Part I: (20 points)

The first part of the homework is theoretical (browsing around Chapters 1 and 2 of the textbook, and with a bit of “symbol-matching”, you should be able to find solutions or hints for most of the points).

Consider the simple linear regression model

(population) \[ Y = \beta_0 + \beta_1 X + \varepsilon, \] with \( \varepsilon \) indep. \( X, E(\varepsilon) = 0, \) var(\( \varepsilon \)) = \sigma^2

(sample) \[ Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \] \( i=1...n \) with \( \varepsilon_i \) iid \( \sim \) \( \varepsilon \)

(normality is not required for the following questions)

A. (2 points) Show that the estimated regression line goes through the point of the sample averages (Av(\( Y \)), Av(\( X \))).

B. (6 points) Find the sample averages and the sample covariance of fitted values \( y_i \), \( i=1...n \) and residuals \( e_i \), \( i=1...n \) (in the process, find the values of \( \Sigma e_i \) and \( \Sigma y_i e_i \)).

C. (6 points) Show that the additive partition of variability SST = SSE + SSR holds.

D. (6 points) Show that the coefficient of determination coincides with the square of the sample correlation between \( X \) and \( Y \): \( R^2 = r_{XY}^2 \)

Part II: (30 points)

The second part of the homework consists of a data analysis. Consider the data scores.mtw (in the collection of example data sets that comes with Minitab). Set:

\( X = \) score for the first test
\( Y = \) score for the second test
Using Minitab, produce:

A. (10 points) Numerical and graphical univariate summaries of X and Y separately. This must include:
- average, st. dev., median and quartiles, extreme values, inter-quartile and max-min range.
- histogram and box-plot
(... plus anything else you might wish to calculate)

B. (10 points) The simple linear regression of Y on X. This must include:
- estimates of the parameters $\beta_0$, $\beta_1$, and of $\sigma^2$.
- the formal expression of the estimated regression line
- the estimate of the mean response on at least one non-sample value of X
- the value of the determination coefficient $R^2$.
- the regression plot (i.e. scatter plot of Y vs X with estimated regression line superimposed)
- the plot of residuals vs X.
(... plus anything else you might wish to calculate)

If you submit Minitab output to answer A. and B., please edit it for readability and highlight the various required elements.

C. (10 points) Write up a page in which, for each of the elements in B. (i.e. only the regression part!) you give a technical and substantive interpretation

(e.g. the slope of the regression line represents [...], its estimated value= [...] can be interpreted by saying that [...], etc).

D. (extra) Write up half a page in which you summarize what you see as the most relevant outcomes of your analysis in both A. and B.

(e.g. in the univariate analysis of, say, Y, average and median are noticeably close/apart, the histogram does/shows not bi-modality, etc.)

(e.g. the simple linear regression has a very high/low explicative power, a linear relationship is noticeably appropriate/inappropriate, the sign of the dependence is/is not surprising, etc.)