Project 2: Simple and Multiple Regression in R and Python

1. **Project description**:

   a) Run the two R programs provided using the data from Lab 2 described in the book " Linear Regression with R – an introduction to data modeling" chapter 3 and 4.
   b) Carry out the analysis of chapter 3 and 4
   c) Implement the R programs in python

2. **Goodness of linear regression fit**

Test1: Correlation coefficient

The accuracy of the least square fit or the strength of the association of the two variables is determined by the correlation coefficient r which is given by the formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

Where $\bar{x}$ and $\bar{y}$ are the mean values of the coordinates of the given data. If the variables are increasing together then $r \to +1$, if one decreases as other increases then $r \to -1$ and if r = 0 then they do not relate to each other.

Test 2: Coefficient of Determination

The coefficient of determination of a linear regression model is the quotient of the variances of the fitted values and observed values of the dependent variable. If we denote yias the observed values of the dependent variable, as its mean, and as the fitted value, then the coefficient of determination is:

$$r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

A common way to summarize how well a linear regression model fits the data is via the coefficient of determination or R^2:

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i-\bar{y})^2}$$

where the summations are over all observations. Thus, it is also the proportion of variation in the dependent (forecast) variable that is accounted for (or explained) by the regression model.

Example in R

#fuel is a data set defined by R, lm is the R function that computes the linear regression approximation of the data

```
x_var <-fuel$City
y_var <- fuel$Carbon
  # apply the cor function
r_cor <-cor(x_var, y_var)
r_cor
fit.lm <- lm(Carbon ~ City, data=fuel)
# apply the r_squared
summary(fit.lm)$r.squared
```