**Regression exercise in R**

**Ex 1:** Given RegEx1.csv data including the independent variable and the dependent variable .

**A simple linear regression** model defines the relationships between and using a line defined by an equation in the following form [Brett Lantz, Machine Learning with R]:

,

subject to minimizing the squared error between the actual value and the predicted value () :

where the errors () are known as residuals.

The solution for depends on the value of . It can be obtained using the following formula:

where and terms indicate the mean value of and , respectively.

The value of that results in the minimum squared error is:

As the variance of can be expressed as:

and the covariance formula between and is:

then the formula for can be rewritten as:

**Q1: What are the values for parameters a and b?**

**Ex2:** The Pearson’s correlation between and is defined by

**Q2: What kind of correlation between and ?**

References: Moore, D. S., Notz, W. I, & Flinger, M. A. (2013). The basic practice of statistics (6th ed.). New York, NY: W. H. Freeman and Company.

The correlation r measures the strength of the linear relationship between two quantitative variables.

• r is always a number between -1 and 1.

• r > 0 indicates a positive association.

• r < 0 indicates a negative association.

• Values of r near 0 indicate a very weak linear relationship.

• The strength of the linear relationship increases as r moves away from 0 toward -1 or 1.

• The extreme values r = -1 and r = 1 occur only in the case of a perfect linear relationship.

Text

Description automatically generated with medium confidence

**Q3: Plot a fitting line using for the data points (, ), using functions abline() and lm().**

**Ex3:** As most real-world analyses have more than one independent variable. **Multiple linear regression** is commonly used for most numeric prediction tasks. It generally follows the form of the following equation [Brett Lantz, Machine Learning with R]:

where represents the residual term.

As multi independent variables used, a condensed formulation defined by bold font matrix calculation is:

To minimise the sum of the squared errors between predicted and actual values**,** the best estimate of the vectorcan be computed as:

**(:** the transpose of matrix **; :** the inverse of matrix **)**

**Given** RegEx2.csv data including two independent variables and the dependent variable .

**Q4: Give the answer for .**

**Q5: Using R function lm() to justify your answer to Q3.**