

The third assignment includes a couple problems about ANOVA and several problems in simple linear regression. Some of problems are best done with statistical software. A forthcoming handout will provide guidance.

- Chapter 5, problem 17 (page 142). This problem asks you to complete a partial ANOVA table.

Solution:

Source	df	SS	MS	F	p
Between	7	35,819	5,117	3.50	0.0099
Within	24	35,088	1,462		
Total	31	70,907			

- Chapter 5, problem 18 (page 142). This is an ANOVA problem with other related questions.

Solution: (a) Find the estimated means for each treatment.

```
> sapply(split(PROTEIN,TREATMNT),mean)
CONTROL CPFA150 CPFA300 CPFA450 CPFA50 CPFA600
185.6000 171.6667 146.6667 151.0000 168.3333 152.3333
```

(b) Fit the model with ten means.

```
> sapply(split(PROTEIN,TRT.DAYGROUP),mean)
GROUP1 GROUP10 GROUP2 GROUP3 GROUP4 GROUP5 GROUP6 GROUP7
168.3333 192.6667 171.6667 146.6667 151.0000 152.3333 157.3333 195.6667
GROUP8 GROUP9
203.3333 179.0000
```

```
> fit <- lm(PROTEIN ~ TRT.DAYGROUP)
> anova(fit)
Analysis of Variance Table
```

```
Response: PROTEIN
      Df Sum Sq Mean Sq F value    Pr(>F)
TRT.DAYGROUP  9 11147.5  1238.6   7.8014 7.154e-05 ***
Residuals    20  3175.3   158.8
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

There is strong evidence that the ten means are different from one another.

To address the question whether or not the model with ten different treatment/day means is better than that with six treatment means, we could fit a two-way ANOVA where the TRT.DAYGROUP variable is nested within the TREATMNT variable (which we will come to in later chapters) in R.

```
> fit <- lm(PROTEIN ~ TREATMNT / TRT.DAYGROUP )
> anova(fit)
Analysis of Variance Table
```

```
Response: PROTEIN
      Df Sum Sq Mean Sq F value    Pr(>F)
```

```
TREATMNT          5 7222.5  1444.5  9.0983 0.0001229 ***
TREATMNT:TRT.DAYGROUP  4 3924.9   981.2  6.1803 0.0020889 **
Residuals         20 3175.3   158.8
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The authors of our textbook prefer an explicit calculation of the ratio of extra sum of squares to the estimate $\hat{\sigma}_{full}^2$, which in this case is identical to the calculation in the last two rows of the ANOVA table above. We could do this also by looking at both of the one-way ANOVA tables separately.

```
> fit <- lm(PROTEIN ~ TRT.DAYGROUP)
> anova(fit)
Analysis of Variance Table
```

Response: PROTEIN

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
TRT.DAYGROUP	9	11147.5	1238.6	7.8014	7.154e-05 ***
Residuals	20	3175.3	158.8		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
> fit <- lm(PROTEIN ~ TREATMNT)
> anova(fit)
Analysis of Variance Table
```

Response: PROTEIN

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
TREATMNT	5	7222.5	1444.5	4.8827	0.003196 **
Residuals	24	7100.3	295.8		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

For this problem, the full model has ten means, so there are $30 - 10 = 20$ degrees of freedom in the full model. The reduced model has six means and $30 - 6 = 24$ degrees of freedom. There are $24 - 20 = 4$ extra degrees of freedom.

We can pull the extra sum of squares off of the two separate ANOVA tables— $7100.3 - 3175.3 = 3925.0$, which is the number in the first ANOVA table.

The F statistic is:

$$F = \frac{(3925)/(24 - 20)}{3175.3/20} = 6.18$$

The p-value is

```
> 1 - pf(6.18,4,20)
[1] 0.002089508
```

There is strong evidence that the means were not the same for the control group each day.

3. Chapter 6, problem 21 (page 171). This open-ended question asks you to make several comparisons.

Solution: Data on the original scale seems to be as good or better than transformed data (in plots), so we will not transform.

Here is a one-way ANOVA.

```
> fit <- lm(TIME ~ COMPOUND)
> anova(fit)
Analysis of Variance Table

Response: TIME
      Df Sum Sq Mean Sq F value    Pr(>F)
COMPOUND  4 401.28  100.32   5.0202 0.001970 **
Residuals 45 899.24   19.98
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is strong evidence that the population means are not identical.

To look further at comparisons between the means, Tukey-Kramer is a good choice because the sample sizes are identical and we are interested on all possible comparisons. We have $\hat{\sigma} = \sqrt{19.98} = 4.47$ and the SE for any difference in means is

$$SE = 4.47 \sqrt{\frac{1}{10} + \frac{1}{10}} = 1.999$$

The 95th percentile of the studentized range distribution (5 groups, 45 degrees of freedom) can be found with R.

```
> qtukey(.95,5,45)
[1] 4.018417
```

The multiplier is this value over the square root of 2, or $4.018/\sqrt{2} = 2.84$. The margin of error is $2.84 \times 1.999 = 5.68$. Any sample means larger than 5.68 are significant at the 5% level.

```
> sapply(split(TIME,COMPOUND),mean)
      I      II      III      IV      V
10.693  6.050  8.636  9.798 14.706
```

Compound V is significantly larger than both compounds II and III. If you create the fit with `aov` instead of `lm`, you can use the R function `TukeyHSD` to do this all automatically.

4. Chapter 7, problem 15 (page 197). This problem asks you to analyze the planet distance from the sun data.

Solution: (a) The regression line is contained in this output.

```
> fit <- lm(log(Distance) ~ Order)
> summary(fit)

Call:
lm(formula = log(Distance) ~ Order)
```

```

Residuals:
      Min       1Q   Median       3Q      Max
-0.1885760 -0.0798477  0.0004328  0.0950997  0.1972653

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.76477    0.09268   8.252 3.49e-05 ***
Order        0.53693    0.01494  35.947 3.93e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1357 on 8 degrees of freedom
Multiple R-Squared:  0.9938,    Adjusted R-squared:  0.9931
F-statistic: 1292 on 1 and 8 DF,  p-value: 3.928e-10
    
```

- (b) In a two-sided test of the null hypothesis that $\beta_1 = \log 2$, $t = -10.49$ and the p-value is small.
- (c) A 95% confidence interval for β_1 is from 0.5025 to 0.5714.

5. Chapter 7, problem 23 (pages 198–199). A classic data set about Old Faithful.

Solution: Here is R output that fits the regression line.

```

> fit <- lm(INTERVAL ~ DURATION)
> summary(fit)

Call:
lm(formula = INTERVAL ~ DURATION)

Residuals:
      Min       1Q  Median       3Q      Max
-14.644  -4.440  -1.088   4.467  15.652

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  33.8282    2.2618   14.96  <2e-16 ***
DURATION     10.7410    0.6263   17.15  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.683 on 105 degrees of freedom
Multiple R-Squared:  0.7369,    Adjusted R-squared:  0.7344
F-statistic: 294.1 on 1 and 105 DF,  p-value: < 2.2e-16
    
```

Scatter plots and residual plots do not show anything unusual. However, it is likely that the points are not independent because they come from successive eruptions of Old Faithful. We will talk about this more later. The `predictPlot` function distributed with the class can be used for estimation and prediction bands.

6. Chapter 7, problem 29 (pages 203–204). An open-ended regression problem about birds.

Solution:

A scatter plot does not indicate the need for a transformation. Here is R code of the regression.

```
> fit <- lm(tcell ~ mass)
> summary(fit)

Call:
lm(formula = tcell ~ mass)

Residuals:
    Min       1Q   Median       3Q      Max
-0.18138 -0.04673  0.01796  0.04219  0.15999

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.08750    0.07868   1.112   0.2800
mass         0.03282    0.01064   3.084   0.0061 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08102 on 19 degrees of freedom
Multiple R-Squared:  0.3336,    Adjusted R-squared:  0.2986
F-statistic: 9.513 on 1 and 19 DF,  p-value: 0.006105
```

There is rather strong evidence that increased mass is associated with higher T-cell response.

7. Read the Chapters 7–8 of the textbook. Write out brief answers to the Conceptual Exercises at the end of each chapter. Compare your responses with those given by the authors a few pages later. You do not need to turn in anything for this question.