# Ma 322: Biostatistics
# Homework Assignment 9

Prof. Wickerhauser

Read Chapter 15, "ANOVA and Regression," pages 263–287 of our text.

NOTE: Machine-readable data for the problems below is in `http://www.math.wustl.edu/~victor/classes/ma322/hw09data.txt`. Cut and paste from that document into a text file, or into an `R` variable by use of the `scan()` function.

1. The following fake data mimics a study of amino acids in six imaginary species of millipedes:

   *Alanine concentration in millipede hæmolymph (mg/100 ml)*

| Species 1 | Species 2 | Species 3 | Species 4 | Species 5 | Species 6 |
|---|---|---|---|---|---|
| 21.5 | 14.5 | 16.0 | 14.8 | 12.1 | 14.4 |
| 19.6 | 17.4 | 20.3 | 15.6 | 11.4 | 14.7 |
| 20.9 | 15.0 | 18.5 | 13.5 | 12.7 | 13.8 |
| 22.8 | 17.8 | 19.3 | 16.4 | 14.5 | 12.0 |

   (a) Test, at the $\alpha = 0.05$ significance level, the hypothesis $H_0$: There is no difference in mean alanine concentration among the species. Use one-factor ANOVA.

   (b) Test, at the $\alpha = 0.05$ significance level, the hypothesis $H_0$: There is no difference in mean alanine concentration between species $A$ and $B$. Use pairwise $t$-tests for every pair $A, B$.

   (c) Test, at the $\alpha = 0.05$ significance level, the hypothesis $H_0$: There is no difference in mean alanine concentration between species $A$ and $B$. Use Tukey's HSD test for every pair $A, B$.

   **Solution:** (a) Perform a one-factor fixed-effects analysis of variance with species number. The variable will be Alanine concentration. The R commands to perform this test are:

```
alanine <- scan()
species <- gl(6,1,24,labels=c("S1","S2","S3","S4","S5","S6"))
anova(lm(alanine ~ species))
```

   The output shows $p = 10^{-6}$, so **Reject** $H_0$ in favor of $H_A$: There is a significant difference in mean alanine concentration among the species.

   (b) Perform all pairwise $t$-tests. The variable will be Alanine concentration. The R commands to perform this test are:

```
alanine <- scan()
species <- gl(6,1,24,labels=c("S1","S2","S3","S4","S5","S6"))
pairwise.t.test(alanine,species))
```

   The output shows a table of $p$-values for the all pairs:

```
   S1      S2      S3      S4      S5
S2 0.00130 -       -       -       -
S3 0.12588 0.17001 -       -       -
S4 0.00016 0.61635 0.02835 -       -
S5 2.2e-06 0.02835 0.00026 0.17001 -
S6 1.3e-05 0.17001 0.00190 0.61635 0.61635
```

Conclude that species pairs (1,2), (1,4), (1,5), (1,6), (2,5), (3,4), (3,5), and (3,6) have significantly different mean Alanine concentrations.

(c) The R commands to perform Tukey's test are:

```
alanine <- scan()
species <- gl(6,1,24,labels=c("S1","S2","S3","S4","S5","S6"))
hsd<-TukeyHSD(aov(alanine ~ species)); hsd; plot(hsd); abline(v=0,lty=3);
```

The output is tabulated below, and is displayed graphically by the `plot()` and `abline()` commands.

```
Tukey multiple comparisons of means
  95% family-wise confidence level

    Fit: aov(formula = alanine ~ species)

    $species
    diff         lwr        upr     p adj
    S2-S1 -5.025  -8.2908722 -1.7591278 0.0014035 *
    S3-S1 -2.675  -5.9408722  0.5908722 0.1471181
    S4-S1 -6.125  -9.3908722 -2.8591278 0.0001535 *
    S5-S1 -8.525 -11.7908722 -5.2591278 0.0000019 *
    S6-S1 -7.475 -10.7408722 -4.2091278 0.0000120 *
    S3-S2  2.350  -0.9158722  5.6158722 0.2492856
    S4-S2 -1.100  -4.3658722  2.1658722 0.8867240
    S5-S2 -3.500  -6.7658722 -0.2341278 0.0316684 *
    S6-S2 -2.450  -5.7158722  0.8158722 0.2132913
    S4-S3 -3.450  -6.7158722 -0.1841278 0.0349479 *
    S5-S3 -5.850  -9.1158722 -2.5841278 0.0002643 *
    S6-S3 -4.800  -8.0658722 -1.5341278 0.0022289 *
    S5-S4 -2.400  -5.6658722  0.8658722 0.2307618
    S6-S4 -1.350  -4.6158722  1.9158722 0.7740767
    S6-S5  1.050  -2.2158722  4.3158722 0.9045504
```

Conclude that species pairs (1,2), (1,4), (1,5), (1,6), (2,5), (3,4), (3,5), and (3,6) have significantly different mean Alanine concentrations. These pairs are indicated by asterisks in the table, and also by the intervals that do not contain the vertical dashed line in the figure. □

2. Test for all factor and interaction effects in the following $3 \times 2$ fixed-effects analysis of variance with equal replication:

| Response to Factors A and B | | | | | | |
|---|---|---|---|---|---|---|
| a1 | | a2 | | a3 | | |
| b1 | b2 | b1 | b2 | b1 | b2 | |
| 34.1 | 35.6 | 38.6 | 40.3 | 41.0 | 42.1 | |
| 36.9 | 36.3 | 39.1 | 41.3 | 41.4 | 42.7 | |
| 33.2 | 34.7 | 41.3 | 42.7 | 43.0 | 43.1 | |
| 35.1 | 35.8 | 41.4 | 41.9 | 43.4 | 44.8 | |
| 34.8 | 36.0 | 40.7 | 40.8 | 42.2 | 44.5 | |

**Solution:** The following R commands may also be used to perform the test:

```
a<-3; b<-2; n<-5; N<-a*b*n;   data <- scan()
A <- gl(a,b, N, labels=c("a1","a2","a3"));
B <- gl(b,1, N, labels=c("b1","b2"));
anova(lm(data ~ A*B))
```

Conclusions:

| Source of $F$ value | num DF | den DF | $F$ | $F_{0.05(1),...}$ | $H_0$? |
|---|---|---|---|---|---|
| (Factor A MS) / (Error MS) | 2 | 18 | 88.580 | 3.55 | Reject |
| (Factor B MS) / (Error MS) | 1 | 18 | 5.162 | 4.41 | Reject |
| (A x B MS) / (Error MS) | 2 | 18 | 0.272 | 3.55 | Do not reject |
| (Factor A MS) / (A x B MS) | 2 | 2 | 326.175 | 19.0 | Reject |
| (Factor B MS) / (A x B MS) | 1 | 2 | 19.008 | 18.5 | Reject |

The first three rows give the fixed-effects of Factors A and B results. Apparently, factors A and B both have a significant effect, but their interaction is insignificant at the 0.05 level.

The last three rows give the random-effects results, assuming random factors A and B. Again, factors A and B have a significant effect, but there is no significant AxB interaction. □

3. Test for all factor and interaction effects in the following $4 \times 3 \times 2$ fixed-effects analysis of variance, where $a_i$ is the level of factor $A$, $b_i$ is the level of factor $B$, and $c_i$ is the level of factor $C$.

| Response to Factors A, B and C | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | a1 | | | a2 | | | a3 | | | a4 | | |
| | b1 | b2 | b3 | b1 | b2 | b3 | b1 | b2 | b3 | b1 | b2 | b3 |
| c1: | | | | | | | | | | | | |
| | 4.1 | 4.6 | 3.7 | 4.9 | 5.2 | 4.7 | 5.0 | 6.1 | 5.5 | 3.9 | 4.4 | 3.7 |
| | 4.3 | 4.9 | 3.9 | 4.6 | 5.6 | 4.7 | 5.4 | 6.2 | 5.9 | 3.3 | 4.3 | 3.9 |
| | 4.5 | 4.2 | 4.1 | 5.3 | 5.8 | 5.0 | 5.7 | 6.5 | 5.6 | 3.4 | 4.7 | 4.0 |
| | 3.8 | 4.5 | 4.5 | 5.0 | 5.4 | 4.5 | 5.3 | 5.7 | 5.0 | 3.7 | 4.1 | 4.4 |
| c2: | | | | | | | | | | | | |
| | 4.8 | 5.6 | 5.0 | 4.9 | 5.9 | 5.0 | 6.0 | 6.0 | 6.1 | 4.1 | 4.9 | 4.3 |
| | 4.5 | 5.8 | 5.2 | 5.5 | 5.3 | 5.4 | 5.7 | 6.3 | 5.3 | 3.9 | 4.7 | 4.1 |
| | 5.0 | 5.4 | 4.6 | 5.5 | 5.5 | 4.7 | 5.5 | 5.7 | 5.5 | 4.3 | 4.9 | 3.8 |
| | 4.6 | 6.1 | 4.9 | 5.3 | 5.7 | 5.1 | 5.7 | 5.9 | 5.8 | 4.0 | 5.3 | 4.7 |

**Solution:** The following R commands may be used to perform the test:

```
a<-4; b<-3; c<-2; n<-4; N<-a*b*c*n;   data <- scan()
A <- gl(a,b, N, labels=c("a1","a2","a3","a4"));
B <- gl(b,1, N, labels=c("b1","b2","b3"));
C <- gl(c,a*b*n, N, labels=c("c1","c2"));
anova(lm(data ~ A*B*C))
```

That analysis yields the following conclusions:

| Source of $F$ value | num DF | den DF | $F$ | $F_{0.05(1),...}$ | $H_0$? |
|---|---|---|---|---|---|
| (Factor A MS) / (Error MS) | 3 | 72 | 127.794 | 2.74 | Reject |
| (Factor B MS) / (Error MS) | 2 | 72 | 47.559 | 3.13 | Reject |
| (Factor C MS) / (Error MS) | 1 | 72 | 53.310 | 3.98 | Reject |
| (A x B MS) / (Error MS) | 6 | 72 | 1.468 | 2.23 | Do not reject |
| (A x C MS) / (Error MS) | 3 | 72 | 7.649 | 2.74 | Reject |
| (B x C MS) / (Error MS) | 2 | 72 | 0.048 | 3.13 | Do not reject |
| (A x B x C MS) / (Error MS) | 6 | 72 | 1.812 | 2.23 | Do not reject |

For the single factor cases, the null hypothesis is $H_0$: there is no significant effect from the factor.

For the multiple factor cases, the null hypothesis is $H_0$: there is no significant interaction between the factors.

It seems that at the 0.05 level, factors A, B, and C all have a strong effect, and there is strong interaction between factors A and C, but there is no significant interaction between factor B and the other two factors. □

4. Given the following data:

| $Y$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|---|
| 51.4 | 0.2 | 17.8 | 24.6 | 18.9 |
| 72.0 | 1.9 | 29.4 | 20.7 | 8.0 |
| 53.2 | 0.2 | 17.0 | 18.5 | 22.6 |
| 83.2 | 10.7 | 30.2 | 10.6 | 7.1 |
| 57.4 | 6.8 | 15.3 | 8.9 | 27.3 |
| 66.5 | 10.6 | 17.6 | 11.1 | 20.8 |
| 98.3 | 9.6 | 35.6 | 10.6 | 5.6 |
| 74.8 | 6.3 | 28.2 | 8.8 | 13.1 |
| 92.2 | 10.8 | 34.7 | 11.9 | 5.9 |
| 97.9 | 9.6 | 35.8 | 10.8 | 5.5 |
| 88.1 | 10.5 | 29.6 | 11.7 | 7.8 |
| 94.8 | 20.5 | 26.3 | 6.7 | 10.0 |
| 62.8 | 0.4 | 22.3 | 26.5 | 14.3 |
| 58.4 | 6.6 | 15.7 | 8.7 | 26.3 |
| 81.6 | 2.3 | 37.9 | 20.0 | 0.5 |

(a) Fit the multiple regression $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$ to the data, computing the sample partial regression coefficients and $Y$ intercept.

(b) Test $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ by ANOVA at the 0.05 level.

(c) Compute the standard error of each partial regression coefficient and test $H_0 : \beta_i = 0$, at the $\alpha = 0.05$ level, individually for each $i = 1, 2, 3, 4$.

(d) Calculate the standard error of estimate and the coefficient of determination.

(e) What is the predicted mean population value $\hat{Y}$ at $X_1 = 5.4$, $X_2 = 20.3$, $X_3 = 18.7$, $X_4 = 11.2$?

(f) What is the 95% confidence interval for $\hat{Y}$ in part (e)?

**Solution:** See `hw09R.txt` for the R commands used to compute the results.

(a) $(b_1, b_2, b_3, b_4) = (2.0735, 2.5798, 0.6407, 1.1014)$; $a = -30.1423$.

(b) $F = 109.1550$ with 4 numerator and 10 denominator degrees of freedom. This has a one-tailed $p$-value of $3.313456 \times 10^{-8}$, so we **reject** $H_0$ at the 0.05 level.

(c) $H_0$ is rejected at the 0.05 significance level in part b. For all parameters, $\nu = 10 = n - m - 1$, giving the following $t$ statistics and their $p$-values:

$$\begin{pmatrix} S_{b_1} \\ S_{b_2} \\ S_{b_3} \\ S_{b_4} \end{pmatrix} = \begin{pmatrix} 0.609 \\ 0.2904 \\ 0.6525 \\ 0.3114 \end{pmatrix} \quad \Rightarrow \begin{pmatrix} t = b_1/S_{b_1} \\ t = b_2/S_{b_2} \\ t = b_3/S_{b_3} \\ t = b_4/S_{b_4} \end{pmatrix} = \begin{pmatrix} 3.336 \\ 6.154 \\ -1.845 \\ -5.26 \end{pmatrix} \quad \Rightarrow P \approx \begin{pmatrix} 0.00536 \\ 0.00003 \\ 0.08790 \\ 0.00015 \end{pmatrix}$$

Hence, at the 0.05 level we **reject** the null hypotheses $\beta_1 = 0$, $\beta_2 = 0$, and $\beta_4 = 0$, but we **do not reject** the null hypotheses $\beta_3 = 0$.

(d) Coefficient of multiple determination: $R^2 = 0.9776$. Adjusted coefficient of multiple determination: $R_a^2 = 0.9687$. Standard error of estimate: $S_{Y \cdot 1,\ldots M} = \sqrt{\text{Residual MS}} = 2.948$.

(e) Substitute the values into the multiple regression equation $\hat{Y} = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4$ to get $\hat{Y} = 57.74$.

(f) Start with the $X$ values used in part (e) to compute $\hat{Y}$ with 95% confidence interval endpoints: $[47.99, 67.49]$. □

5. Perform a stepwise regression analysis of the data in Problem 4.

**Solution:** See `hw09R.txt` for the R commands used to compute the results. The first step is already done in Problem 4, part (c); variable $X_3$ has the greatest $P$ value for $H_0 : \beta = 0$, so remove it from the regression.

In Step 2, repeat parts (a,b,d,c) of Problem 1 on the remaining data set $(Y, X_1, X_2, X_4)$. This gives $(b_1, b_2, b_4) = (1.4740, 1.7031, 0.1720)$, $a = 18.1039$. For all parameters, $\nu = 11 = n - m - 1$, giving the following $t$ values and their likelihoods:

$$\begin{pmatrix} S_{b_1} \\ S_{b_2} \\ S_{b_4} \end{pmatrix} = \begin{pmatrix} 0.1540 \\ 0.3968 \\ 0.3792 \end{pmatrix} \quad \Rightarrow \begin{pmatrix} t = b_1/S_{b_1} \\ t = b_2/S_{b_2} \\ t = b_4/S_{b_4} \end{pmatrix} = \begin{pmatrix} 9.569 \\ 4.292 \\ 0.454 \end{pmatrix} \quad \Rightarrow P \approx \begin{pmatrix} 1 \times 10^{-6} \\ 0.00127 \\ 0.65884 \end{pmatrix}$$

Hence, at the 0.05 level we **reject** the null hypotheses $\beta_1 = 0$ and $\beta_2 = 0$, but we **do not reject** the null hypotheses $\beta_4 = 0$.

In Step 3, repeat parts (a,b,d,c) of Problem 1 on the remaining data set $(Y, X_1, X_2)$. This gives $(b_1, b_2) = (1.4752, 1.5297)$, and $a = 24.8652$. For all parameters, $\nu = 12 = n - m - 1$, giving the following $t$ values and their likelihoods:

$$\begin{pmatrix} S_{b_1} \\ S_{b_2} \end{pmatrix} = \begin{pmatrix} 0.1488 \\ 0.1030 \end{pmatrix} \quad \Rightarrow \begin{pmatrix} t = b_1/S_{b_1} \\ t = b_2/S_{b_2} \end{pmatrix} = \begin{pmatrix} 9.912 \\ 14.846 \end{pmatrix} \quad \Rightarrow P \approx \begin{pmatrix} 4 \times 10^{-7} \\ 4 \times 10^{-9} \end{pmatrix}$$

Hence, at the 0.05 level we **reject** the null hypotheses $\beta_1 = 0$ and $\beta_2 = 0$, concluding that $Y = \alpha + \beta_1 X_1 + \beta_2 X_2$ with $\alpha \approx 24.8652$, $\beta_1 \approx 1.4752$, and $\beta_2 \approx 1.5297$. □

6. Analyze the five variables in Problem 4 as a multiple correlation.

   (a) Compute the simple correlation coefficient for each pair of variables.

   (b) Compute the multiple correlation coefficient $R$ for each variable in terms of the other 4, and test $H_0 : R = 0$ at the 0.05 level in each case.

   (c) Compute the partial correlation coefficients for the five variables.

   **Solution:** See `hw09R.txt` for the `R` commands used to compute the results.

   (a) With the $Y$ variable in the first column (as $X_0$, so to speak), the simple correlation matrix is:

$$\begin{aligned} r &= \frac{1}{\sqrt{\operatorname{diag} SSCP}} SSCP \frac{1}{\sqrt{\operatorname{diag} SSCP}} \\ &= \begin{pmatrix} 1.0000000 & 0.6791849 & 0.86282623 & -0.45558037 & -0.82482122 \\ 0.6791849 & 1.0000000 & 0.25194242 & -0.80577226 & -0.23873447 \\ 0.8628262 & 0.2519424 & 1.00000000 & -0.09089558 & -0.96534792 \\ 0.4555804 & -0.8057723 & -0.09089558 & 1.00000000 & -0.04742297 \\ 0.8248212 & -0.2387345 & -0.96534792 & -0.04742297 & 1.00000000 \end{pmatrix} \end{aligned}$$

   where $SSCP$ is the "sum of squares and cross products" matrix.

   (b) Read the $F$ statistic for these tests from the `R` output:

$$F = \left( \frac{R^2}{1 - R^2} \right) \left( \frac{\text{Residual DF}}{\text{Regression DF}} \right)$$

   For the 5 linear models, we have:

$$\begin{pmatrix} R^2(Y \sim X_1 + X_2 + X_3 + X_4) \\ R^2(X_1 \sim Y + X_2 + X_3 + X_4) \\ R^2(X_2 \sim X_1 + Y + X_3 + X_4) \\ R^2(X_3 \sim X_1 + X_2 + Y + X_4) \\ R^2(X_4 \sim X_1 + X_2 + X_3 + Y) \end{pmatrix} = \begin{pmatrix} 0.9776 \\ 0.9674 \\ 0.9917 \\ 0.9314 \\ 0.9863 \end{pmatrix} \Rightarrow F = \begin{pmatrix} 109.2 \\ 74.25 \\ 298.9 \\ 33.97 \\ 180 \end{pmatrix} \Rightarrow P \approx \begin{pmatrix} 3 \times 10^{-8} \\ 2 \times 10^{-7} \\ 2 \times 10^{-10} \\ 9 \times 10^{-6} \\ 3 \times 10^{-9} \end{pmatrix}.$$

   Each $F$ statistic has 4 and 10 degrees of freedom in the numerator and denominator, respectively. Thus in all cases, **reject** the null hypothesis.

   (c) The pairwise partial correlation coefficient matrix is given by the off-diagonal terms (the diagonals are all $-1$ with the simplified formula that we use):

$$\begin{aligned} p &\stackrel{\text{def}}{=} -\frac{1}{\sqrt{\operatorname{diag} r^{-1}}} r^{-1} \frac{1}{\sqrt{\operatorname{diag} r^{-1}}} \\ &= \begin{pmatrix} -1.0000000 & 0.8360126 & 0.7581710 & 0.4242227 & 0.4338247 \\ 0.8360126 & -1.0000000 & -0.9121198 & -0.8219524 & -0.7644270 \\ 0.7581710 & -0.9121198 & -1.0000000 & -0.8196893 & -0.9100163 \\ 0.4242227 & -0.8219524 & -0.8196893 & -1.0000000 & -0.8916415 \\ 0.4338247 & -0.7644270 & -0.9100163 & -0.8916415 & -1.0000000 \end{pmatrix} \end{aligned}$$

   Notice how this uncovers the weak dependence of $Y$, the first column variable, on $X_3$ and $X_4$ in rows 4 and 5. The simple correlation matrix $R$ shows a too-strong correlation, in positions (1,4) and (1,5), between $Y$ and $X_3$ and between $Y$ and $X_4$, respectively. □

7. Each of five research papers was read by each of six reviewers. Each reviewer then marked the quality of the five papers as follows:

| | Paper | | | | |
|---|---|---|---|---|---|
| Reviewer | 1 | 2 | 3 | 4 | 5 |
| A | 5 | 4 | 3 | 1 | 2 |
| B | 4 | 5 | 3 | 2 | 1 |
| C | 5 | 4 | 1 | 2 | 3 |
| D | 5 | 3 | 2 | 4 | 1 |
| E | 4 | 5 | 2 | 3 | 1 |
| F | 5 | 4 | 1 | 3 | 2 |

(a) Calculate the Kendall coefficient of concordance.

(b) Test, at the $\alpha = 0.01$ significance level, whether the rankings by the six reviewers are in agreement.

**Solution:** See `hw09R.txt` for the R commands used to compute the results.

(a) Compute the rank sums and the Kendall concordance coefficient for $m = 6$ judges and $n = 5$ ranked items:

$$\begin{pmatrix} R_1 \\ R_2 \\ R_3 \\ R_4 \\ R_5 \end{pmatrix} = \begin{pmatrix} 28 \\ 25 \\ 12 \\ 15 \\ 10 \end{pmatrix}; \qquad W = \frac{\sum_{i=1}^{n} R_i^2 - \frac{1}{n}\left[\sum_{i=1}^{n} R_i\right]^2}{m^2(n^3 - n)/12} = 0.7167$$

(b) For $W = 0.7167$, the Friedman chi-squared value is $\chi_r^2 = m(n-1)W = 17.2$. This has a significant $p < 0.01$, so we **reject** the null hypothesis $H_0$: the six reviewers disagree, in favor of $H_A$: the six reviewers are in agreement.

□