Statistics 411/511
Lab 10

Lab Instructions: If you want to work along with the TA in lab, please sit near the front. If you prefer to go to lab but work at your own pace, please sit near the back, and wait for the appropriate time to ask any questions.

Lab 10: One-way ANOVA and ANOVA for Regression

Objectives for this Lab:

- Explore the ANOVA table for a simple linear regression.
- Compare simple linear regression and separate means models in a lack of fit test.

1. Start up RStudio. Load the Sleuth3 package. If you are working in Bexell, you'll have to install the packages again as described in item 5(a) of Lab 1.

   ```
   > library(Sleuth3)
   ```

2. We will be working with the electrical insulating fluid of case study 8.1.2. Get a summary of the data.

   ```
   > summary(case0802)
   ```

   The data frame contains three variables named Time, Voltage, and Group. The summary tells us Time and Voltage are numeric, since R reports a six-number summary. Group is categorical (a "factor" variable in R). The summary for the factor variable lists the names of the different groups and how many observations are in each. There are seven groups, one for each unique value of Voltage.

3. ANOVA for simple linear regression.

   (a) Display 8.4 on page 211 of the textbook puts Time on a log scale. To be consistent with this, we will fit a regression of log(Time) on Voltage, then get a summary of the analysis.

   ```
   > case0802.lm <- lm(log(Time)~Voltage,data=case0802)
   > summary(case0802.lm)
   ```

   The bottom of Display 8.4 gives the estimated regression line and $\hat{\text{SD}}$ of log(BDT). Find these elements in your R output.

   (b) In item 2(f) of Lab 5, we used R's `anova()` function to output an ANOVA table associated with an aov object, the output of `aov()`. The ANOVA table compiled and organized the information needed to calculate the F-statistic which tested $H_0 : \mu_1 = \ldots = \mu_I$. It turns out there's an F-test to do in simple linear regression, and so an ANOVA table is useful here too. We get it by applying `anova()` function to the lm object created in item 3.

   ```
   > anova(case0802.lm)
   ```

   You should see the familiar columns for degrees of freedom (Df), sum of squares (Sum Sq), and mean squares (Mean Sq). Compare the numbers in the table with table (a) in Display 8.8.

(c) We have seen (Homework 6, October 28 lecture notes) that the ANOVA F-test of $H_0 : \mu_1 = \ldots = \mu_I$ was an "extra sum of squares test" to compare two models, the separate means model and the equal means model. The test statistic is a scaled version of the "extra sum of squares," the difference in residual sums of squares between reduced and full models.

$$\text{F-statistic} = \frac{[\text{resSS(reduced)} - \text{resSS(full)}]/[\text{df(reduced)} - \text{df(full)}]}{\hat{\sigma}_{\text{full}}}. \quad (1)$$

We have seen that the residual sum of squares is the sum of the squared residuals where each residual is the difference between a data value (Y) and its "fitted value" $\hat{Y}$:

$$\text{Residual Sum of Squares} = \sum_{\text{all obs.}} (Y - \hat{Y})^2.$$

The separate means model's fitted values $\hat{Y}$ are the observed group means. The equal mean model's fitted values are equal to the average of all the observations.

In the regression setting, the observations Y come with an associated X, and we have seen that the fitted value is the estimated mean of Y for a given X:

$$\hat{\mu}\{Y|X\} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Therefore, a residual of the simple linear regression is

$$\log(\text{Time}) - (\hat{\beta}_0 + \hat{\beta}_1 \text{Voltage})$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ come from the output in item 3(a).

In fact, the lm object contains the residuals. View the first few.

```
> head(case0802.lm$residuals)
```

(d) The residual sum of squares in your ANOVA table of item 3(b) is the sum of squared residuals. Verify this by calculating it directly.

```
> sum(case0802.lm$residuals^2)
```

(e) Also verify the other two numbers in the residual row of your ANOVA table. The degrees of freedom should be the sample size minus the number of mean parameters.

```
> nrow(case0802)-2
```

and the mean square is the sum of squares divided by the degrees of freedom.

(f) Let's turn to the row labeled "Voltage" in the ANOVA table. As with the ANOVA tables we've seen in Chapter 5 and Homework 6, the sum of squares in this row are "extra sums of squares," the difference in residual sums of squares between reduced and full models. In Chapter 5 and Homework 6, the reduced model was the equal means model

$$\mu\{Y_{ij}\} = \mu$$

and the full model was the separate means model

$$\mu\{Y_{ij}\} = \mu_i.$$

In the ANOVA table of item 3(b), the Voltage row gives the difference between reduced and full models where the full model is the simple linear regression model

$$\mu\{Y_i|X_i\} = \beta_0 + \beta_1 X_i$$

and the reduced model is again the equal means model, but the customary notation is different:

$$\mu\{Y_i|X_i\} = \beta_0.$$

As in problem 2(c) of Homework 6, we can get the residual sum of squares for the equal means model:

```
> case0802.equalmeans<-aov(log(Time)~1,data=case0802)
> anova(case0802.equalmeans)
```

You should get a residual sum of squares of 370.23. Take the difference between this number and the residual sum of squares from the ANOVA table in item 3(b) to confirm that the sum of squares for Voltage is the difference in residual sums of squares between the equal means and simple linear regression models.

(g) The degrees of freedom for Voltage are similarly "extra degrees of freedom," the difference in residual degrees of freedom between the reduced and full models. Since the residual degrees of freedom are always the sample size minus the number of mean parameters, the extra degrees of freedom are exactly how many *more* parameters the full model has. Confirm that this works for the ANOVA table of item 3(b).

(h) The ANOVA F-test in Chapter 5 and Homework 6 compared full model $\mu\{Y_{ij}\} = \mu_i$ with reduced model $\mu\{Y_{ij}\} = \mu$. The null hypothesis was $H_0 : \mu_1 = \ldots = \mu_I$, exactly the restriction on the parameters of the full model that yielded the reduced model. Using the same logic, what is the null hypothesis tested by $F = 78.141$ in item 3(b)?

4. Lack of linear fit test It may be that the simple linear regression model is inadequate–the means don't fall on a straight line. For some data sets, we can test the fit of the linear regression model by comparing it to the separate means model.

(a) If we think of the seven voltage levels as groups rather than a numeric explanatory variable, we can fit a separate means model to the insulating fluid data.

```
> case0802.aov<-aov(log(Time)~Group,data=case0802)
> anova(case0802.aov)
```

Find the residual sum of squares (173.75) and residual degrees of freedom (69) on the R output. Compare the output to the second ANOVA table in Display 8.8.

(b) The residual sum of squares for the equal means model is not shown in the R output from `anova(case0802.aov)`, but it shows up as the "Total" sum of squares in *both* ANOVA tables in Display 8.8, and we calculated it explicitly in item 3(f).

(c) We now have three plausible models for the breakdown time data: separate means, equal means, and simple linear regression. As we briefly discussed in class, simple linear regression is intermediate between separate means and equal means in terms of complexity and flexibility. From most complex (i.e. most flexible and containing the largest number of parameters) to simplest (i.e. least flexible and containing the smallest number of parameters), they are:

Name	Model	Residual SS	Residual df
Separate means	$\mu\{Y_{ij}\|X_i\} = \mu_i$		
Simple linear regression	$\mu\{Y_{ij}\|X_i\} = \beta_0 + \beta_1 X_i$		
Equal means	$\mu\{Y_{ij}\|X_i\} = \mu$		

Look back at your R output to find the three residual sums of squares and degrees of freedom, and write them in the appropriate places in the table above. (You're not going to hand this lab in. Noting these quantities is just for your convenience.)

(d) So far, we have used the extra sum of square F-statistic in (1) to compare the separate means and equal means models and to compare simple linear regression and equal means models. In general, whenever the normality, equal variance, and independence assumptions are met, we can use an extra sum of squares F test to compare a more complex ("full") model with a simpler ("reduced") one, provided the simpler model is a special case of the more complex model. "Special case" means that the simpler model places restrictions on the parameters of the more complex model. These restrictions dictate the null hypothesis of the test, as observed in item 3(h).

Note that the simple linear regression model can be expressed as a special case of the separate means model if all the means μ_i fall on the line determined by $\beta_0 + \beta_1 X_i$. Therefore, we can test this hypothesis with an F-statistic of the form given in (1).

Use the residual SS and df's from item 4(c) to calculate the F-statistic. You should get approximately 0.502.

(e) Use `pf()` to calculate the p-value of the lack of fit test. You'll need numerator and denominator degrees of freedom. The numerator degrees of freedom are the extra degrees of freedom [df(reduced) − df(full)], and the denominator degrees of freedom are the residual degrees of freedom for the full model. Since the separate means model has seven parameters and the simple linear regression model has two, the numerator degrees of freedom are 5. You should have residual df = 69 in item 4(c).

```
> 1-pf(0.502,5,69)
```

The p-value is about 0.78, as shown in Display 8.10. What does this tell us about median breakdown times?

(f) We can use R's `anova()` function to compare full and reduced models. We've fit the equal means, simple linear regression, and separate means models in R. Giving two of these "objects" to `anova()` in order from simplest to most complex will calculate extra sum of squares F-statistics and p-values for the test comparing the two models:

```
> anova(case0802.lm,case0802.aov)
```