Stat 333: Applied Regression Analysis
Spring 2015
Due Date: Friday, February 13 in class

Relevant text chapters: Ch. 2

Instructions: You may (and are encouraged to) discuss homework problems with other students, but the solutions that you should provide should be your own and not directly copied from another student. Show your work wherever possible, and write out the formulas you used to arrive at your solutions. If you have any questions or difficulties, you are more than welcome to come see me in my office. Note that datasets for textbook problems can be found on the CD that came with the textbook or can be downloaded from

https://netfiles.umn.edu/users/nacht001/www/nachtsheim/index.html

1. Prove the following

$$E(\hat{\beta}_0) = \beta_0$$

$$Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{SS_{XX}} \right)$$

In your proof, make sure that you can demonstrate that the mean response and the estimated slope coefficient are independent, that is

$$Cov(\bar{Y}, \hat{\beta}_1) = 0$$

2.In simple linear regression, the *F* test in the ANOVA table and the *t*-test for the slope coefficient both test the same hypotheses, that is H_0: $\beta_1 = 0$ vs. H_1: $\beta_1 \neq 0$ When the null hypothesis is true, show that $F = t^2$ where

$$F = \frac{SSR/1}{SSE/n-2} = \frac{MSR}{MSE}$$

and

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MSE}{SS_{XX}}}}$$

3. Ch.2, Problem 2.10 (p.91)

4. Ch. 2, Problem 2.13 (p. 91).

5. Ch. 2, problem 2.23 (p.93), parts a) – e). In part a), compute the ANOVA table by hand, showing your calculations. Verify that this matches the output from the function anova() in R. What sample statistic is equivalent to a "Mean Squares Total" and what is the value of this statistic?

6. The idea of sampling distributions in the context of simple linear regression corresponds to the notion of drawing samples with values of the predictor staying fixed from sample to sample, but values of the response changing from sample to sample. In this exercise, suppose that we know a true regression relationship given by

$$Y_i = 125 - 5X_i + \varepsilon_i$$

where $\varepsilon_i \sim N(0, 4.5)$

Also, let us assume that this relationship holds for $5 \leq X \leq 20$.

A script is available at Learn@UW that creates sample data for Y_i using different values of X_i . Specifically, the script generates 2 samples:

Sample 1: Using the values X of 5, 7, 9, 11, 13, 15, 17, 19, a sample of 20 observations of Y for each value of X, giving a total of *n* = 160 observations.

Sample 2:Using the values X of 9, 10, 11, 12, 13, 14, 15, 16, a sample of 20 observations of Y for each value of X, giving a total of *n* = 160 observations.

Part A:

Using this script, generate a "Sample 1" and a "Sample2". For each sample, regress *Y* on *X* and print out the results of the summary() and anova() functions on your output from lm(). Which samples had estimates of β_0, β_1 , and σ^2 that were closer to the true values of these parameters? What difference do you notice in the standard errors for $\hat{\beta}_0$and $\hat{\beta}_1$between these samples?

Next, use the function predict() to generate a 95% confidence interval for $E(Y_h)$ and a 95% prediction interval for $Y_{h(new)}$, when $X_h = 15$. Print out your output and comment on what you find. Which interval is wider? What are the values of the standard errors associated with these intervals?

Part B:

In the script for this exercise, the sample sizes for each sample are set by the code

```
n <- c(20,20,20,20,20,20,20,20)
```

Change these values so that you draw samples with 2000 observations for each value of X. Now repeat Part A using these new "Sample1" and "Sample2"s. What differences do you notice in your parameter estimates β_0, β_1 , and σ^2 and standard errors for or $\hat{\beta}_0$and $\hat{\beta}_1$ as compared to part A? What difference do you notice in your confidence interval for $E(Y_h)$ and prediction interval for $Y_{h(new)}$ as compared to your results in Part A?