STAT 462 – Applied Regression Analysis
Fall 2017, Homework 2

Theory/numerical part (30 points)

1. (10 points) Suppose you have the following output in R.

```
> y=c(3,-2,7,4)
> x1=c(5,5,-5,-5)
> x2=c(2,0,2,0)
> X=model.matrix(y~x1+x2)
> solve(t(X)%*%X)
            (Intercept)   x1    x2
(Intercept)        0.50 0.00 -0.25
x1                 0.00 0.01  0.00
x2                -0.25 0.00  0.25
> t(X)%*%y
            [,1]
(Intercept)   12
x1           -50
x2            20
```

   Compute (by hand) the least square estimates $\widehat{\boldsymbol{\beta}}$. Use your estimates to compute (by hand) the fitted values $\widehat{\boldsymbol{y}}$.

2. (20 points) Consider again a simple linear regression model, with $\hat{\beta}_0$ and $\hat{\beta}_1$ the least square estimators. Prove that

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$$

   and

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left[\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]\right)$$

   [Hint: for the variance of the intercept, you need to prove and use the fact that $Cov(\bar{y}, \hat{\beta}_1) = 0$*]*.

Applied part (70 points)

Consider again the dataset contained in "BODY_FAT.TXT". Please read carefully the description of this dataset and the guidelines for this part of the homework before performing the analyses in R and writing your results (file Homework_dataset_guide.pdf). Do not exceed 5 pages.

In the following analyses, use the body fat percentage you re-computed in Homework1. If in Homework1 you found some units with erroneous measurements, you can remove them or present results with and without them (provide an appropriate justification for removing units if you do).

A. (25 Points) Consider again the simple regression models for body fat percentage versus weight, height and abdomen circumference. For each of them:
   - (10 points) Perform a test to see if the predictor has an impact on the body fat percentage, and interpret the result. Provide null and alternative hypotheses, value of the test statistics used, the distribution of the test statistics under the null hypothesis and the p-value of the test.
   - (10 points) Build and interpret a 99% confidence interval for the slope (Use language such as: for every […] increase in […] we are […]% confident that […], etc).
   - (5 points) Compare the three regressions, considering the tests and confidence interval results, as well as the coefficients of determination R^2.

   [Hint: use functions `lm`*,* `summary` *and* `confint`*]*.

B. (15 Points) Does the data contain evidence that, on average, for each additional cm of abdomen circumference the body fat percentage increases by more than 0.5 points? Perform a test to answer this question (provide null and alternative hypotheses, value of the test statistics used, the distribution of the test statistics under the null hypothesis and the p-value of the test).
   [Hint: use functions `summary` *and the argument* `coefficients` *in its output]*.

C. (15 Points) Consider the residuals from the simple linear regression of body fat percentage on abdomen circumference. Draw a scatterplot of these residuals versus each of the other quantitative variables in the dataset that might be used as additional predictors: Weight, Height, NeckC, ChestC, HipC, ThighC, KneeC, AnkleC, BicepsC, ForearmC, WristC. Is there any visual evidence that other variables might capture something that is left out by abdomen circumference? Do these make sense – i.e. are they proxies for factors useful in predicting body fat, which are not conveyed by abdomen circumference?

A. (15 Points) Fit a multiple linear regression model for body fat percentage, with predictors AbdomenC, Weight, Height, NeckC, ChestC, HipC, ThighC, KneeC, AnkleC, BicepsC, ForearmC, WristC. Make sure you provide:
   - Equation of the population model employed;
   - Least square estimates of the parameters $\beta_0, \beta_1, \dots, \beta_{p-1}$ and σ^2;
   - The equation of the estimated regression model;
   - The value of the determination coefficient R^2.

   Perform a test to determine if at least one of the predictors is significantly different from zero, and interpret the results. Provide null and alternative hypotheses, value of the test statistics used, the distribution of the test statistics under the null hypothesis and the p-value of the test. Do you think this model is better than the one that only considered AbdomenC as predictor?
   [Hint: to fit a model with multiple predictors, write all of them in the model formula of `lm` function, separated by `+`; e.g. `y~x1+x2+x3`*]*.