Homework set #2 (100 points total)

Theory Part I: (30 points)

1. (10 points) Prove that

\[ SSR = b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \]

2. (10 points) Prove that

\[ F = \frac{MSR}{MSE} = \left( \frac{b_1}{se(b_1)} \right)^2 = t^2 \]

This has to do with equivalence of the F and t statistics to test \( H_0: \beta_1 = 0 \) vs \( H_1: \beta_1 \neq 0 \). Hint: you can use the equality in (1).

3. (10 points) Prove that, for the simple linear regression of \( y \) on \( x \)

\[ R^2 = (corr(y, x))^2 \]

Applied Part; “body fat” analysis: (70 points)

Please refer to the data description and general guidelines file when performing the following analyses and preparing your write-up.

A. Consider again the simple linear regression of body fat percentage versus abdomen circumference, and the simple linear regression of body fat percentage versus weight/height. For each

- (10 points) Build and interpret 95% and 99% confidence intervals for the slope.
- (5 points) Test and interpret \( \beta_1 = 0 \) vs \( H_1: \beta_1 \neq 0 \) (you can find the p-values in the regression output).

(10 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? (you have to set up and perform a test of hypothesis to answer this question).

B. (5 points) For both the regression of body fat percentage versus abdomen circumference, and the simple linear regression for body fat percentage versus weight/height, produce the fitted line plot with 95% confidence interval for the mean response and 95% prediction interval “bands” superimposed. Comment and interpret.

(10 points) In addition, compute 95% confidence intervals for mean body fat percentage and 95% prediction intervals for body fat percentage in correspondence of the 10, 25, 50, 75 and
90 percentiles of abdomen circumference, and of weight/height ratio. Comment and interpret.

C. (10 points) For both the regression of body fat percentage versus abdomen circumference, and the simple linear regression for body fat percentage versus weight/height, write down and interpret the ANOVA decomposition $SSTot = SSR + SSE$. Relate this to the value of the determination coefficient $R^2$ in each regression.

D. (10 points) For both the regression of body fat percentage versus abdomen circumference, and the simple linear regression for body fat percentage versus weight/height produce the plot of (regular) residuals versus the predictor variable. Use these in conjunction with the “Unusual observations” part of the regression output to comment on whether there are unusual observations in the data (this may depend on whether you have or have not deleted units with anomalous values from your data set).

(10 points) In addition, for each regression, produce the histogram of the (regular) residuals, and use this to comment on whether the assumption of normal errors for the underlying simple regression model seems adequate. In light of the analyses in (A) and (B), why would we want to check this?