# SOEN6811 – winter term 2022

## Assignment 2 on Empirical Validation of Software Measurements

### (3 points in total, weight: 5%, individual work)

Posted on 19/05/2022, due on 28/05/2022

**Link between Assignment 1 and Assignment 2:**

In Assignment 1, measurement data on Length and Effort was provided and the students derived productivity values for each pair of data on effort and length. Productivity data was analyzed and students investigated whether the average productivity can be reliably used for further analysis:

> the values of a productivity were plotted on a chart, average productivity and the control limits where calculated and depicted on the chart; if all productivity data points are within the control limits we can conclude that there are no programmers with an unusually low or high productivity thus the average productivity can be used for the future analysis of productivity.

In Assignment 2 the students will use the measurement data to: 1) validate the measurement data empirically, 2) derive an effort estimation model, and 3) evaluate the prediction power of your model using test data.

**Detailed instructions:**

**PART 1 (1 pts).** In Part 1 of Assignment 2 you will practice three useful simple data analysis techniques for validating empirically software measurement data.
*NOTE: Please update first A2-data-2022 excel file with your own Length-Work Effort data, add your data as row #P86.*

**1.1** Determine the relationship among SLOC data points describing data of one variable (length), which are provided in the updated *A2-data-2022* excel file. Apply the following techniques and record the results:

a) **(0.15 pts)** Averaging – mean, median, standard deviation

b) **(0.6 pts)** Box plot – a summary of the range of a set of data about one variable (where most of the data are clustered, outlier data). Derive the range of acceptable values, the range of values that need a quick review, and the range for outliers. Interpret the results.

**1.2: (0.25 pts)** Apply Bar chart analysis technique to Effort data provided in the *A2-data-2022* updated excel file. Interpret the results.

**PART 2 (1 pts)** Investigate the relationship across two variables in **A2-data-2022** dataset: length (independent variable) and effort (dependent variable). based on the historical data provided in the *A2-data-2022* excel file. The resulting formula is your effort estimation model.

**2.1** For this purpose apply the following techniques:

a) **(0.25 pts)** Scatter plot (or scatter diagram) – visually determine the likelihood of an underlying relationship between two variables: Length and Effort. Identify atypical data (not organized the same way as the other points)

b) **(0.25 pts)** Correlation analysis – use statistical methods to confirm whether there is a true relationship between the variables Length and Effort.

c) **(0.25 pts)** Regression Analysis – understand the type of the relation between the independent variable (Length) and the dependent variable (Effort).

The expected outcome from Part 2.1 is a regression model that will be used to predict future effort from estimated program length (that is, a programmer's effort estimation from length).

**2.2 (0.25 pts)** Effort estimation modeling involves several assumptions. What are the assumptions your had to make in your effort estimation model?

**Part 3 (1 pts).** validate empirically the prediction power of your estimation model using the TEST-A2-data dataset in an excel file included with Assignment 2.

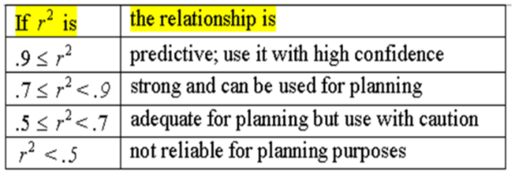
**3.1 (0.25 pts)** Firstly, estimate the effort for each SLOC in the TEST-A2-data data set using the regression model derived in part 2. For example, assume your effort estimation model is Effort = A x SLOC + B where A and B are calculated according to the formulas provided in the corresponding slide of Lecture 3. If Length reported on a line in TEST file is 100SLOC, then your estimated effort will be Effort = A x 100 + B. Write the numerical value of the estimated effort in the corresponding cell of the excel file.

**3.2 (0.5 pts)** Apply Coefficient of Determination R-square technique using regression outputs (the effort estimations calculated in 3.1).:
R-square will give you an estimate of the relationship between movements of a dependent variable (effort) based on an independent variable's (length) movements.

You can use the following formula:
Coefficient of Determination = MSS / TSS
where

TSS – Total Sum of Squares = $\Sigma$ (Yi – Ym)$^2$
MSS – Model Sum of Squares = $\Sigma$ (Y^ – Ym)$^2$
Y^ is the predicted effort value of using the effort estimation model,
Yi is the ith value of ith work effort reported by the programmer, and Ym is the mean value of all work effort data points reported by the programmers.

**(0.25 pts)** Interpret the results. You can use the table below:



| If $r^2$ is | the relationship is |
|---|---|
| $.9 \leq r^2$ | predictive; use it with high confidence |
| $.7 \leq r^2 < .9$ | strong and can be used for planning |
| $.5 \leq r^2 < .7$ | adequate for planning but use with caution |
| $r^2 < .5$ | not reliable for planning purposes |