Stat 230, Spring 2016
Homework 3: Simulation
due Thursday 2/11/16 at 11:59pm on bcourses.

Part 1: Problems from the book
Ch 4: Discussion #1-5 read and understand, answers are in the back (don't hand in).
Ch 4: Discussion #17. Write out solution.

Part 2:

1. Create a 100x4 data frame with columns WT, HT, BMI, and ϵ with the following conditions, roughly based on current US statistics:

   - Weight is normal with mean 180 and SD 40.
   - Height is normal with mean 66 and SD 3, correlation between height and weight is 0.7.
   - These conditions will determine how you generate ϵ, think about bivariate normal distribution and OLS model. It may be useful to first generate X,Y, and Z as standard normal with desired correlations, then transform to get weight, height, and ϵ. Do this from scratch, not using anything fancier than rnorm(). However, transform so that the averages, SDs, and r are exact.
   - BMI = (Weight in Pounds / (Height in inches x Height in inches)) x 703

2. Use OLS to get coefficient estimates for the regression model predicting weight based on height and BMI, including an intercept term. You can either use the function(s) you wrote for HW 2 or use `lm()`. What assumptions of the model are violated, if any? (See page 42 of text, in particular.)

3. What is the actual value of β_1, the coefficient corresponding to HT?

4. Is the estimate biased? Why or why not?

5. Plot residuals vs each column of X. (Four plots - note that in practice you would not be able to plot vs ϵ but since this is a simulation we can.) Are *eps* and e orthogonal and/or independent to HT and WT? (That's a total of 8 questions.)

6. Repeat 1 and 2, 1000 times (`replicate()`).

7. Plot a histogram of the 1000 values of $\hat{\beta}_1$, and include a vertical line through the corresponding parameter. Comment on whether $\hat{\beta}_1$ looks biased.