Bios 312: Modern Regression Analysis
February 1, 2012
Lab 2: Linear Regression #1

In this lab, we will use salary data from the year 1995. We will focus on the variables: **salary, sex, rank, startyr, and yrdeg**

Part 1: Initial dataset manipulation

1. Read in the salary dataset
2. Remove all observations that are not from 1995
3. Describe the dataset

Part 2: Salary and sex (unadjusted analysis; binary predictor variable)

1. Create an indicator variable (MALE) for male gender
2. Create an indicator variable (FEMALE) for female gender
3. Create a descriptive plot of the association between salary and sex
4. Fit the following two simple linear regression models for salary. Compare the output
    a. A model with an intercept and MALE
    b. A model with an intercept and FEMALE
5. Using summary statistics by sex, calculate the pooled estimate of the standard deviation. Compare this estimate to the root mean square error obtained in the models from (4.a) and (4.b)
6. Fit the following two regression models. Interpret each of the regression coefficients and compare the output
    a. A model with an intercept, MALE, and FEMALE
    b. A model with MALE and FEMALE (but no intercept); this is called 'cell means coding'
7. Use the 'xi:' command (Stata) or the 'factor()' function in R to fit the following model. Interpret the regression coefficients
    a. Stata: 'xi: regress salary i.sex'
    b. R: 'lm(salary ~ factor(sex))'
8. Suppose that we are interested in the yearly (rather than monthly) salary
    a. Create a new outcome variable representing yearly salary (YRSAL)
    b. Fit a regression model using YRSAL as the outcome and MALE as the predictor (with an intercept). Compare the results from this output to the output obtained in 4.a

Part 3: Salary and rank (unadjusted analysis; categorical predictor variable)

1. Create indicator variables for
    a. Rank of Assistant Professor (ASSIST)
    b. Rank of Associate Professor (ASSOC)
    c. Rank of Full Professor (Full)
2. Generate a box plot of salary by rank
3. Fit the following three regression models (using salary as the outcome) and compare the output
    a. A model with an intercept, ASSIST, and ASSOC
    b. A model with an intercept, ASSOC, and FULL
    c. A model with an intercept, ASSIST, and FULL
4. Use the 'xi:' command (or factor() function) to fit the following model
    a. Stata: 'xi: salary i.rank'
    b. R: 'lm(salary ~ factor(rank)'
5. Using summary statistics by rank, find the pooled standard deviation. Compare this estimate to the root mean square error obtained in models 3.a, 3.b, 3.c, and 4
6. Fit a regression model using cell means coding by including ASSIT, ASSOC, and FULL as predictors but no intercept. Interpret the regression coefficients and tests given in the output.

Part 4: Salary by starting year and year of degree (continuous predictor variables startyr and yrdeg)

1. Create a scatter plot of yrdeg by startyr
    a. Add a lowess (or other smooth) line to the plot
    b. Add a regression line to the plot
2. Estimate the association between yrdeg and startyr using the following three approaches. Then, compare the results from each approach. What are the similarities and differences?
    a. Calculate the Pearson correlation between yrdeg and startyr
    b. Regress yrdeg (outcome) on startyr (predictor)
    c. Regress startyr (outcome) on yrdeg (predictor)
3. Examine the association between salary and experience using unadjusted models. Interpret the output from the following models
    a. Regress salary on startyr
    b. Regress salary on yrdeg
4. Fit a multivariable model that includes both yrdeg and startyr as predictors of salary. What is the interpretation of the regression coefficients in this model?