STAT 462 – Applied Regression Analysis
Fall 2017, Homework 3

<u>Theory part</u> (30 points)

1. (15 points) Consider a simple linear regression model, with $\hat{\beta}_0$ and $\hat{\beta}_1$ the least square estimators and $\hat{y}_i$ the fitted values. Prove that the coefficient of determination R^2 is equal to the square of the sample correlation coefficient $\left(corr(x,y)\right)^2$.

2. (15 points) Consider a simple linear regression model. Show that the F statistics and the t statistics to test $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$ are equivalent, i.e. show that

$$F = \frac{SS_{reg}}{RSS/(n-2)} = \left(\frac{\hat{\beta}_1}{se\left(\hat{\beta}_1\right)}\right)^2 = t^2$$

<u>Applied part</u> (70 points)

Consider again the dataset contained in “BODY_FAT.TXT”.

In the following analyses, <u>use the body fat percentage you re-computed</u> in Homework1. In addition, please <u>remove</u> the observation with <u>erroneous measurement</u> (observation 42).

Consider the simple regression model for body fat percentage versus abdomen circumference.

A. (15 Points) Compute the 0.10, 0.25, 0.50, 0.75 and 0.90 quantiles of abdomen circumference on the data, and use the fitted regression to produce a point estimate of the mean body fat percentage at each of these quantiles. What can you say about these estimates? Which are the least reliable and why?

B. (30 Points) For each of the 0.10, 0.25, 0.50, 0.75 and 0.90 quantiles of abdomen circumference on the data, produce a 95% confidence interval for the mean body fat percentage. Draw a scatterplot of abdomen circumference and body fat percentage, adding the fitted regression line and the confidence intervals. Also, produce 95% prediction interval for the body fat percentage, and draw a scatterplot superimposing the fitted regression line and the prediction intervals. How can you interpret the two types of intervals? Why are they different?
[Hint: use function `predict`, `plot`, `abline` and `lines(c(x0,x0),c(lwr,upr),lwd=2)` where lwr and upr are the lower and upper bounds of the interval to be drawn*].*

Consider now both the simple regression model for body fat percentage versus abdomen circumference, and the multiple linear regression model you fitted in Homework2, with response body fat percentage, and predictors AbdomenC, Weight, Height, NeckC, ChestC, HipC, ThighC, KneeC, AnkleC, BicepsC, ForearmC, WristC.

C. (25 Points) For both models, draw the scatterplot of residuals versus the fitted values and the Q-Q plot of residuals. In addition, perform a Shapiro-Wilk test for the normality of the errors. Provide null and alternative hypothesis and p-value of the test. Use these plots and test to comment on whether the assumptions on the errors for the two regression models are adequate.
   [Hint: use function `qqnorm`, `qqline` and `shapiro.test`)