Homework set #1 (100 points total)

**Theory Part I:** (30 points)

Consider fitting a simple linear regression model with least squares, to obtain

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

slope

\[
    b_0 = \bar{y} - b_1 \bar{x}
\]

intercept

Recall fitted values and residuals from the fitted regression line are defined as

\[
    \hat{y}_i = b_0 + b_1 x_i \quad \text{fitted value}
\]

\[
    e_i = y_i - \hat{y}_i \quad \text{residual}
\]

Use algebra to show that the following equalities hold (6 points per equality):

\[
    \sum_{i=1}^{n} e_i = 0
\]

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

\[
    \sum_{i=1}^{n} x_i e_i = 0
\]

\[
    \sum_{i=1}^{n} \hat{y}_i e_i = 0
\]

\[
    \bar{y} = b_0 + b_1 \bar{x}
\]

**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.

Recompute body fat percentage using the “Density” variable and Siri's equation: 
\((495/\text{Density}) - 450\) (to do this, you will need Calc > Calculator). Are there any erroneous values in “SiriBF%”? If so, employ your recomputed variable in further analyses.

Along with body fat percentage, consider the variables weight, height and abdomen circumference. Produce a scatter plot matrix (to do this, you will need Graph > Matrix Plot (Simple)). 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.
A. (10 points) Produce numerical and graphical summaries for body fat percentage, weight, height and abdomen circumference. Is there anything notable about the data? Do the distributions appear symmetric and bell-shaped, or are they skewed? 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? (You will have to set up and perform tests of hypothesis to answer these questions).

B. (30 points) Fit three separate simple regression models for body fat percentage versus weight, height and abdomen circumference. Organize the output in appropriate tables and figures, making sure you include:
- estimates of the parameters $\beta_0$, $\beta_1$, and of $\sigma^2$.
- 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 fitted regression line superimposed)

What can you say about the three regressions? Do they indicate positive or negative relationships? (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).

C. (20 points) Form a new predictor variable as the ratio of weight on height (to do this, you will need Calc > Calculator), and fit another simple regression model for body fat percentage versus weight/height. Is the ratio a better predictor than weight or height separately? (Again, use the coefficients of determination $R^2$).

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

D. (10 points) Compute the 10, 25, 50, 75 and 90 percentiles of abdomen circumference on the data, and use the fitted regression for body fat percentage on abdomen circumference to estimate the mean body fat percentage at each of these percentiles. What can you say about these estimates? Which are the least reliable and why?