For this lab, we will be using salary data from the 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

Salary and sex (unadjusted analysis; binary predictor variable)

1. Create an indicator variable for male gender
2. Fit the following regression model using classical linear regression
   1. A model with an intercept and male
3. Using model 2.1, extract the following from the regression model output
   1. The fitted values, saved as 'fitted'
   2. The residuals, saved as 'epsilon'
4. List the variables salary, sex (or male), fitted, and epsilon.
   1. What are the fitted values for males and females?
   2. What assumptions are needed for the fitted values to be interpreted as means (expected values) or as predictions of new observations?
   3. Add (fitted + epsilon) for each of the first 5 subjects and compare the answer to the observed data
5. Calculate the standard deviation of the residuals. Compare this result to the regression output from model 2.1
6. Create a plot of the residuals (y-axis) versus fitted values (x-axis)
   1. Jitter the fitted values to visualize overlapping points more easily
   2. Note the average of the residuals
   3. Does the variance of the residuals appear to be similar in the two groups?
   4. Conduct a F-test for the equality of variance for the residuals by sex. Explain why this test is not useful for model building.
7. Compare the results of the following analyses
   1. Classic linear regression of salary on male gender
2. Linear regression with robust standard errors of salary on male gender
3. A two-sample t-test assuming unequal variances of salary on male gender

8. Generate a new outcome variable: ln(salary) = log(salary)
   1. Fit a regression of ln(salary) on male gender
   2. Interpret the regression coefficients (on the original scale)
   3. Exponentiate the coefficients, $e^b$, and interpret the output
   4. Create a 95% confidence for $e^b$ and interpret

Salary by year of degree
1. Create a scatter plot of the yrdeg versus salary
   1. Add a lowess (or other smooth) line to the plot
   2. Add a regression line to the plot
2. Examine the correlation between yrdeg and salary
   1. Using Pearson's correlation
   2. Using classical linear regression
3. Using model 2.1, extract the following from the regression model output
   1. The fitted values, saved as 'fitted2'
   2. The residuals, saved as 'epsilon2'
4. List the variables salary, yrdeg, fitted, and epsilon.
   1. What is the equation to obtain the fitted values?
   2. What assumptions are needed for the fitted values to be interpreted as means (expected values) or as predictions of new observations?
   3. Add (fitted + epsilon) for each of the 5 subjects and compare to the observed data
5. Calculate the standard deviation of the the residuals. Compare this result to the regression output from model 2.1
6. Create a plot of the residuals versus fitted values. Visually inspect for signs of homoscedasticity or heteroscedasticity
   ○ Note: I am not advocating that you visually inspect your data for signs of homoscedasticity or heteroscedasticity to decide which model to fit. It may be used to diagnose if the model you chose a priori was appropriate or not.
7. Create a plot of the residuals versus yrdeg. Visually inspect for signs of homoscedasticity or heteroscedasticity
8. Compare the results of the following analyses
   1. Classic linear regression of salary on yrdeg
   2. Linear regression with robust standard errors of salary on yrdeg