*REPORT*

*MULTIPLE REGRESSION ANALYSIS*

*NAME: ADITI NIMBARTE*

*PRN: 23060641029*

*CLASS: MSc I Div A  
SUBJECT: LINEAR MODELS*

*GROUP NO. 11: LINEAR REGRESSION PROJECT FOR SALES FORECASTING*

***INTRODUCTION***

**Multiple Linear Regression:**

Simple linear regression enables statisticians to predict the value of one variable using the available information about another variable. Linear regression attempts to establish the relationship between the two variables along a straight line.

Multiple regression is a type of regression where the dependent variable shows a **linear** relationship with two or more independent variables. It can also be **non-linear**, where the dependent and independent variables do not follow a straight line.

Both linear and non-linear regression track a particular response using two or more variables graphically. However, non-linear regression is usually difficult to execute since it is created from assumptions derived from trial and error.

**Formula for Multiple Linear Regression:**



Where:

* **yi** is the dependent or predicted variable.
* **β0** is the y-intercept, i.e., the value of y when both xi and x2 are 0.
* **β1** and **β2** are the regression coefficients representing the change in y relative to a one-unit change in **xi1** and **xi2**, respectively.
* **βp** is the slope coefficient for each independent variable.
* **ϵ** is the model’s random error (residual) term.

**Assumptions of Multiple Linear Regression:**

Multiple linear regression is based on the following assumptions:

**1. A linear relationship between the dependent and independent variables**

The first assumption of multiple linear regression is that there is a linear relationship between the dependent variable and each of the independent variables. The best way to check the linear relationships is to create scatterplots and then visually inspect the scatterplots for linearity. If the relationship displayed in the scatterplot is not linear, then the analyst will need to run a non-linear regression or transform the data using statistical software, such as SPSS.

**2. The independent variables are not highly correlated with each other**

The data should not show multicollinearity, which occurs when the independent variables (explanatory variables) are highly correlated. When independent variables show multicollinearity, there will be problems figuring out the specific variable that contributes to the variance in the dependent variable. The best method to test for the assumption is the Variance Inflation Factor method.

**3. Homoscedasticity- The variance of the residuals is constant**

Multiple linear regression assumes that the amount of error in the residuals is similar at each point of the linear model. This scenario is known as homoscedasticity. When analysing the data, the analyst should plot the standardized residuals against the predicted values to determine if the points are distributed fairly across all the values of independent variables. To test the assumption, the data can be plotted on a scatterplot or by using statistical software to produce a scatterplot that includes the entire model.

**4. Independence of observation**

The model assumes that the observations should be independent of one another. Simply put, the model assumes that the values of residuals are independent. To test for this assumption, we use the Durbin Watson statistic.

The test will show values from 0 to 4, where a value of 0 to 2 shows positive autocorrelation, and values from 2 to 4 show negative autocorrelation. The mid-point, i.e., a value of 2, shows that there is no autocorrelation.

**5. Multivariate normality**

Multivariate normality occurs when residuals are normally distributed. To test this assumption, look at how the values of residuals are distributed. It can also be tested using two main methods, i.e., a histogram with a superimposed normal curve or the Normal Probability Plot method.

**Goodness of Fit of the Model:**

From the multiple linear regression model output, you can determine the fitted multiple linear regression equation. This equation is useful to make predictions about the mpg value for new observations. You can go through a few metrics as discussed below to evaluate how “good” the multiple regression model in R fits the data:

* Multiple R-Squared:

This metric of **multiple regression in R**measures the strength of the linear relationship between the response variable and the predictor variables. A multiple R-squared of 1 shows a perfect linear relationship, whereas a multiple R-squared of 0 shows that no linear relationship exists.

Multiple R alternatively denotes the square root of R-squared.  It is the variance proportion in the response variable which can be explicated by the predictor variables.

* Residual Standard Error:

This metric of **multiple regression in R**calculates the average distance that the observed values fall from the regression line.

**How to choose a good linear model?**

A model that fulfils the application’s conditions is the minimum requirement. However, we may find different models that fulfil this criterion. So, we might get confused about how to select between various valid models of multiple regression plot in R.

The 3 most common tools to choose a good linear model are:

* The coefficient of determination R2
* The p-value associated with the model
* The Akaike Information Criterion

It is important to note that the first two approaches are suitable for both simple and multiple linear regression; the third one applies to multiple linear regression only.

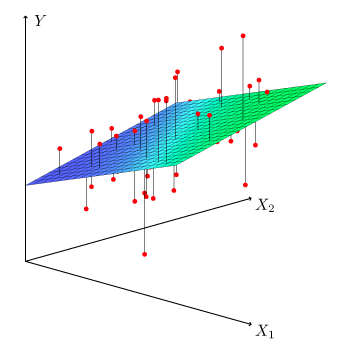
**Multiple linear regression models are defined by the equation:**

Y=β0+β1X1+β2X2+⋯+βpXp+ϵ

It is similar than the equation of simple linear regression, except that there is more than one independent variables (X1, X2,…,Xp).

Estimation of the parameters β0,…,βp by the method of least squares is based on the same principle as that of simple linear regression, but applied to p dimensions. It is thus no longer a question of finding the best line (the one which passes closest to the pairs of points (yi,xi)), but finding the p-dimensional plane which passes closest to the coordinate points (yi,xi1,…,xip).

This is done by ***minimizing* the sum of the squares of the deviations of the points on the plane**:



***PROBLEM STATEMENT***

To predict the sales of coffee depending on the factors like Cost of Goods Sold (COGS), Profit Difference, Inventory Margin, Market Size, Marketing, Profit, Total Expenses, etc by developing a multiple linear regression model that can:

* Predict daily sales for each coffee shop based on historical data and the provided attributes.
* Improve the accuracy of sales predictions compared to traditional forecasting methods.
* Provide insights into the factors that most significantly influence daily sales.

This will allow the coffee chain to:

* Optimize inventory management by aligning stock levels with predicted demand.
* Effectively plan staffing schedules to meet customer needs and avoid unnecessary labour costs.
* Make data-driven decisions about marketing campaigns and promotions.
* Gain a deeper understanding of customer behaviour and preferences.

By achieving these goals, the coffee chain can improve its operational efficiency, profitability, and customer satisfaction.

***DATASET***

***Source data link:***

[***https://www.kaggle.com/datasets/amruthayenikonda/coffee-chain-sales-dataset***](https://www.kaggle.com/datasets/amruthayenikonda/coffee-chain-sales-dataset)

***Dataset reduced to required variables:***

Marketing, Target profit, Margin, Target cogs, Total expenses (1063 observations)

******

The Coffee Sales Data dataset provides valuable insights into the performance of a coffee chain across various locations.

***Key Attributes:***

1. Area Code: A unique identifier for different geographical areas or regions where the coffee chain operates.

2. COGS (Cost of Goods Sold): The total cost incurred by the coffee chain in producing or purchasing the products it sells.

3. Difference between Actual and Target Profit: This attribute indicates how well the company performed in terms of profit compared to its target. It reflects the financial performance against predefined goals.

4. Date: The date of sales transactions, which allows for time-based analysis of sales trends and patterns.

5. Inventory Margin: The difference between the cost of maintaining inventory and the revenue generated from selling those inventory items.

6. Margin: The profit margin, which is the percentage of profit earned from sales. It's a critical financial metric.

7. Market Size: Information about the size of the market in each area, helping to understand the potential customer base and market dynamics.

8. Profit: financial gain achieved by the company after deducting the cost of goods sold (COGS) and other expenses from the revenue generated through sales.

9. Sales: represent the revenue generated from the coffee chain's products, reflecting its financial performance and customer demand.

***STEPS***

1. Load Necessary Libraries and Datasets:

* + We load the required libraries: readxl for reading Excel files, ggplot2 for creating plots, and car for some advanced statistical functions.
  + We read the dataset from a CSV file. The file.choose() function opens a dialog box to select the file interactively.

2. Splitting Dataset into Training and Testing Sets

* + 70% of the data is randomly sampled to create the training dataset (train\_data), and the remaining 30% is assigned to the testing dataset (test\_data).

3. Exploratory data analysis in excel

* We found the correlation between Sales and different factors on which sales are dependent.
* After finding out the correlation, we choose 5 predictor variables Marketing, Target profit, Margin, Target cogs, Total expenses which had the best adjusted R2.

4. Fitting Multiple Linear Regression Model

* + We fit a multiple linear regression model where Sales is the response variable and Marketing, Target profit, Margin, Target cogs, Total expenses are the predictor variables, using the training data.
  + The summary () function provides detailed information about the fitted model, including coefficients, standard errors, t-values, and p-values.

5. Model Adequacy

* + We conduct a model adequacy test to assess the assumptions of the multiple linear regression model.

6. Actual vs. Fitted Plots for Training and Testing Sets

* + We create scatter plots comparing actual sales values with predicted sales values for both training and testing datasets.
  + The red line represents the trend line fitted by the model.
  + Legends are added to indicate the data points (blue) and the trend line (red).

*Activities in brief:*

* + We load the necessary libraries, read the dataset, and split it into training and testing sets as before.
  + EDA is done to select the predictor variables.
  + The linear regression model is fitted to the training data, and a summary of the model is printed.
  + QQ plot, Actual vs Fitted plots for both training and testing sets, and Residuals vs Fitted plots are generated and saved.
  + Residuals vs Actual plot is created to assess the model's adequacy.
  + All plots are saved to the specified location.
  + Legends are added to the plots to improve readability and interpretation.
  + Detailed comments are provided throughout the code for clarity and understanding.

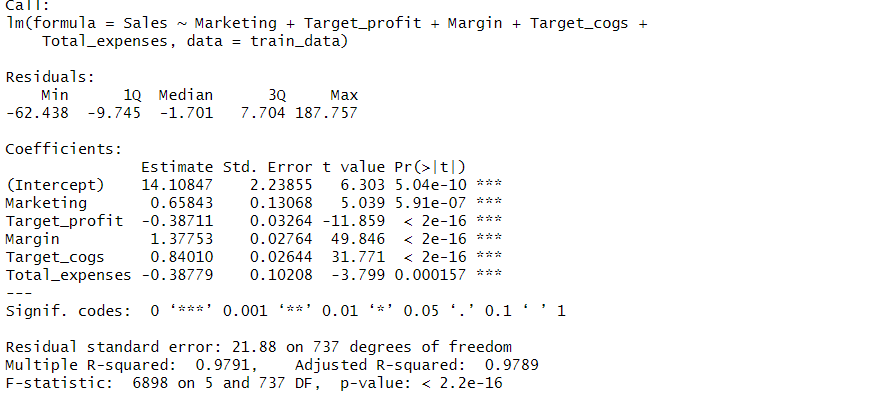
***EXPLORATORY DATA ANALYSIS***



***CODE FILE***



***OUTPUT & INTERPRETATION***

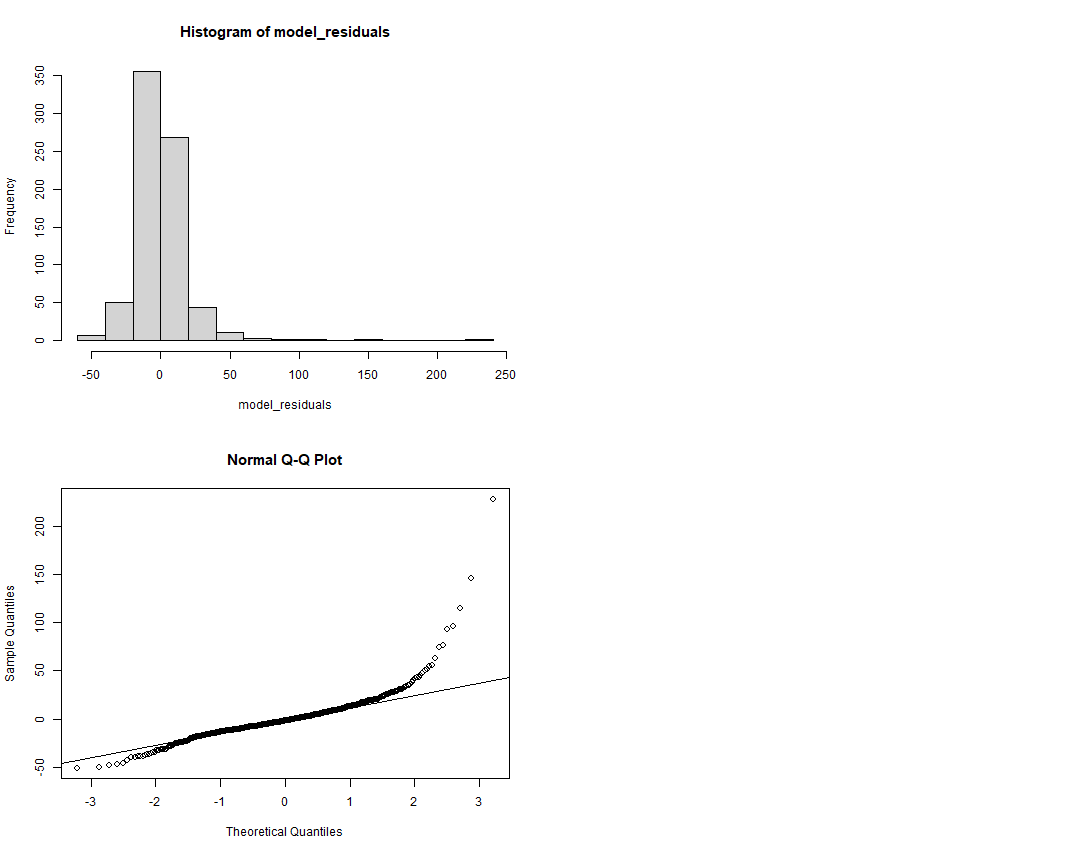


*fig(1)*

* From the summary, we get our multiple linear regression equation as:

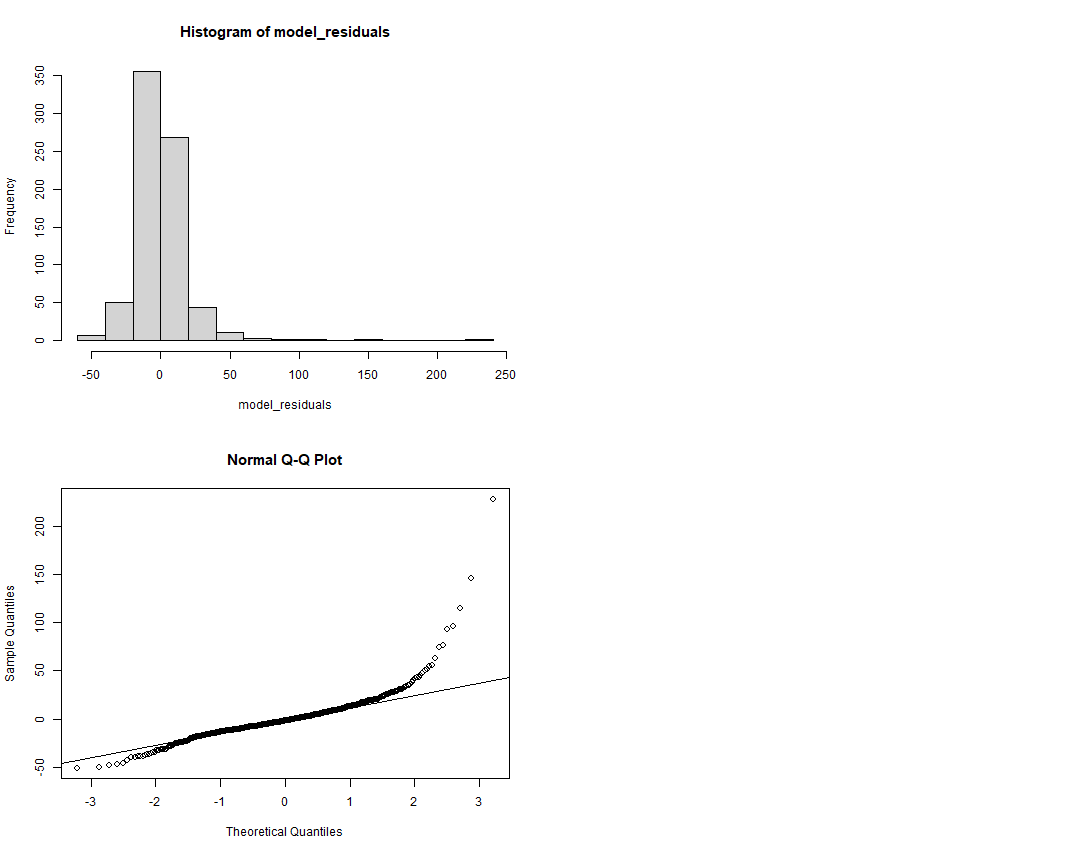
Y = 14 + (-0.65\*marketing) + (-0.38711\*Target\_profit) +1.37\*Margin +0.84\*Target\_cops +(-0.38779\*Total\_expenses) ± e

* Multiple R square is 0.97 shows that there is almost a perfect linear relationship between the response variable & the predictor variables.



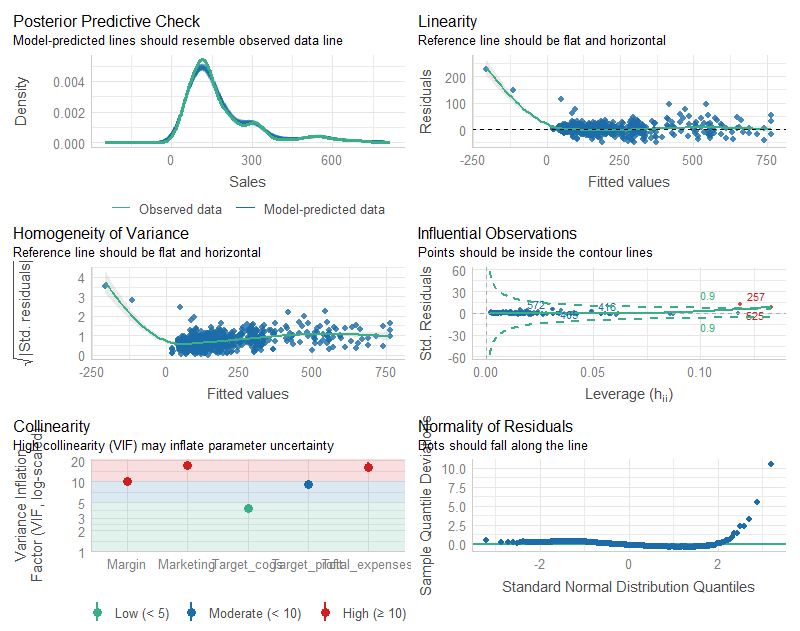
*fig(2)*

* This histogram shows the frequency distribution of the residuals from a model. Residuals are the differences between the observed and predicted values.
* A large concentration of residuals is near 0, indicating that for many observations, the model's predictions were close to the actual values.
* The distribution of residuals appears to be right-skewed, with a tail extending towards the positive values. This skewness suggests that there might be a systematic overprediction by the model for some data points.



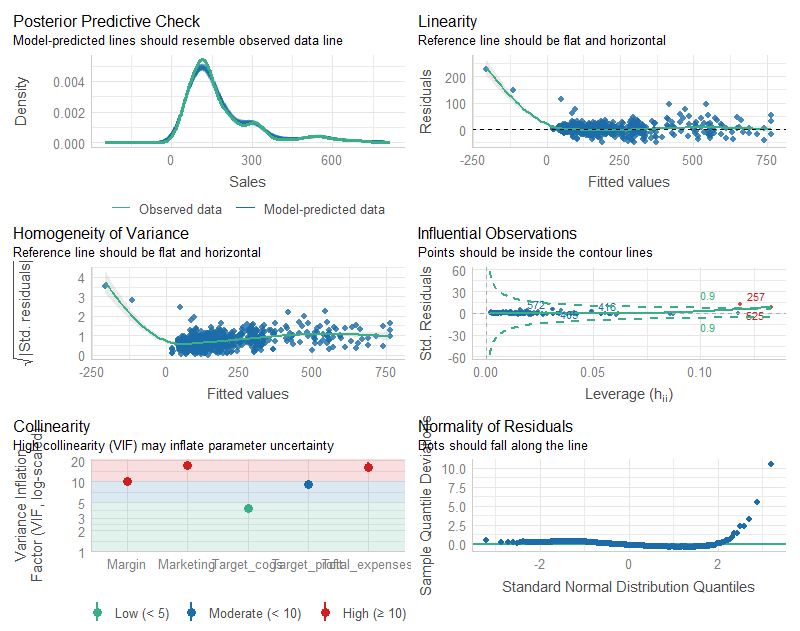
*fig(3)*

* + A Q-Q (quantile-quantile) plot is used to compare the distribution of a dataset with a theoretical normal distribution.
  + The data points form a curve rather than lying on the reference line, which suggests that the residuals do not follow a normal distribution.
  + The curve’s shape, being higher on the ends and lower in the middle, indicates a "heavy-tailed" distribution—there are more extreme values than would be expected in a normal distribution.



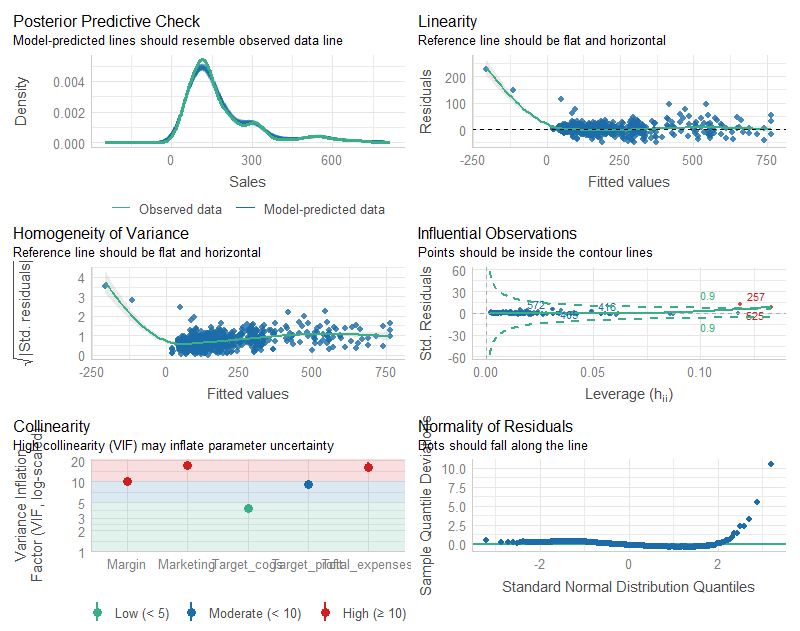
*fig(4)*

* + This plot compares the observed data with the data predicted by the model.
  + The model-predicted data should resemble the observed data line for the model to be considered a good fit.
  + There appears to be a discrepancy between the observed and predicted data around the peak of the sales density, which suggests that the model might not be perfectly capturing the underlying distribution of sales data.



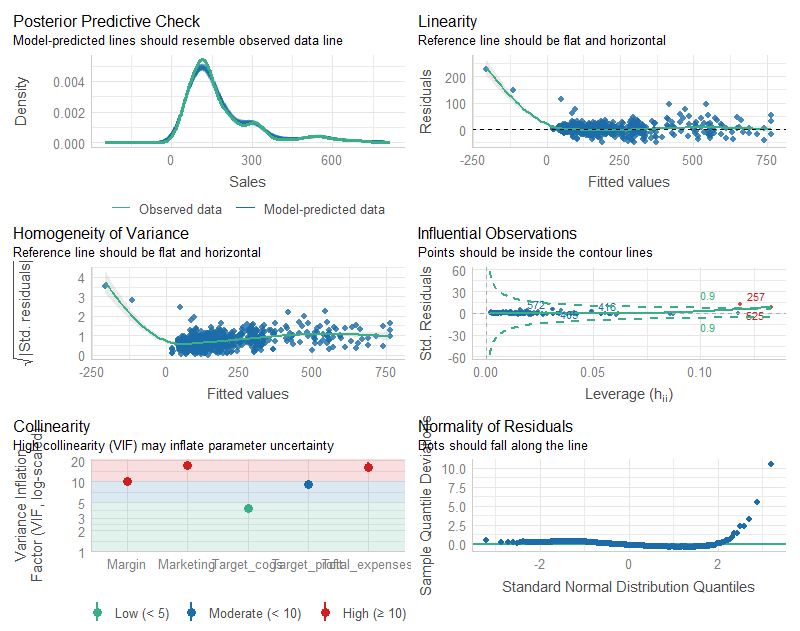
*fig(5)*

* + This scatter plot assesses the assumption of linearity in a regression model.
  + The reference line (ideally flat and horizontal) represents what the relationship would look like if it were perfectly linear.
  + The pattern of the residuals (deviations from the reference line) shows a clear curve, indicating that the relationship is not linear. This suggests that a simple linear model may not be the best fit for the data.



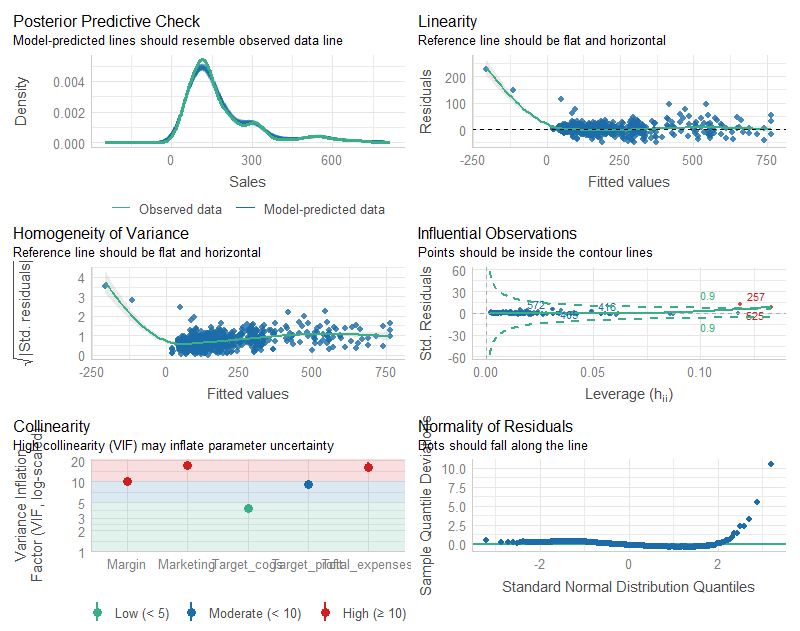
*fig(6)*

* + This plot is used to check the assumption of equal variance (homoscedasticity) in the residuals of a regression model.
  + The residuals should be spread randomly around the horizontal reference line, with no discernible pattern.
  + However, the plot indicates an increasing spread in residuals as the fitted values increase, which suggests heteroscedasticity – the variance of the residuals is not constant across levels of the independent variable.



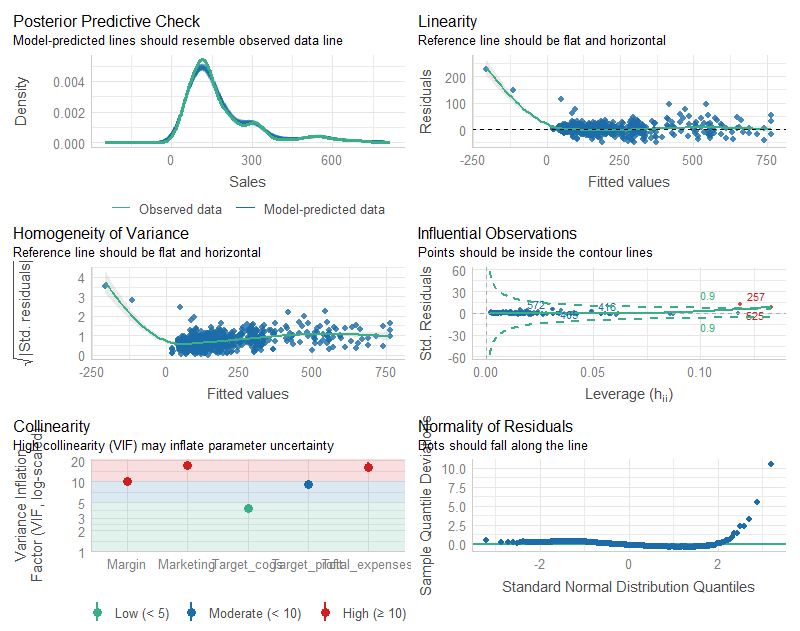
*fig(7)*

* + This plot shows the standardized residuals against the leverage of each observation to identify points that have a significant influence on the model's calculations.
  + Points outside of the Cook's distance contours are considered to be influential. Here, there are a few points outside the contours, with two points very far out, suggesting they have high leverage and are potentially having a disproportionately large impact on the model.



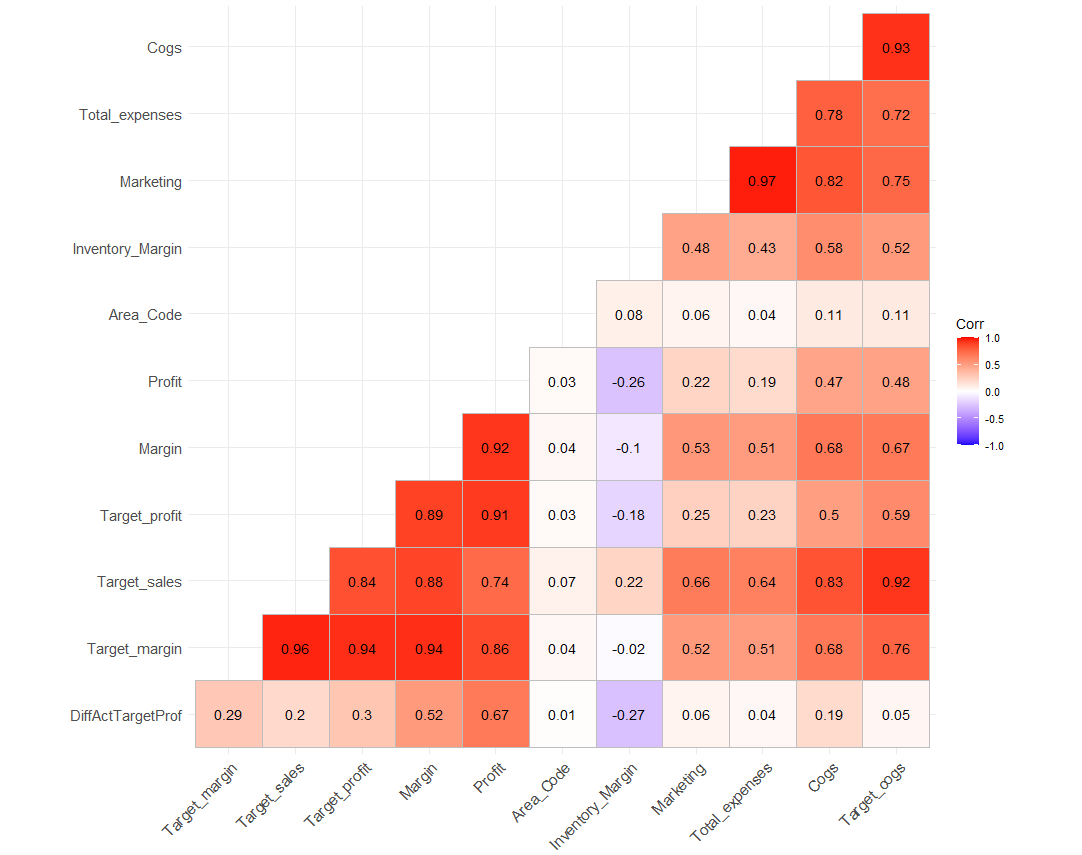
*fig(8)*

* + This plot is used to assess the normality of the residuals. Ideally, the points should fall along the reference line if the residuals are normally distributed.
  + The curve of the points in this plot indicates a deviation from normality, especially with a few extreme values in the tails. This suggests that the residuals have a non-normal distribution, which can impact the validity of some statistical tests.



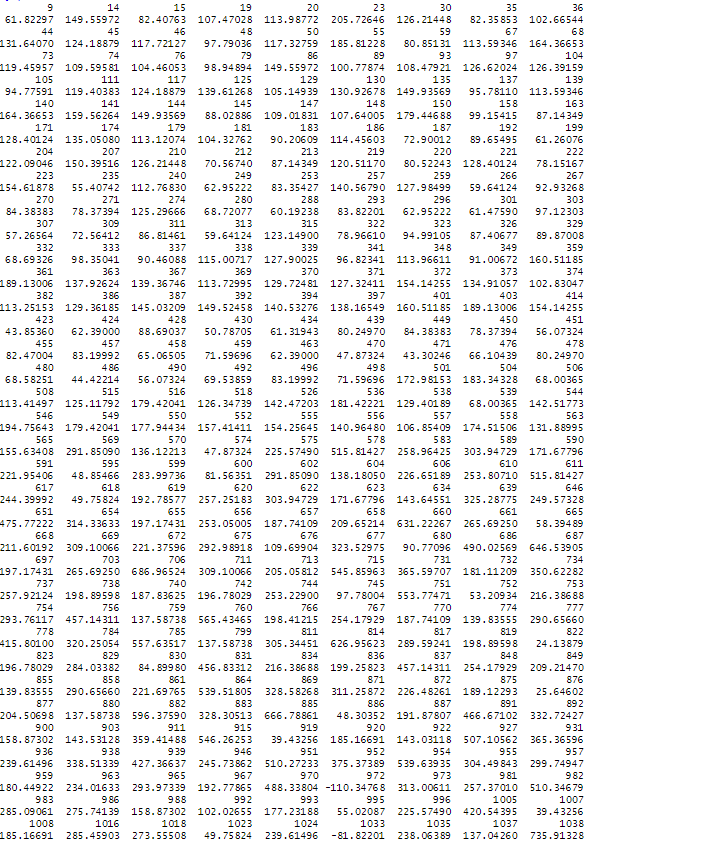
*fig(9)*

* + The VIF values are used to detect the presence of multicollinearity among the independent variables in a regression model. Multicollinearity can inflate the variance of the parameter estimates and make them unstable.
  + VIF values above 10 (marked in red) indicate high multicollinearity. In this plot, two variables seem to have a VIF greater than 10, suggesting that they may be collinear with other independent variables in the model.



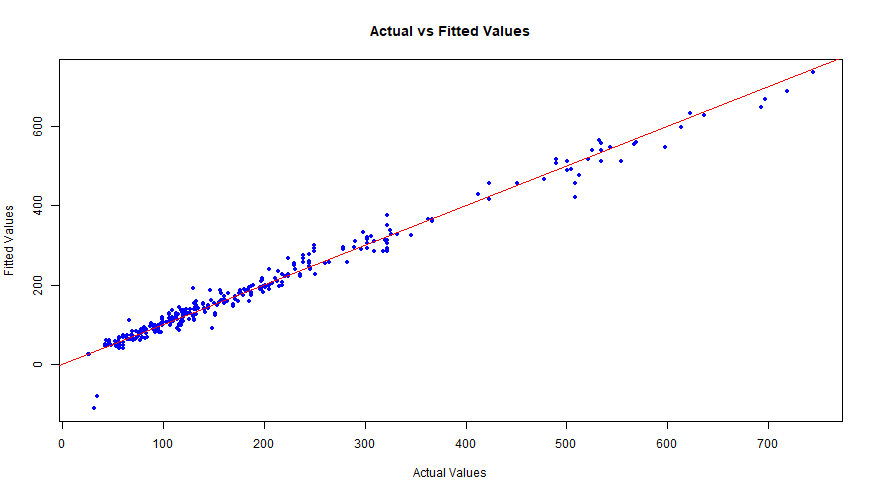
*fig(10)*

* + This heatmap shows the correlation coefficients between different variables.
  + The colors represent the strength and direction of the correlation: red for a strong positive correlation, blue for a strong negative correlation, and white for no correlation.
  + Some pairs, like Target\_profit and Profit or Cogs (Cost of Goods Sold) and Total\_expenses, show very strong positive correlations (close to 1).
  + Variables such as DiffActTargetProf seem to have a low to moderate positive correlation with several variables and a negative correlation with Area\_Code.



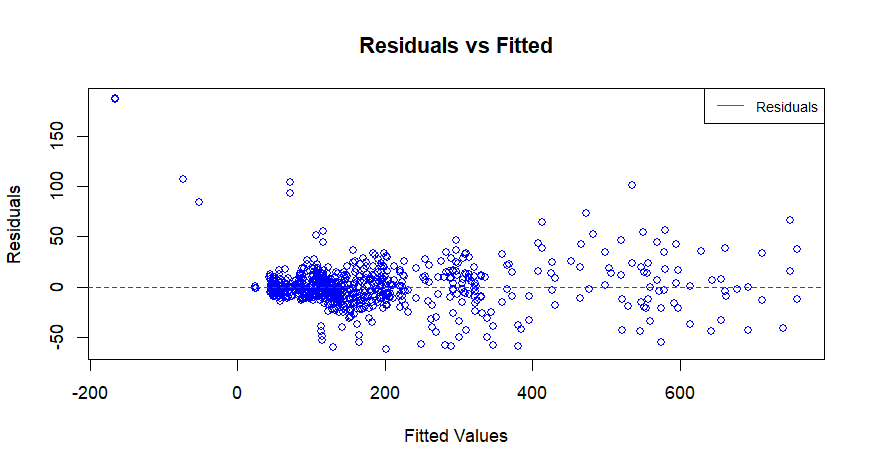
*fig(11)*

The fig() shows the predicted values of Sales based on the tested model.



*fig(12)*

The graph fig() suggests that the model's predictions closely match the actual observed values for the coffee sales. This indicates a good fit between the model and the test data.



*fig(13)*

* The residuals "bounce randomly" around the 0 line. This suggests that the assumption that the relationship is linear is reasonable.
* The residuals roughly form a "horizontal band" around the 0 line. This suggests that the variances of the error terms are equal.
* No one residual "stands out" from the basic random pattern of residuals. This suggests that there are no outliers.

***CONCLUSION***

* + The multiple linear regression model seems to be appropriate for predicting sales based on marketing, target profit, Margin, Target cogs, Total expenses as indicated by the model summary.
  + Model adequacy tests suggest that the assumptions of multiple linear regression are reasonably met.
  + The actual vs. fitted plots show that the model captures the relationship between sales and all the predictor variables well, both for the training and testing datasets.
  + The residuals vs. fitted plot roughly form a "horizontal band" around the 0 line. This suggests that the variances of the error terms are equal.

***REFERENCES***

* [www.kaggle.com](http://www.kaggle.com)
* [www.dataworld.com](http://www.dataworld.com)
* [www.datacamp.com](http://www.datacamp.com)
* [www.linkedin.com](http://www.linkedin.com)