# STAT 462 – Applied Regression Analysis
## Fall 2017, Homework 1

**Theory part (30 points)**

1. (15 points) Consider fitting a simple linear regression model with least squares, obtaining the estimators $\hat{\beta}_0, \hat{\beta}_1$ for the intercept and the slope, respectively. Let $\hat{y}_i$ be the fitted values, and $e_i = \hat{\varepsilon}_i$ the residuals. Show that the following equalities (geometric properties of the least square line) holds (5 points each equality):

$$\sum_{i=1}^{n} e_i = 0$$

$$\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i$$

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

2. (15 points) Consider the linear model without any predictor $Y = \beta_0 + \varepsilon$, where $Y$ is the response and $\varepsilon$ the random error term. Write the least square objective function for this model and find the least square estimator for the parameter $\beta_0$, i.e. find the $\hat{\beta}_0$ that minimizes the least square objective function. Is it a known quantity?

**Applied part (70 points)**

Consider 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).

*A.* (5 points) Re-compute body fat percentage using the "Density" variable and Siri's equation: (495/Density) – 450 (note that the values are percentages, so they must be between 0 and 100; change any negative value to 0, and any value >100 to 100). Are there any erroneous values in the variable "SiriBFPerc"? If so, employ your recomputed variable in further analyses. *[Hint: the variable "SiriBFPerc" has only one decimal place. Before comparing it to your recomputed variable, you will need to round your recomputed variable. You can do so with the command* `round(variable,digits=1)`*]*
Along with body fat percentage, consider the variables weight, height and abdomen circumference. Subsample the dataset in order to have only the variables of interest *[Hint: use the command* `dataset[,columns_to_keep]`*]*. Produce a scatter plot matrix of the subsampled dataset *[Hint: use the command* `plot(subsampled_dataset)`*]*. Looking at this will help you see if there are units (men) for which some of the measurements contain obvious mistakes. Are there any? If so, remove those units in further analyses.

B. (20 points) Produce numerical and graphical summaries for body fat percentage, weight, height and abdomen circumference. Do the distributions appear symmetric and bell-

shaped, or are they skewed? What can you observe about the variability of the different variables? Are there some very extreme values in any of the distributions?
Based on these data, is there evidence that the average body fat percentage in the male population exceeds 20%? How about evidence that the average weight in the male population exceeds 180 pounds? Perform hypothesis tests to answer these questions, and provide the p-values.

C. (30 points) Compute the correlation between the body fat percentage and each of the variables weight, height and abdomen. Then, fit three separate simple regression models for body fat percentage versus weight, height and abdomen circumference. For each regression, make sure you provide:

- Model employed;
- Least square estimates of the parameters $\beta_0$, $\beta_1$ and $\sigma^2$;
- The equation of the estimated regression line;
- The value of the determination coefficient R$^2$;
- The regression plot (i.e. scatter plot of Y vs X with the fitted regression line superimposed).

*[Hint: use functions* `lm`*,* `summary`*,* `plot` *and* *`abline`]*.
What can you say about the relationship between the body fat percentage and each of the other three variables? What can you say about the three regressions? (Use language such as: the slope of the regression line represents [...], its estimated value= [...] can be interpreted by saying that [...], etc). Which among weight, height and abdomen circumference appears to be the best predictor for body fat percentage? (Address this comparing the coefficients of determination R$^2$ of the three regressions).

D. (10 points) Form a new predictor variable as the ratio of weight on height. Compute the correlation between the body fat percentage and weight/height ratio, and fit another simple regression model for body fat percentage versus the weight/height ratio (again, provide population model, parameter estimates…). Is the ratio a better predictor than weight or height separately? (Again, use the coefficients of determination R$^2$).

E. (5 points) Do weight/height ratio and abdomen circumference seem to “capture” the same underlying information? (To address this, you will have to look at the relationship between the weight/height ratio and abdomen circumference, in particular their correlation, and possibly the regression of one on the other).