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.

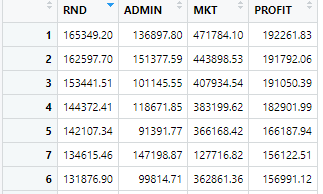
Sample Dataset: **Startup Profit Dataset**

**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 R Software; |
| 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.

**Linearity**

It’s important to evaluate whether linearity assumptions of linear regression is met. One useful diagnostic tool for this purpose is the Residuals vs. Fitted plot, which helps us check the linearity and homoskedasticity assumptions in our reduced model.

|  |
| --- |
| Using R Software; |
|  |
|  |

The Residuals vs. Fitted plot for the reduced regression model shows that the residuals are randomly scattered around the zero line with no obvious patterns. This suggests that the linearity assumption is met, meaning the model captures the linear relationship well. The residuals also exhibit consistent spread across fitted values, indicating constant variance (homoskedasticity). While some residuals deviate slightly, there are no significant outliers. Overall, the plot suggests the model satisfies key assumptions, though additional statistical tests could further confirm these findings.

**Independence of Residuals**

In linear regression, the assumption of independence of residuals is crucial to ensure that the residuals (errors) from the model are not correlated. When residuals are correlated, it indicates that the model might be missing important patterns in the data, which can lead to biased estimates and incorrect conclusions. The Durbin-Watson test is commonly used to check for autocorrelation in the residuals, particularly in time-series data or data with ordered observations.

|  |
| --- |
| Using R Software; |
|  |

Interpretation: The Durbin-Watson statistic of , combined with a very small p-value of , indicates significant positive autocorrelation in the residuals. This suggests that the assumption of independence of residuals is violated. The presence of autocorrelation implies that the residuals are not independent and that there may be underlying patterns or trends in the data that the model has not captured.

**Homoscedasticity**

In linear regression, the assumption of homoscedasticity requires that the residuals have constant variance across all levels of the fitted values. When this assumption is violated (heteroscedasticity), it can lead to inefficient estimates and misleading standard errors. To check for constant variance, a Scale-Location plot (also known as a spread-location plot) is commonly used. This plot visualizes whether the residuals are spread evenly across the range of fitted values.

|  |
| --- |
| Using R Software; |
|  |
|  |

The Scale-Location plot shows a generally horizontal red line with no clear upward or downward trend, indicating that the spread of the residuals remains relatively consistent across the fitted values. This suggests that the assumption of homoscedasticity is reasonably met. There are no clear patterns or trends that would suggest heteroscedasticity, as the residuals appear to be evenly distributed.

Overall, the plot supports the conclusion that the residuals have constant variance, meeting the homoscedasticity assumption of linear regression. However, additional statistical tests, such as the Breusch-Pagan test, could be used to confirm this observation more objectively.

**Normality of Residuals**

The assumption of normality of residuals is essential in linear regression to ensure the validity of hypothesis tests and confidence intervals. This assumption can be checked using visual methods like the Q-Q plot and statistical tests such as the Shapiro-Wilk test. The Q-Q plot provides a graphical assessment, while the Shapiro-Wilk test offers a formal statistical approach.

|  |
| --- |
| Using R Software; |
|  |
|  |
|  |

The Q-Q plot shows that most residuals align well with the theoretical quantiles, suggesting approximate normality, although slight deviations at the tails indicate potential departures from normality. However, the Shapiro-Wilk test, with a W value of 0.93708 and a p-value of 0.01034, leads to the rejection of the null hypothesis of normality at the 5% significance level. This confirms that the residuals are not perfectly normally distributed.

Overall, the normality assumption is violated, as indicated by the Shapiro-Wilk test, despite the Q-Q plot suggesting approximate normality. The statistical evidence clearly points to a departure from normality.

**Multicollinearity**

Since the reduced model has only one predictor, multicollinearity is not a concern.

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.

The analysis identified several key violations of regression assumptions. The Durbin-Watson test indicated significant autocorrelation in the residuals, violating the independence assumption. This suggests that the residuals are not independent, which may lead to biased estimates and inaccurate inferences. The Scale-Location plot further indicated potential heteroscedasticity, showing that the residuals do not have constant variance across the range of fitted values. This non-constant variance can impact the efficiency of the regression estimates and lead to incorrect standard errors, affecting hypothesis tests and confidence intervals.

Additionally, the Shapiro-Wilk test revealed a violation of the normality assumption, with residuals not following a normal distribution, despite the Q-Q plot showing only slight deviations from normality. This violation suggests that statistical tests relying on the normality of residuals may be less reliable. To address these issues, potential remedies include applying transformations to the variables, using robust regression methods to account for heteroscedasticity and non-normality, employing generalized least squares (GLS) to handle autocorrelation, or using bootstrapping techniques to improve the reliability of statistical inferences. While these violations are noted, no remedial actions are performed in this analysis.

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

The table below provides an overview of the assumptions tested, the methods used, the results, and whether each assumption was satisfied, violated, or not applicable. This summary offers a quick reference to understand the overall model diagnostics and highlights areas that may need further attention or remedial action.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ASSUMPTIONS** | **Method of Detection** | | | **Satisfied (✔) or Violated (✖)** | **Possible Remedial Measures** |
| **Graphical** | **Statistical** | **Both** |
| **Linearity** | **✔** |  |  | **✔** |  |
| **Independence of Residuals** |  | **✔** |  | **✖** | Possible approaches include adding lagged variables, considering omitted variables, or using advanced regression techniques like generalized least squares that account for autocorrelation |
| **Homoscedasticity** | **✔** |  |  | **✔** |  |
| **Normality of Residuals** |  |  | **✔** | **✖** | Transforming the dependent or independent variables using logarithmic, square root, or inverse transformations |
| **Multicollinearity** |  |  |  | N/A |  |

REFERENCES:

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