ASSUMPTIONS OF MULTIPLE LINEAR REGRESSION

Provide one sample dataset that you will use for this activity. Ensure that your sample data is

distinct from those of your classmates.

Here are the key points for developing a detailed handout:

**1. Introduce your sample data, and provide a brief discussion about its purpose, structure, and a description of each variable.**

Assumptions of Multiple Linear Regression: Detailed Handout

Sample Dataset: Startup Profit Dataset

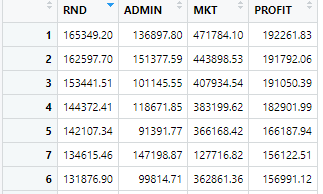
Source: Rahul. (2020). Startup dataset [50startup.csv]. Kaggle. https://www.kaggle.com/datasets/rahul1301/startup-dataset/data

**Purpose of the Dataset**

The purpose of this dataset is to investigate how expenditures in different operational areas affect a startup's profitability. The analysis is intended to identify key factors influencing profit and assist decision-makers in optimizing budget allocations for maximum return on investment.

**Structure of the Dataset:**

The dataset is organized as a table with 50 rows x 4 columns where each row represents a startup, and each column corresponds to a specific variable. The dependent variable in the dataset is Profit, recorded under the PROFIT column. There are three independent variables considered: the first is Research and Development (R&D) expenditure, listed in the RND column; the second is Administration expenditure, found in the ADMIN column; and the third is Marketing expenditure, represented in the MKT column. Each of these variables contributes to understanding the factors influencing the profitability of startups.

**

*Figure 1. First six rows of the dataset*

**Description of each Variable:**

Dependent Variable:

* Profit (PROFIT) :  Represents the net profit generated by the startup.

Independent Variables:

* Research and development expenditure (RND): Represents the amount of money allocated to research and development activities.
* Administration expenditure (ADMIN): Reflects the operational costs associated with administrative activities.
* Marketing Expenditure (MKT):  Denotes the expenditure on marketing and promotional activities.

**2. Build a multiple linear regression model using all the independent variables, regardless of  their significance based on p-values, and name this model “fullmodel.” Then, create a second model named “reducedmodel” that includes only the significant independent variables. Discuss each variable that is significant in predicting the dependent variable, explaining its impact. For example, discuss the sign of the beta coefficient and reference relevant articles that support your discussion.**

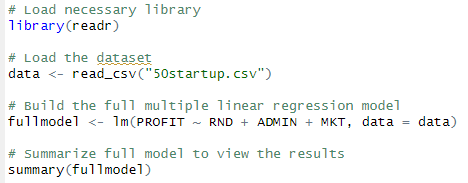
To create the equation for the full model, we use the coefficients from the regression output:

Where:

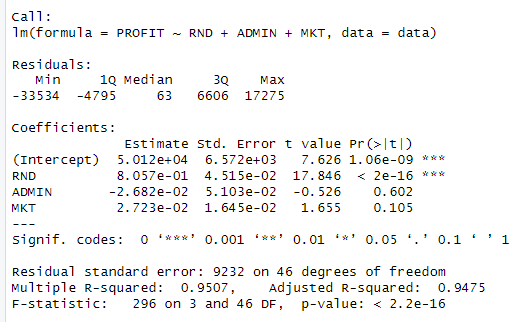
is the intercept or the constant term.

are the coefficients for the predictors.

Using the R software;*R script;*

****

*Result;*



[1]

The analysis identifies **RND (Research and Development)** as a highly significant predictor of **PROFIT**, with a p-value of less than [1]. The beta coefficient for RND is positive , indicating that for every one-unit increase in RND investment, profit increases by 8.057 units, holding all other variables constant. This strong and positive relationship highlights the critical role of R&D in driving profitability. Firms that allocate more resources to R&D often benefit from innovation, improved products, and enhanced competitiveness, which directly contribute to higher profits.

On the other hand, the variables **ADMIN (Administration)** and **MKT (Marketing)** are not statistically significant predictors of profit, with p-values of 0.602 and 0.105, respectively. Although ADMIN has a slightly negative coefficient (-0.02682), suggesting a potential minor negative impact of increased administrative expenses on profitability, the effect is not strong enough to draw meaningful conclusions. Similarly, the positive coefficient for MKT (0.02723) suggests a potential link between marketing efforts and profit, but the lack of statistical significance indicates that the effect is inconsistent or negligible in this model.

**3. Use R software to determine if the assumptions of multiple linear regression are satisfied for the “reducedmodel.” Discuss each test used and include the relevant code (you may present a screenshot of the software to support your discussion). It is also better to present some graphical methods to determine whether the assumptions are satisfied.**

**4. If there is any violation of the assumptions, indicate this in your paper and suggest**

**possible remedial measures, but you do not need to perform the remedial actions.**

**5. At the end of the detailed handout, before the references, include the following table as a summary.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ASSUMPTIONS** | **Method of Detection** | | | **Satisfied (✔) or Violated (X)** | **Possible Remedial Measures** |
| **Graphical** | **Statistical** | **Both** |
| **Linearity (TV & Radio)** |  |  |  |  |  |
| **Normality of Residuals** |  |  |  |  |  |
| **Homoscedasticity** |  |  |  |  |  |
| **Multicollinearity** |  |  |  |  |  |
| **Independence of Residuals** |  |  |  |  |  |