ASSUMPTIONS OF MULTIPLE LINEAR REGRESSION:

DETAILED HANDOUT

**GROUP 5-BSM 4-3**

GALINDO, DIVINE MERCY

ROLDAN, JHUSTINE RAE

ARRIEGA, ALEXANDER JR.

**I. Introduction to the Sample Dataset**

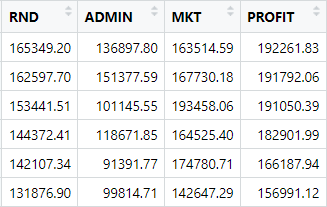
The sample dataset chosen is called **Startup Profit Dataset** consists of fifty observations and four variables collected from various startups. It is a valuable resource for analyzing the relationship between operational expenditures and profitability in startups. The primary goal of this dataset is to identify key drivers of profitability and provide actionable insights for decision-makers aiming to optimize their budget allocations for maximum return on investment.

**A. 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.

**B. 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*

**C. 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.

**II. Building and Analyzing Linear Regression Models**

This section outlines the development of two multiple linear regression models: **fullmodel**, which includes all independent variables, and **reducedmodel**, which incorporates only significant predictors. The significance of each variable is analyzed, including the direction and magnitude of its impact on the dependent variable, supported by relevant literature.

**A. 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; |
|  |

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

Full Model:

|  |
| --- |
|  |

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 approximately .
* **RND**: Highly significant , indicating it strongly predicts PROFIT. The positive coefficient suggests that as R&D expenditure increases by 1 unit, PROFIT increases by units on average. This highlights the crucial role of R&D investment in enhancing profitability, consistent with literature emphasizing the importance of innovation in driving financial performance.
* **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. This aligns with some studies that find administrative costs may not directly contribute to profitability.
* **MKT**: Significant , with a positive coefficient , suggesting a meaningful positive relationship with PROFIT. As marketing expenditure increases by 1 unit, PROFIT increases by units on average. This indicates that marketing efforts positively impact profitability, supporting research that emphasizes the role of marketing in revenue growth and customer acquisition.

**B. Reduced Regression Model**

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

Here, only **RND** and **MKT** is retained, as ADMIN is 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 only on R&D and Marketing expenditure, which has been statistically proven to have a significant impact on profitability. 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 $on average, holding other factors constant. The coefficient Indicates that for every additional unit increase in MKT expenditure, the profit increases by $on average, holding other factors constant.

**III. Assessing Assumptions of the Multiple Linear Regression for the Reduced Model**

**A. 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.

**B. 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. The null hypothesis of the Durbin-Watson test is that residuals are independent. A indicates no evidence of autocorrelation.

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

The **Durbin-Watson test** result indicates **positive autocorrelation** in the residuals. This **violates the assumption of independence of residuals** in multiple linear regression, as residuals are not randomly distributed and follow a systematic pattern.

**C. Homoscedasticity**

Homoscedasticity is a fundamental assumption in multiple linear regression, requiring that the variance of residuals remains constant across all levels of the fitted values. Violations of this assumption, known as heteroscedasticity, can lead to inefficient estimates and unreliable hypothesis tests. To assess homoscedasticity, we use visual diagnostics, such as the Scale-Location plot, which reveals patterns in the spread of residuals, and formal statistical methods like the Breusch-Pagan test, which tests for non-constant variance. Together, these methods provide a comprehensive evaluation of whether the homoscedasticity assumption holds in the regression model.

|  |
| --- |
| Using R Software to perform Scale-Location plot; |
|  |
|  |

In this plot, the standardized residuals are plotted against the square root of the fitted values, with a red trend line included to highlight any patterns. Upon examining the plot, the residuals appear to be scattered relatively evenly around the red trend line, without any noticeable funnel or cone-shaped pattern. This suggests that the variance of the residuals remains fairly constant across the range of fitted values.

While there are a few labeled points (such as observations 50, 15, and 16) indicating potential outliers, these do not seem to contribute to any systematic changes in the variance of the residuals. In conclusion, the Scale-Location plot does not indicate a significant violation of the homoscedasticity assumption. This suggests that the variance of the residuals is reasonably consistent, supporting the validity of the multiple linear regression model in this respect.

|  |
| --- |
| Using R Software to perform **Breusch-Pagan test**; |
|  |

The **Breusch-Pagan test** result indicates that the p-value is greater than , so we fail to reject the null hypothesis of homoscedasticity. This means the assumption of constant variance of residuals is **not violated** in the reduced model, and homoscedasticity is satisfied.

**D. 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 reveals deviations from the reference line, particularly in the tails, suggesting potential non-normality of the residuals. This is confirmed by the Shapiro-Wilk test , where the p-value is less than , leading to the rejection of the null hypothesis of normality. Therefore, the assumption of normality of residuals is **violated**. To address this, transformations of the dependent variable or the use of robust regression methods may be considered.

**E. Multicollinearity**

The assumption of multicollinearity is crucial to ensure the reliability of the model estimates. Multicollinearity occurs when two or more predictors in the model are highly correlated, leading to difficulty in determining the unique effect of each predictor on the dependent variable. A common method to test for multicollinearity is by calculating the Variance Inflation Factor (VIF) for each predictor.

The VIF quantifies how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors.

* A VIF value of **1** indicates no multicollinearity.
* Values between **1 and 5** suggest low to moderate multicollinearity.
* A VIF greater than **10** is typically considered indicative of high multicollinearity, warranting further investigation.

By assessing VIF values, we can determine whether multicollinearity is a concern and take appropriate corrective measures if needed.

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

The results reveal that both predictors, RND and MKT, exhibit high VIF values exceeding the threshold of **10**, which is a clear indication of significant multicollinearity. This result confirms that the multicollinearity assumption has been **violated**. This suggests that these variables are highly correlated with each other, leading to inflated standard errors for their coefficients. Consequently, the reliability and interpretability of the regression coefficients for RND and MKT are compromised.

**IV. Violations of Regression Assumptions and Remedial Measures**

In multiple linear regression, ensuring that key assumptions are met is essential for producing valid and reliable results. Based on the assumption check results, certain violations were identified in the model. The Durbin-Watson test result ) revealed positive autocorrelation in the residuals, indicating a violation of the independence assumption. This suggests that the residuals are not randomly distributed but follow a systematic pattern. To address this issue, possible remedies include incorporating lagged variables to account for autocorrelation, applying generalized least squares (GLS), or using a time series model such as ARIMA if the data has temporal dependencies.

The assumption of normality of residuals was also found to be violated. The Q-Q plot showed deviations from the reference line, particularly in the tails, and the Shapiro-Wilk test confirmed non-normality. Remedial measures for this violation include transforming the dependent variable using log, square root, or Box-Cox transformations, applying non-parametric or robust regression techniques, or increasing the sample size, which can reduce sensitivity to non-normality.

Additionally, multicollinearity was detected between the predictors RND and MKT, as indicated by high VIF values exceeding the threshold of 10. This compromises the reliability and interpretability of the regression coefficients. To mitigate multicollinearity, one can remove or combine highly correlated predictors, standardize the predictors, or use advanced techniques such as ridge regression or partial least squares regression.

While the assumptions of linearity and homoscedasticity were satisfied, the identified violations of independence, normality, and multicollinearity require attention. The suggested remedial measures provide a structured approach to addressing these issues, ensuring the model’s validity and robustness.

**V. 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** |  | **✔** |  | **✖** |  |

**REFERENCES:**

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