ASSUMPTIONS OF MULTIPLE LINEAR REGRESSION : DETAILED HANDOUT

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

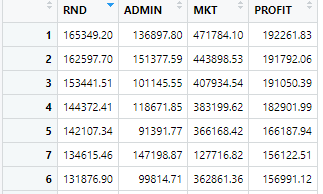
Sample Dataset: **Startup Profit Dataset**

**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.

**Full Regression Model**

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

Where: is the intercept or the constant term and are the coefficients for the predictors.

Using the R software:

|  |  |
| --- | --- |
| *R script* | *Result* |
| **https://lh7-rt.googleusercontent.com/docsz/AD_4nXcvoWnY1vEtesD5lY7vvRU0j7HWiluy9s68Mnd1m_IzggUhmOZyZyw2daL14Ih1f-oCz9f_IpDn1YkERv7FiSN5mJTbjrbeWaKND2GD1pH4ylWut2OCPImz8JgmBgLPCHtvcx92_A?key=HQm2tuXXRIOMCssi7MGEOkN9** | https://lh7-rt.googleusercontent.com/docsz/AD_4nXfhB1YBngA4dQo8tnvcMSo-LGRfmlPMglVqfPFW6vF45d4VPg5U34r4UtBUAuKJXncG7B076NnzBrU4fhAOH4dpBws9c02WuKR6ICYNnvj4DFMeMmoFEWvGFtXTm4rfJ3ZlZWoebg?key=HQm2tuXXRIOMCssi7MGEOkN9 |

Substituting the coefficients from the results into the equation , the full model becomes:

**Full Model:**

|  |
| --- |
|  |

**Discussion of results:**

* : This means of the variation in PROFIT is explained by the independent variables (RND, ADMIN, MKT).
* Adjusted A slightly adjusted value that accounts for the number of predictors.
* Residual standard error: on degrees of freedom.

**Significance of Variables**:

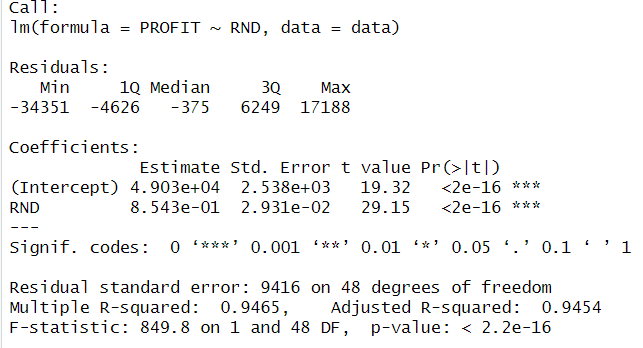
* : The y-intercept, represents the predicted value of PROFIT when RND, ADMIN, and MKT are all zero. In this case, if a company has no investment in R&D, administrative expenses, or marketing, it is expected to have a baseline profit of .
* **RND**: Highly significant , indicating it strongly predicts PROFIT. The positive coefficient suggests that as RND increases by 1 unit, PROFIT increases by 0.8057 units on average.
* **ADMIN**: Not significant , meaning its contribution to predicting PROFIT is not statistically meaningful. Although ADMIN has a slightly negative coefficient , suggesting a potential minor negative impact of increased administrative expenses on profitability, the effect is not strong enough to draw meaningful conclusions.
* **MKT**: Marginally significant , with a positive coefficient , suggesting a weak positive relationship with PROFIT. Results suggest a potential benefits of marketing effort to profitability but the lack of statistical significance indicates that the effect is inconsistent or negligible in this model.

**Reduced Regression Model**

The reduced model should include only significant variables. Based on the p-values:

Here, only **RND** is retained, as ADMIN and MKT are not significant at the typical threshold.

Using R software:



Substituting the coefficients from the results into the equation P, the reduced model becomes:

**Reduced Model:**

|  |
| --- |
|  |

The reduced model simplifies the prediction of profit by focusing solely on R&D expenditure, which has been statistically proven to have a strong and positive impact on profitability. This aligns with prior research that highlights the importance of innovation-driven investments. The new baseline profit is is the expected profit when no investments on the predictors. The coefficient Indicates that for every additional unit increase in R&D expenditure, the profit increases by **$0.8543** on average, holding other factors constant.

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** |  |  |  |  |  |

REFERENCES:

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