Stat 333: Applied Regression Analysis
Spring 2015
Due Date: Friday, February 20 in class

Relevant text chapters: Ch. 3

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 that
a)

$$E(MSR) = \sigma^2 + {\beta_1}^2 \sum_i (X_i - \bar{X})^2$$

where

$$MSR = \frac{{SS_{XY}}^2}{SS_{XX}} = {\hat{\beta}_1}^2 (SS_{XX})$$

b) Given that $SSE/\sigma^2 \sim \chi^2(n-2)$, prove that

$$E(MSE) = \sigma^2$$

Note: If you are not familiar with the chi-square (χ^2) distribution, please review appendix A.4 (p.647)

2. Prove that for the Pearson product-moment correlation coefficient that

$$r = \frac{SS_{XY}}{\sqrt{SS_{XX} SS_{YY}}} = \left(\frac{1}{n-1}\right) \sum_i \left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right)$$

3. Ch. 3, Problem 3.3 (p. 146)

4. Ch. 3, Problem 3.15 (p. 150)

For this problem, plot the residuals against the predictor (X). Describe the pattern you see and suggest how might you remedy this problem.

5. Ch. 3, Problem 3.23

6. In 1973, Francois Anscombe published a paper describing 4 data sets (11 observations each of 4 sets of response variables $Y_1 - Y_4$ and corresponding predictor variables $X_1 - X_4$). The data have the interesting property that when a simple linear regression model is fit (i.e. $Y_1 = \beta_0 + \beta_1 X_1 + \varepsilon$) each regression gives nearly identical parameter estimates and summary statistics, despite the fact that the data themselves are quite different. The goal of this exercise is for you to recognize the importance in producing scatter plots of the data before fitting and the importance in producing residual plots after fitting a linear regression model. The data Anscombe created are built in with R and can be accessed by simply typing anscombe at the > prompt in the R console.

For this exercise I want you to do the following for each pair of X, Y:

i) Produce a scatter plot of each pair. Do you think a linear regression model will fit the data well? Explain your reasoning.

ii) Fit the simple linear regression model using lm() and print out your results.

iii) Write out the fitted regression line and state the value of $R^2$and *MSE*

iv) Print out the ANOVA table .

v) Produce a scatter plot of the residuals against the predictor X.

vi) Describe the residual plot you obtained. What features do you notice?