STA 471 Practice Problems 1

1. Consider the simple linear regression model given by $Y_i = \beta_0 + \beta_1 X_i + e_i$, $i = 1, ..., n$.

a. If the two variables under study are height and weight, which one is X and which one is Y?
b. List the model assumptions.
c. Why are distributional assumptions made on the e_i's?
d. What is the mean value Y_i? Does it depend on X_i?
e. What is the variance of Y_i? Does it depend on X_i?
f. After the data has been collected, is $Var(\widehat{\beta}_1)$ known?
g. Give two reasons why one may be interested in estimating $Var(\widehat{\beta}_1)$.
h. What is the relationship between the sample correlation coefficient and the estimate of the slope?

2. a. After a regression model has been fit to a data set and the investigator is satisfied the model adequately summarizes the relationship between the variables in question, why bother performing diagnostics using the residuals?
b. Why does one check for constant variance (or normality or other assumptions) based on the residuals? Why not use the Y data?

3. a. Why do we make a distinction between confidence intervals and prediction intervals?
b. Suppose an investigator built a confidence interval for β_0 and a confidence interval for β_1. Given a specific value of X, say, x_*, could he/she use the above intervals to construct a confidence interval for $\beta_0 + \beta_1 x_*$?
c. Why is a prediction interval for a new value of Y wider than a confidence interval for the mean value of Y?

4. How would you determine the "added value" a second predictor variable brings to a regression model that already contains one predictor variable? Please explain.

5. Why is it useful to use matrix notation to describe a regression model?