Professor Notes About Linear Models Foundations Homework 2

- Tables (and figures) should be labeled as described in the homework format description. Table labels go ABOVE the table and figure labels go BELOW the table. Tables (and figures) should be referred to in your answers. See the key below for a model of this.
- Use complete sentences to answer questions.
- Use an R appendix to show the code you used to produce results. Do not include R code in any of your other answers.
- Keep "many" decimals in intermediate calculations ... i.e., don't round until the final answer.
- Note explanations in the key below.

pH in Two Rivers

Table 1. Results from 2-sample t-test of diastolic blood pressure by diet type.

```
t = 6.9849, df = 18, p-value = 1.599e-06
95 percent confidence interval:
 1.576042 2.931958
sample estimates:
mean in group A mean in group B
          8.662           6.408
```

Table 2. Analysis of variance table for the diastolic blood pressure by diet type.

```
          Df  Sum Sq Mean Sq F value    Pr(>F)
river      1 25.4026 25.4026  48.789 1.599e-06
Residuals 18  9.3719  0.5207
```

Table 3. Coefficient results from the one-way ANOVA fit of diastolic blood pressure by diet type.

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.6620     0.2282  37.961  < 2e-16
riverB       -2.2540     0.3227  -6.985  1.6e-06
---
Residual standard error: 0.7216 on 18 degrees of freedom
Multiple R-squared: 0.7305,Adjusted R-squared: 0.7155
F-statistic: 48.79 on 1 and 18 DF,  p-value: 1.599e-06
```

1. The overall p-value from the ANOVA table (1.6e-06; Table 2), for the slope coefficient (1.6e-06; Table 3), and for the two-sample t-test (1.6e-06; Table 1) are equal. These p-values are all equivalent because the 2-sample t-test hypothesis of equal means (or difference in means equals zero) is the same as the hypothesis for the slope (see below about the slope representing the difference in means) which is the same as the hypothesis for the ANOVA table (i.e., simple model of one mean representing both groups).
2. With these p-values, very strong evidence to reject the null hypothesis exists. Thus, the mean pH appears to differ between the two rivers.
3. The intercept coefficient for the linear model (8.66; Table 3) is the same as the mean of the first (A) group in the 2-sample t-test (8.66; Table 1). This occurs because an intercept is defined as the "value of Y ("pH") when X=0, on average" and river A is coded with a zero in `lm()` (because the levels are code alphabetically).

4. The difference in the means (i.e., 6.41-8.66 = -2.25; Table 1) is the same as the slope coefficient in the linear model coefficients (i.e., -2.25; Table 3). This is because the slope coefficient shows the change in Y ("pH") for a one "unit" change in X ("river"), which is a change from river A (coded with a zero) to river B (code with a one) because of the coding in `lm()`.
5. The df from the two-sample t-test (18; Table 1) and the within-group df from the ANOVA table(18; Table 2) are identical. The within-group df are equal to the total number of individuals ($n = n_1 + n_2$) minus the number of groups ($I = 2$), which is the same as for the 2-sample t-test (i.e., $n_1 + n_2 - 2$.
6. The F test statistic (48.79) is equal to the square of the t test statistic (6.98^2=48.79). This relationship occurs when the numerator df for the ANOVA is equal to one (i.e., there are only two groups).
7. The SE for the difference in means is equal to $\frac{\bar{x}_1 - \bar{x}_2}{t} = \frac{8.662-6.408}{6.9849} = 0.3227$. The pooled variance is then equal to this value squared and divided by the sum of the reciprocals of the sample sizes – i.e., $\frac{SE^2_{\bar{x}_1-\bar{x}_2}}{\frac{1}{n_1}+\frac{1}{n_2}} = \frac{0.3227^2}{\frac{1}{10}+\frac{1}{10}} = 0.5207$.
8. The pooled variance computed in the previous question is the same as MS_{within} (Table 1).

R Appendix.

```
library(NCStats)
setwd("T:/Biometry/")
d <- read.csv("pHrivers.csv")
rvr.t <- t.test(pH~river,data=d,var.equal=TRUE)
lm1 <- lm(pH~river,data=d)
anova(lm1)
summary(lm1)
```