Bios 312
Modern Regression Analysis
Lab #5: Multivariable regression
February 22, 2012

For this lab, we will be using salary data from they year 1995. We will focus on the variables: **salary, sex, and yrdeg**

Initial dataset manipulations

1. Read in the salary dataset
2. Remove any observations that are not from 1995 (use the 'year' variable)
3. Describe the dataset
4. Create an indicator variable for male gender

Another way to think about regression is the amount of variablity in the outcome (Y) that is explained by the predictors (X). In simple linear regression, we regress X on Y because we believe that X will explain some of the variability in Y. This leads to an alternate way of thinking about statistical tests for regression coefficients in terms of the variability of the outcome. Specifically, the null hypothesis is that X explains none of the variability in Y, and the alternative hypothesis is that X explains more variability than would be expected by chance alone. In this lab we will consider the interpretation of statistical tests in a multivariable model that adds another predictor (W) to the model. Salary will be outcome (Y), male gender the predictor of interest (X), and year of degree the additional covariate (W).

1. Model 1: Fit a simple linear regression model with salary as the outcome and male as the predictor. Save the residuals from this model. Interpret these residuals in terms of the unexplained variability in salary.

2. Model 2: Fit a simple linear regression model with yrdeg as the outcome and male as the predictor. Save the residuals from this model. Interpret these residuals in terms of the unexplained variability in yrdeg.

3. Plot the residuals from model 2 (X-axis) versus the residuals from model 1 (y-axis). Describe any association you see. It may be helpful to add a lowess smooth or other smooth line to the plot.

4. Fit a simple linear regression model using the residuals from model 1 as the outcome and the residuals from model 2 as the predictor. Interpret the slope coefficient from this model.

5. Fit a multivariable linear regression model with salary as the outcome using predictors male and yrdeg. What is the interpretation of the male coefficient in this model? What is the interpretation of the yrdeg coefficient?

6. Compare the slope estimate for yrdeg from model 5 to the slope estimate obtained in question 4. Explain your findings.

7. What implications does this have for fitting multivariable models with even more predictors? Consider the implications in relation to the association between (1) X and W and (2) the residuals of Y given X and residuals of X given W.

Additional problem

1. Create a new variable FULL that takes on the value 1 for full professors and 0 for Assistant or Associate Professors.

2. Determine if FULL explains some of the variability in salary after adjusting for year of degree and gender by fitting the multivariable regression model and by regressing residuals from “Model A” on the residuals from “Model B” other as was done previously (you will need to figure out what “Model A” and “Model B” should be). Compare the results from the two models.