**Homework** **1 – Ch.’s 1 & 2** – due Wednesday, September 17, 2014

As a reminder, you need to attach to an email your R code for these problems. You should also attach a word document with your answers, along with appropriate output, graphs, and comments from the R program that helps to answer each question. Please let me know if you have any questions.

**1.)**

a. Compute descriptive statistics for 2 quantitative variables. Looking at their mean and median, what type of skewness do you think the variables have and why?

b. Compute a box plot, histogram and a (normal) qq plot for your 2 variables above. Comment on the shape of the distribution.

c. Take a square root transformation of 1 of the variables and make a side by side histogram of the original variable and the square root one. Comment on the plots.

d. Take a (natural) log transformation of the other variable and make boxplots of the original variable and the log one, one on top of the other. Comment on the plots.

e. Evaluate a 3rd variable graphically and comment on that variable’s distribution.

f. Make boxplots of your 3rd variable by a categorical variable. Comment on whether there is a difference in your 3rd variable by categories.

Take a logical subset of your original dataset.

g. Repeat part (b).

Here is an example of how to take a subset:

> health <- read.csv(choose.files(), header=TRUE)

> health1 <- subset(health, GENDER==1)

> attach(health1)

Here I am taking only observations in the data where GENDER was 1 and ignoring all other observations.

2.)

a. Calculate the correlations among 3 of your quantitative variables. Comment on your results.

b. Make scatterplots between all 3 pairs of variables in part a. Comment on whether you feel the correlations and the plots coincide.

c. Fit a basic linear regression model using 1 variable as the outcome variable and another as the explanatory variable. Comment on the fit using *R*² and the *t*-statistic. Interpret R².

d. Write out the equation for part (c) and interpret b0 and b1.

e. Fit a basic linear regression model using the same variable from part (c) as the outcome variable and another as the explanatory variable. Comment on the fit using *R*² and the p-value. Which model fits better, this one or the one in part (c)?

f. Do a hypothesis test on you model in part (e). Provide hypothesis statements, formula work, t scores, p-value, and comment on your decision.

g. Find a 99% confidence interval corresponding to the slope in part (e). Interpret this interval.

h. Suppose your X value in part (e) is 5, solve for your Y. Also calculate the residual.

i. Obtain a 95% prediction interval for your prediction in part (h). Interpret this interval.

j. Repeat part (f), but test whether your slope is significantly different from 2.