Homework #2
Due: January 22 in lab

1. Calculate the ANOVA tables for problem 1 of HW 1. Perform the appropriate hypothesis tests for assessing the relationship between NOx and mortality. Write up an explanation for a non-statistical audience. (Note: Do not just copy the F-statistic or t-statistic from the R output, show how the test is constructed)

2. Using the data from HW 1, problem 1. Reassess the relationship between NOx and mortality, adjusting for rainfall using multiple linear regression.

   (a) Create the appropriate plots to visually assess the relationship between these three variables and comment on what you see.

   (b) What is an appropriate model for assessing rainfall as a confounder of the relationship between NOx and mortality? Fit this model using least squares.

   (c) Perform appropriate hypothesis tests for assessing the relationship between NOx and mortality.

   (d) Write-up your results as though you were explaining any potential risks associated with NOx levels to the general public.

3. The number of pounds of steam used per month at a plant is thought to be related to the average number monthly ambient temperature. The past year's usages and temperatures are in a data set called hw2.3.dat on the course website. Usage is recorded as lbs/1000.

   (a) Fit a simple linear regression model to the data.

   (b) Test for significance of the relationship between steam and temperature.

   (c) Plant management believes that an increase in average ambient temperature of 1 degree will increase average monthly steam consumption by 10,000 lbs. Do the data support this statement?

   (d) Construct a 99% prediction interval on steam usage in a month with average ambient temperature of $58^o$.

4. Consider the simple linear regression model $y = 50 + 10x + \epsilon$, where $\epsilon \sim_{iid} N(0, 16)$. Suppose that $n = 19$ pairs of observations are used to fit this model. Generate 500 samples of 19 observations, drawing one observation for each level of $x = 1, 1.5, 2, \ldots, 10$ for each sample.

   (a) For each sample, compute the least-squares estimates of the slope and intercept. Construct histograms of the sample values of $\hat{\beta}_0$ and $\hat{\beta}_1$. Discuss the shapes of these histograms.

   (b) For each sample, compute an estimate of $E(y|x = 3.5)$. Construct a histogram of the estimates you obtained. Discuss the shape of the histogram.

(c) For each sample, compute a 95% confidence interval on the slope. How many of these intervals contain the true value $\beta_1 = 10$? Is this what you would expect? Why?

(d) For each estimate of $E(y|x = 3.5)$ in part (b), compute the 95% confidence interval. How many of these confidence intervals contain the true value of $E(y|x = 3.5)$? Is this what you would expect? Why?

5. Suppose we have fit the straight-line regression model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$, but the response is affected by a second variable $x_2$ such that the true regression function is

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

(a) Is the least-squares estimator of the slope in the original simple linear regression model unbiased?

(b) Show the bias in $\hat{\beta}_1$.

(c) How does this affect your interpretation of $\hat{\beta}_1$?