normal except in one sex-species group. In addition, the variability in crest widths for M.f. fuliginosus is smaller than the variability seen in the other two species. However, given that the outliers present are not extreme, and that the other deviations from assumptions are minor, I would think that inferences from the standard analysis would hold.

10 pts



(g) (10 pts) Summarize your findings on the effect of sex and species on the crest widths.

Solution:

Our analysis shows significant main effects for sex and species, but no significant interaction between sex and species. On average, females are approximately 25.3mm larger than males regardless of species whereas kangaroos in M. f. fuliginosus are approximately 3mm larger than the M. giganteus (0) and M.f. melanops, regardless of sex.

10 pts