MA386 Statistical Programming HW 5

Due Tuesday , April 29, beginning of class

1: In this problem you will investigate speeding up your simulation code for problem 2 of HW 4. In problem 2, you investigated the significance level of `permtest` both when the null hypothesis H_o: $\mu_x = \mu_y$ was true and when it was false. In this problem you will redo the H_o true simulation twice: once using your original code and once using a different, hopefully faster, version of `permtest`.

More specifically, you are to do the following:

1. Write a simple function - call it `permsim` - to generate two samples of size 20 from a standard normal distribution **500** times using `rnorm()`. For each of these 500 pairs of samples, compute the corresponding p-value for testing H_o: $\mu_x = \mu_y$ vs. H_a: $\mu_x \neq \mu_y$ using `permtest`.
2. Load/source `persim` into R then time how long it takes to run by issuing the command `system.time(permsim)`.
3. Write a new version of `permtest` - call it `permtest2` - which does what `permtest` does but without using a loop. (Hint: compute the N values of $|\bar{X}_1^* - \bar{X}_2^*|$ by using `apply` and a custom-made function with an N row by 1 column matrix.)
4. Compare the output of `permtest2` with that of `permtest` on some data to make sure `permtest` is logically correct.
5. Repeat step 2 with a version of `permsim` which calls `permtest2` instead of `permtest`.

By avoiding the "dreaded for loop," (see section 14.2 in your book) `permtest2` should run substantially faster than `permtest`. Send me a text document containing your interpretation of the two `system.time()` calls and your estimate of the degree of speedup of `permtest2` over `permtest`. Also include copies of your `permsim` and `permtest2` code.

2: In this problem you will conduct a regression analysis similar to the ones you did in MA223 or MA382. Since we haven't discussed in any detail how to conduct a regression analysis using R a major purpose of this exercise is to have you learn how to use R to do a familiar data analysis by consulting your book, R's online help, and the internet. Follow the analysis outline below and write up your results, including R text and graphical output, in a Word document which you will print out and turn in. Try to make your Word document as professional looking as possible, i.e., use full sentences, use appropriate labels in your graphs, reduce your graphs to an appropriate size, etc. As much as possible, try to do everything using built-in R functions and try to do this problem on your own. Note: you may want to duplicate your analyses in Minitab to check your work. In general checking your analyses by using familar statistical software is a good way to check your work when using a new statistical language/package.

Recall the *Simple Linear Model* which you learned in MA223/MA382:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, \ldots, n$$

where the random errors $\{\epsilon_i\}$ satisfy the following assumptions:

1. The errors $\epsilon_1, \ldots, \epsilon_n$ all have mean 0, i.e., $\mu_{\epsilon_i} = 0$ for all i.
2. The errors $\epsilon_1, \ldots, \epsilon_n$ all have the same variance σ^2, i.e., $\sigma^2_{\epsilon_i} = \sigma^2$ for all i.
3. The errors $\epsilon_1, \ldots, \epsilon_n$ are independent random variables.
4. The errors $\epsilon_1, \ldots, \epsilon_n$ are normally distributed.

Fit the model and check these assumptions as follows:

i. Read the CSV file `mpgweight.csv` into R and fit the simple linear model where the response variable y is city miles per gallon (`citympg`) and the predictor variable x is curb weight in pounds (`wt`). Recall that you can fit a regression model using `lm()`.

ii. Using the regression model object created by `lm()` in part i and the function `summary()` answer/do the following:

1. What are the least squares estimates of β_0 and β_1?
2. What is the regression estimate of the random error variance σ^2?
3. What is the t-statistic for testing H_0: $\beta_1 = 0$ vs. H_1: $\beta_1 \neq 0$, what is the corresponding p-value, and what do you conclude at $\alpha = 0.05$?
4. Construct a 95% confidence interval for β_1.
5. Determine the value of R^2 and interpret it.

iii. In order for the t-test and confidence interval you computed in ii to be reliable, the random errors need to meet - at least approximately - the four assumptions above. Recall that the residuals provide estimates of the standard errrors. Check these assumptions via residual analysis as follows:

1. Create a scatter plot of the residuals vs the predictor variable `wt`. Save your plot in a ".png" file using `png()` and then put a copy in your Word document. Based on this plot, do you feel the first two assumptions are met?
2. Create a normal probability plot of the residuals using `qqnorm` and save it in a ".png" file and put a copy of it in your Word document. Based on this plot, do you feel the normality assumption is met?