Market Share Statistical Analysis

Multiple Linear Regression Analysis Using R Programming Language

Fathieh Qakei

2023-12-09

Contents

I. Abstract

This statistical analysis uses R Programming Language to understand and model the market share of a specific product over a 36-month period using a dataset named "Market Share" collected from a national database (Nielsen). The response variable, **marketshare** , is the key variable of interest, and the study aims to identify and analyze factors that influence variations in market share. The predictor variables include average monthly price, Gross Nielsen rating points (gnrpoints), discount presence, promotion presence, month, and year.

The hypotheses tested are based on linear relationships between independent variables and the response variable. The null hypothesis H_0 posits that none of the independent variables have a significant linear relationship with **marketshare**, while the alternative hypothesis H_a asserts that at least one coefficient is not equal to zero, indicating a significant linear relationship.

After conducting automatic variable selection methods, criterion-based statistics, residuals diagnostics, and model validation, two variables, **gnrpoints** and **year**, were dropped as they did not exhibit a significant linear association with **marketshare**. The final multiple linear regression model includes **price**, **discount**, **promotion**, and **month** as predictors.

To assess the proportion of variation explained by each predictor variable, the coefficient of multiple determination R^2 is calculated. The ANOVA table indicates the percentage of variation explained by each variable: price (3.614%), discount (65.066%), promotion (6.47%), and month (5.586%).

II. Introduction

A. Study Design.

Market share refers to the proportion or percentage of total sales in the market that a specific product represents during a given period. It is a measure of how well the product is performing relative to other products in the same market. Market share is a key indicator for companies to understand their competitive position and track the success of their products (Hayes, 2022).

In a dataset called Market Share that was collected from a national database (Nielsen) for 36 consecutive months from September, 1999, through August, 2002, **marketshare** is the response variable, meaning it is the variable of interest that the company executives want to understand and model. The executives want to identify and analyze the factors (independent variables) that influence or contribute to variations in market share over the 36 consecutive months. This analysis could help them make informed decisions on marketing strategies, product development, or other aspects of their business to improve market share.

Each line of the dataset has an identification number and provides information on six other variables for each month. As shown in the table below, these other six variables are: **price**: the average monthly price of product in U.S. dollars, **gnrpoints** : Gross Nielson rating points which is an index of the amount of advertising exposure that the product recieved, **discount**: Presence or absence of discount price during period: 1 if discount, 0 otherwise, **promotion** : Presence or absence of package promotion during period: 1 if promotion present, 0 otherwise, **month**: Month (Jan-Dec), and **year**: Year (1999 - 2002). All of the predictor variables are numeric except for the **month** variable which is a character.

Table 1 : Description of Variables

Variable	Variable Name	Description
1. idnum	Identification number	$1 - 36$
2. marketshare	Market share	Average monthly market share for product (percent)
3. price	Price	Average monthly price of product (dollars)
4. gnrpoints	Gross Nielson rating points	An index of the amount of advertising exposure that the product received

Variable	Variable Name	Description
5. discount	Discount price	Presence or absence of discount price during period: 1 if discount, 0 otherwise
6. promotion	Package promotion	Presence or absence of package promotion during period: 1 if promotion present, 0 otherwise
7. month	Month	Month (Jan-Dec)
8. year	Year	Year (1999 - 2002)

B. Aims/Title.

The ultimate purpose of the study is to determine which factors have the most influence on the market share of a certain product.

III. Methods

Preliminary Model.

A multiple linear regression model is represented by the following equation:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \varepsilon_i$$

, Where:

Y_i is the value of the response variable in the i-th observation.

β_0 is the fixed and unknown Y-intercept parameter that represents the expected mean value of the response variable when all predictor variables are zero.

$\beta_1, \beta_2, \dots, \beta_6$ are the coefficients or the fixed and unknown slope parameters for $X_{i1}, X_{i2}, \dots, X_{i6}$ respectively, indicating the change in the response variable for a one-unit change in each predictor variable, holding other predictors constant.

ε_i is a random error term associated with the ith observation and indicates the degree to which Y_i remains unexplained after accounting for the model. $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$, which is read as: the random error $\varepsilon_i$is independently and identically distributed as a normal distribution with mean 0 and variance σ_ε^2. The variance σ_ε^2 specifies how much individual values of the random variable ε_i deviate from the mean (0, in this case).

X_{i1} = Average monthly price of product for the i^{th} observation, X_{i2} = Discount price for the i^{th} observation, X_{i3} = Package promotion for the i^{th} observation, X_{i4} = Month for the i^{th} observation, X_{i5} = Gross Nielson rating points for the i^{th} observation, X_{i6} = year for the i^{th} observation.

Hypotheses

The following hypotheses are to be tested by the end of the analysis:

The null hypothesis (H_0) is defined as follows:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_6 = 0$$

The null hypothesis states that none of the independent variables have significant linear relationship with the response variable.

The corresponding alternative hypothesis (H_a) is:

$$H_a : \text{At least one } \beta_i \text{ is not equal to 0, where } i = 1, 2, \dots, 6$$

The alternative hypothesis asserts that at least one of the coefficients for the independent variables is not equal to zero, which indicates a significant linear relationship with the response variable.

Explanation of OLS Estimators

In linear regression, the Ordinary Least Squares (OLS) method aims to find the best-fitting line through a set of data points. The OLS estimators are the values assigned to the coefficients (β_i) in the regression model, and they are chosen to minimize the least-squares criterion.

The least-squares criterion is defined as the sum of the squared residuals, where the residual for each observation i is the difference between the observed value (Y_i) and the predicted value ($\hat{Y}_i$):

$$\text{Residual}(e_i) = Y_i - \hat{Y}_i$$

The least-squares criterion is formulated as:

$$\text{SSE} = \sum_{i=1}^{n} (e_i)^2$$

The goal of OLS is to find the values of the coefficients that minimize this sum across all observations. The resulting OLS estimators represent the coefficients that make the model the best fit to the observed data, striking a balance between accurately representing the data and avoiding overfitting. The OLS estimated model can be expressed as follows:

$\hat{Y}_i = b_0 + b_1 \cdot X_{i1} + b_2 \cdot X_{i2} + b_3 \cdot X_{i3} + b_4 \cdot X_{i4} + b_5 \cdot X_{i5} + b_6 \cdot X_{i6}$

where:

$\hat{Y}_i$ is the predicted or estimated value for the ith observation, $b_0, b_1, b_2, \ldots, b_6$ are the OLS coefficient estimators, $X_{i1}, X_{i2}, \ldots, X_{i6}$ are same values of the independent variables for the ith observation defined previously.

OLS Assumptions The effectiveness of OLS relies on several key assumptions as mentioned by Valchanov in his article *Exploring the 5 OLS Assumptions for Linear Regression Analysis* (2021):

The relationship between the independent variables and the dependent variable is assumed to be linear. Additionally, each independent variable is uncorrelated with the error term. This ensures that each predictor contributes independently to the model's predictions.The error term has a population mean of zero, and observations of the error term are uncorrelated with each other. No pattern exists in the residuals that could be exploited for predicting other residuals. Also, the error term has a constant variance across all levels of the independent variables. This assumption ensures that the spread of residuals remains consistent. No independent variable is perfectly collinear with other variables. This prevents issues with estimating unique coefficients. Finally, the error term is assumed to be normally distributed for valid statistical inference.

Analysis of Dataset Variables

A quick look at the variables of the dataset is provided below:

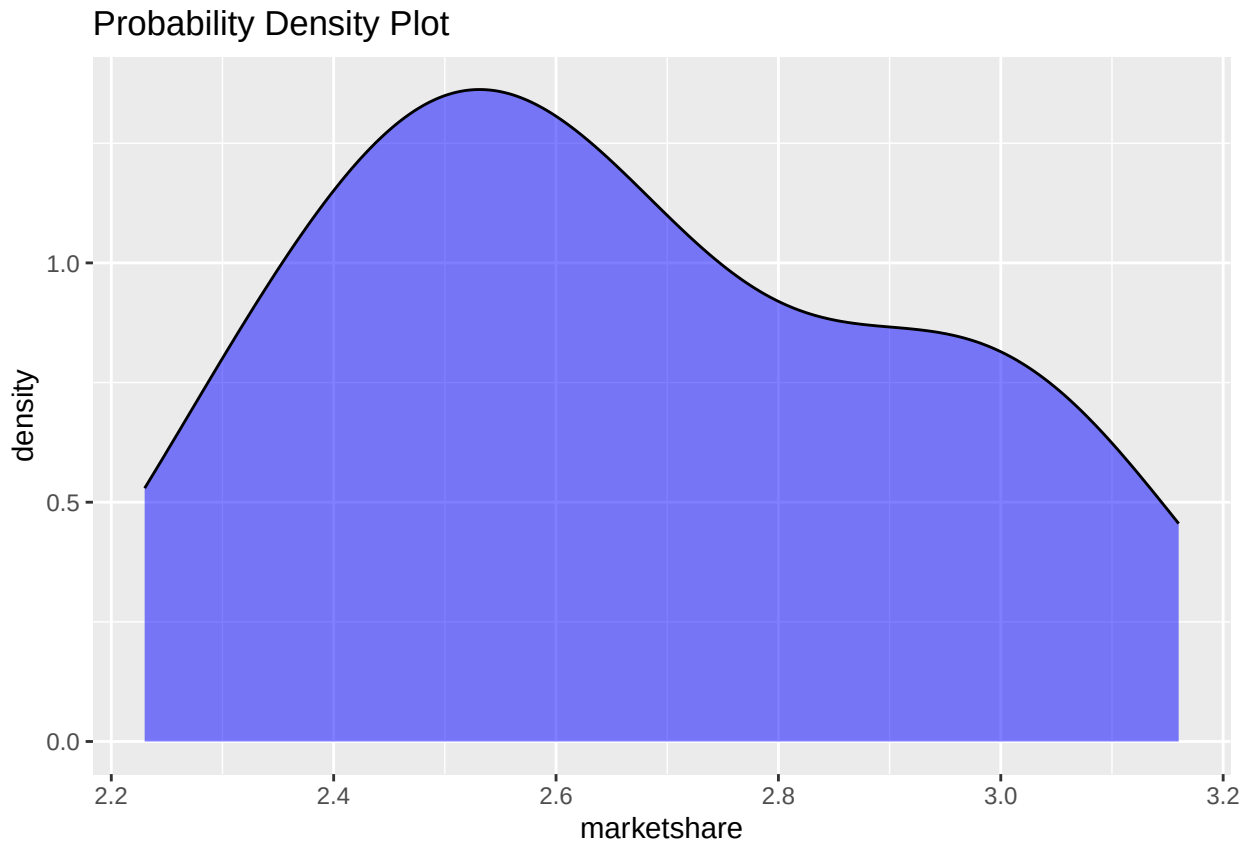
```
## # A tibble: 6 x 8
##   idnum marketshare price gnrpoints discount promotion month  year
##   <dbl>       <dbl> <dbl>     <dbl>    <dbl>     <dbl> <chr> <dbl>
## 1     1        3.15  2.20       498        1         1 Sep    1999
## 2     2        2.52  2.19       510        0         0 Oct    1999
## 3     3        2.64  2.29       422        1         1 Nov    1999
## 4     4        2.55  2.42       858        0         1 Dec    1999
## 5     5        2.69  2.18       566        1         0 Jan    2000
## 6     6        2.38  2.21       536        0         0 Feb    2000
```

The **Summary Table** below shows name, type, and rang for each variable in the dataset. It is shown that all variables are numeric except for the **month** variable which is character. It has to be converted into ordered factors then numeric values. No missing values are found in the dataset.

Table 2: Basic Statistics for Dataset Variables

Variable	Type	Range
idnum	Numeric	Min.: 1.00, 1st Qu.: 9.75, Median: 18.50, Mean: 18.50, 3rd Qu.: 27.25, Max.: 36.00
marketshare	Numeric	Min.: 2.230, 1st Qu.: 2.473, Median: 2.640, Mean: 2.664, 3rd Qu.: 2.880, Max.: 3.160
price	Numeric	Min.: 2.124, 1st Qu.: 2.200, Median: 2.280, Mean: 2.324, 3rd Qu.: 2.420, Max.: 2.781
gnrpoints	Numeric	Min.: 72.0, 1st Qu.: 268.0, Median: 412.0, Mean: 388.1, 3rd Qu.: 499.5, Max.: 858.0
discount	Numeric	Min.: 0.0000, 1st Qu.: 0.0000, Median: 1.0000, Mean: 0.5833, 3rd Qu.: 1.0000, Max.: 1.0000
promotion	Numeric	Min.: 0.0000, 1st Qu.: 0.0000, Median: 1.0000, Mean: 0.5556, 3rd Qu.: 1.0000, Max.: 1.0000
month	Character	Length: 36, Class: character, Mode: character
year	Numeric	Min.: 1999, 1st Qu.: 2000, Median: 2001, Mean: 2001, 3rd Qu.: 2001, Max.: 2002

A probability density plot of the response variable is shown below. No skewness is indicated.



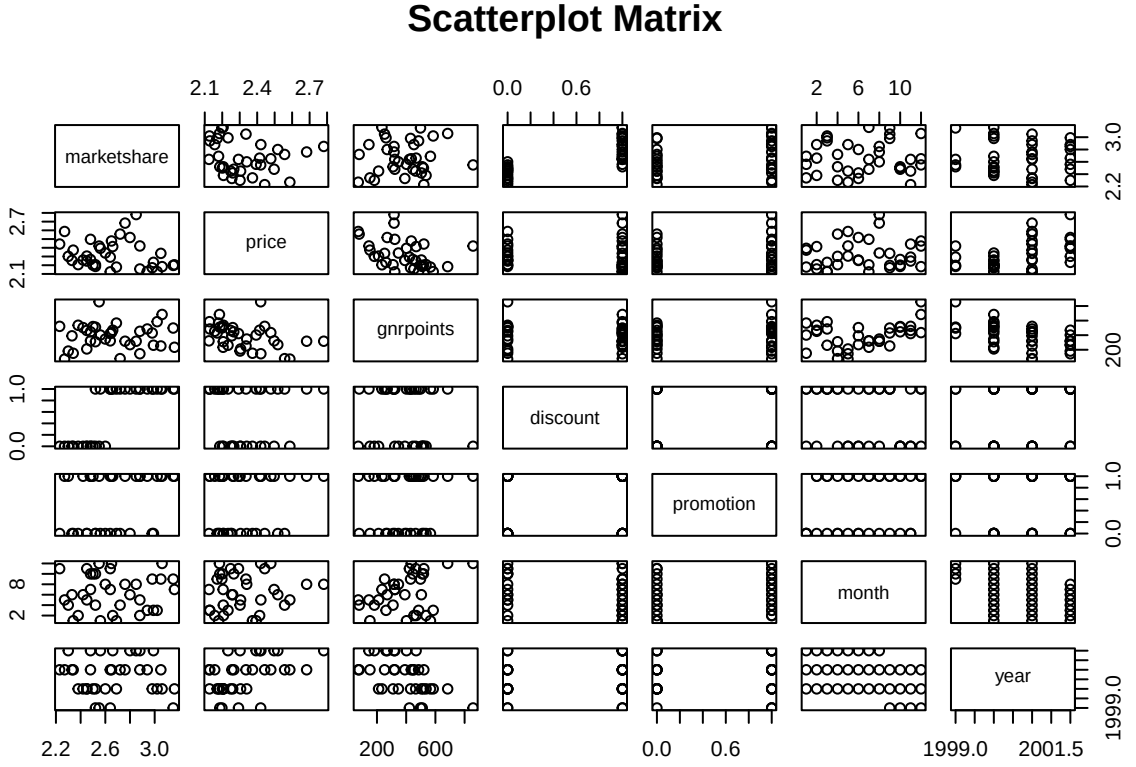
The summary table below shows the preliminary model containing all 6 predictor variables with their coefficients' estimates, standard errorr, t values, and probability values.

Table 3: Coefficients Summary for Initial Full Model

Variable	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	45.5090	74.6339	0.610	0.5468
price	-0.3553	0.2007	-1.770	0.0872
discount	0.4186	0.0531	7.887	1.07×10^{-8}
promotion	0.1042	0.0546	1.911	0.0660
month	0.0108	0.0091	1.188	0.2446
gnrpoints	-0.0001	0.0002	-0.613	0.5447
year	-0.0212	0.0374	-0.566	0.5757

Diagnostics for Predictors

The following scatterplot matrix indicates **gnrpoints** variable is positively associated with **month** and **promotion**, and negatively associated with **price** and **year**; **year** is positively associated with **price** and **discount**, and negatively associated with **gnrpoints** and **month**; **marketshare** response variable is highly positively associated with **discount** and **promotion**. **gnrpoints** and **year** variables show multicollinearity with other variable.

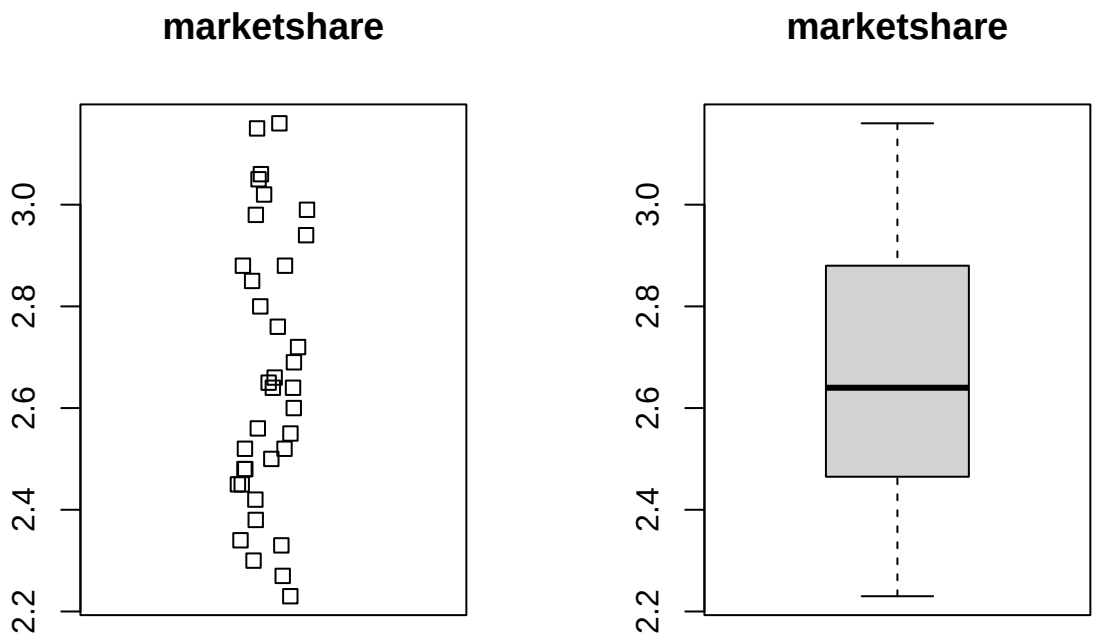


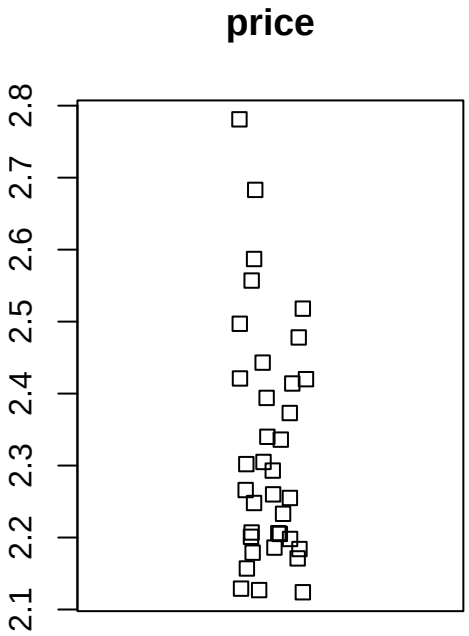
2. Correlation Matrix The Pearson correlation coefficients for all pairwise association are shown in the correlation dataframe below. It shows that **discount** variable highly associated with the response variable **marketshare** with a Pearson's correlation coefficient r= .791, **promotion** is moderately associated with the response variable r=.305, Gross Nielson rating points **gnrpoints** variable is moderately associated with **month** and **year** variables, and **year** is moderately associated with the **price** variable.

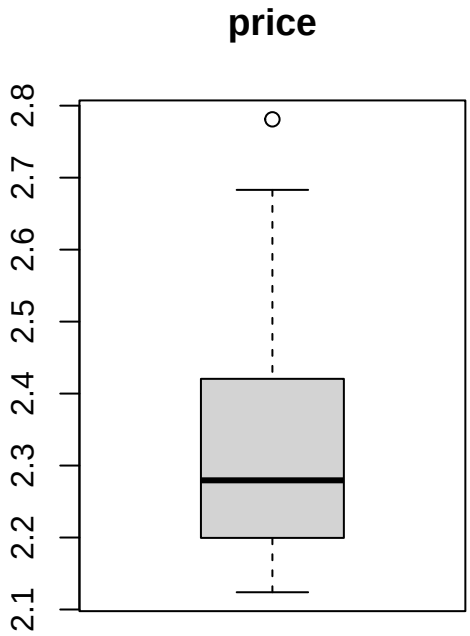
```
##             marketshare        price   gnrpoints     discount  promotion
## marketshare  1.00000000 -0.188506951  0.07256395  0.790730217 0.30499506
## price       -0.18850695  1.000000000 -0.38777648 -0.008473014 0.15696106
## gnrpoints    0.07256395 -0.387776478  1.00000000 -0.072632893 0.14666874
```

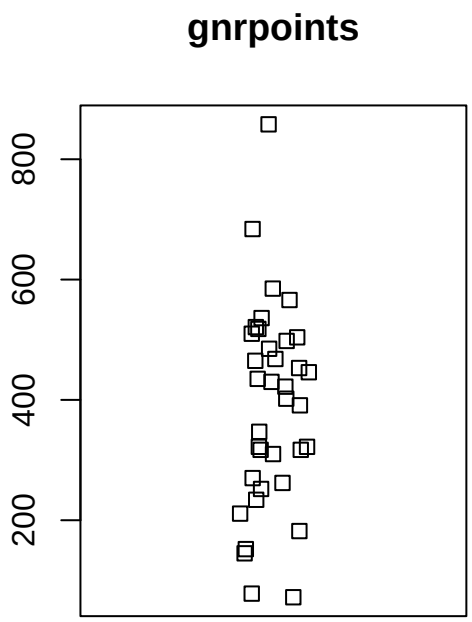
```
## discount     0.79073022 -0.008473014 -0.07263289  1.000000000 0.15118579
## promotion    0.30499506  0.156961061  0.14666874  0.151185789 1.00000000
## month        0.03982082  0.099243070  0.34877846 -0.171378614 0.24290635
## year        -0.05236820  0.481381933 -0.52162061  0.179284291 0.03952847
##                   month        year
## marketshare  0.03982082 -0.05236820
## price        0.09924307  0.48138193
## gnrpoints    0.34877846 -0.52162061
## discount    -0.17137861  0.17928429
## promotion    0.24290635  0.03952847
## month        1.00000000 -0.40967325
## year        -0.40967325  1.00000000
```

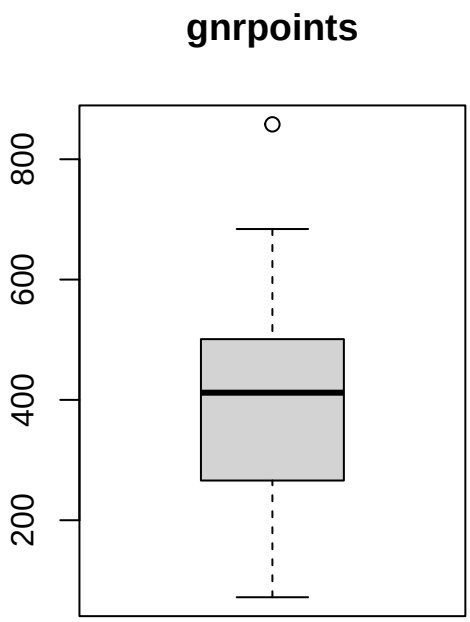
3. Strip plots/Boxplots Strip plots for predictors and the dependent variable are shown next to boxplots of the same data:

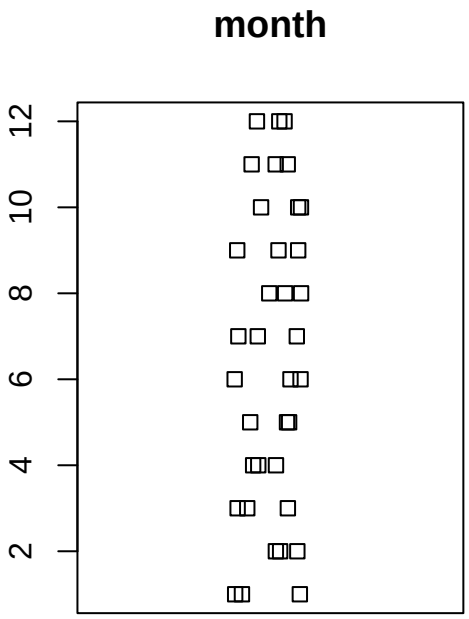


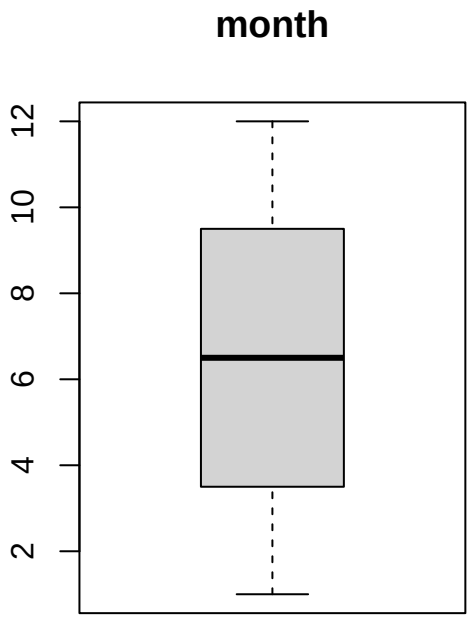

price
2.1
2.2
2.3
2.4
2.5
2.6
2.7
2.8

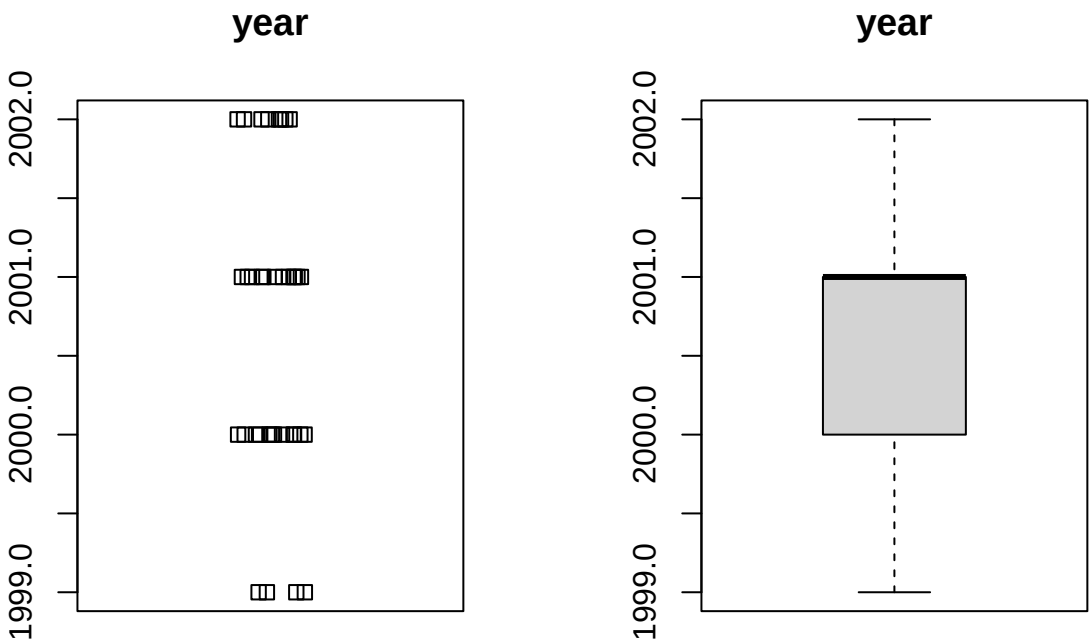

price
2.1
2.2
2.3
2.4
2.5
2.6
2.7
2.8


gnrpoints
800
600
400
200


gnrpoints
800
600
400
200

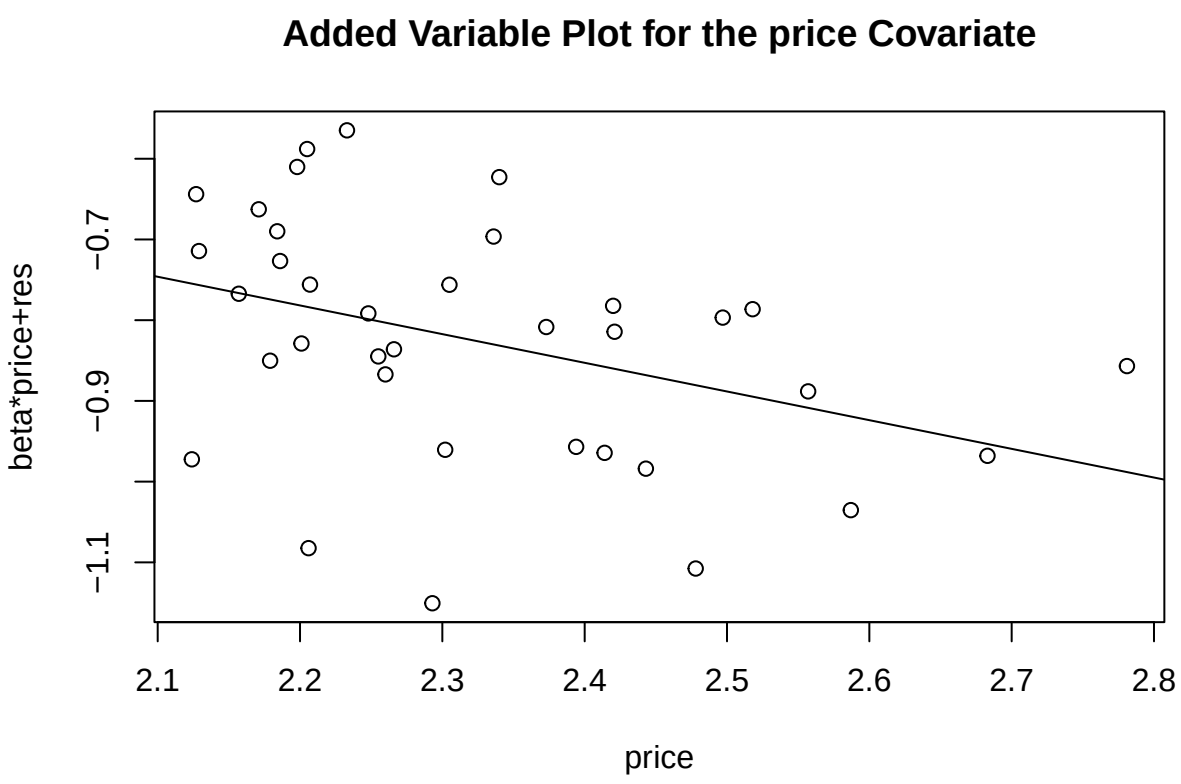

month
2
4
6
8
10
12

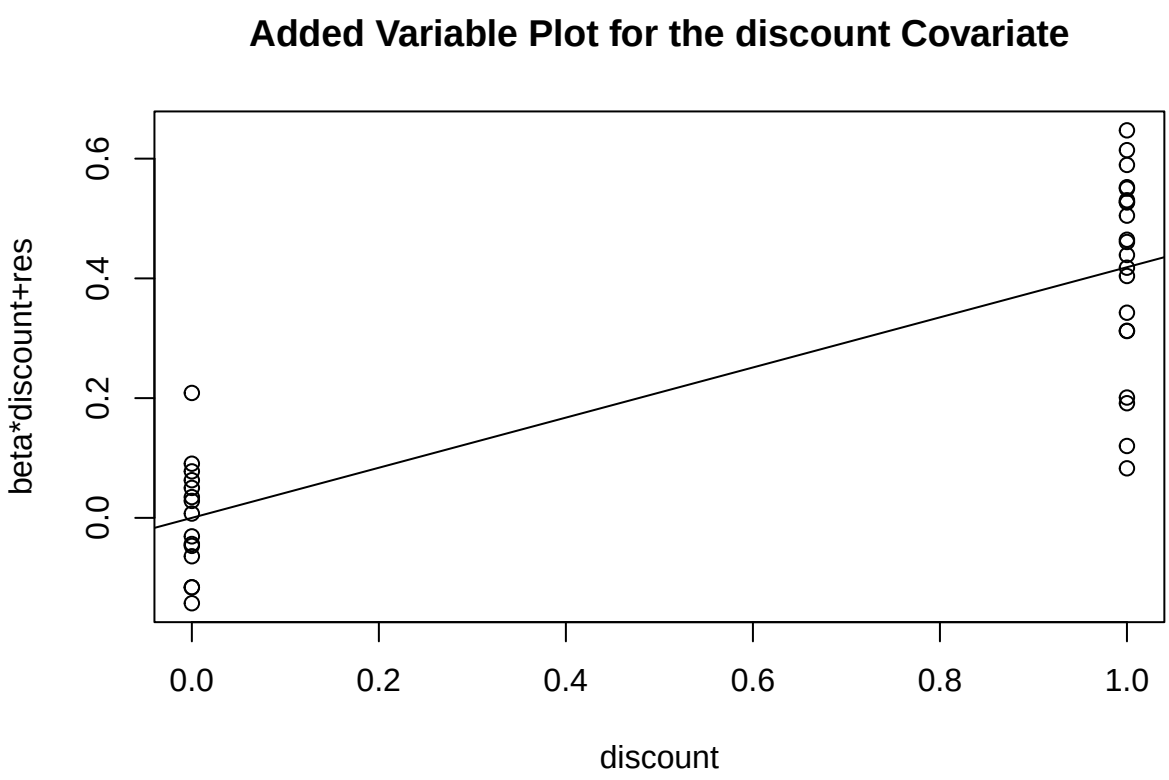

month
2
4
6
8
10
12

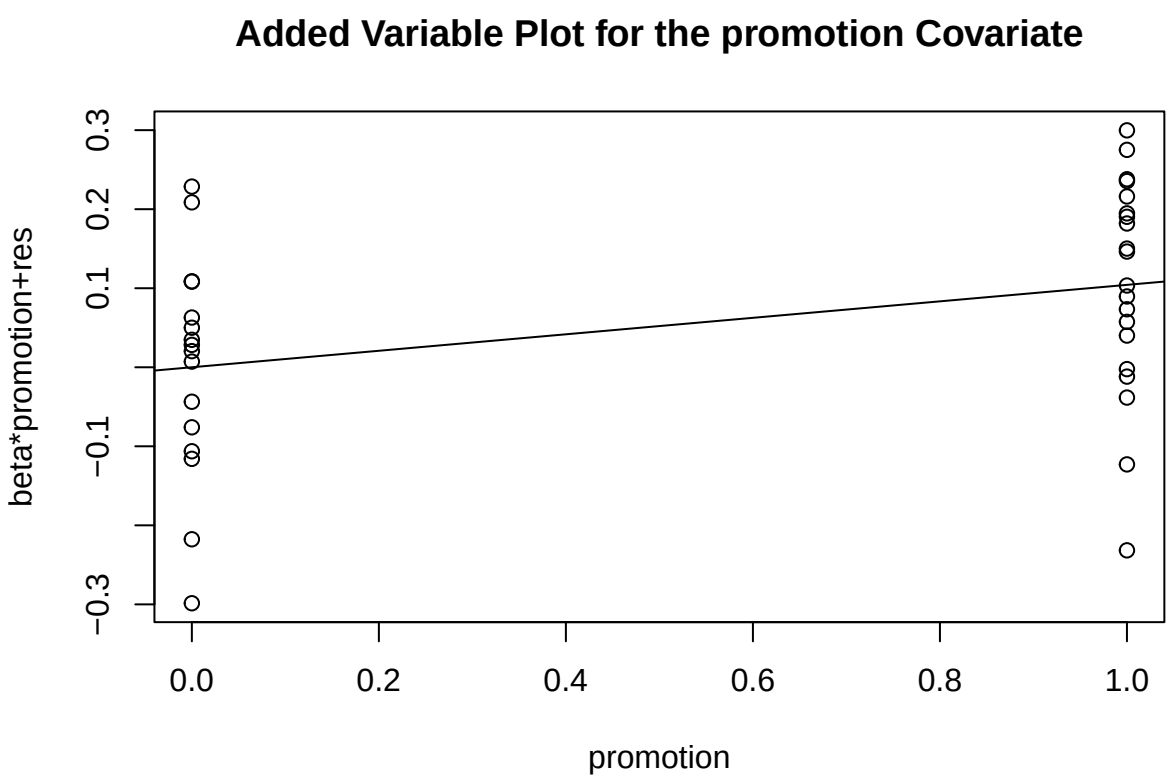


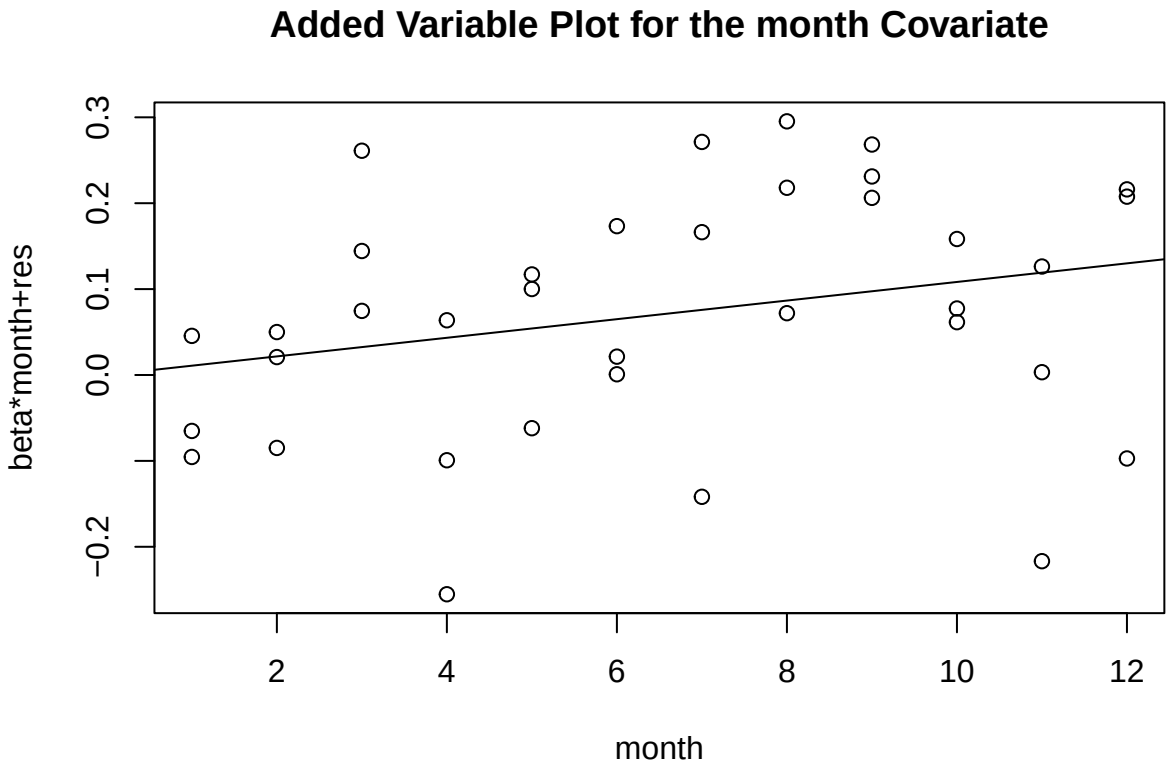
Screening of Predictors

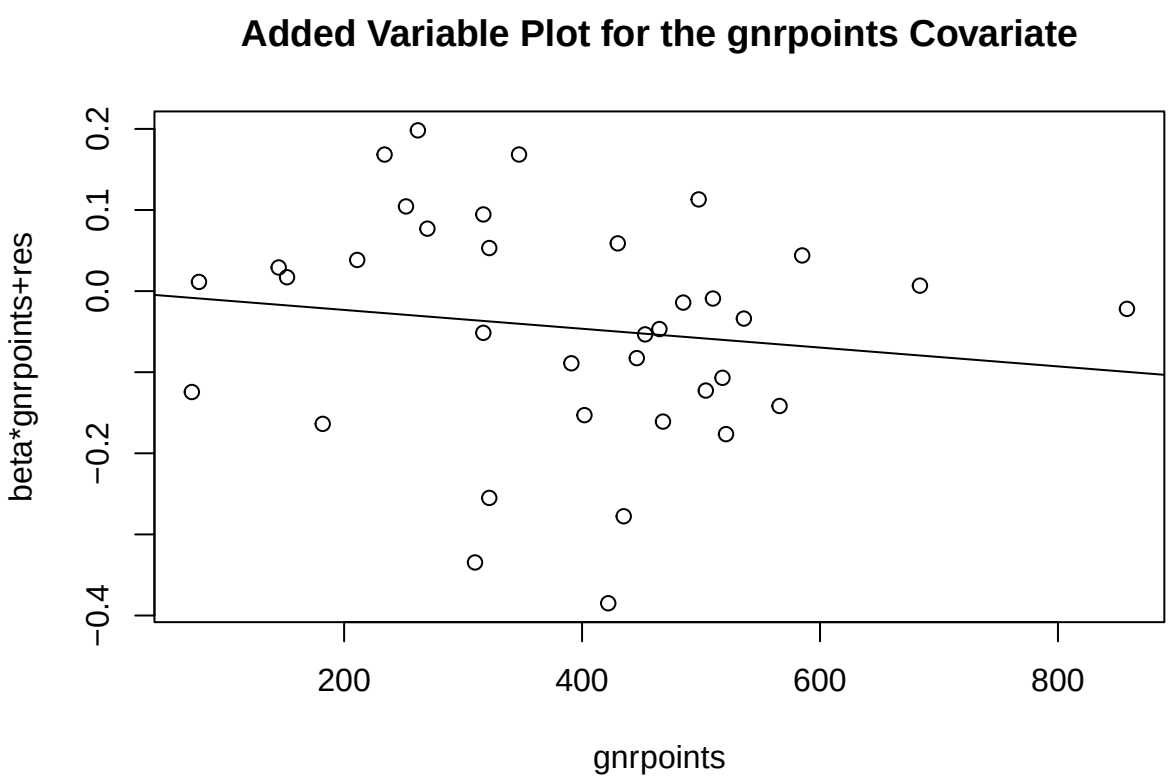
1. Added variable plots Added variable plots for each of the covariates are shown below. Added variable plots (also known as partial residual plots or adjusted variable plots) provide evidence of the importance of a covariate given the other covariates already in the model. They also display the nature of the relationship between the covariate and the outcome (i.e., linear, curvilinear, transformation necessary, etc.) and any problematic data points with respect to the predictor. The plots below show that **year** and **gnrpoints** covariates seem to not provide any added value to the model that already includes all other covariates; the slopes of their linear relationship with the **marketshare** outcome is close to zero. All other variables seem to provide some added value to the model that already includes all other covariates in it. The slopes of their linear relationship with the **marketshare** outcome is either positive or negative.

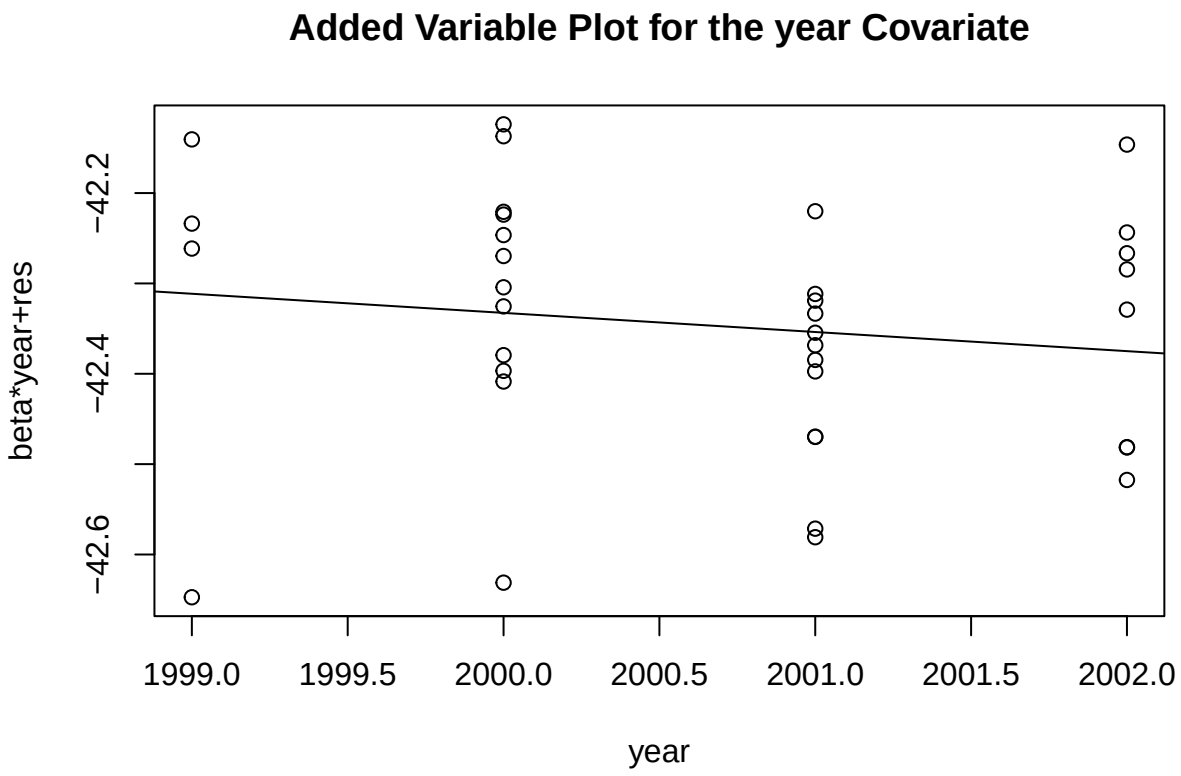

Added Variable Plot for the price Covariate
beta*price+res
−0.7
−0.9
−1.1
2.1
2.2
2.3
2.4
2.5
2.6
2.7
2.8
price


Added Variable Plot for the discount Covariate
beta*discount+res
0.0
0.2
0.4
0.6
discount
0.0
0.2
0.4
0.6
0.8
1.0


Added Variable Plot for the promotion Covariate
beta*promotion+res
promotion
0.0
0.2
0.4
0.6
0.8
1.0
−0.3
−0.1
0.1
0.2
0.3


Added Variable Plot for the month Covariate
beta*month+res
0.3
0.2
0.1
0.0
−0.2
2
4
6
8
10
12
month


Added Variable Plot for the gnrpoints Covariate
beta*gnrpoints+res
0.2
0.1
0.0
−0.2
−0.4
200
400
600
800
gnrpoints



2. Multicollinearity Multicollinearity can create instability in estimation and so it should be avoided. **Variance inflation factors (VIF)** measure how much the variances of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related. A maximum VIF in excess of 10 is a good rule of thumb for multicollinearity problems (Kutner and Nachtsheim, 2014). Based on the maximum VIF, 1.98, there do not appear to be any issues that need remediation.

```
##     price  discount promotion     month gnrpoints      year
##  1.658791  1.091792  1.171636  1.579570  1.578312  1.981352
```

Automatic variable selection methods

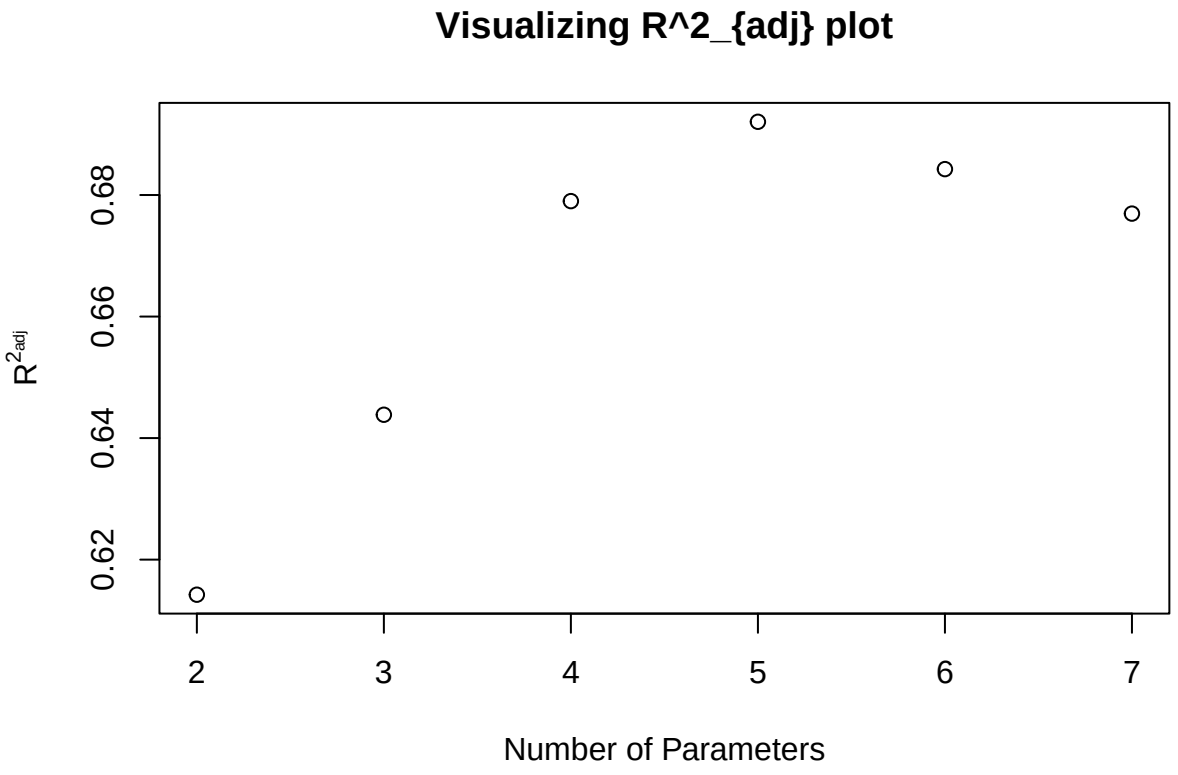
Automatic variable selection methods can be a useful starting point in eliminating redundant variables. It is used as a guide to the screening and removal (or addition) of predictors. The best fitting model will be chosen based on different statistical criteria followed by model validation. These criteria are R^2_{adj} (larger values are better), Bayes Information Criterion BIC (smaller values are better), and Mallow's C_p statistic (values of C_p close to p (number of beta coefficients) are better). An explanation of each criterion and its results and visualization plots will be discussed separately. Here, sequential replacement method is used and the following shows the possible six subset regression models with their predictors:

```
## Subset selection object
## Call: regsubsets.formula(marketshare ~ price + discount + promotion +
##     month + gnrpoints + year, data = market_share, method = "seqrep")
## 6 Variables  (and intercept)
##           Forced in Forced out
## price         FALSE      FALSE
## discount      FALSE      FALSE
## promotion     FALSE      FALSE
```

```
## month         FALSE      FALSE
## gnrpoints     FALSE      FALSE
## year          FALSE      FALSE
## 1 subsets of each size up to 6
## Selection Algorithm: 'sequential replacement'
##          price discount promotion month gnrpoints year
## 1  ( 1 ) " "   "*"      " "       " "   " "       " "
## 2  ( 1 ) " "   "*"      " "       " "   " "       "*"
## 3  ( 1 ) "*"   "*"      "*"       " "   " "       " "
## 4  ( 1 ) "*"   "*"      "*"       "*"   " "       " "
## 5  ( 1 ) "*"   "*"      "*"       "*"   "*"       " "
## 6  ( 1 ) "*"   "*"      "*"       "*"   "*"       "*"
```

1. The adjusted coefficient of multiple determination R^2_{adj}. The coefficient of multiple determination R^2 represents the proportion of the total variance in the dependent variable explained by the independent variables, R^2_{adj} adjusts this value to penalize models with a larger number of predictors, thus providing a more accurate measure of the model's goodness of fit. Larger value represents better predictive model. Here it is the model that has 4 predictor variables which are: **price**, **discount**, **promotion**, and **month**.

```
## [1] 0.6142323 0.6438419 0.6789944 0.6920464 0.6842594 0.6769416
```



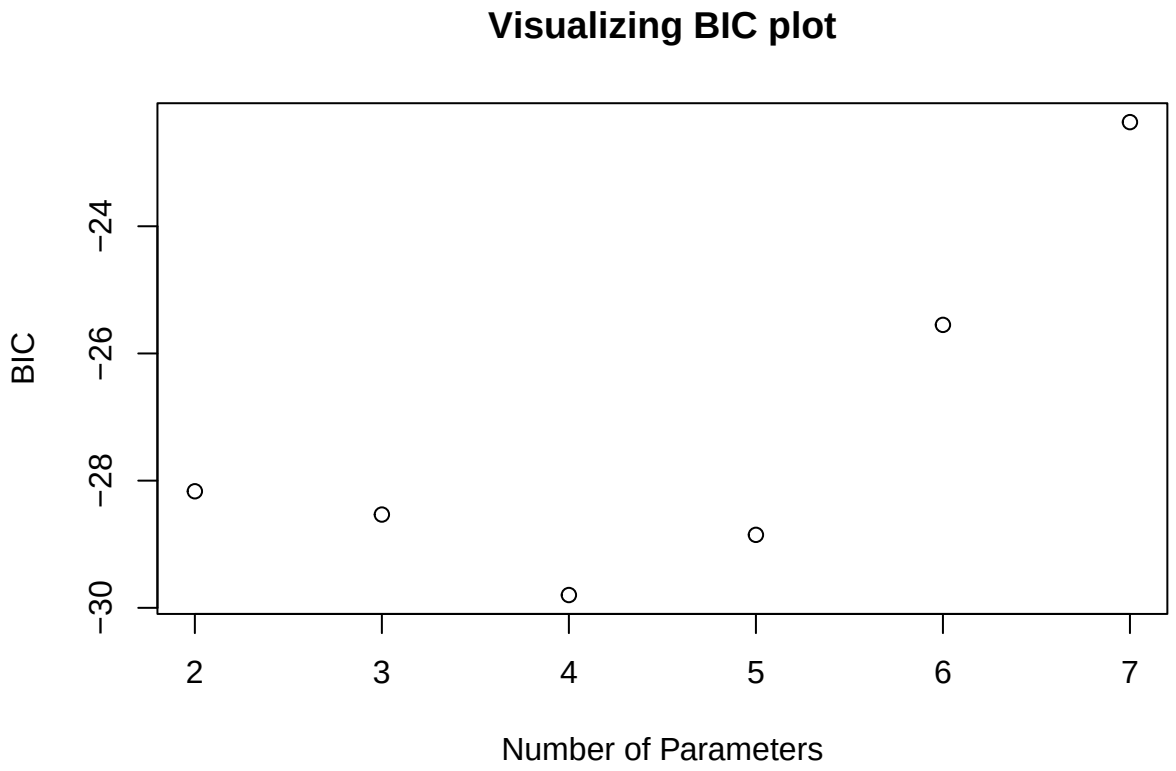
2. The Bayesian Information Criterion (BIC) The Bayesian Information Criterion (BIC) is another model selection method that penalizes the inclusion of additional predictors in the model. It is given by the formula:

$$BIC = n \cdot \ln(SSE/n) + k \cdot \ln(n)$$

.

Smaller BIC value represents a better predictive model. Here it is the model that has 3 predictor variables which are: **price**, **discount**, and **promotion**.

```
## [1] -28.16723 -28.53340 -29.79864 -28.85241 -25.55035 -22.36245
```

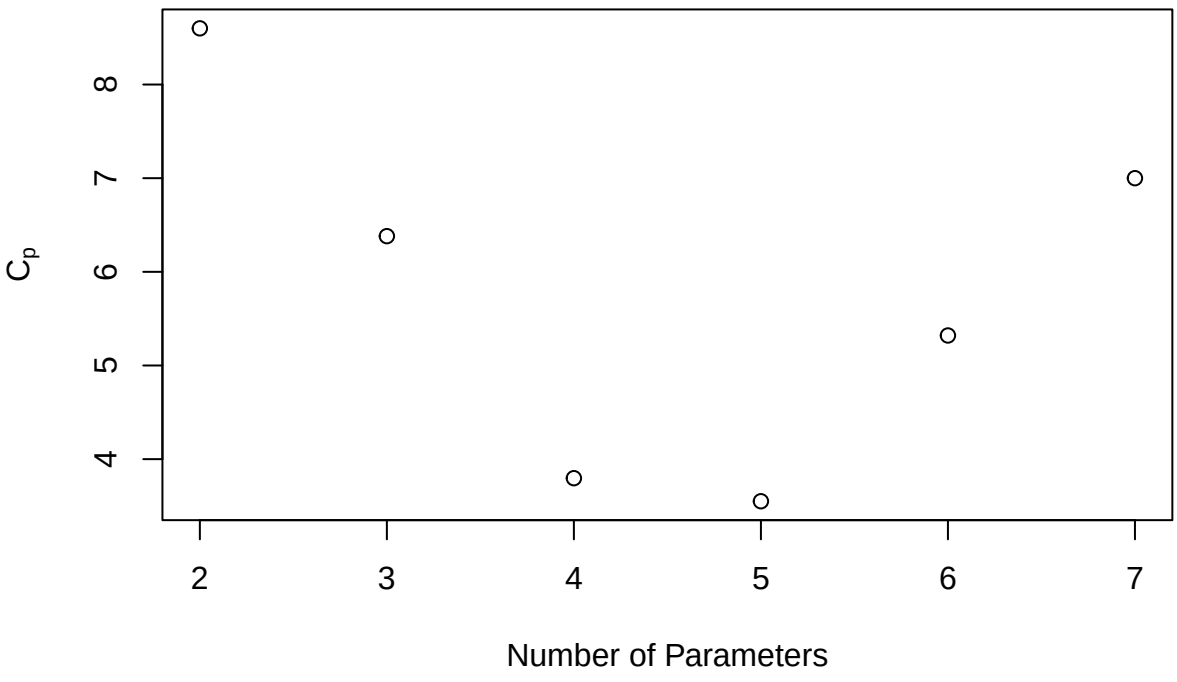


3. Mallows' C_p criterion Mallows' C_p criterion measure the total mean squared error of the fitted values for each subset regression model. It helps balance the model's goodness of fit and model simplicity . It is given by the formula:

$$C_p = \frac{SSE_p}{MSE_{\text{full}}} - (n - 2p)$$

where SSE_p is the sum of squared errors for the model with p predictors, MSE_{full} is the mean squared error for the full model, n is the number of observations, and p is the number of predictors. C_p value that is close to p (number of beta coefficients) represents better predictive model. Here, it is the model that has 3 predictor variables which are: **price**, **discount**, and **promotion**.

```
## [1] 8.599787 6.381096 3.796670 3.550583 5.320452 7.000000
```




```
##
## Call:
## lm(formula = marketshare ~ price + discount + promotion, data = market_share)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.286376 -0.100465 -0.002259  0.104174  0.240020
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.18527    0.36505   8.726  5.7e-10 ***
## price       -0.35269    0.15738  -2.241   0.0321 *
## discount     0.39914    0.05125   7.787  7.0e-09 ***
## promotion    0.11803    0.05149   2.292   0.0286 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1498 on 32 degrees of freedom
## Multiple R-squared:  0.7065, Adjusted R-squared:  0.679
## F-statistic: 25.68 on 3 and 32 DF,  p-value: 1.191e-08
```

```
##
## Call:
## lm(formula = marketshare ~ price + discount + promotion + month,
##     data = market_share)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.30355 -0.08138  0.02486  0.09287  0.23021
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.144380   0.358544   8.770 6.68e-10 ***
## price       -0.366445   0.154408  -2.373   0.0240 *
## discount     0.416139   0.051409   8.095 3.85e-09 ***
## promotion    0.096772   0.052298   1.850   0.0738 .
## month        0.011502   0.007493   1.535   0.1349
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1467 on 31 degrees of freedom
## Multiple R-squared:  0.7272, Adjusted R-squared:  0.692
## F-statistic: 20.66 on 4 and 31 DF,  p-value: 2.205e-08
```

Criterion-Based Statistics

Based on the previous automatic variable selection methods, it is hard to determine if the model with 4 or 5 parameters is the best predictive model. Therefore, other criteria not produced by the `regsubsets` function will be compared. These criteria are Akaike's information criterion (AIC) and Prediction Sum of Squares ($PRESS$). These statistics are compared for the best two potential subset models resulted from automatic variable selection.

1. Akaike Information Criterion (AIC) AIC is a model selection criteria that penalize models having large numbers of predictors, and given by : The AIC_P (Corrected Akaike Information Criterion) is calculated using the formula:

$$\text{AIC}_P = n\ln(SSE_2) - n\ln(n) + 2p$$

From the following, the model with 5 parameters has AIC= -133.57, while the one with 4 parameters has AIC= -132.94. The model that has smaller AIC value is a better predictive model.

```
## [1]    4.0000 -132.9353
```

```
## [1]    5.0000 -133.5726
```

2. Prediction Sum of Squares ($PRESS$) $PRESS$ assesses the predictive power of a regression model by measuring the sum of squared differences between actual and predicted responses when each observation is excluded from the model. From the following, the model with 5 parameters has $PRESS$= 0.872, while the one with 4 parameters has $PRESS$= 0.889. The model that has smaller $PRESS$ value is a better predictive subset model.

```
## [1] 0.8886516
```

```
## [1] 0.8728642
```

Model Validation

Model validation can help us select the model that has the best predictive performance in a hold-out sample. There are several approaches to model validation, two of which are shown here.

1. Leave-one-out cross validation The process of "Leave-one-out cross validation" works as follows: 1. One data point will be left out and the model is built using the remaining data. 2. The model is tested against the data point removed in Step 1 and the prediction error is recorded. 3. The process is repeated

for all data points. 4. The overall prediction error is computed by averaging the prediction errors. 5. When comparing models, the model with lowest MSPE should be selected.

MSPE is a measure of the average squared difference between the observed values and the values predicted and given by the formula : $MSPE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. RMSE value that is shown below is the square root of the mean squared differences and given by $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$. Lower values for MSPE or RMSE tell that the model's predictions align more with the actual data.

The best two subset models are trained using the process of "Leave-one-out cross validation" as shown below, and the model that has **price**, **discount**, **promotion**, and **month** predictors has the lower MSPE.

```
## Linear Regression
##
## 36 samples
##  3 predictor
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 35, 35, 35, 35, 35, 35, ...
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.1571139  0.6384616  0.1318772
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
## Linear Regression
##
## 36 samples
##  4 predictor
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 35, 35, 35, 35, 35, 35, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   0.155712  0.6458495  0.1277597
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

2. K-fold cross validation The method of "K-fold cross validation" performs better for larger datasets where training and testing data are available/feasible. This method involves:

1. The data are randomly split into k subsets. One of the subsets is reserved for testing.
2. The model is built(trained) on the remaining $k-1$ subsets.
3. The model is tested on the reserved subset and the mean squared prediction error is recorded.
4. The process is repeated with changing the testing subset each time until all k subsets have served as the testing set.
5. The average of the k mean squared prediction errors is calculated.
6. when comparing models, the model with the lowest MSPE is chosen.

The model that has **price**, **discount**, **promotion**, and **month** predictors has the lower MSPE.

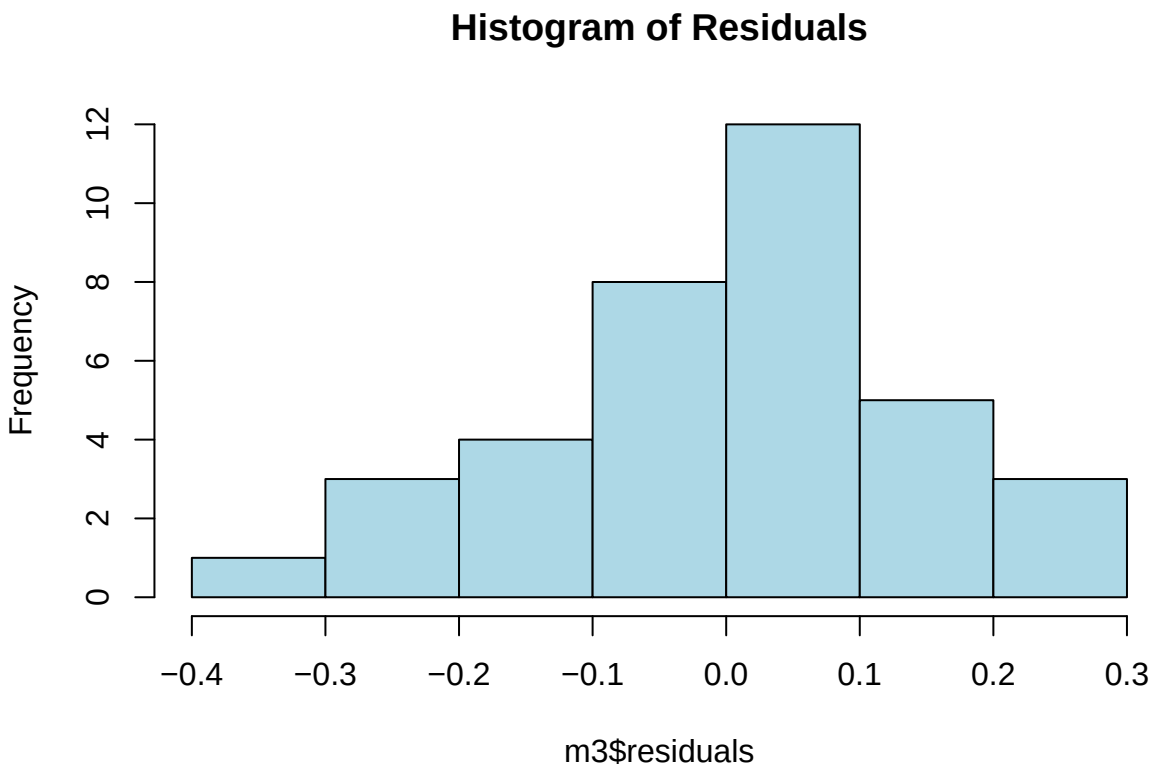
```
## Linear Regression
##
## 36 samples
```

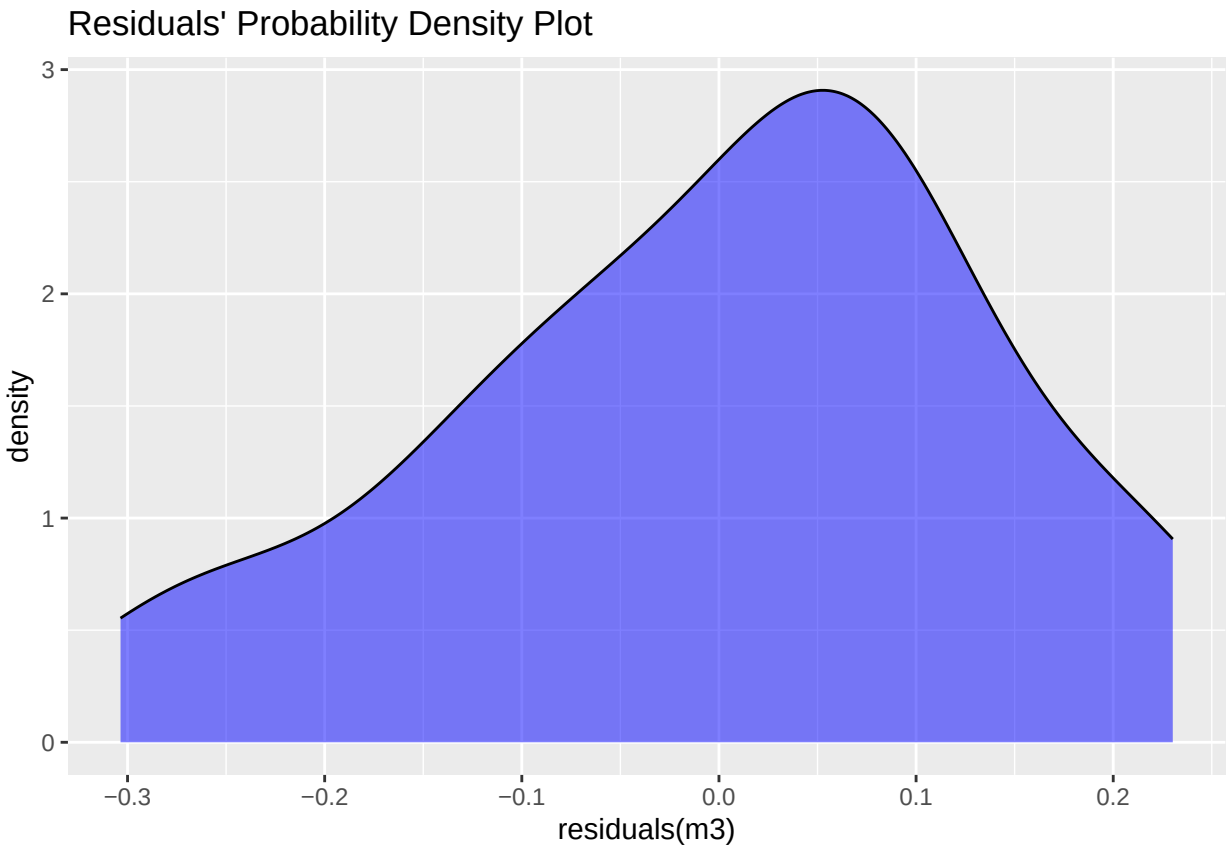
```
##  3 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 32, 32, 32, 33, 32, 33, ...
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.1467486  0.7135301  0.1322593
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
## Linear Regression
##
## 36 samples
##  4 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 33, 34, 32, 32, 32, 32, ...
## Resampling results:
##
##   RMSE       Rsquared  MAE
##   0.1510346  0.759817  0.132907
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

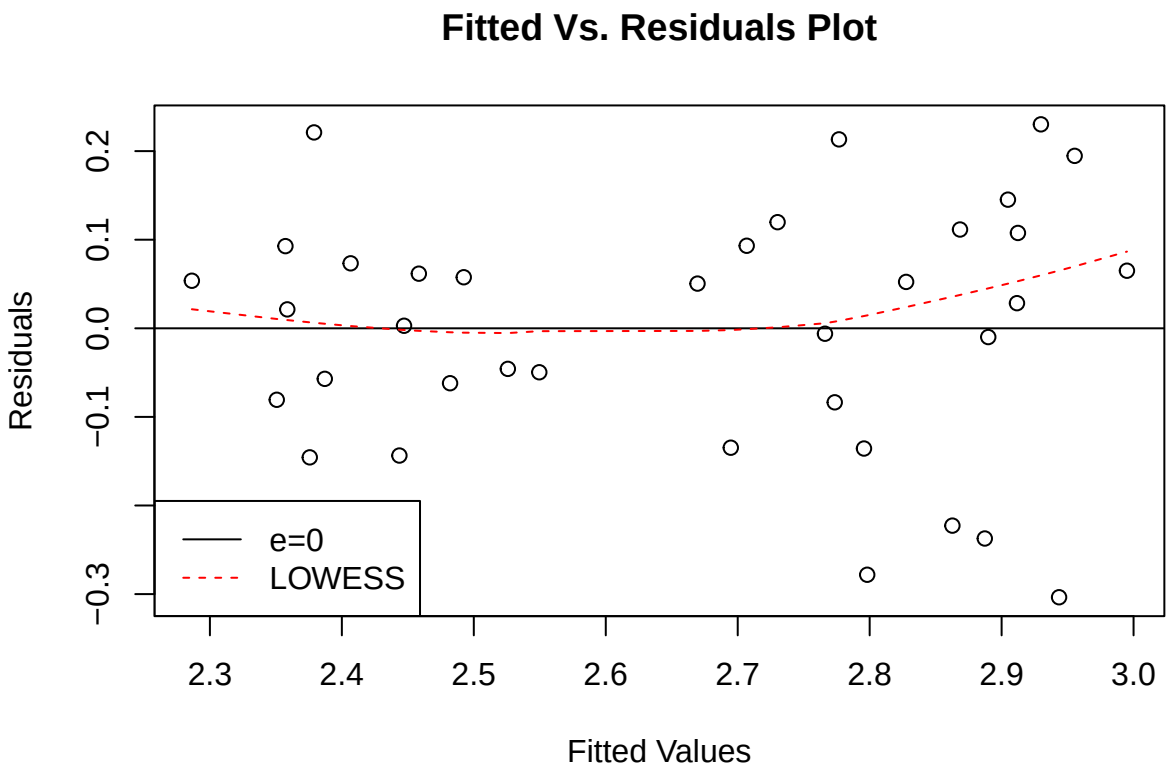
Residual Diagnostics

1. Model Completeness The histogram of residuals along with the residuals' probability density plot are shown below; they support the normality assumption of the residuals to some extent. However, there is slight left skewness in the residual's distribution, which needs further investigations. The reason for the slight skewness could be due to the small sample size of data points or the possibility of having outliers.


Histogram of Residuals
Frequency
0
2
4
6
8
10
12
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
m3$residuals



The fitted-versus-residual plot looks like a “cone” shape for the upper half of the fitted values, and this shows that variance of the residuals might not be constant,and possibly heteroscedasticity is present.



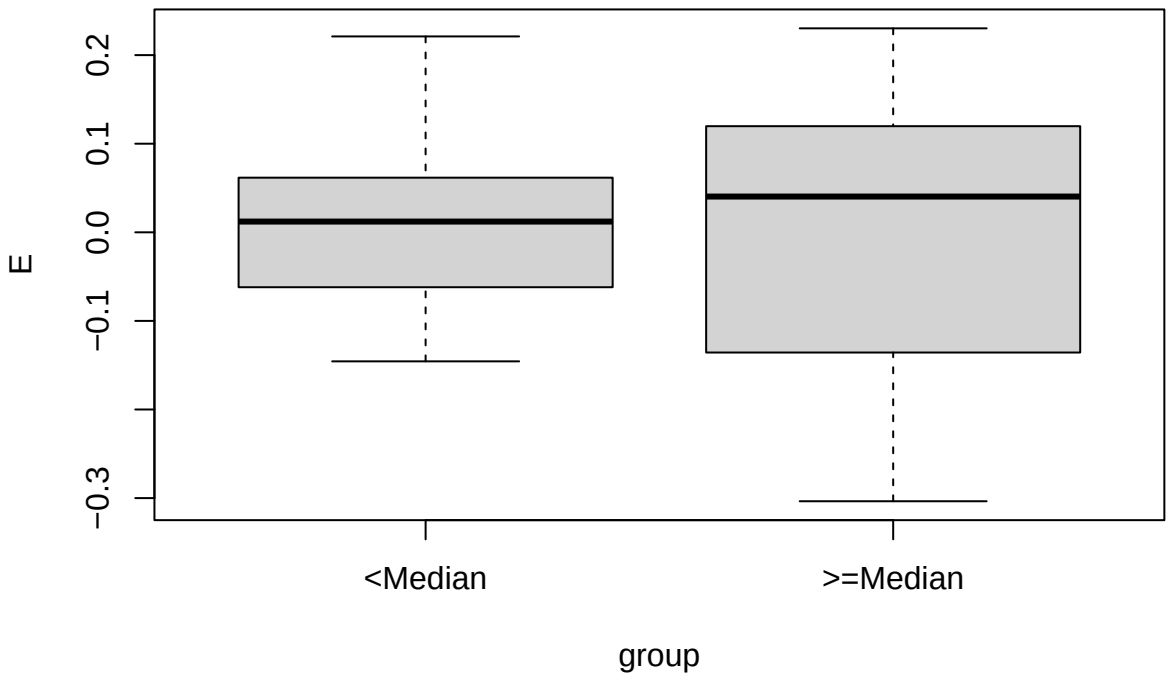
The "cone" shape for the upper half of the fitted values in the previous plot could be due to the lack of data points in the middle of the plot. However, to formally test for non-constant variance, a test called Brown-Forsythe test will be performed with a specified significance level of .05. The Brown-Forsythe test evaluates if variances differ between two groups based on the level of variable X. It divides the data into low and high X groups, using absolute deviations around medians for robustness. The test simplifies to a two-sample t-test on these absolute deviations. Despite deviations not necessarily being normally distributed, the test is robust and applicable when variance is constant, and sample sizes aren't extremely small. It's a useful tool for checking homogeneity of variances in cases where the assumption of constant variance might be compromised (Kutner et. al., 2014).

Null hypothesis H0: The error variance is constant .

Alternative hypothesis HA:The error variance is not constant .

If the p-value from Brown-Forsythe test is less than the significance level .05, we reject the null hypothesis. Otherwise, we fail to reject and conclude that the error variance is constant.

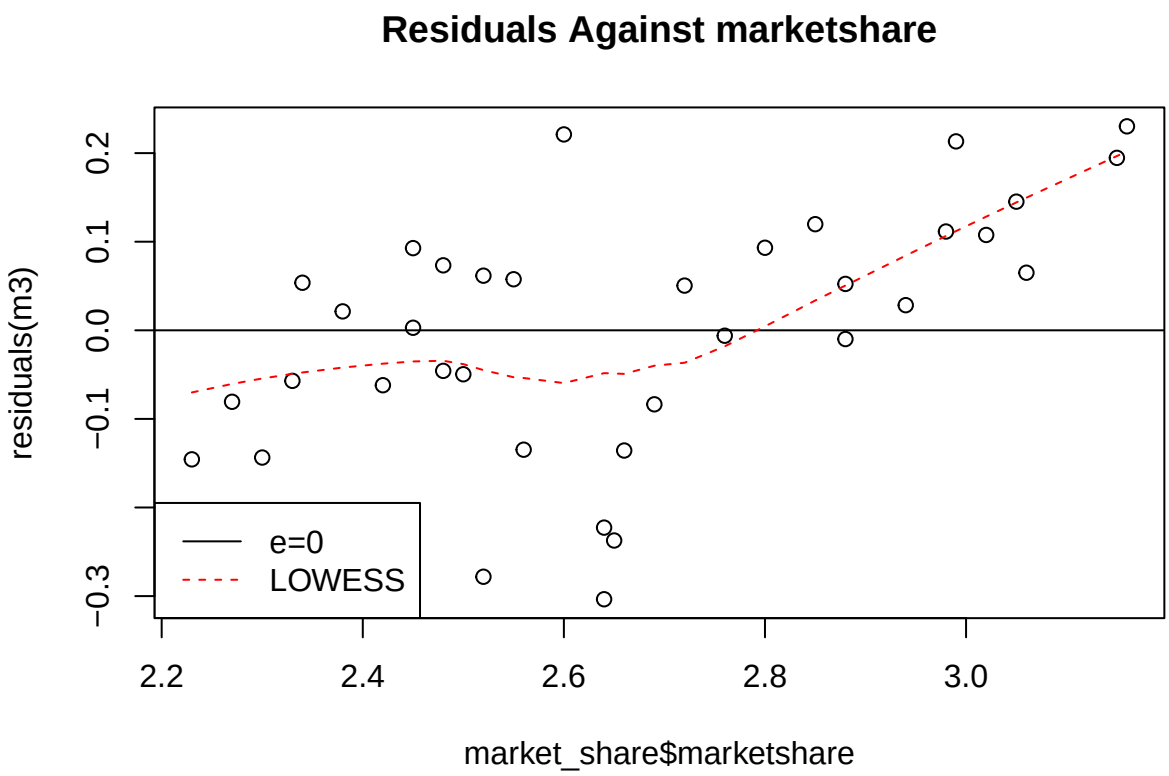
Since the P_value = 0.983 from the results of the test below, which is greater than 0.05, Brown-Forsythe test results in failing to reject H0 and conclude that the error variance is constant.

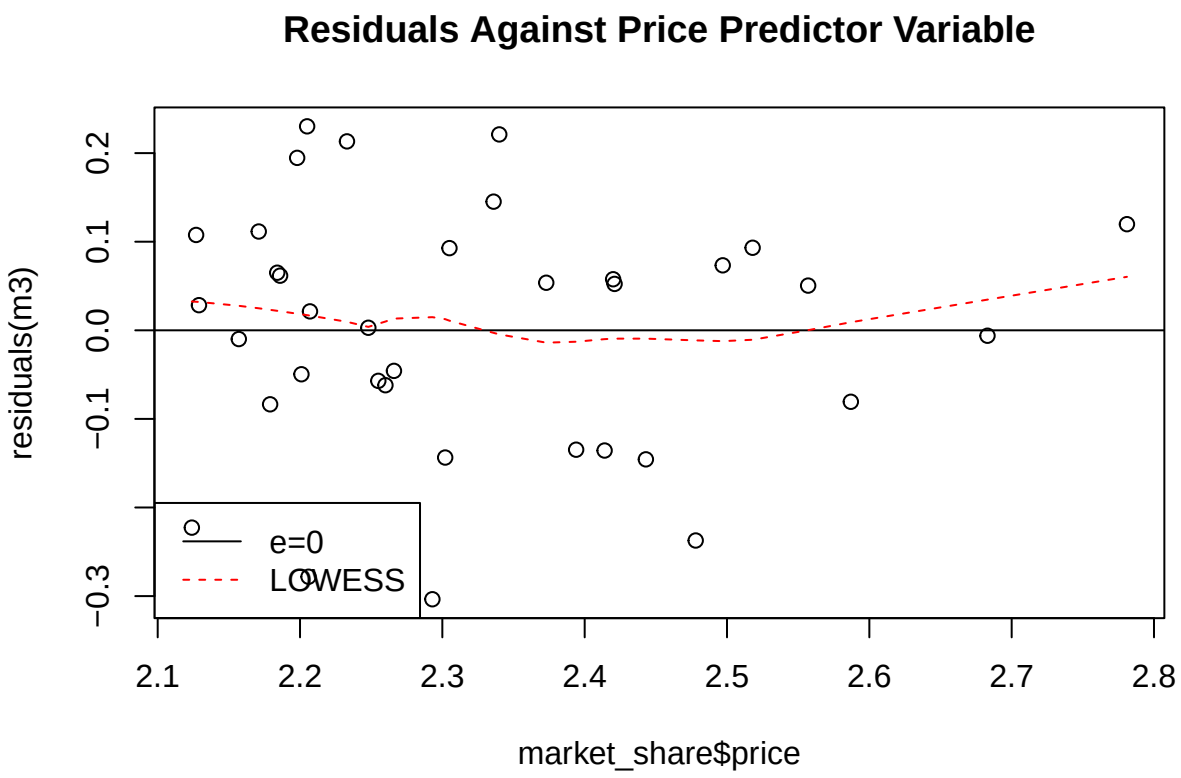


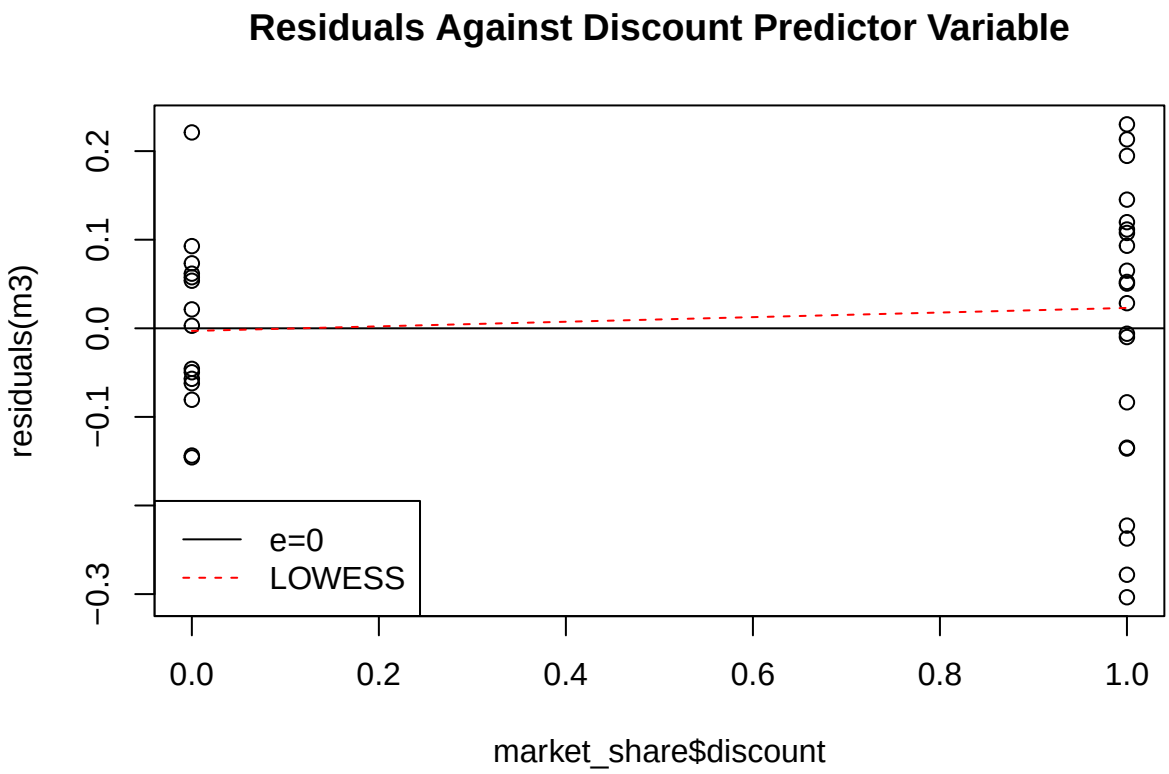

```
## # A tibble: 2 x 2
##   group        var
##   <fct>      <dbl>
## 1 <Median  0.00957
## 2 >=Median 0.0297
```

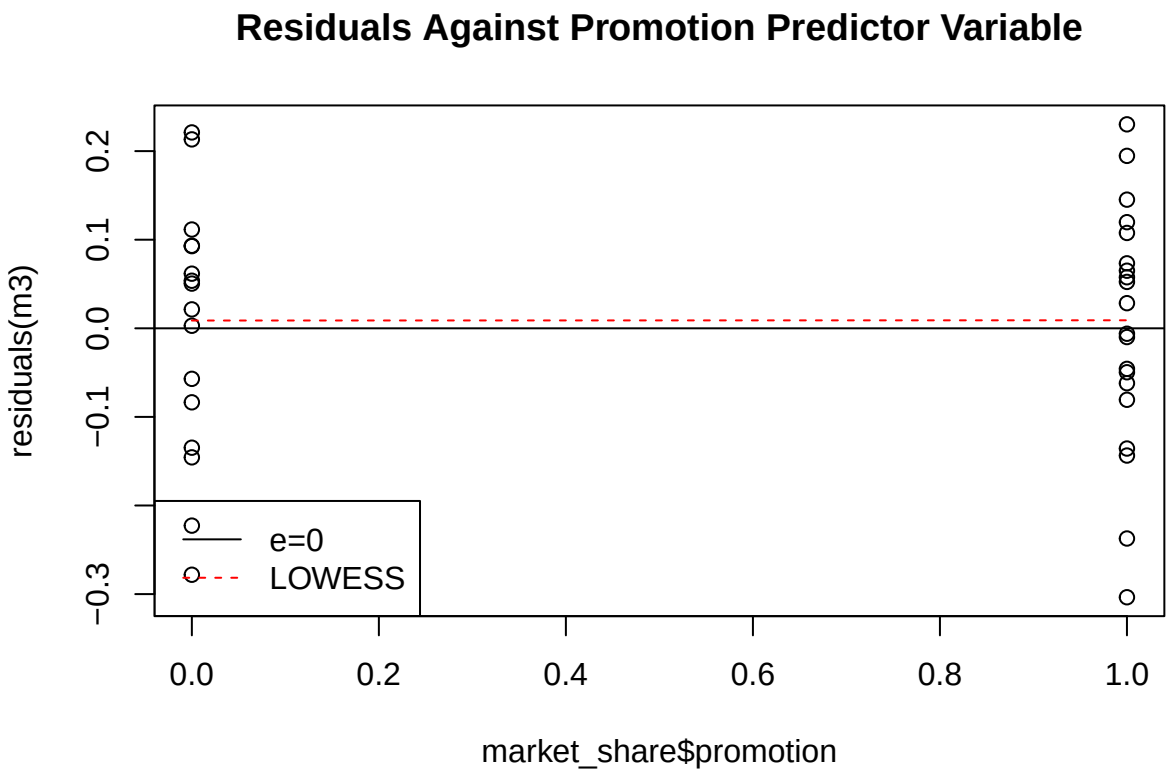
```
##
##   Brown-Forsythe Test (alpha = 0.05)
## ------------------------------------------------------------
##   data : E and group
##
##   statistic  : 0.0004490343
##   num df     : 1
##   denom df   : 26.93328
##   p.value    : 0.98325
##
##   Result     : Difference is not statistically significant.
## ------------------------------------------------------------
```

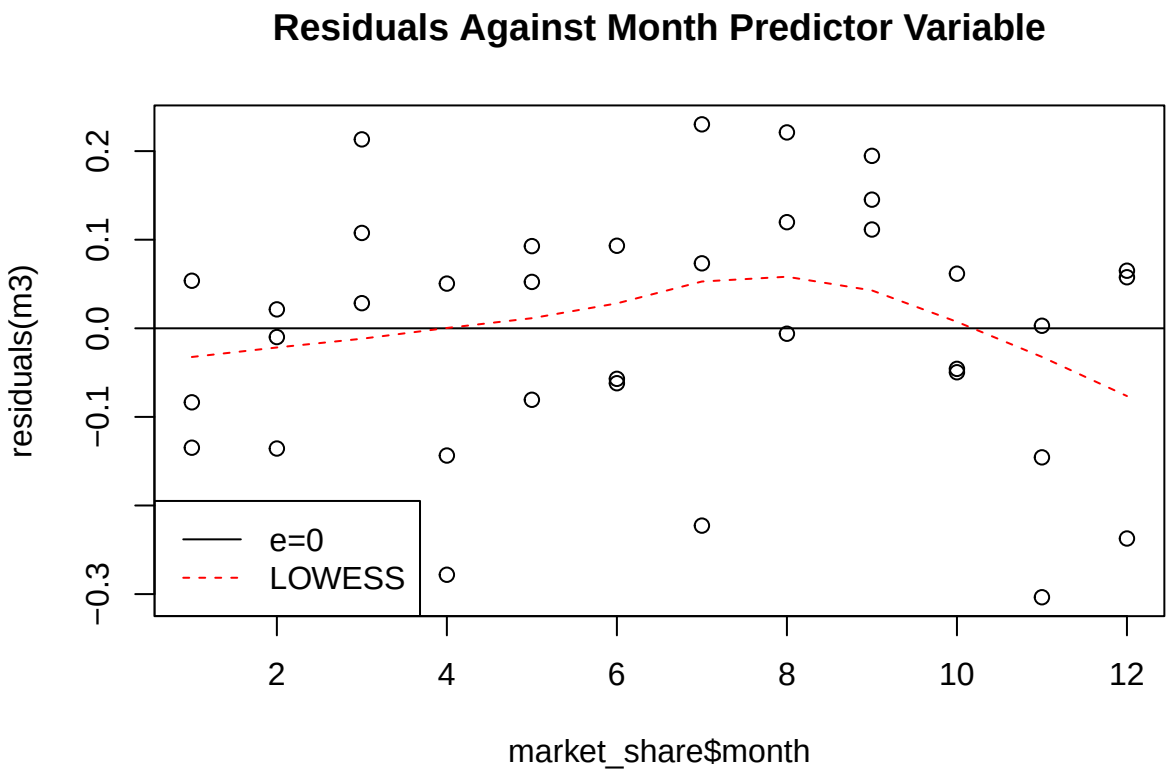
Residuals against the response variable **marketshare** and against every predictor variable are plotted below.Residuals against the response variable plot shows that, for the most part, residuals are increasing as the response variable is increasing. However, it was proven by Brown-forsythe test that the error variance is constant (difference is not statistically significant). Also, logarithmic and square root transformations of the response variable were conducted, but did not change the distribution of the error terms. The issue again could be caused by the small sample size of observations. Residuals against each of the predictor variables seem to be randomly distributed with no systemic patterns.


Residuals Against marketshare
residuals(m3)
0.2
0.1
0.0
−0.1
−0.3
e=0
LOWESS
2.2
2.4
2.6
2.8
3.0
market_share$marketshare

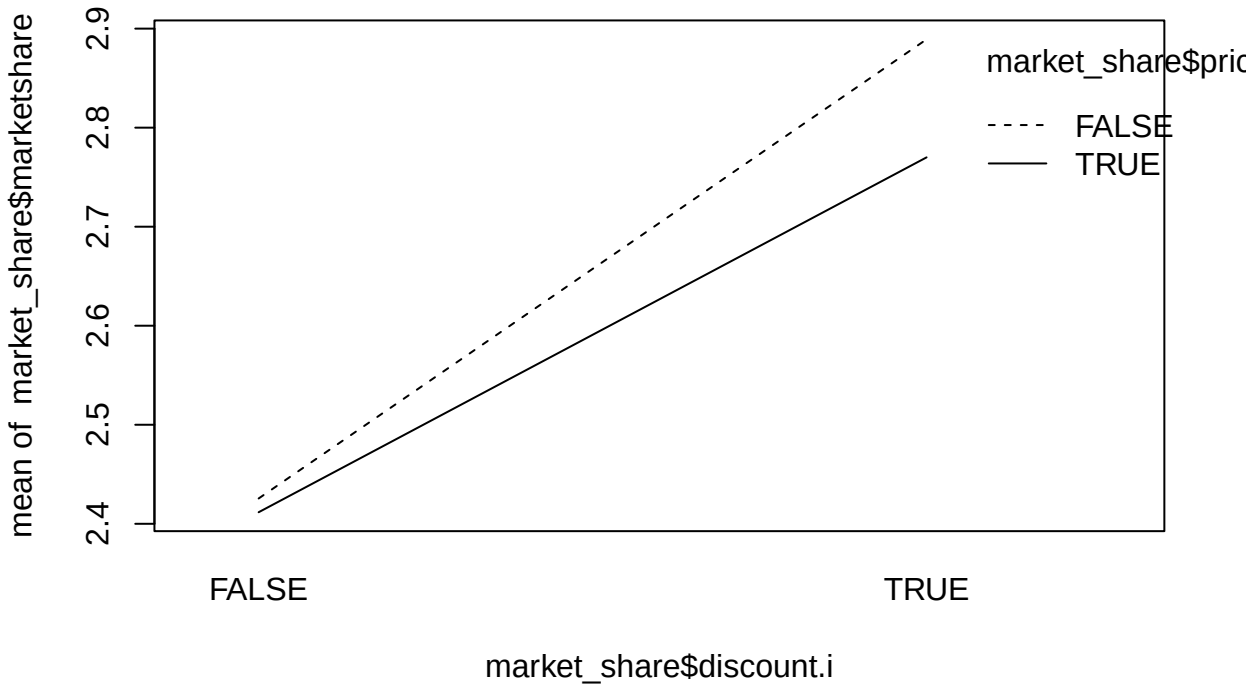

Residuals Against Price Predictor Variable
residuals(m3)
0.2
0.1
0.0
−0.1
−0.3
e=0
LOWESS
2.1
2.2
2.3
2.4
2.5
2.6
2.7
2.8
market_share$price

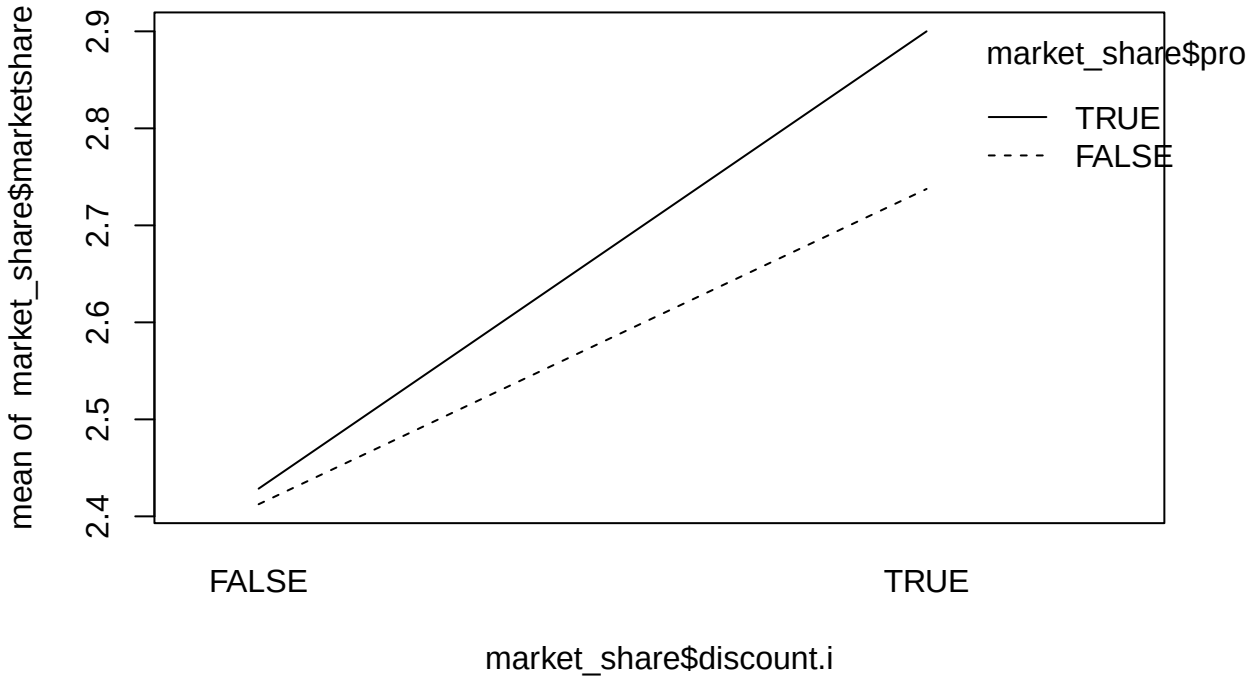

Residuals Against Discount Predictor Variable
residuals(m3)
0.2
0.1
0.0
−0.1
−0.3
e=0
LOWESS
0.0
0.2
0.4
0.6
0.8
1.0
market_share$discount

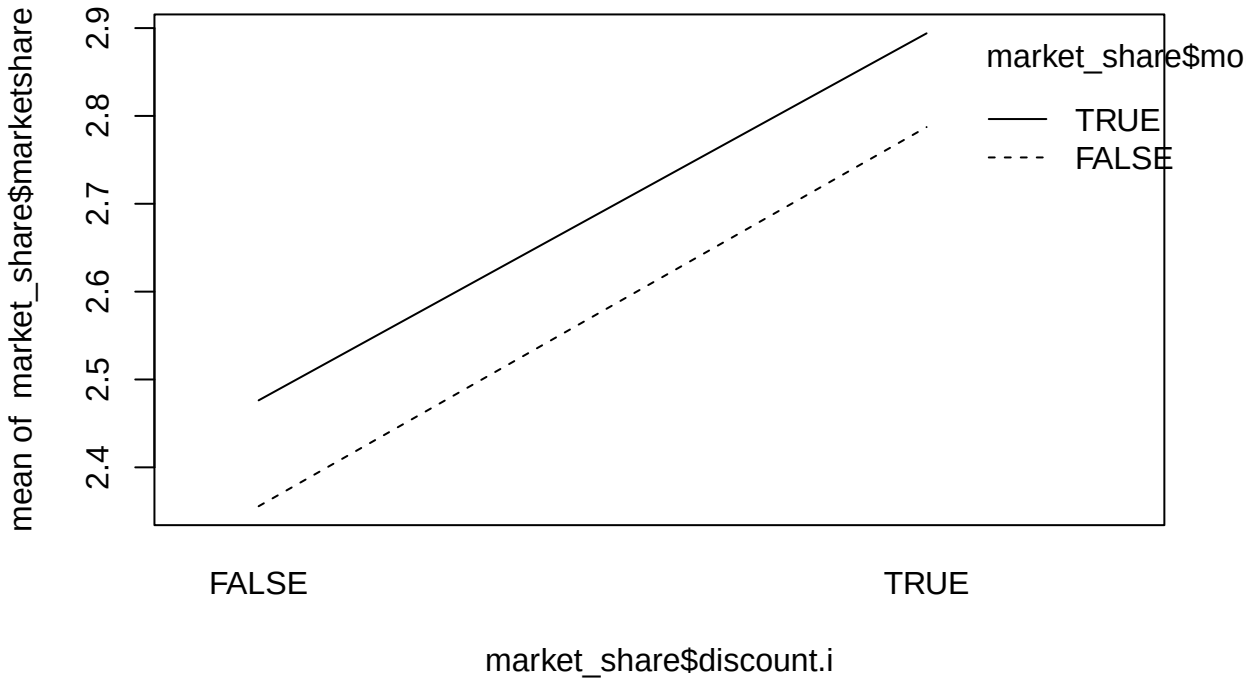

Residuals Against Promotion Predictor Variable
residuals(m3)
0.2
0.1
0.0
−0.1
−0.3
e=0
LOWESS
0.0
0.2
0.4
0.6
0.8
1.0
market_share$promotion



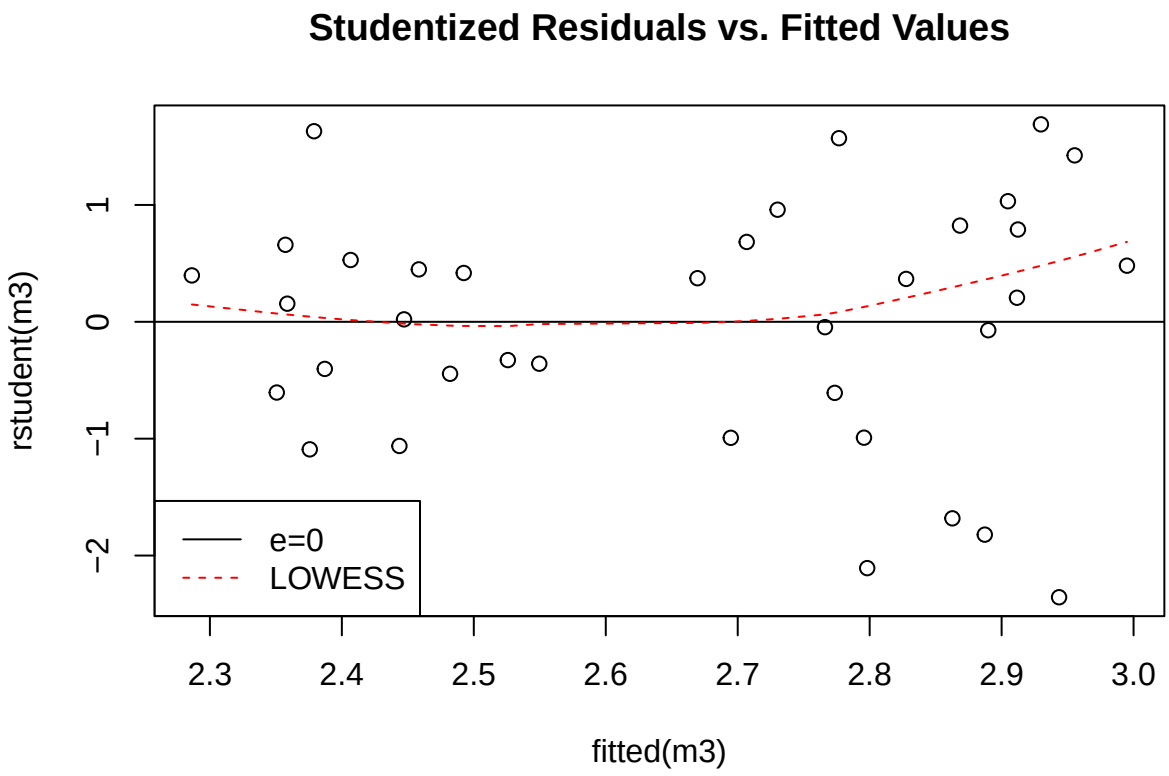
Check if any variable interaction terms improve the predictive measures of the regression model. The plots below don't show any interaction between the covariates.


mean of market_share$marketshare
2.4
2.5
2.6
2.7
2.8
2.9
FALSE
TRUE
market_share$discount.i
market_share$pric
FALSE
TRUE

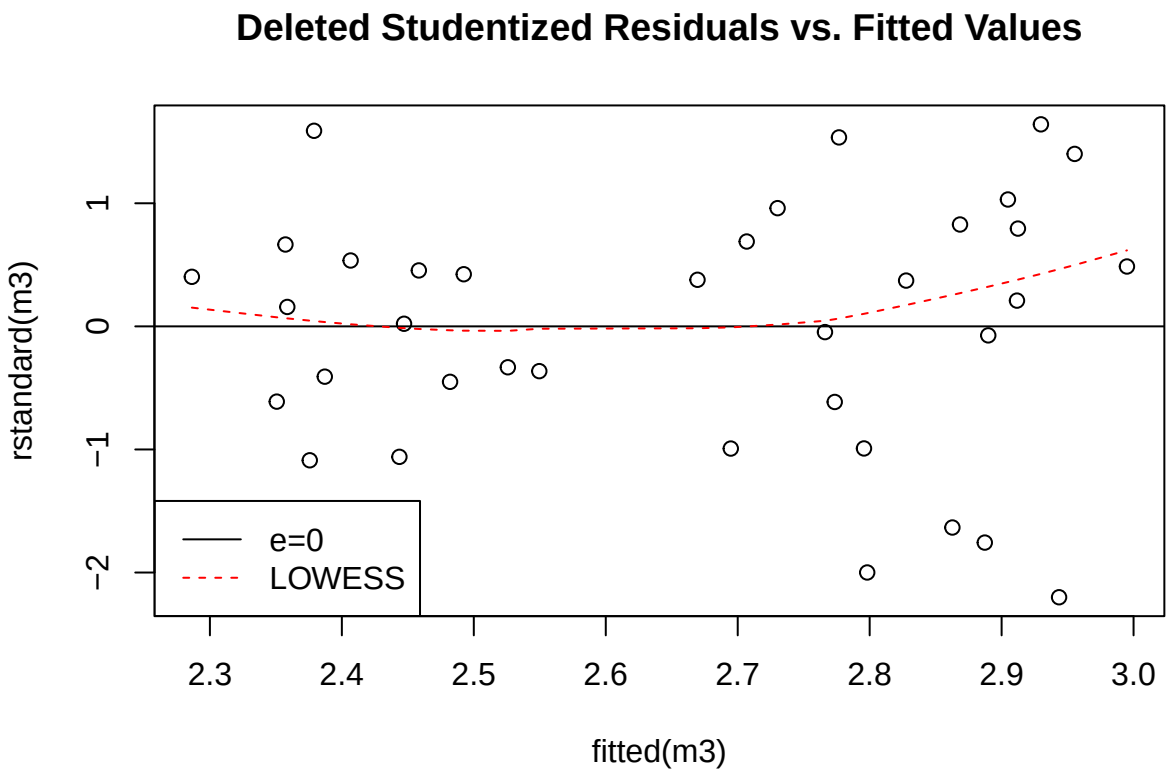

mean of market_share$marketshare
2.4
2.5
2.6
2.7
2.8
2.9
FALSE
TRUE
market_share$discount.i
market_share$pro
TRUE
FALSE

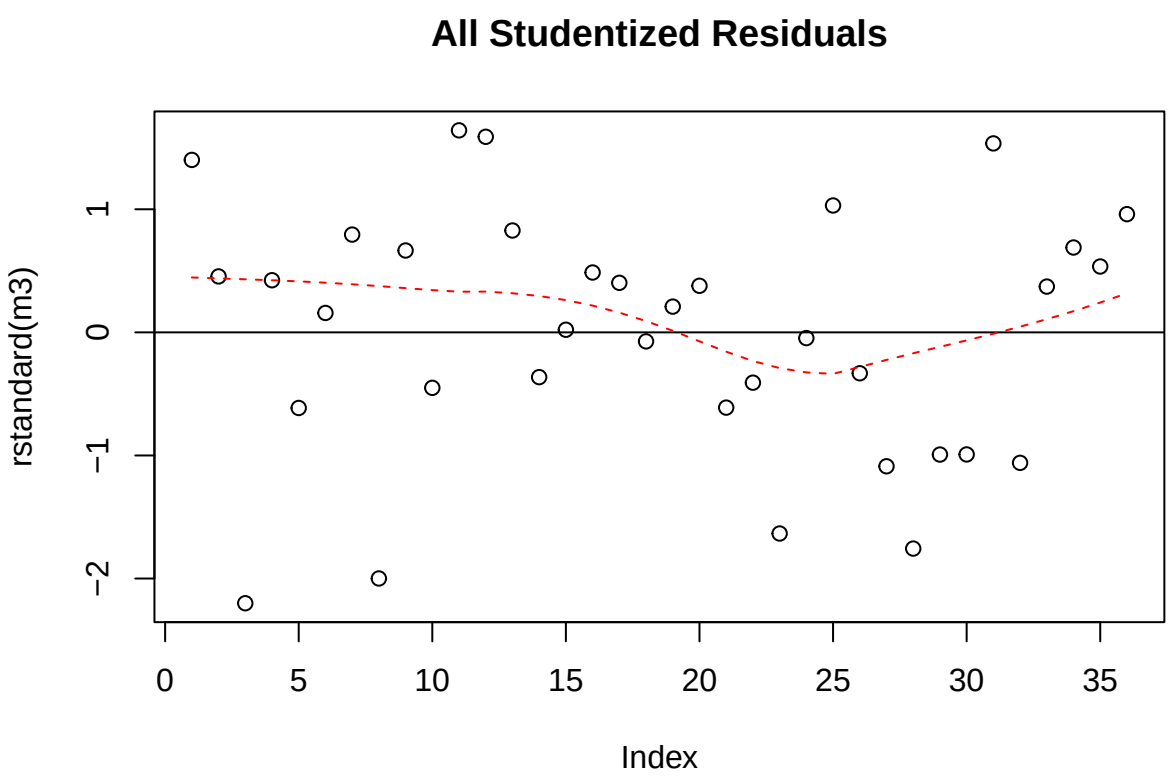


2. Investigating Outliers and Influential Points Plots of studentized residuals, and deleted studentized residual are shown below. Other than the lack of data points in the middle, the assumption of constant variance (homoscedasticity) is met.There are no large deleted studentized residuals indicated as outlying with respect to the X and Y.


Studentized Residuals vs. Fitted Values
rstudent(m3)
fitted(m3)
e=0
LOWESS
2.3
2.4
2.5
2.6
2.7
2.8
2.9
3.0
−2
−1
0
1

```
## integer(0)
```


Deleted Studentized Residuals vs. Fitted Values
rstandard(m3)
fitted(m3)
e=0
LOWESS
1
0
−1
−2
2.3
2.4
2.5
2.6
2.7
2.8
2.9
3.0




```
## named integer(0)
```

Another way of identifying outlying cases with respect to the X and Y is the hat matrix. The hat matrix, denoted by H, maps the observed values of the dependent variable to the predicted values and is defined as:

$$H = X(X^T X)^{-1} X^T$$

where:

X is the design matrix of the model, containing the values of the predictor variables. X^T is the transpose of the design matrix. $(X^T X)^{-1}$ is the inverse of the product of the transpose of X and X. The diagonal elements of the hat matrix are often denoted as H_{ii}, representing the leverage values.if H_{ii} is greater than 2p/n where p (number of parameters)= 5 and n (sample size)= 36, then it is a high leverage point that potentially influences the model fit (Kutner et. al., 2014). As shown below, no cases were identified as outlying with regard to their X values.

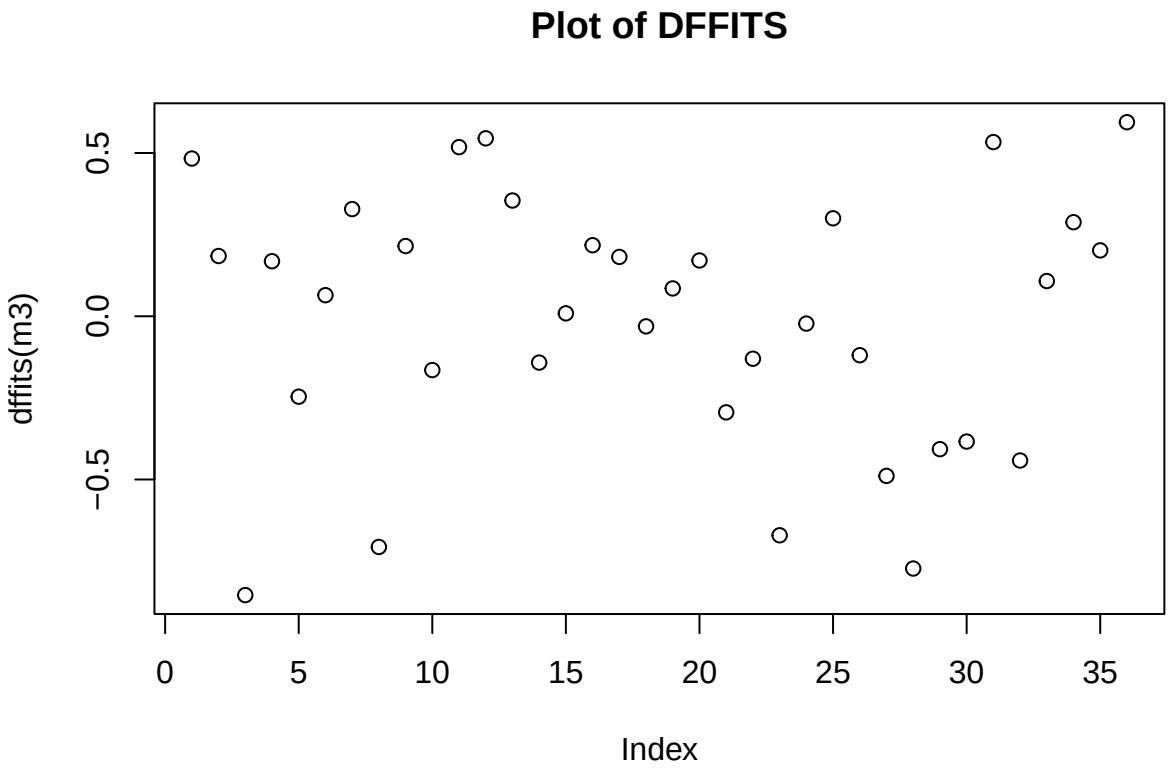
```
## named integer(0)
```

TO test if a case/observation is influencing the fitted values and estimated regression coefficients, DFFITS, Cook's distance, and DFBETAS measures will be investigated.

DFFITS "Difference in Fits" is a measure that considers the influence of the ith case on the fitted value $\hat{Y}_i$ for this case. The DFFITS for the i-th observation is given by:

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(-i)}}{\sqrt{\text{MSE}_i \cdot h_{ii}}}$$

where: $\hat{Y}_i$ is the predicted value when the i-th observation is included in the model. $\hat{Y}_{i(-i)}$ is the predicted value when the i-th observation is excluded from the model. MSE_i is the Mean Squared Error associated with the i-th observation. h_{ii} is the leverage of the i-th observation. The rule of thumb for identifying influential cases: "a case is influential if the absolute value of DFFITS exceeds 1 for small to medium data sets and 2*sqrt(p/n) for large data sets" (Kutner et. al., 2014). The following cases were identified as influential:




```
##  3 28
##  3 28
```

`DFBETAS` measures the standardized effect of deleting each individual observation on each coefficient.. For the k-th coefficient, `DFBETAS` is calculated as:

$$\text{dfbetas}_{ik} = \frac{b_k - b_{k(i)}}{\sqrt{\text{MSE}_i^{(Ckk)}}}$$

where: - b_k is the estimated coefficient for the k-th variable when all observations are included. - $b_{k(i)}$ is the estimated coefficient for the k-th variable when the i-th observation is excluded. - $\text{MSE}_i^{(Ckk)}$ is the Mean Squared Error associated with the i-th observation for the k-th coefficient.

The numerator represents the change in the estimated coefficient when the i-th observation is excluded, and the denominator is the square root of the Mean Squared Error for the k-th coefficient associated with the i-th observation. A good rule of thumb for identifying influential cases: "consider a case influential if the absolute value of DFBETAS exceeds 1 for small to medium data sets and 2/sqrt(n) for large data sets" (Kutner et. al., 2014) . The following orders of the DFBETAS exceed the threshold of 0.333:

```
##  [1]  36  72  75  80  95 100 116 131 147 172
```

Some of the DFBETAS for cases 3 and 28 exceed the threshold of .333; the values are shown below:

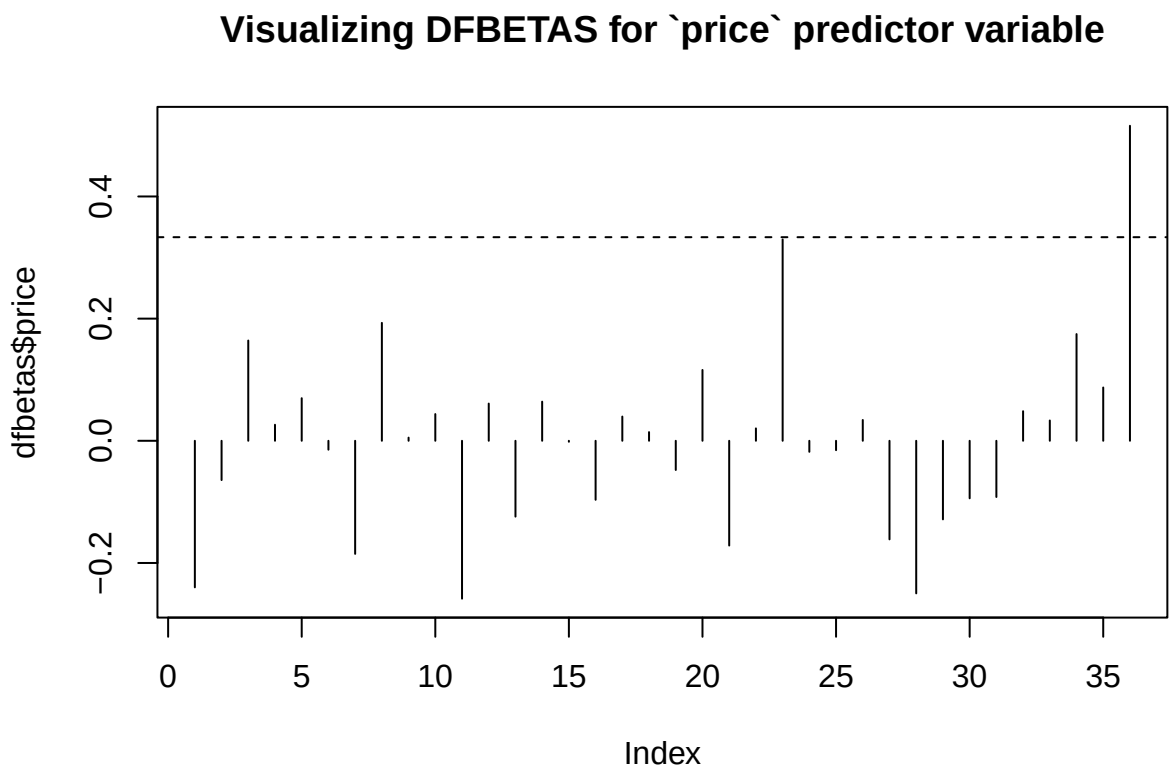
DFBETAS for case 3:

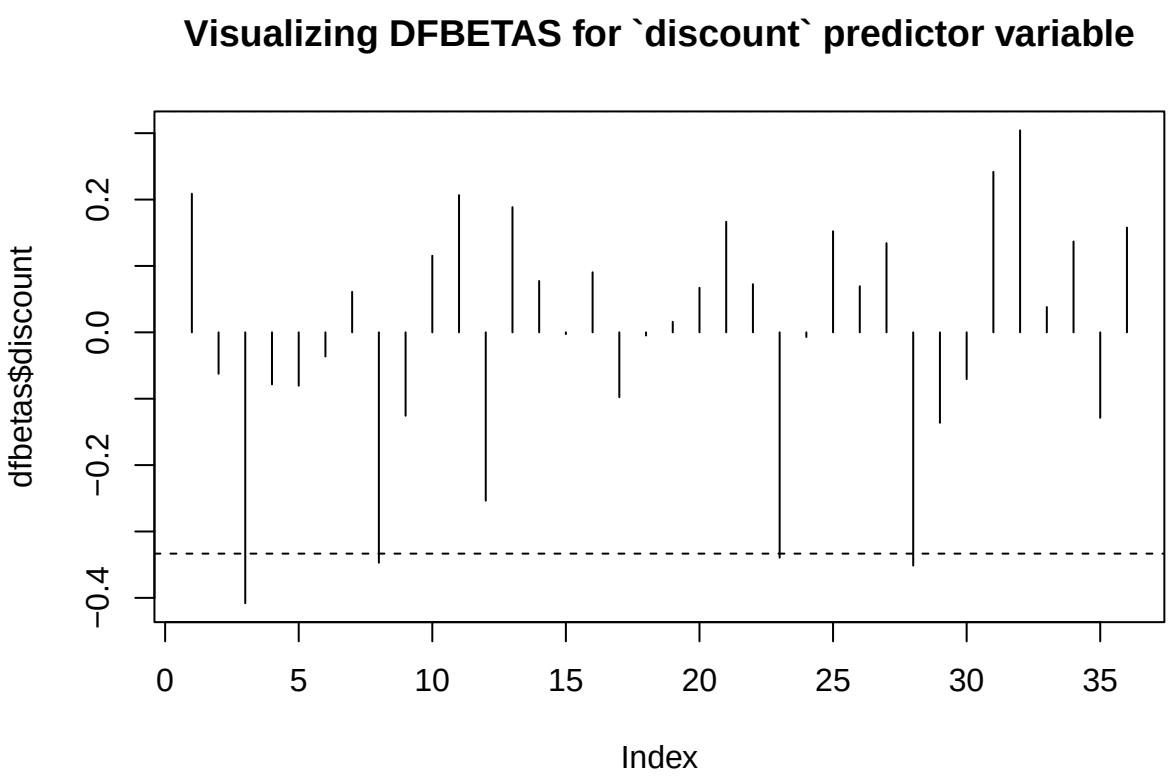
```
## (Intercept)       price    discount   promotion       month
##  0.06860998  0.16416456  0.40808943  0.18195002  0.55460522
```

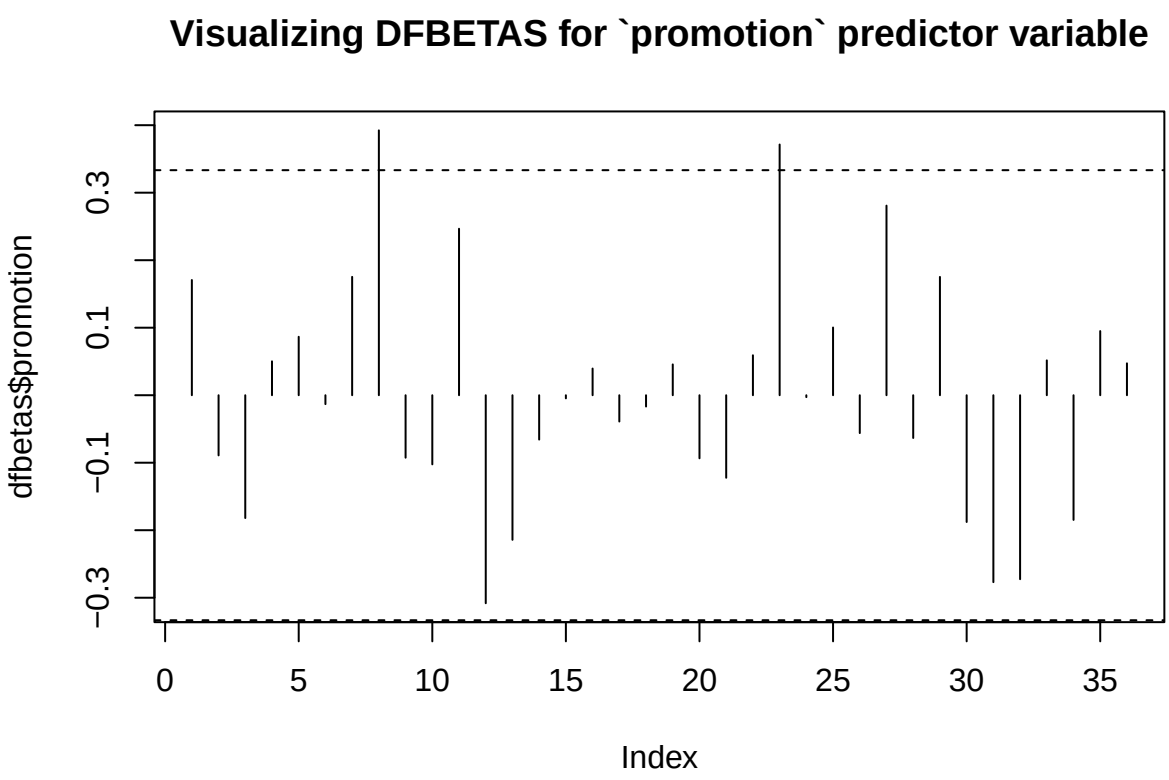
DFBETAS for case 28:

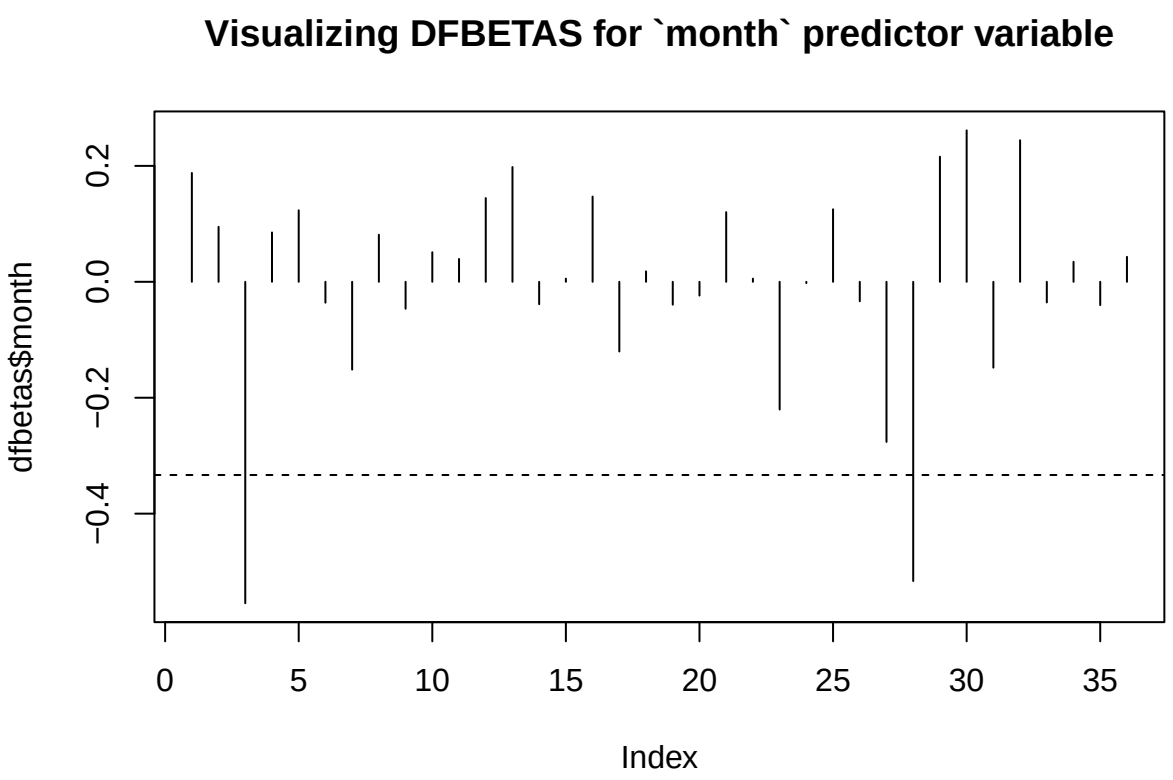
```
## (Intercept)       price    discount   promotion       month
##  0.33219567  0.24976243  0.35138976  0.06339212  0.51635110
```

Visualize the DFBETAS for each covariate :

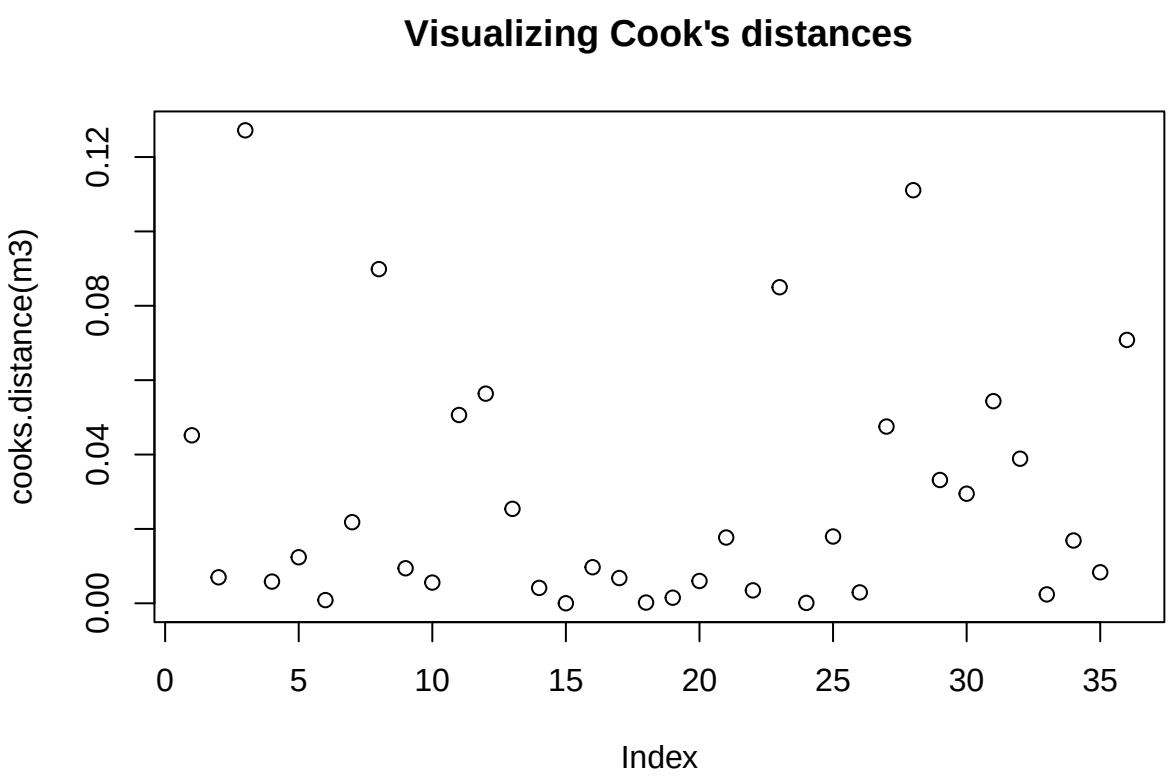



Visualizing DFBETAS for `discount` predictor variable
dfbetas$discount
0.2
0.0
−0.2
−0.4
0
5
10
15
20
25
30
35
Index


Visualizing DFBETAS for `promotion` predictor variable
dfbetas$promotion
0.3
0.1
−0.1
−0.3
0
5
10
15
20
25
30
35
Index



Finally, Cook's distance measure considers the influence of the ith case on all n fitted values. Cases 3 and 28 have percentiles that are higher than 10th percentile. However their cooks distance are in the 12.7th percentile and 11.1th percentile respectively, which are slightly larger than 10th percent; the extent of the influence may not be large enough to call for consideration of remedial measures.



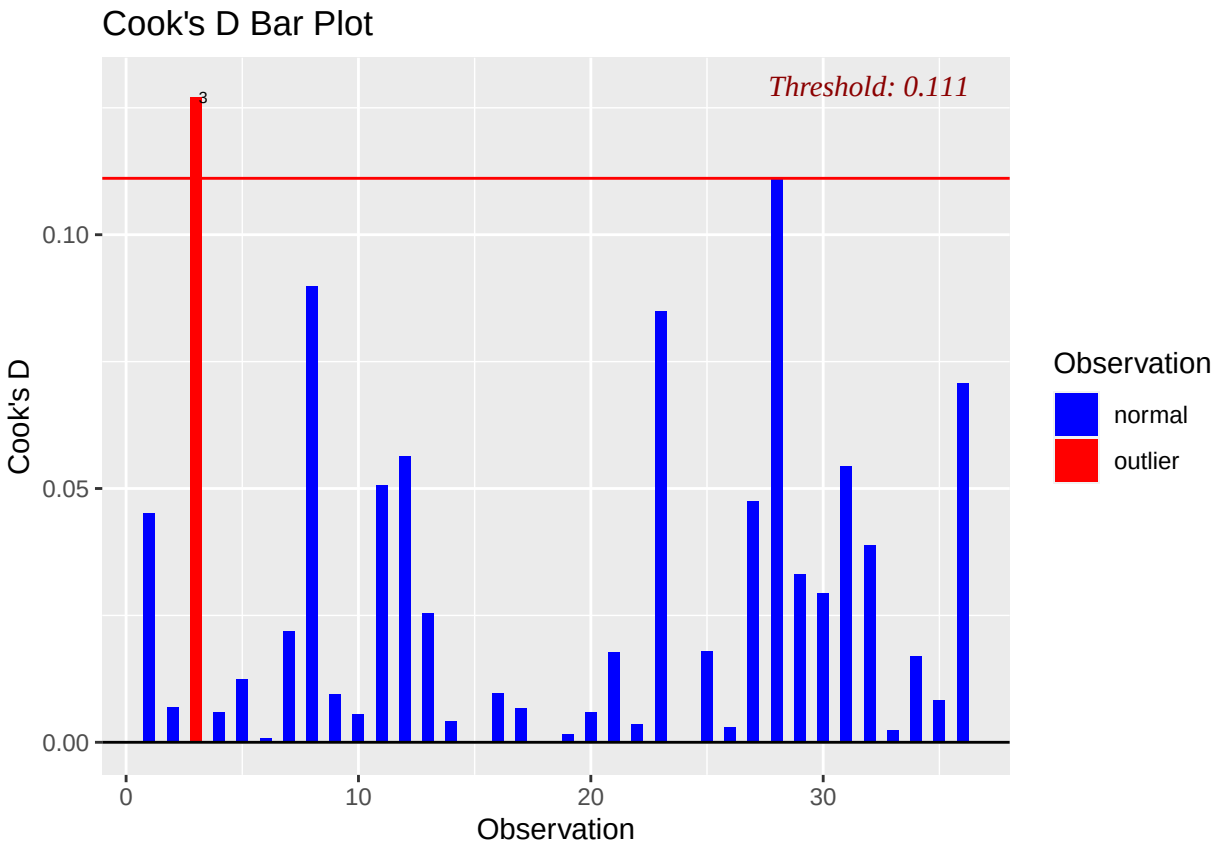

```
##  3 28
##  3 28
```

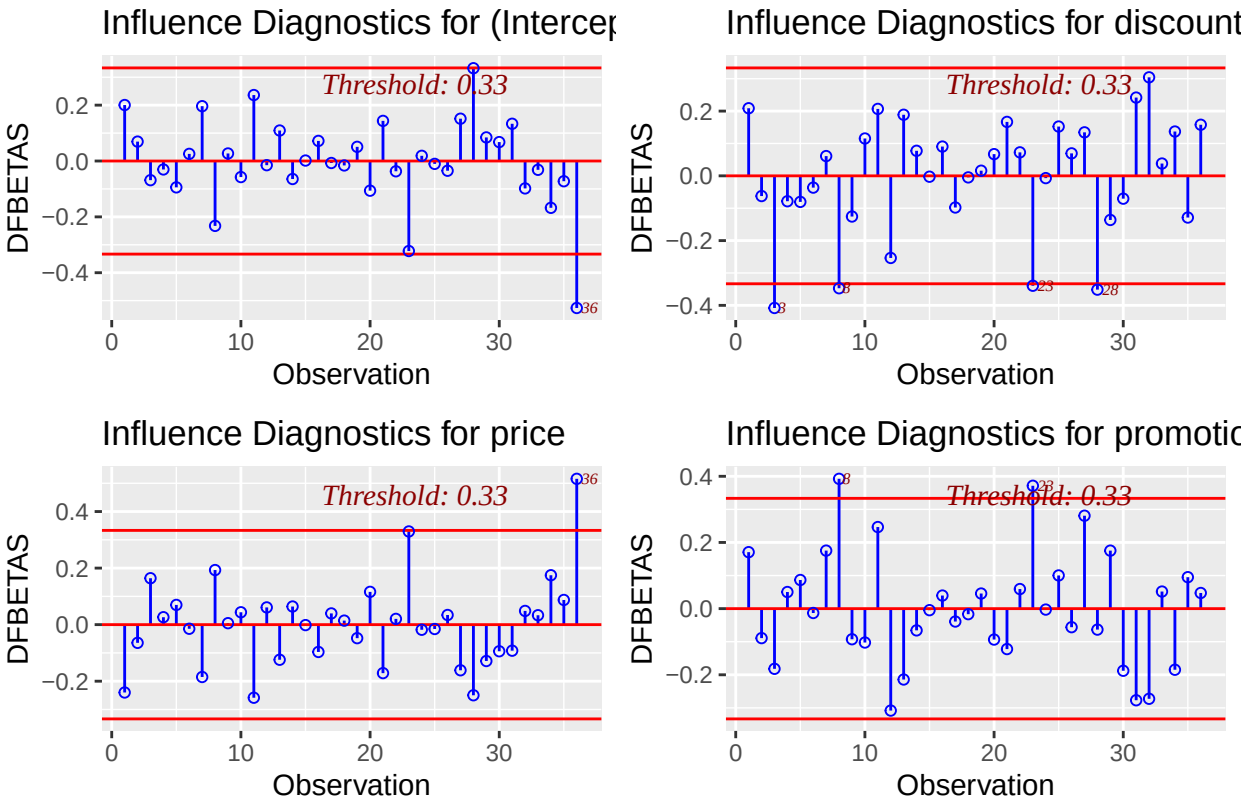
```
## named integer(0)
```

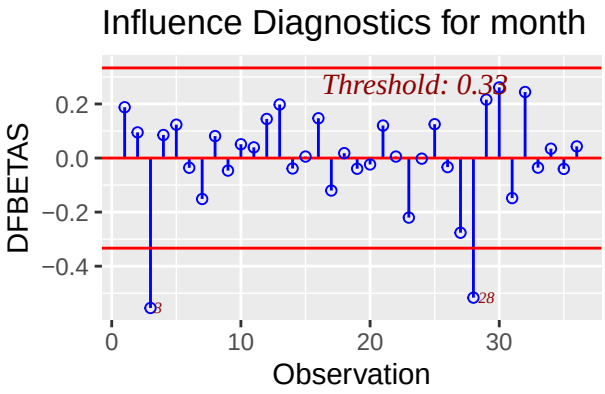
```
## Cook's distance for case 3 is: 0.1271727
```

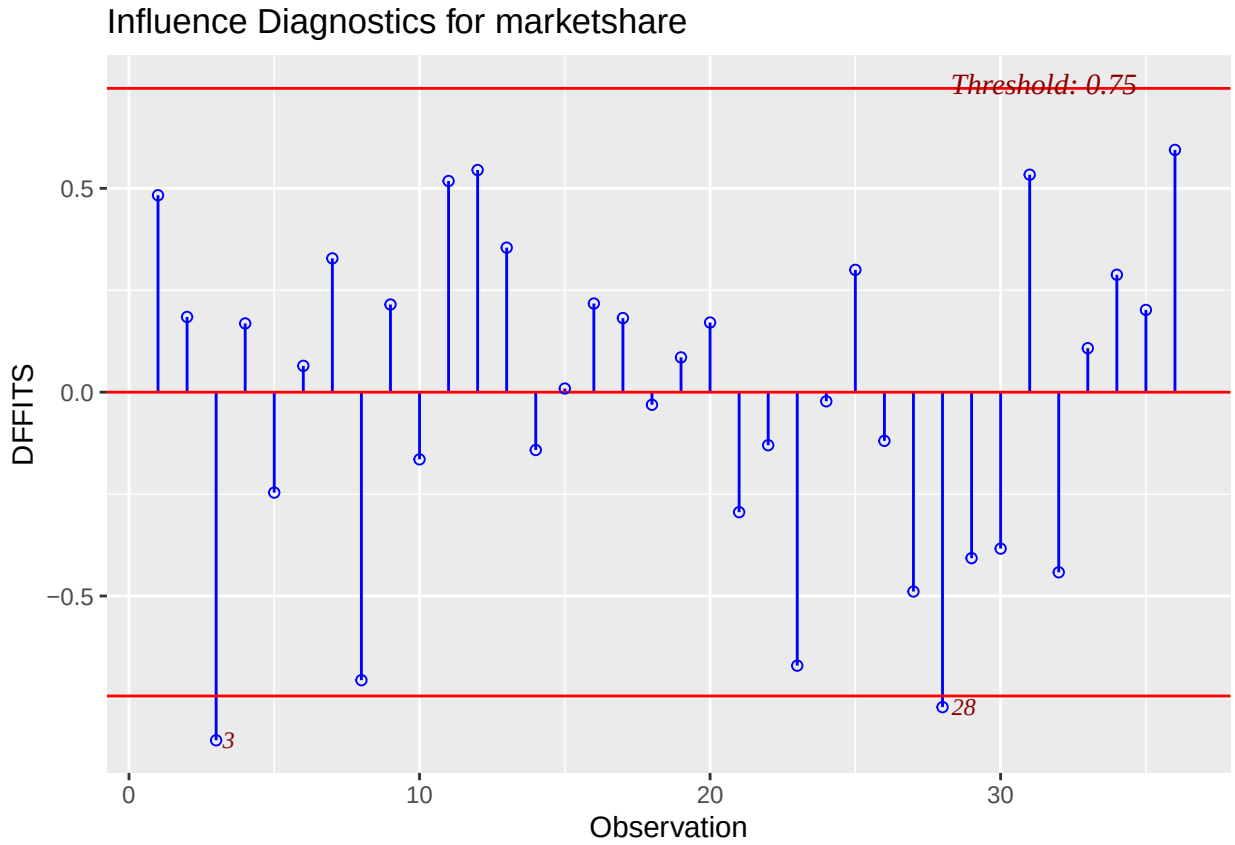
```
## Cook's distance for case 28 is: 0.1110754
```

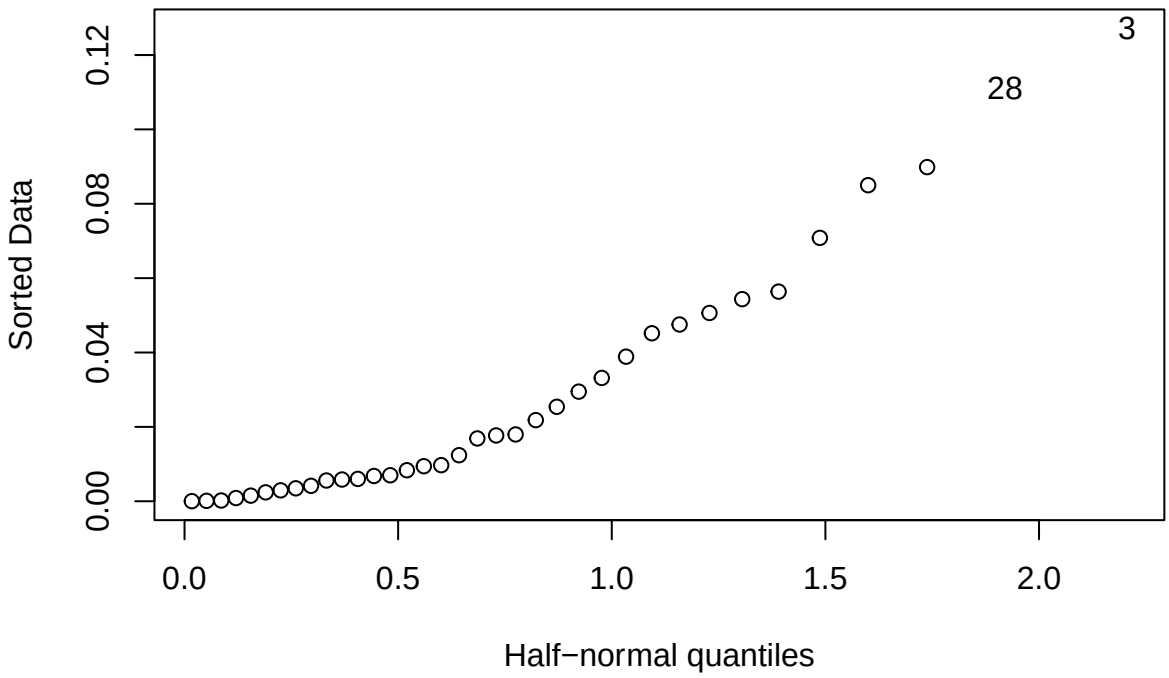
Another ways of visualizing DFFITS, DFBETAS, and Cook's Distances:


Cook's D Bar Plot
Threshold: 0.111
3
0.10
0.05
0.00
Cook's D
0
10
20
30
Observation
Observation
normal
outlier


Influence Diagnostics for (Intercep
Influence Diagnostics for discount
Influence Diagnostics for price
Influence Diagnostics for promotic
Threshold: 0.33
DFBETAS
Observation
0.4
0.2
0.0
−0.2
−0.4
0
10
20
30
36
8
23
28


Influence Diagnostics for month
Threshold: 0.33
DFBETAS
0.2
0.0
−0.2
−0.4
3
28
0
10
20
30
Observation


Influence Diagnostics for marketshare
Threshold: 0.75
0.5
0.0
−0.5
DFFITS
3
28
0
10
20
30
Observation



Model validation steps are repeated to compare a model with outliers to the one without outliers. As shown from the following two model validation steps, deleting the outliers (cases 3 and 28) improved the predictive performance of the regression model. The model without the outliers have lower MSPE value.

```
## Linear Regression
##
## 36 samples
##  4 predictor
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 35, 35, 35, 35, 35, 35, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   0.155712  0.6458495  0.1277597
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
## Linear Regression
##
## 34 samples
##  4 predictor
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 33, 33, 33, 33, 33, 33, ...
```

```
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.1351745  0.7473042  0.1100311
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
## Linear Regression
##
## 36 samples
##  4 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 32, 32, 32, 33, 32, 33, ...
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.1479296  0.6889859  0.1293582
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
## Linear Regression
##
## 34 samples
##  4 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 31, 32, 30, 31, 30, 30, ...
## Resampling results:
##
##   RMSE       Rsquared  MAE
##   0.1272179  0.819875  0.1075379
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
##
## Call:
## lm(formula = marketshare ~ price + discount + promotion + month,
##     data = market_share_without)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28419 -0.07231  0.01254  0.07290  0.22024
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.033892   0.316631   9.582 1.72e-10 ***
## price       -0.348938   0.135578  -2.574  0.01544 *
## discount     0.458864   0.046358   9.898 8.32e-11 ***
## promotion    0.110567   0.045616   2.424  0.02182 *
## month        0.020297   0.006999   2.900  0.00705 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.1275 on 29 degrees of freedom
## Multiple R-squared:  0.8074, Adjusted R-squared:  0.7808
## F-statistic: 30.38 on 4 and 29 DF,  p-value: 5.407e-10
```

Also, from the comparison of the two summary statistics tables below, it is shown that the standard error, which is a measure of the variability of the coefficient estimates that represents the average amount by which the estimate is expected to deviate from the true population parameter, is reduced for each coefficient estimate. If an investigator is to use the t test, this makes each of the predictor variables more important in the model because removing the outliers increased the t value for each coefficient estimate and reduced the probability of having a false positive finding. Therefore, if the final model without outliers meets the constant variance and normality assumptions, then it is said to be the best multiple linear regression subset model.

Summary Statistics Table for the Model without Outliers

Coefficient	Estimate	Std. Error	t value	Pr(>
(Intercept)	3.033892	0.316631	9.582	1.72e-10 ***
price	-0.348938	0.135578	-2.574	0.01544 *
discount	0.458864	0.046358	9.898	8.32e-11 ***
promotion	0.110567	0.045616	2.424	0.02182 *
month	0.020297	0.006999	2.900	0.00705 **

Residual standard error: 0.1275 on 29 degrees of freedom
Multiple R-squared: 0.8074, Adjusted R-squared: 0.7808
F-statistic: 30.38 on 4 and 29 DF, p-value: 5.407e-10

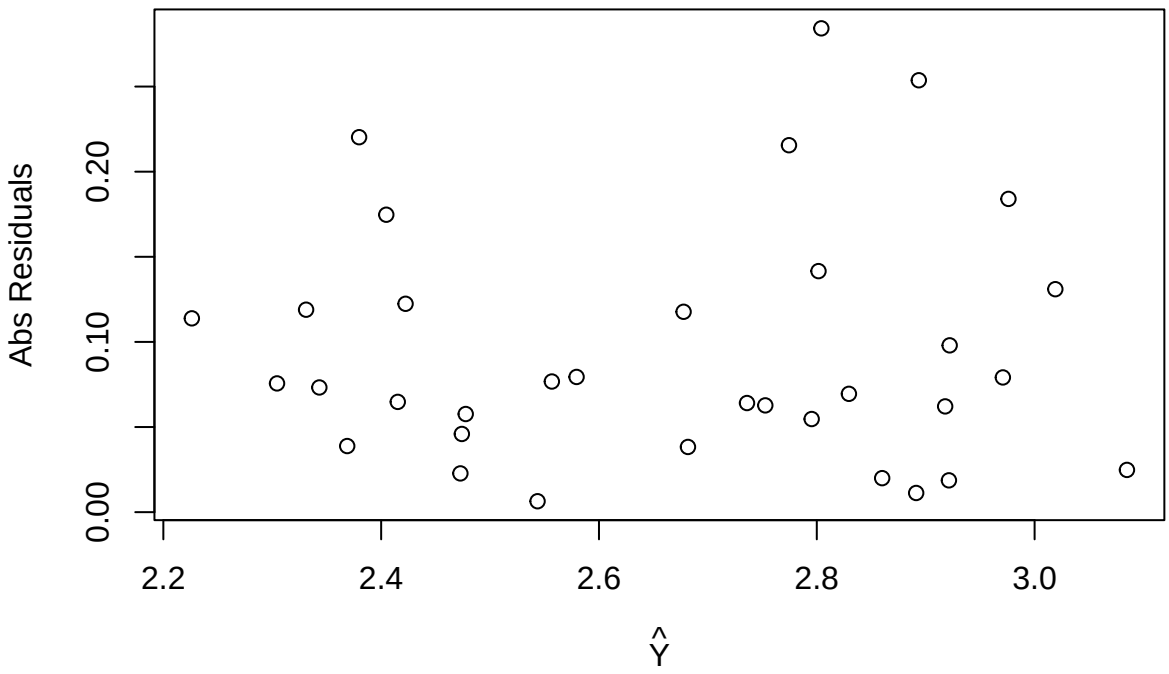
Summary Statistics Table for the Model with Outliers

Coefficient	Estimate	Std. Error	t value	Pr(>
(Intercept)	3.144380	0.358544	8.770	6.68e-10 ***
price	-0.366445	0.154408	-2.373	0.0240 *
discount	0.416139	0.051409	8.095	3.85e-09 ***
promotion	0.096772	0.052298	1.850	0.0738 .
month	0.011502	0.007493	1.535	0.1349

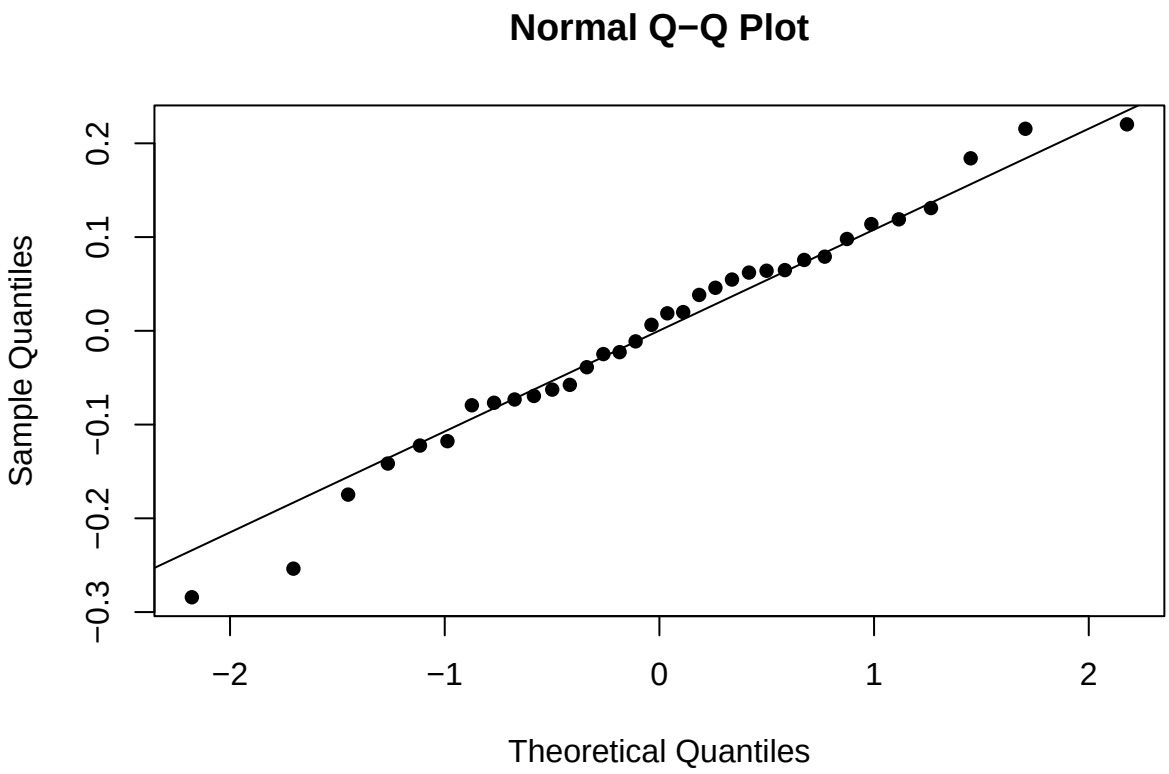
Residual standard error: 0.1467 on 31 degrees of freedom
Multiple R-squared: 0.7272, Adjusted R-squared: 0.692
F-statistic: 20.66 on 4 and 31 DF, p-value: 2.205e-08

The model without outliers will be considered the final model if the check for constant variance and normality assumptions is met.

3. Constant Variance There are no apparent issues with non-constant variance in the model. However, as mentioned previously, there is a lack of data points in the middle.



4. Normal quantile plot Also, the following normal quantile plot supports the normality of the final model's residuals:



IV. Results/Discussions

After conducting automatic variable selection methods, which are (in principle) based on partial F-tests, criterion-based statistics, residuals diagnostics, and model validation, β_5 and β_6 along with their corresponding x_i which are **gnrpoints** and **year** did not have significant linear association with the response variable **marketshare**, thus got dropped from the model. The final model is now given by:

$$Y_i = \beta_0 + \beta_1 * \text{ price} + \beta_2 * \text{ discount} + \beta_3 * \text{ promotion} + \beta_4 * \text{ month} + \varepsilon_i$$

, which supports the alternative hypothesis,mentioned previously, that not all β_i are zero. The estimated (fitted) multiple linear regression model is given by:

$$\hat{Y}_i = 3.034 - 0.349 \cdot \text{price} + 0.459 \cdot \text{dicount} + 0.111 \cdot \text{promotion} + 0.02 \cdot \text{month}$$

Partial residual plots, residual-versus-fitted plots, and measures of influence were investigated ; cases number 3 and 28 were identified as high influence points. No issues with linearity, constant variance, independence, or normality were identified. The lack of data in the center of all plots was confused with the possibility of having heteroscedasticity or non-constant variance of the residuals, though Brown-Forsythe test formally proved that error variance is constant.

To calculate the proportion of variation in **marketshare** response variable that is explained by each predictor variable, the coefficient of multiple determination R^2 for each predictor variable is calculated as the ratio of the regression sum of squares (SSR) to the total sum of squares (SSTO) , and it is given by the formula:

$$R^2 = \frac{SSR}{SSTO}$$

where: SSR is the regression sum of squares extracted from the Analysis of Variance (ANOVA) table below for each coefficient estimate, $SSTO$ is the total sum of squares that is calculated by adding the sum of squares for all coefficient estimates and sum of squares for the residuals shown in the ANOVA table below.

R^2 expresses the proportion of the total variance in the dependent variable that is explained by the independent variables in the model.

From the implications of the Analysis of Variance (ANOVA) table below , **price** variable explains 3.614% of the total variation in **marketshare** response variable,**discount** explains 65.066% , **promotion** explains 6.47% ,and lastly **month** explains 5.586%. Details are shown in the Methods section.

```
## Analysis of Variance Table
##
## Response: marketshare
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## price      1 0.08838 0.08838  5.4404  0.026823 *
## discount   1 1.59115 1.59115 97.9477 8.346e-11 ***
## promotion  1 0.15821 0.15821  9.7393  0.004059 **
## month      1 0.13661 0.13661  8.4094  0.007048 **
## Residuals 29 0.47110 0.01624
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

V. Conclusion

In conclusion, this comprehensive statistical analysis of the market share dataset offers valuable insights into the dynamics influencing the performance of a specific product over a 36-month period. The final multiple linear regression model, derived through meticulous variable selection and validation methods, includes significant predictors such as price, discount, promotion, and month. The findings reject the null hypothesis, confirming that at least one independent variable has a substantial linear relationship with marketshare. Notably, the analysis highlights the pivotal role of discount in driving market share, explaining a significant 65.066% of the variation. The model's robustness is reinforced by thorough diagnostics, including partial residual plots and influence assessments. Overall, this study equips company executives with a nuanced understanding of the factors shaping market share, enabling informed decisions on marketing strategies and product development to enhance their competitive position in the market.

VI. References

Hayes, A. , (2022, August 23).*Market Share: What It Is and the Formula for Calculating It.* Investopedia. https://www.investopedia.com/terms/m/marketshare.asp

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2014), Applied Linear Statistical Models, 5th ed., McGraw-Hill Irwin.

Valchanov, I. (2021, October 20). *Exploring the 5 OLS Assumptions for Linear Regression Analysis.* 365 Data Science. https://365datascience.com/tutorials/statistics-tutorials/ols-assumptions/.

VII. Appendix

```
# Load packages
library(tidyverse)
library(caret)
library(asbio)
library(olsrr)
library(xtable)
library(shiny)
```

```r
library(knitr)
library(DT)
require(scatterplot3d)
require(Hmisc)
require(rgl)
require(faraway)
library(car)
library(vroom)
library(leaps)
library(corrplot)
library(ggplot2)
library(r02pro)
library(tibble)
library(lmtest)
library(onewaytests)
library(readxl)

#load Data
market_share <- read_excel("C:/Users/amjad ijaq/OneDrive/Desktop/fall2023/STAT823/final project/market_s
    na = "NA")
view(market_share)
display_output <- function(dataset, out_type, filter_opt = 'none') {

  if (out_type == "pdf") {
    out_table <- DT::datatable(dataset, filter = filter_opt)
  } else {
    out_table <- knitr::kable(dataset)
  }

  out_table
}

# Function to calculate predicted sum of squares (PRESS)
PRESS <- function(linear.model) {
  #' calculate the predictive residuals
  pr <- residuals(linear.model)/(1-lm.influence(linear.model)$hat)
  #' calculate the PRESS
  PRESS <- sum(pr^2)

  return(PRESS)
}

head(market_share)

sum_tab <- summary(market_share)

# Select only character type variable
char_var <- market_share %>%
  select(where(is.character))
#find unique levels of the character vector

unique(market_share$month)
```

```r
# Convert character variable to unordered factor
market_share <- market_share %>%
  mutate_at(vars(month), as.factor)


# Convert `month` variable to ordered factor with specified levels
market_share$month <- factor(market_share$month, ordered = TRUE, levels = c( "Jan", "Feb", "Mar", "Apr"

# Convert month variable to numeric
market_share$month <- as.numeric(market_share$month)

#the total number of NA values in the market_share data set is zero.
NAs <- sum(is.na(market_share))

# Create a probability density plot of the `marketshare` response variable using ggplot2, no need for a

ggplot(data.frame(x = market_share$marketshare), aes(x)) +
  geom_density(fill = "blue", color = "black", alpha = 0.5) +
  labs(title = "Probability Density Plot", x = "marketshare")

m1 <- lm(marketshare~ price + discount+ promotion+ month+gnrpoints+ year, data=market_share)

# Create scatterplot matrix
pairs(market_share[, 2:8], main="Scatterplot Matrix")

as.data.frame (cor(market_share[, 2:8]))

for (i in 2:4){
  par(mfrow=c(1,2))
  stripchart(market_share[,i], main = names(market_share)[i],
                          vertical = T, method = "jitter")
  boxplot(market_share[,i], main = names(market_share)[i])
  par(mfrow=c(1,1))
}


for (i in 7:8){
  par(mfrow=c(1,2))
  stripchart(market_share[,i], main = names(market_share)[i],
                          vertical = T, method = "jitter")
  boxplot(market_share[,i], main = names(market_share)[i])
  par(mfrow=c(1,1))
}

#Added variable plots
prplot(m1,1)
title("Added Variable Plot for the price Covariate")


prplot(m1,2)
title("Added Variable Plot for the discount Covariate")

prplot(m1,3)
```

```
title("Added Variable Plot for the promotion Covariate")

prplot(m1,4)
title("Added Variable Plot for the month Covariate")


prplot(m1,5)
title("Added Variable Plot for the gnrpoints Covariate")


prplot(m1,6)
title("Added Variable Plot for the year Covariate")

vif(m1)

ma <- regsubsets(marketshare~ price+  discount+ promotion+ month+ gnrpoints+ year, data= market_share, n

(sma <- summary(ma))

sma$adjr2 # Adjusted R^2, biggest value is better predicting model
plot(2:7,sma$adjr2, xlab = "Number of Parameters", ylab = expression(R^2[adj]), main = "Visualizing R^2_

sma$bic # BIC smaller is better predicting model
plot(2:7, sma$bic, xlab = "Number of Parameters", ylab = expression(BIC), main = "Visualizing BIC plot"

sma$cp # Cp = p is better
plot(2:7, sma$cp, xlab = "Number of Parameters", ylab = expression(C[p]),main = "Visualizing C_p plot")

#best two selected model subsets
m2 <- lm(marketshare~ price+ discount+ promotion, data=market_share)
summary(m2)


m3 <- lm(marketshare~ price+ discount+ promotion+ month, data=market_share)
summary(m3)

#compare between models
# Extract AIC

extractAIC(m2)
extractAIC(m3) # the smallest AIC is model m3 with 5 parameters(4 predictor variables)

# Extract PRESS (Predicted Error Sum of Squares, it is a validation statistic for measuring the predict

PRESS(m2)
PRESS(m3) #has the smallest PRESS value

# Define the training method
tr <- trainControl(method="LOOCV")

# Train the first reduced model (m2)
mreduced.1 <- train(marketshare~ price+ discount+ promotion, data=market_share, method = "lm", trControl
print(mreduced.1)
```

```r
# Train the second reduced model (m3)
mreduced.2 <- train(marketshare~ price+ discount+ promotion+ month,data=market_share, method = "lm", tr
print(mreduced.2)

# Define the training method
set.seed(123)
tr <- trainControl(method = "cv", number = 10)


# Train the first reduced model (m2)
mreduced.1 <- train(marketshare~ price+ discount+ promotion, data=market_share, method = "lm", trContro
print(mreduced.1)

# Train the second reduced model (m3)
mreduced.2 <- train(marketshare~ price+ discount+ promotion+ month, data=market_share, method = "lm", t
print(mreduced.2)

# Histogram of residuals: checks for normality
hist(m3$residuals, col = "lightblue", main = "Histogram of Residuals") # the normality assumption is no

#Residuals' Probability Density Plot
ggplot(data.frame(x = residuals(m3)), aes(x)) +
  geom_density(fill = "blue", color = "black", alpha = 0.5) +
  labs(title = "Residuals' Probability Density Plot", x = "residuals(m3)")

plot(fitted(m3), residuals(m3), main = "Fitted Vs. Residuals Plot",
     xlab = "Fitted Values", ylab = "Residuals")

abline(h=0)
lines(lowess(residuals(m3)~fitted(m3)),lty=2,col="red")
legend("bottomleft",c("e=0","LOWESS"),col=c("black","red"),lty=c(1,2))

market_share$group <- fitted(m3) < median(fitted(m3))
market_share$E <- residuals(m3)
market_share$group <- factor(market_share$group,levels=c(TRUE,FALSE),labels=c("<Median",">=Median"))
boxplot(E~group,data=market_share)

market_share %>%
  group_by(group) %>%
  summarise(var=var(E))


#perform Broum-Forsythe test
bf.test(E~group,data=market_share)

#plot residuals against the response variable and against covariates of the chosen model(price, discoun

#residuals against the response variable plot shows that, for the most part, residuals are increasing a
plot(market_share$marketshare, residuals(m3), main = "Residuals Against marketshare")
abline(h=0)
lines(lowess(residuals(m3)~ market_share$marketshare),lty=2,col="red")
legend("bottomleft",c("e=0","LOWESS"),col=c("black","red"),lty=c(1,2))
```

```r
plot(market_share$price, residuals(m3), main = "Residuals Against Price Predictor Variable")
abline(h=0)
lines(lowess(residuals(m3)~ market_share$price),lty=2,col="red")
legend("bottomleft",c("e=0","LOWESS"),col=c("black","red"),lty=c(1,2))

plot(market_share$discount,residuals(m3), main = "Residuals Against Discount Predictor Variable")
abline(h=0)
lines(lowess(residuals(m3)~ market_share$discount),lty=2,col="red")
legend("bottomleft",c("e=0","LOWESS"),col=c("black","red"),lty=c(1,2))

plot(market_share$promotion, residuals(m3), main = "Residuals Against Promotion Predictor Variable" )
abline(h=0)
lines(lowess(residuals(m3)~ market_share$promotion),lty=2,col="red")
legend("bottomleft",c("e=0","LOWESS"),col=c("black","red"),lty=c(1,2))

plot(market_share$month, residuals(m3),main = "Residuals Against Month Predictor Variable")
abline(h=0)
lines(lowess(residuals(m3)~ market_share$month),lty=2,col="red")
legend("bottomleft",c("e=0","LOWESS"),col=c("black","red"),lty=c(1,2))

#price, discount, promotion, and month

market_share$price.i <- market_share$price > mean(market_share$price)
market_share$discount.i <- market_share$discount > mean(market_share$discount)
market_share$promotion.i <- market_share$promotion > mean(market_share$promotion)
market_share$ month.i <- market_share$ month > mean(market_share$ month)

#no apparent interaction is necessary to be added to the model.
interaction.plot( market_share$discount.i,market_share$price.i,market_share$marketshare )
interaction.plot(market_share$discount.i, market_share$promotion.i,market_share$marketshare )
interaction.plot(market_share$discount.i,market_share$month.i, market_share$marketshare )

# Plot studentized residuals against fitted values
plot(rstudent(m3) ~ fitted(m3), main = "Studentized Residuals vs. Fitted Values")
abline(h = 0)
lines(lowess(fitted(m3), rstudent(m3)), lty = 2, col = "red")
legend("bottomleft", c("e=0", "LOWESS"), col = c("black", "red"), lty = c(1, 2))

# Identify studentized residuals
identify(rstudent(m3) ~ fitted(m3))

# Plot deleted studentized residuals against fitted values
plot(rstandard(m3) ~ fitted(m3), main = "Deleted Studentized Residuals vs. Fitted Values")
abline(h = 0)
lines(lowess(fitted(m3), rstandard(m3)), lty = 2, col = "red")
legend("bottomleft", c("e=0", "LOWESS"), col = c("black", "red"), lty = c(1, 2))

# Plot all deleted studentized residuals
plot(rstandard(m3), main = "All Studentized Residuals")
abline(h = 0)
lines(lowess(rstandard(m3)), lty = 2, col = "red")

# Identify observations with absolute deleted studentized residuals greater than 3
```

```r
which(abs(rstandard(m3)) > 3)  # No unusual residuals

#The hat matrix is also useful after the model has been selected and fitted for determining whether an

#If h_new,new is well within the range of leverage values hii for the cases in the data set, no extrapo

#hat values are the diagonal elements of the hat matrix
#if hii is greater than 2p/n, then it is a high leverage point (potentially influence the model fit)
which(hatvalues(m3)> 2*5/36) # there are no high leverage data points, this indicates no outlying cases

# DFFITS measure considers the influence of the ith case on the fitted value Y.hat for this case,
plot(dffits(m3), main= "Plot of DFFITS")

which(abs(dffits(m3))> 2*sqrt(5/36) ) # a guideline for identifying influential cases: a case is influe


#dfbetas tell the standardized effect of deleting each individual observation on each coefficient. The


#check which DFBETAS are larger than 2/sqrt(36)
which(abs(dfbetas(m3)) > 2/sqrt(36))

abs(dfbetas(m3)[3,])

abs(dfbetas(m3)[28,])

#price, discount, promotion, and month

dfbetas <- as.data.frame(dfbetas(m3))

thresh <- 2/sqrt(36)

#plot DFBETAS for `price`  with threshold lines
plot(dfbetas$price, type='h')
abline(h = thresh, lty = 2)
abline(h = -thresh, lty = 2)
title(main = "Visualizing DFBETAS for `price` predictor variable")


#plot DFBETAS for `discount`  with threshold lines
plot(dfbetas$discount, type='h')
abline(h = thresh, lty = 2)
abline(h = -thresh, lty = 2)
title(main = "Visualizing DFBETAS for `discount` predictor variable")

#plot DFBETAS for ` promotion`  with threshold lines
plot(dfbetas$ promotion, type='h')
abline(h = thresh, lty = 2)
abline(h = -thresh, lty = 2)
title(main = "Visualizing DFBETAS for `promotion` predictor variable")

#plot DFBETAS for `month`  with threshold lines
plot(dfbetas$month, type='h')
```

```r
abline(h = thresh, lty = 2)
abline(h = -thresh, lty = 2)
title(main = "Visualizing DFBETAS for `month` predictor variable")

#Cook's distance measure considers the influence of the ith case on all n fitted values.

plot(cooks.distance(m3)) # Compare percentile F(p,n-p) to 10 or 20 percent. If the percentile value is
title(main = "Visualizing Cook's distances")



p <- pf(cooks.distance(m3),5,36-5)
q <- qf(p, 5, 36-5)
which(q>.1) # it appears that cases 3 and 28  does influence the regression fit,however their cooks dis
which(q>.2)# no cases with quantile higher than 20th percentile seem to have apparent influence on the


cat("Cook's distance for case 3 is:", cooks.distance(m3)[3], "\n")

cat("Cook's distance for case 28 is:", cooks.distance(m3)[28], "\n")

#Visualizing DFFITS, DFBETAS, and Cook's Distances

ols_plot_cooksd_bar(m3) # One way to visualize Cook's distance
ols_plot_dfbetas(m3) # Visualize influence on estimation of betas
ols_plot_dffits(m3) # Visualize influence on estimation of Y

# Another approach to getting influence statistics
m2i <- influence(m3) # Save influence stats
halfnorm(cooks.distance(m3)) # Another approach to visualize Cook's distance

#consider a model without cases 3 and 28
market_share_without <- market_share[-c(3,28),]

#using Leave_one_out cross validation
# Define the training method
tr <- trainControl(method="LOOCV")



mreduced <- train(marketshare~ price+ discount+ promotion+ month, data=market_share, method = "lm", trC
print(mreduced)

m_subset <- train(marketshare~ price+ discount+ promotion+ month, data=market_share_without, method = "

print(m_subset)

#using K-fold cross validation
# Define the training method
set.seed(123)
tr <- trainControl(method = "cv", number = 10)

mreduced <- train(marketshare~ price+ discount+ promotion+ month, data=market_share, method = "lm", trC
```

```r
print(mreduced)

m_subset <- train(marketshare~ price+ discount+ promotion+ month, data=market_share_without, method = "

print(m_subset)

#Final best model
m_final <- lm(marketshare~ price+ discount+ promotion+ month, data=market_share_without)
summary(m_final)

sum_fin <- summary(m_final)
sum_m3 <- summary(m3)

plot(abs(residuals(m_final))~predict(m_final), xlab = expression(hat(Y)), ylab = "Abs Residuals")

qqnorm(residuals(m_final),pch=16)
qqline(residuals(m_final))


(anova_m_final <- anova(m_final))
(SSTO <- sum(anova_m_final$`Sum Sq`))

SSR_price  <- anova_m_final$`Sum Sq`[1]
(R_sq_price <- SSR_price / SSTO)

SSR_discount <- anova_m_final$`Sum Sq`[2]
(R_sq_discount <- SSR_discount / SSTO)

SSR_promotion <- anova_m_final$`Sum Sq`[3]
(R_sq_promotion <- SSR_promotion / SSTO)

SSR_month <- anova_m_final$`Sum Sq`[4]
(R_sq_month <- SSR_month / SSTO)
```