**Contents**

**Title** 1

**Abstract** 1

**Introduction** 2

Primary Analysis Objective2

**Materials and Methods** 3

Data Source 3

Statistical Analysis 3

Model Assumptions 4

Primary Objective Analysis 4

Model Selection 10

Goodness of Fit 16

**Result 17**

**Title**

To determine the influence of factors affecting the market share of food product.

**Abstract**

For a food product manufacturing firm, It is important to account influencing factors which directly affects the success of product and market share associated to it. Analysis of sales data collected over a period, can play pivotal role in enhancing the product and this can also lead to increase in sales as well as profit.

The objective of this analysis is to determine the influencing factors (independent Variables) for a better market share of food product. For this study, Data collected by National Database (Nielsen) over period of 36 consecutive months (From Sep,1999 to Aug,2002) is used as data set.

The Average monthly market share of product (in percentage) is monitored against the input features from data set. Statistical analysis carried out on given data to better understand the influence of factors. Input features from the data set are average monthly price of product in dollars (Price), Gross Nielsen Rating points(gnrpoints), Discount Price(discount), Package promotion(promotion), Month and year.

**Introduction**

A food product can only be successful if the manufacturing firm knows exactly about the influencing factors associated to the market share. Better Understanding of demand and supply towards target audience could lead to product domination in the market. Also, A comprehensive analysis of product data could suggest the better marketing strategies for increasing the market share. For this study, Data collected by National Database (Nielsen) is used. This data has been collected for 36 consecutive months starting from Sept,1999 to end of Aug,2002. The factors taken in account for this study are Market Share, Identification no. for each month, average monthly price of product in dollars(Price), Gross Nielsen Rating points(gnrpoints) which refers to an index of the amount of advertisement exposure that the product received, Discount price(discount) which refers to presence or absence of discount price(1 if discount,0 otherwise), Package promotion(promotion) which refers to presence or absence of package promotion during period(1 if promotion,0 otherwise), Months and year (1999-2002).

In this study, each input feature from the data set has been analyzed against the market share with the help several statistical methods to determine its influence over market share. Finally, a best model was selected based on R^2(proportion of variation in the outcome that is explained by the predictor variables) and CP (A parameter that penalizes inclusion of addition variables to model) for combination of input features.

**Primary Analysis Objectives**

To investigate the linear association between market share (in percentage) and factors accounted in the study for influencing the market share of the food product.

**Materials and Methods**

**Data Source**

The is collected from 1999 to 2002 from a national database (Nielsen) for 36 consecutive months. The data consists of factors influencing the market share of a particular product produced by a large packaged food manufacturer. Each record in the data contains 6 variables i.e., Month (January to December), Year (1999 to 2002), Package Promotion (1 if package promotion is present, 0 otherwise), Discount Price (1 if discount is present, 0 otherwise), Gross Nielson rating points (An index of the amount of advertising exposure that the product received) and Price (Average monthly price of product in dollars) which collectively affect the market share of the product.

**Statistical Analysis**

We have got the data in excel (.xlsx) format which consisted of 36 records. We are going to use multiple linear regression to find the correlation between the market share and the rest of the factors and we have analyzed the data keeping this goal in mind. The data doesn’t have any null values in the preliminary analyses. We have performed linear regression for each of the predictor variable with the target to confirm the individual relationship between them. We have also changed the type of predictor variable Month from character to numerical value so we can fit the model easily.

**Primary Objective Analysis**

We are going to explore the individual predictors and the impact that they can have over the response variable before selecting the best model. It will help us in determining the importance of each factor and also detecting outliers or skewness in data so that we can transform the data accordingly. After exploring the dataset, we will check for linearity between the factors and the target variable individually.

**Model Selection**

To eliminate the redundant and less relevant variables for the response variable from the model, Automatic variable selection method is used. Package “leap” in R contains the function “regsubsets” which can be used for the same purpose. Based on Mallow’s Cp, BIC and R2, best model can be selected. The best model should have the lower values of Mallow’s Cp and BIC and the higher values of R2. In our case, the best model should have variables price, discount, promotion, month and year. The values of Mallow’s Cp, BIC and R2 are shown in Figure (EDIT).

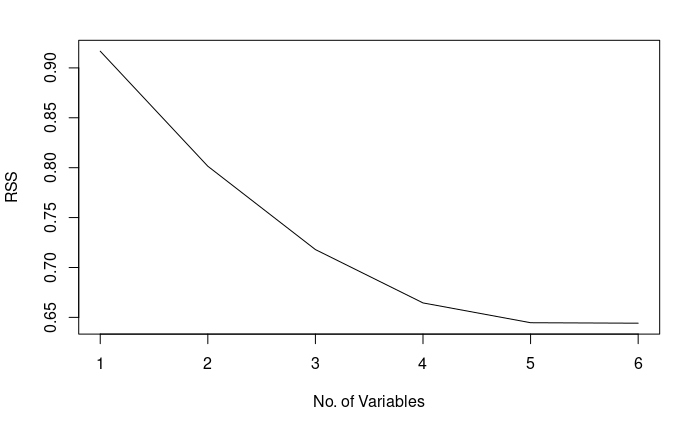
 

Figure 1 R-square values of variables Figure 2 BIC values of variables

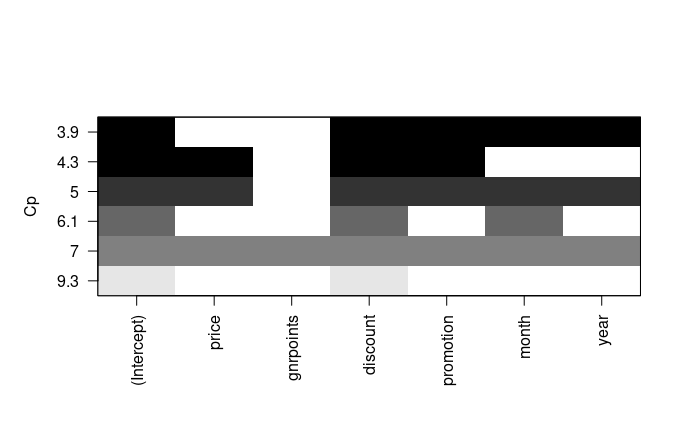


Figure 3 Cp values of variables.

Thus, the final model can be expressed as:

*Yi = β0 + β1X1 + β2X2 + β3X3 + β4X4 + β5X5 +€i*

where,

* *Y*i is the Market Share output
* *X1* is the Price
* *X2* is the Discount
* *X3* is the Promotion
* *X4* is the Month
* *X5* is the Year
* €i is the error term; €i ∼ *iidN*(0, σ2 )
* *i* = 1,2,3, · · ·, 36.
* β0, β1, β3, β4, β5 and σ 2 are the unknown parameters to be estimated.

**Multicollinearity between Predictor Variables:**

We can see the correlation coefficients r between the variables present in the dataset in Figure (EDIT) Our target variable Market Share seems to be more related to price, discount, promotion, and month and less related to gnrpoints and year. The collinearity between the variables can be observed based on coefficient of Correlation r. The closer the value of r to 1 or -1, the higher is the collinearity between the variables. Highest collinearity is shown between Market Share and Discount. Highest collinearity between the predictors is shown between Gnr points and Year. The VIF (Variation Inflation Factor) value is also considered to check the multicollinearity between the variables which are shown in the Table (EDIT). If VIF values are more than 10, then it is assumed that there is multicollinearity. In our dataset, no variables have a VIF value higher than 10, thus we can assume the absence of multicollinearity in the final model.

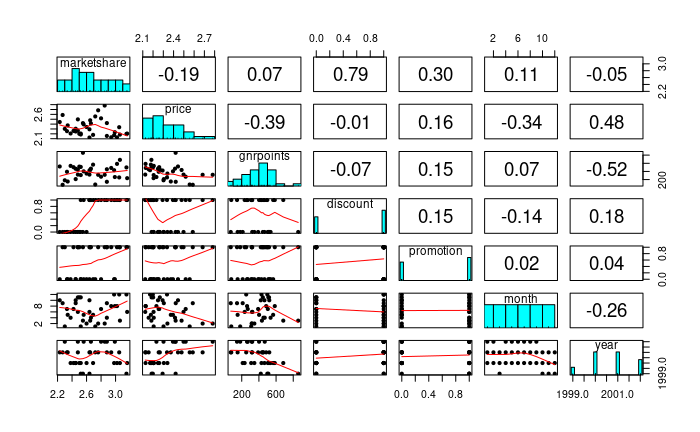


Figure 4 Pairwise Pearson’s correlations coefficients for all variables present in dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | Price | Discount | Promotion | Month | Year |
| **VIF value** | 1.579528 | 1.102519 | 1.136714 | 1.202841 | 1.663702 |

Table 1 VIF values of variables present in the dataset for multicollinearity

**Outliers and Influential points:**

Standardized residual plot can be considered to detect outliers in the data. In the residual vs leverage plot shown in Figure (EDIT), no data points are lying outside the Cook’s distance line (The red dashed line). This means that the data we have doesn’t contain any outliers in it. Hence it is found that the final model is robust with including all the records from the data.

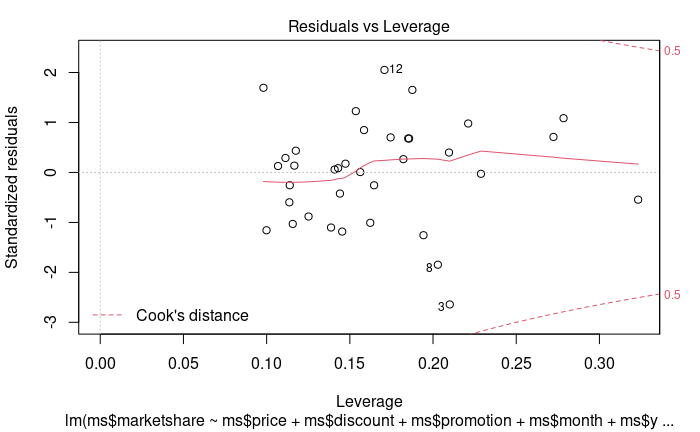


Figure 5 Residuals vs Leverage Plot

**Residual diagnostics**

Figure (EDIT) shows the Residual diagnostic through which the model assumptions can be checked. In (a), equal variance can be observed which satisfies the homoskedasticity. Normality of error terms can be seen in (b) and (c). Independence of error term can be observed from (d).

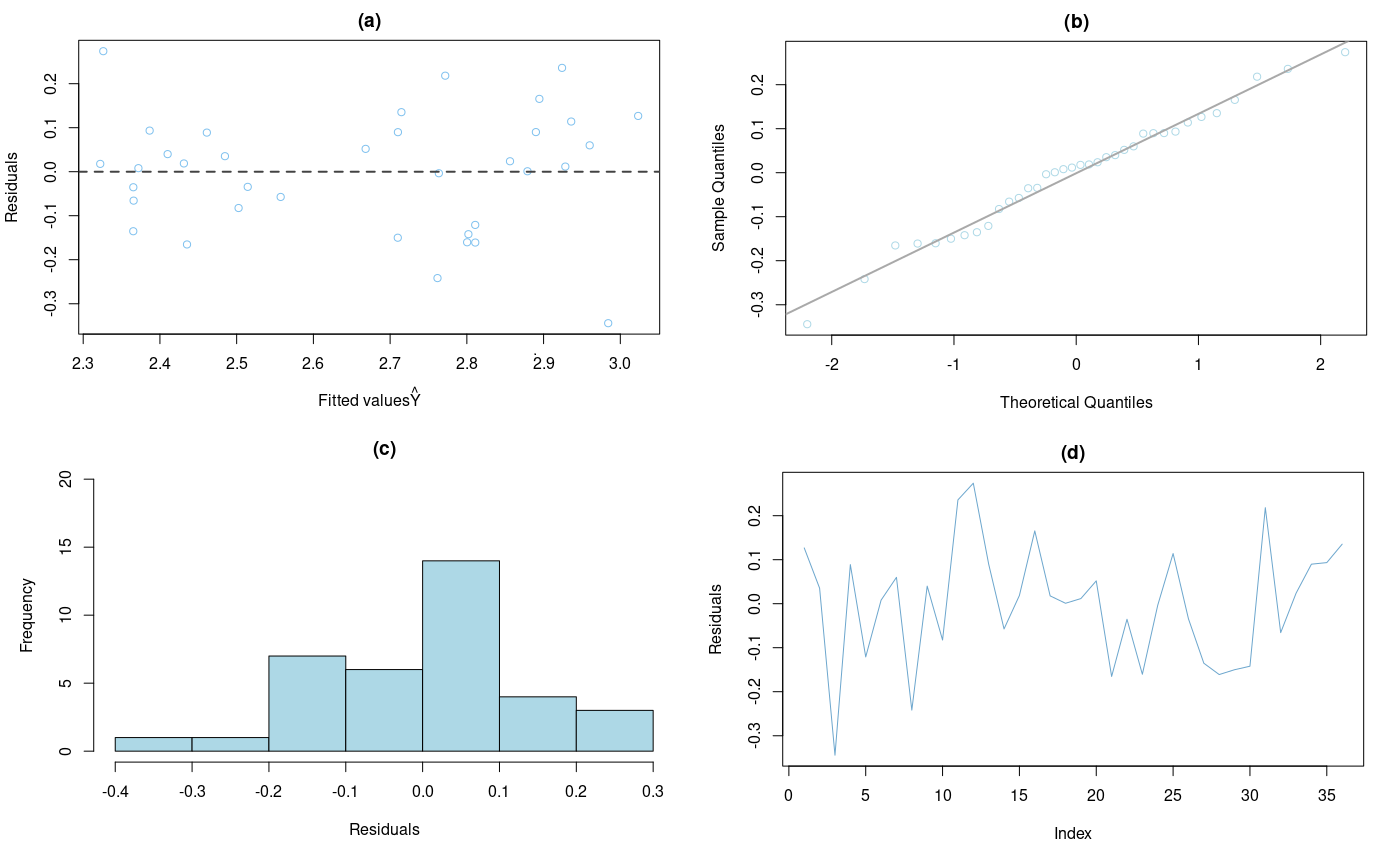


Figure 6Residuals Diagnostics

**Goodness of Fit**

Edit

**Results**

The individual results for the relation between the target variables and response variable are shown in the earlier section. The coefficient of correlation was also calculated and Discount variable was found to be most correlated variable to the Market Share followed by promotion, price, month, gnr points and year respectively. On the basis of BIC, R2 and Cp values, the best model was found. To validate the model, 5 random samples are taken from the database and predicted values are calculated with the help of the fitted Multiple Linear Regression model. All the predicted values lie in the 95% range of the model prediction and the actual values lie in the predicted range of the model as shown in Table (EDIT).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Observation | Price | Discount | Promotion | Month | Year | Actual  Market Share | Predicted  Market  Share | Predicted Range |
| 1 | 2.198 | 1 | 1 | Sep | 1999 | 3.15 | 3.02 | 2.88 to 3.16 |
| 2 | 2.186 | 0 | 0 | Oct | 1999 | 3.52 | 2.48 | 2.35 to 2.61 |
| 3 | 2.373 | 0 | 0 | Jan | 2000 | 2.34 | 2.35 | 2.24 to 2.45 |
| 4 | 2.373 | 0 | 0 | Jan | 2001 | 2.34 | 2.32 | 2.22 to 2.42 |
| 5 | 2.302 | 0 | 1 | Apr | 2002 | 2.3 | 2.36 | 2.19 to 2.53 |

Table 2 Actual vs Predicted values of Market share using the estimated Multiple Linear regression model.