Factors affecting sales at retail Stores

Akshat Jain
Kunal Agrawal
Rishabh Jain

Econometrics

Executive Summary

This project is a study to find the reasons affecting the sales at a multi-outlet retail brand. The dataset contains the item and outlet related information of more than 8000 products.

The company started as a single brick and mortar but now had managed to establish 4 retail outlets under the 'X' brand and by 2015 X had created its presence across the country with its retail stores in 24 states within the country.

Environmental changes in the industry, such as changing demands of the consumers, changing preferences, etc. led to a 30% decrease in their sales margin.

The store managers raised concerns that because of sales fluctuating, they are not able to assess the amount of orders that they should make, which would convert to sales. The questions that we will be exploring are What are the factors affecting sales performance? How sales are differentiating between different types of stores? On what factors or attributes should the company focus to increase sales?

Describing the Data

The data of this project comprises sales and 11 other aspect of 8523 products across the country.

- **Item_Identifier**: Item ID
- **Item_Weight**: Weight of the Item
- **Item_Fat_Content**: The Fat content in item categorized as Low Fat and Regular
- **Item_Visibility**: Visibility on a scale of 0-1
- **Item_Type**: Type of the Item categorized as Dairy, Soft drinks, Baking goods, breads, frozen foods etc.
- **Item_MRP**: Maximum Retail Price of the Item
- **Outlet_Identifier**: Outlet ID
- **Outlet_Establishment_Year**: V/S
- **Outlet_Size**: The size of outlet where the item is being sold classified into small, medium, high depending on size of the store
- **Outlet_Location_Type**: The location of outlet where the item is being sold
- **Outlet_Type**: Type of outlet where the item is being sold categorized as supermarket1, supermarket2, supermarket3, grocery store which is basically departmental store, specialty store, convenience store, and grocery store
- **Item_Outlet_Sales**: Sales data of the selected item ## Data Loading

Packages Used

```
library(ggplot2)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(tidyr)
library(car)

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

library(lmtest)

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':


library(sandwich)
library(knitr)
library(stargazer)
```

```r
sales <- read.csv2("Data_EatEasy_text.txt",header = T, strip.white=TRUE,sep = '\t',stringsAsFactors = T)
str(sales)
```

```
## 'data.frame':    8523 obs. of  12 variables:
##  $ Item_Identifier          : Factor w/ 1559 levels "DRA12","DRA24",..: 157 9 663 1122 1298 759 697 739 441 991 ...
##  $ Item_Weight              : Factor w/ 416 levels "10","10.1","10.195",..: 408 181 84 101 389 542 34 71 101 ...
##  $ Item_Fat_Content         : Factor w/ 5 levels "LF","low fat",..: 3 5 3 5 3 5 5 3 5 5 ...
##  $ Item_Visibility          : Factor w/ 7880 levels "0","0.003574698",..: 665 881 716 1 1 1 396 6772 708 5782 ...
##  $ Item_Type                : Factor w/ 16 levels "Baking Goods",..: 5 15 11 7 10 1 14 14 6 6 ...
##  $ Item_MRP                 : Factor w/ 5938 levels "100.0016","100.0042",..: 3850 4669 1159 2485 4823 4759 4940 267 5834 2678 ...
##  $ Outlet_Identifier        : Factor w/ 10 levels "OUT010","OUT013",..: 10 4 10 1 2 4 2 6 8 3 ...
##  $ Outlet_Establishment_Year: int  1999 2009 1999 1998 1987 2009 1987 1985 2002 2007 ...
##  $ Outlet_Size              : Factor w/ 3 levels "Large","Medium",..: 2 2 2 2 1 2 1 2 3 1 ...
##  $ Outlet_Location_Type     : Factor w/ 3 levels "Tier 1","Tier 2",..: 1 3 1 3 3 3 3 3 2 2 ...
##  $ Outlet_Type              : Factor w/ 4 levels "Grocery Store",..: 2 3 2 1 2 3 2 4 2 2 ...
##  $ Item_Outlet_Sales        : Factor w/ 3493 levels "1000.6974","1001.3632",..: 1977 2317 913 3186 3489 2772 1811 2126 72 2445 ...
```

```r
sum(is.na(sales))
```

```
## [1] 0
```

```r
sum(!complete.cases(sales))
```

```
## [1] 0
```

```r
unique(sales$Item_Fat_Content)
```

```
## [1] Low Fat Regular low fat LF      reg
## Levels: LF low fat Low Fat reg Regular
```

```r
unique(sales$Item_Type)
```

```
##  [1] Dairy                 Soft Drinks           Meat
##  [4] Fruits and Vegetables Household             Baking Goods
##  [7] Snack Foods           Frozen Foods          Breakfast
## [10] Health and Hygiene    Hard Drinks           Canned
## [13] Breads                Starchy Foods         Others
```

```
## [16] Seafood
## 16 Levels: Baking Goods Breads Breakfast Canned Dairy ... Starchy Foods
```

```r
unique(sales$Outlet_Identifier)
```

```
##  [1] OUT049 OUT018 OUT010 OUT013 OUT027 OUT045 OUT017 OUT046 OUT035 OU
T019
## 10 Levels: OUT010 OUT013 OUT017 OUT018 OUT019 OUT027 OUT035 OUT045 ... OU
T049
```

```r
unique(sales$Outlet_Size)
```

```
## [1] Medium Large  Small
## Levels: Large Medium Small
```

```r
unique(sales$Outlet_Location_Type)
```

```
## [1] Tier 1 Tier 3 Tier 2
## Levels: Tier 1 Tier 2 Tier 3
```

```r
unique(sales$Outlet_Identifier)
```

```
##  [1] OUT049 OUT018 OUT010 OUT013 OUT027 OUT045 OUT017 OUT046 OUT035 OU
T019
## 10 Levels: OUT010 OUT013 OUT017 OUT018 OUT019 OUT027 OUT035 OUT045 ... OU
T049
```

Data Cleaning

The Item_Fat_Content contains five different codes for the two categories, so let us encode them to ‘LF’ and ‘Reg’

```r
sum(sales[,3]=="low fat")
```

```
## [1] 112
```

```r
sales[which(sales[,3]=="low fat"),3]<- "LF"
sum(sales[,3]=="low fat")
```

```
## [1] 0
```

```r
sum(sales[,3]=="Low Fat")
```

```
## [1] 5089
```

```
sales[which(sales[,3]=="Low Fat"),3]<- "LF"
sum(sales[,3]=="Low Fat")

## [1] 0

sum(sales[,3]=="Regular")

## [1] 2889

sales[which(sales[,3]=="Regular"),3]<- "reg"
sum(sales[,3]=="Regular")

## [1] 0

sum(sales[,3]=="LF")

## [1] 5517

sum(sales[,3]=="reg")

## [1] 3006

levels(sales$Item_Fat_Content)

## [1] "LF"      "low fat" "Low Fat" "reg"     "Regular"

sales$Item_Fat_Content<-as.factor(sales$Item_Fat_Content)
sales$Item_Fat_Content<-droplevels(sales$Item_Fat_Content,"low fat")
levels(sales$Item_Fat_Content)

## [1] "LF"  "reg"

table(sales$Item_Fat_Content)

##
##   LF  reg
## 5517 3006
```

Transforming the numerical variables and creating new data frame with variables of interest

```
sales$Item_Weight<-as.numeric(as.character(sales$Item_Weight))
sales$Item_Visibility<-as.numeric(as.character(sales$Item_Visibility))
sales$Item_MRP<-as.numeric(as.character(sales$Item_MRP))

salesData<-sales
salesData<-salesData[,-c(1,3,5)]
salesData<-salesData[,-c(4:9)]
```

Dummy Variable Generation

When predictor variables are qualitative in nature and is a non-metric variable then we use dummy variable. We cannot ignore the qualitative variable in the model when these qualitative variable having limited categorical values has a good correlation with the dependent or response variable. But since the nature is qualitative, we cannot calculate mean, SD or variance statistic for comparisons. Hence, we use dummy variables. We have five dummy variables, which are as follows: - Item_Fat_Content_LF : indicates whether the fat content is ‘Low Fat’ or not - Outlet_Size_L: indicates whether the store size is Large(1) or not (0) - Outlet_Size_L: indicates whether the store size is Medium(1) or not (0) - Outlet_Location_Type_T1: indicates whether the store location is Tier 1(1)or not (0) - Outlet_Location_Type_T2: indicates whether the store location is Tier 2(1)or not (0) Outlet_Type_SM1: : indicates whether the store type is Supermarket 1(1)or not (0) Outlet_Type_SM2: : indicates whether the store type is Supermarket 2(1)or not (0) Outlet_Type_SM3: : indicates whether the store type is Supermarket 3(1)or not (0)

```
Item_Fat_Content_LF<- rep(0, length(sales$Item_Fat_Content))
Item_Fat_Content_LF[which(sales[,3]=="LF")]<- 1
salesData$Item_Fat_Content_LF<-Item_Fat_Content_LF


unique(sales$Outlet_Size)

## [1] Medium Large  Small
## Levels: Large Medium Small

Outlet_Size_L<- rep(0, length(sales$Outlet_Size))
Outlet_Size_M<- rep(0, length(sales$Outlet_Size))
Outlet_Size_L[which(sales[,9]=="Large")]<- 1
Outlet_Size_M[which(sales[,9]=="Medium")]<- 1
salesData$Outlet_Size_M<-Outlet_Size_M
salesData$Outlet_Size_L<-Outlet_Size_L




unique(sales$Outlet_Location_Type)

## [1] Tier 1 Tier 3 Tier 2
## Levels: Tier 1 Tier 2 Tier 3
```

```
Outlet_Location_Type_T1<- rep(0, length(sales$Outlet_Location_Type))
Outlet_Location_Type_T2<- rep(0, length(sales$Outlet_Location_Type))
Outlet_Location_Type_T1[which(sales[,10]=="Tier 1")]<- 1
Outlet_Location_Type_T2[which(sales[,10]=="Tier 2")]<- 1
salesData$Outlet_Location_Type_T1<-Outlet_Location_Type_T1
salesData$Outlet_Location_Type_T2<-Outlet_Location_Type_T2


unique(sales$Outlet_Type)
```

```
## [1] Supermarket Type1 Supermarket Type2 Grocery Store     Supermarket Type3
## 4 Levels: Grocery Store Supermarket Type1 ... Supermarket Type3
```

```
levels(sales$Outlet_Type)
```

```
## [1] "Grocery Store"     "Supermarket Type1" "Supermarket Type2"
## [4] "Supermarket Type3"
```

```
Outlet_Type_SM1<- rep(0, length(sales$Outlet_Type))
Outlet_Type_SM2<- rep(0, length(sales$Outlet_Type))
Outlet_Type_SM3<- rep(0, length(sales$Outlet_Type))
Outlet_Type_SM1[which(sales[,11]=="Supermarket Type1")]<- 1
Outlet_Type_SM2[which(sales[,11]=="Supermarket Type2")]<- 1
Outlet_Type_SM3[which(sales[,11]=="Supermarket Type3")]<- 1
salesData$Outlet_Outlet_Type_SM1<-Outlet_Type_SM1
salesData$Outlet_Outlet_Type_SM2<-Outlet_Type_SM2
salesData$Outlet_Outlet_Type_SM3<-Outlet_Type_SM3


unique(sales$Outlet_Identifier)
```

```
##  [1] OUT049 OUT018 OUT010 OUT013 OUT027 OUT045 OUT017 OUT046 OUT035 OU
T019
## 10 Levels: OUT010 OUT013 OUT017 OUT018 OUT019 OUT027 OUT035 OUT045 ... OU
T049
```

```
levels(sales$Outlet_Identifier)
```

```
##  [1] "OUT010" "OUT013" "OUT017" "OUT018" "OUT019" "OUT027" "OUT035" "OUT04
5"
##  [9] "OUT046" "OUT049"
```

```r
Outlet_Identifier_13<- rep(0, length(sales$Outlet_Identifier))
Outlet_Identifier_17<- rep(0, length(sales$Outlet_Identifier))
Outlet_Identifier_18<- rep(0, length(sales$Outlet_Identifier))
Outlet_Identifier_19<- rep(0, length(sales$Outlet_Identifier))
Outlet_Identifier_27<- rep(0, length(sales$Outlet_Identifier))
Outlet_Identifier_35<- rep(0, length(sales$Outlet_Identifier))
Outlet_Identifier_45<- rep(0, length(sales$Outlet_Identifier))
Outlet_Identifier_46<- rep(0, length(sales$Outlet_Identifier))
Outlet_Identifier_49<- rep(0, length(sales$Outlet_Identifier))
Outlet_Identifier_13[which(sales[,7]=="OUT013")]<- 1
Outlet_Identifier_17[which(sales[,7]=="OUT017")]<- 1
Outlet_Identifier_18[which(sales[,7]=="OUT018")]<- 1
Outlet_Identifier_19[which(sales[,7]=="OUT019")]<- 1
Outlet_Identifier_27[which(sales[,7]=="OUT027")]<- 1
Outlet_Identifier_35[which(sales[,7]=="OUT035")]<- 1
Outlet_Identifier_45[which(sales[,7]=="OUT045")]<- 1
Outlet_Identifier_46[which(sales[,7]=="OUT046")]<- 1
Outlet_Identifier_49[which(sales[,7]=="OUT049")]<- 1
salesData$Outlet_Identifier_13<-Outlet_Identifier_13
salesData$Outlet_Identifier_17<-Outlet_Identifier_17
salesData$Outlet_Identifier_18<-Outlet_Identifier_18
salesData$Outlet_Identifier_19<-Outlet_Identifier_19
salesData$Outlet_Identifier_27<-Outlet_Identifier_27
salesData$Outlet_Identifier_35<-Outlet_Identifier_35
salesData$Outlet_Identifier_45<-Outlet_Identifier_45
salesData$Outlet_Identifier_46<-Outlet_Identifier_46
salesData$Outlet_Identifier_49<-Outlet_Identifier_49
salesData$sales<-as.numeric(as.character(sales$Item_Outlet_Sales))

str(salesData)
```

```
## 'data.frame':    8523 obs. of  21 variables:
##  $ Item_Weight           : num  9.3 5.92 17.5 19.2 8.93 ...
##  $ Item_Visibility       : num  0.016 0.0193 0.0168 0 0 ...
##  $ Item_MRP              : num  249.8 48.3 141.6 182.1 53.9 ...
##  $ Item_Fat_Content_LF    : num  1 0 1 0 1 0 0 1 0 0 ...
##  $ Outlet_Size_M         : num  1 1 1 1 0 1 0 1 0 0 ...
##  $ Outlet_Size_L         : num  0 0 0 0 1 0 1 0 0 1 ...
##  $ Outlet_Location_Type_T1: num  1 0 1 0 0 0 0 0 0 0 ...
##  $ Outlet_Location_Type_T2: num  0 0 0 0 0 0 0 0 1 1 ...
##  $ Outlet_Outlet_Type_SM1 : num  1 0 1 0 1 0 1 0 1 1 ...
```

```
##  $ Outlet_Outlet_Type_SM2 : num  0 1 0 0 0 1 0 0 0 0 ...
##  $ Outlet_Outlet_Type_SM3 : num  0 0 0 0 0 0 0 1 0 0 ...
##  $ Outlet_Identifier_13   : num  0 0 0 0 1 0 1 0 0 0 ...
##  $ Outlet_Identifier_17   : num  0 0 0 0 0 0 0 0 0 1 ...
##  $ Outlet_Identifier_18   : num  0 1 0 0 0 1 0 0 0 0 ...
##  $ Outlet_Identifier_19   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Outlet_Identifier_27   : num  0 0 0 0 0 0 0 1 0 0 ...
##  $ Outlet_Identifier_35   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Outlet_Identifier_45   : num  0 0 0 0 0 0 0 0 1 0 ...
##  $ Outlet_Identifier_46   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Outlet_Identifier_49   : num  1 0 1 0 0 0 0 0 0 0 ...
##  $ sales                  : num  3735 443 2097 732 995 ...
```

Models of Regression Analysis

Let's run some tests to compare the sales. Let's fitting all parameters of salesData.

```
fitall <- lm(sales ~ ., salesData)
summary(fitall)
```

```
##
## Call:
## lm(formula = sales ~ ., data = salesData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4308.6  -672.5   -90.4   572.8  7915.9
##
## Coefficients: (5 not defined because of singularities)
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1737.9288    82.1498 -21.156  < 2e-16 ***
## Item_Weight               -0.7181     2.8958  -0.248  0.80417
## Item_Visibility         -290.2203   247.7663  -1.171  0.24149
## Item_MRP                  15.5605     0.1964  79.220  < 2e-16 ***
## Item_Fat_Content_LF      -51.6809    25.6204  -2.017  0.04371 *
## Outlet_Size_M            -42.2953    61.7117  -0.685  0.49313
## Outlet_Size_L            -66.4084    54.6623  -1.215  0.22444
## Outlet_Location_Type_T1  -22.3323    77.5669  -0.288  0.77342
## Outlet_Location_Type_T2 -189.9551   106.8130  -1.778  0.07538 .
## Outlet_Outlet_Type_SM1  2027.8447    87.7891  23.099  < 2e-16 ***
## Outlet_Outlet_Type_SM2  1631.3801    72.1919  22.598  < 2e-16 ***
```

```
## Outlet_Outlet_Type_SM3   3358.5562    72.1933  46.522  < 2e-16 ***
## Outlet_Identifier_13     -64.8345   109.3959  -0.593  0.55343
## Outlet_Identifier_17     171.6621    52.4187   3.275  0.00106 **
## Outlet_Identifier_18           NA         NA      NA       NA
## Outlet_Identifier_19           NA         NA      NA       NA
## Outlet_Identifier_27           NA         NA      NA       NA
## Outlet_Identifier_35     172.4849    63.0597   2.735  0.00625 **
## Outlet_Identifier_45           NA         NA      NA       NA
## Outlet_Identifier_46    -140.2116    80.9106  -1.733  0.08315 .
## Outlet_Identifier_49           NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1128 on 8507 degrees of freedom
## Multiple R-squared:  0.5636, Adjusted R-squared:  0.5628
## F-statistic: 732.3 on 15 and 8507 DF,  p-value: < 2.2e-16
```

Reading data, here we could see that the 56% of variation could be explained by this model. Also, we could observe that the 'Item MRP', 'Outlet_Outlet_Type_SM1', 'Outlet_Outlet_Type_SM2', 'Outlet_Outlet_Type_SM3' are significant at 99.9 level of confidence. While 'Outlet_Identifier_17, Outlet_Identifier_35', 'Item_Fat_Content_LF', and 'Outlet_Identifier_46, Outlet_Location_Type_T2' are significant at 99%, 95%, and 90% level of confidence respectively.

Omitted variable bias

Omitted-variable bias is observed in a model when we leave out one or more relevant variables. The bias results in the model attributing the effect of the missing variables to those that were included. To fix omitted variable bias, we need to keep adding control variables and to keep an eye on the adjusted R-square and F-stat.

```
fit1 <- lm(sales ~Item_MRP, salesData)
summary(fit1)
```

```
##
## Call:
## lm(formula = sales ~ Item_MRP, data = salesData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3871.2  -770.1   -64.0   696.4  9443.6
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.5751    37.6712  -0.307    0.759
## Item_MRP     15.5530     0.2444  63.635   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1405 on 8521 degrees of freedom
## Multiple R-squared:  0.3221, Adjusted R-squared:  0.3221
## F-statistic:  4049 on 1 and 8521 DF,  p-value: < 2.2e-16
```

```
fit2 <- lm(sales ~Item_MRP+ Outlet_Outlet_Type_SM1, salesData)
summary(fit2)
```

```
##
## Call:
## lm(formula = sales ~ Item_MRP + Outlet_Outlet_Type_SM1, data = salesData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3682.2  -801.9  -117.3   678.7  9693.8
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -258.4313    42.6633  -6.057 1.44e-09 ***
## Item_MRP                 15.5388     0.2424  64.106  < 2e-16 ***
## Outlet_Outlet_Type_SM1  380.3133    31.7382  11.983  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1393 on 8520 degrees of freedom
## Multiple R-squared:  0.3334, Adjusted R-squared:  0.3332
## F-statistic:  2130 on 2 and 8520 DF,  p-value: < 2.2e-16
```

```
fit3 <- lm(sales ~Item_MRP+ Outlet_Outlet_Type_SM1 + Outlet_Outlet_Type_SM2, salesData)
summary(fit3)
```

```
##
## Call:
## lm(formula = sales ~ Item_MRP + Outlet_Outlet_Type_SM1 + Outlet_Outlet_Type_SM2,
##     data = salesData)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3681.9  -804.8  -125.6   676.1  9718.2
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -282.2000    45.9860  -6.137  8.8e-10 ***
## Item_MRP                15.5365     0.2424  64.098  < 2e-16 ***
## Outlet_Outlet_Type_SM1  404.4171    36.1985  11.172  < 2e-16 ***
## Outlet_Outlet_Type_SM2   76.5142    55.2672   1.384    0.166
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1393 on 8519 degrees of freedom
## Multiple R-squared:  0.3335, Adjusted R-squared:  0.3333
## F-statistic:  1421 on 3 and 8519 DF,  p-value: < 2.2e-16
```

```
fit4 <- lm(sales ~Item_MRP+ Outlet_Outlet_Type_SM1 + Outlet_Outlet_Type_SM2+Outlet_Outlet_Type_SM3, salesData)
summary(fit4)
```

```
##
## Call:
## lm(formula = sales ~ Item_MRP + Outlet_Outlet_Type_SM1 + Outlet_Outlet_Type_SM2 +
##     Outlet_Outlet_Type_SM3, data = salesData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4298.5  -672.5   -76.7   568.0  7911.6
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -1843.3843    44.0244  -41.87   <2e-16 ***
## Item_MRP                15.5616     0.1965   79.20   <2e-16 ***
## Outlet_Outlet_Type_SM1  1962.0483    37.5102   52.31   <2e-16 ***
## Outlet_Outlet_Type_SM2  1634.1338    50.5296   32.34   <2e-16 ***
## Outlet_Outlet_Type_SM3  3361.8803    50.4270   66.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1130 on 8518 degrees of freedom
## Multiple R-squared:  0.562,  Adjusted R-squared:  0.5618
## F-statistic:  2733 on 4 and 8518 DF,  p-value: < 2.2e-16
```

```
fit5 <- lm(sales ~Item_MRP+Outlet_Outlet_Type_SM1 + Outlet_Outlet_Type_SM2+Outlet_Outlet_Type_SM3+Outlet_Identifier_17, salesData)
summary(fit5)
```

```
##
## Call:
## lm(formula = sales ~ Item_MRP + Outlet_Outlet_Type_SM1 + Outlet_Outlet_Type_SM2 +
##     Outlet_Outlet_Type_SM3 + Outlet_Identifier_17, data = salesData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4298.9  -671.7   -79.5   567.6  7911.3
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -1843.8271    44.0217 -41.884   <2e-16 ***
## Item_MRP                 15.5648     0.1965  79.214   <2e-16 ***
## Outlet_Outlet_Type_SM1  1951.6131    38.1098  51.210   <2e-16 ***
## Outlet_Outlet_Type_SM2  1634.1294    50.5255  32.343   <2e-16 ***
## Outlet_Outlet_Type_SM3  3361.8819    50.4229  66.674   <2e-16 ***
## Outlet_Identifier_17      62.8306    40.6473   1.546    0.122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1130 on 8517 degrees of freedom
## Multiple R-squared:  0.5622, Adjusted R-squared:  0.5619
## F-statistic:  2187 on 5 and 8517 DF,  p-value: < 2.2e-16
```

```
fit6 <- lm(sales ~Item_MRP+ Outlet_Outlet_Type_SM1 + Outlet_Outlet_Type_SM2+Outlet_Outlet_Type_SM3+Outlet_Identifier_17+Outlet_Identifier_35, salesData)
summary(fit6)
```

```
##
## Call:
## lm(formula = sales ~ Item_MRP + Outlet_Outlet_Type_SM1 + Outlet_Outlet_Type_SM2 +
##     Outlet_Outlet_Type_SM3 + Outlet_Identifier_17 + Outlet_Identifier_35,
##     data = salesData)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4298.2  -671.9   -88.3   573.1  7911.8
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1843.0353    43.9999 -41.887   <2e-16 ***
## Item_MRP                15.5591     0.1964  79.222   <2e-16 ***
## Outlet_Outlet_Type_SM1  1925.7785    38.9785  49.406   <2e-16 ***
## Outlet_Outlet_Type_SM2  1634.1372    50.4995  32.359   <2e-16 ***
## Outlet_Outlet_Type_SM3  3361.8791    50.3970  66.708   <2e-16 ***
## Outlet_Identifier_17      88.6603    41.4600   2.138   0.0325 *
## Outlet_Identifier_35     129.2369    41.3889   3.123   0.0018 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1129 on 8516 degrees of freedom
## Multiple R-squared:  0.5627, Adjusted R-squared:  0.5624
## F-statistic:  1826 on 6 and 8516 DF,  p-value: < 2.2e-16
```

```
fit7 <- lm(sales ~Item_MRP+ Item_Visibility +Outlet_Outlet_Type_SM1 + Outlet_Outlet_Type
_SM2+Outlet_Outlet_Type_SM3, salesData)
summary(fit7)
```

```
##
## Call:
## lm(formula = sales ~ Item_MRP + Item_Visibility + Outlet_Outlet_Type_SM1 +
##     Outlet_Outlet_Type_SM2 + Outlet_Outlet_Type_SM3, data = salesData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4277.8  -670.2   -78.7   567.6  7899.1
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1816.3015    51.1097 -35.537   <2e-16 ***
## Item_MRP                15.5616     0.1965  79.196   <2e-16 ***
## Item_Visibility       -258.2163   247.5381  -1.043    0.297
## Outlet_Outlet_Type_SM1  1950.6508    39.0689  49.928   <2e-16 ***
## Outlet_Outlet_Type_SM2  1622.8116    51.6819  31.400   <2e-16 ***
## Outlet_Outlet_Type_SM3  3349.9384    51.7099  64.783   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1130 on 8517 degrees of freedom
## Multiple R-squared:  0.5621, Adjusted R-squared:  0.5618
## F-statistic:  2187 on 5 and 8517 DF,  p-value: < 2.2e-16
```

```
fit8 <- lm(sales ~Item_MRP+ Item_Fat_Content_LF +Outlet_Outlet_Type_SM1 + Outlet_Outlet_Type_SM2+Outlet_Outlet_Type_SM3, salesData)
summary(fit8)
```

```
##
## Call:
## lm(formula = sales ~ Item_MRP + Item_Fat_Content_LF + Outlet_Outlet_Type_SM1 +
##     Outlet_Outlet_Type_SM2 + Outlet_Outlet_Type_SM3, data = salesData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4331.1  -671.1   -85.5   569.2  7929.7
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -1809.9551    47.1260 -38.407   <2e-16 ***
## Item_MRP                 15.5593     0.1965  79.196   <2e-16 ***
## Item_Fat_Content_LF     -50.8447    25.6040  -1.986   0.0471 *
## Outlet_Outlet_Type_SM1  1961.8549    37.5038  52.311   <2e-16 ***
## Outlet_Outlet_Type_SM2  1633.8029    50.5211  32.339   <2e-16 ***
## Outlet_Outlet_Type_SM3  3361.6803    50.4184  66.676   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1129 on 8517 degrees of freedom
## Multiple R-squared:  0.5622, Adjusted R-squared:  0.562
## F-statistic:  2188 on 5 and 8517 DF,  p-value: < 2.2e-16
```

```
fit9 <- lm(sales ~Item_MRP+ Outlet_Size_L +Outlet_Outlet_Type_SM1 + Outlet_Outlet_Type_SM2+Outlet_Outlet_Type_SM3, salesData)
summary(fit9)
```

```
##
## Call:
## lm(formula = sales ~ Item_MRP + Outlet_Size_L + Outlet_Outlet_Type_SM1 +
```

```
##     Outlet_Outlet_Type_SM2 + Outlet_Outlet_Type_SM3, data = salesData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4298.9  -672.8   -84.5   572.2  7911.3
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1830.8687    44.5326 -41.113   <2e-16 ***
## Item_MRP                  15.5644     0.1965  79.219   <2e-16 ***
## Outlet_Size_L            -56.1251    30.2701  -1.854   0.0638 .
## Outlet_Outlet_Type_SM1  1966.7430    37.5902  52.321   <2e-16 ***
## Outlet_Outlet_Type_SM2  1621.2258    50.9998  31.789   <2e-16 ***
## Outlet_Outlet_Type_SM3  3348.9776    50.8977  65.798   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1129 on 8517 degrees of freedom
## Multiple R-squared:  0.5622, Adjusted R-squared:  0.562
## F-statistic:  2188 on 5 and 8517 DF,  p-value: < 2.2e-16
```

```
fit10 <- lm(sales ~Item_MRP+ Outlet_Size_M +Outlet_Outlet_Type_SM1 + Outlet_Outlet_Type_SM2+Outlet_Outlet_Type_SM3, salesData)
summary(fit10)
```

```
##
## Call:
## lm(formula = sales ~ Item_MRP + Outlet_Size_M + Outlet_Outlet_Type_SM1 +
##     Outlet_Outlet_Type_SM2 + Outlet_Outlet_Type_SM3, data = salesData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4298.9  -670.5   -78.5   571.0  7911.3
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1846.9093    44.3185 -41.674   <2e-16 ***
## Item_MRP                  15.5644     0.1965  79.191   <2e-16 ***
## Outlet_Size_M             22.7601    32.8362   0.693    0.488
## Outlet_Outlet_Type_SM1  1959.4065    37.7045  51.967   <2e-16 ***
## Outlet_Outlet_Type_SM2  1614.5011    57.9280  27.871   <2e-16 ***
```

```
## Outlet_Outlet_Type_SM3  3342.2529    57.8348  57.790   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1130 on 8517 degrees of freedom
## Multiple R-squared:  0.5621, Adjusted R-squared:  0.5618
## F-statistic:  2186 on 5 and 8517 DF,  p-value: < 2.2e-16
```

```r
fit11 <- lm(sales ~Item_MRP+ Outlet_Location_Type_T1+Outlet_Outlet_Type_SM1 + Outlet_Outlet_Type_SM2+Outlet_Outlet_Type_SM3, salesData)
summary(fit11)
```

```
##
## Call:
## lm(formula = sales ~ Item_MRP + Outlet_Location_Type_T1 + Outlet_Outlet_Type_SM1 +
##     Outlet_Outlet_Type_SM2 + Outlet_Outlet_Type_SM3, data = salesData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4298.5  -672.5   -76.8   567.9  7911.6
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1843.3051    46.2644 -39.843   <2e-16 ***
## Item_MRP                  15.5616     0.1965  79.191   <2e-16 ***
## Outlet_Location_Type_T1    -0.1618    29.0686  -0.006    0.996
## Outlet_Outlet_Type_SM1   1962.0234    37.7786  51.935   <2e-16 ***
## Outlet_Outlet_Type_SM2   1634.0549    52.4821  31.135   <2e-16 ***
## Outlet_Outlet_Type_SM3   3361.8014    52.3835  64.177   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1130 on 8517 degrees of freedom
## Multiple R-squared:  0.562,  Adjusted R-squared:  0.5618
## F-statistic:  2186 on 5 and 8517 DF,  p-value: < 2.2e-16
```

```r
fit12 <- lm(sales ~Item_MRP+ Outlet_Location_Type_T1 +Outlet_Outlet_Type_SM1 + Outlet_Outlet_Type_SM2+Outlet_Outlet_Type_SM3, salesData)
summary(fit12)
```

```
##
## Call:
```

```
## lm(formula = sales ~ Item_MRP + Outlet_Location_Type_T1 + Outlet_Outlet_Type_SM1 +
##     Outlet_Outlet_Type_SM2 + Outlet_Outlet_Type_SM3, data = salesData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4298.5  -672.5   -76.8   567.9  7911.6
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1843.3051    46.2644 -39.843   <2e-16 ***
## Item_MRP                  15.5616     0.1965  79.191   <2e-16 ***
## Outlet_Location_Type_T1    -0.1618    29.0686  -0.006    0.996
## Outlet_Outlet_Type_SM1   1962.0234    37.7786  51.935   <2e-16 ***
## Outlet_Outlet_Type_SM2   1634.0549    52.4821  31.135   <2e-16 ***
## Outlet_Outlet_Type_SM3   3361.8014    52.3835  64.177   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1130 on 8517 degrees of freedom
## Multiple R-squared:  0.562,  Adjusted R-squared:  0.5618
## F-statistic:  2186 on 5 and 8517 DF,  p-value: < 2.2e-16
```

```
fit13 <- lm(sales ~Item_MRP+ Item_Weight +Outlet_Outlet_Type_SM1 + Outlet_Outlet_Type_SM2+Outlet_Outlet_Type_SM3, salesData)
summary(fit13)
```

```
##
## Call:
## lm(formula = sales ~ Item_MRP + Item_Weight + Outlet_Outlet_Type_SM1 +
##     Outlet_Outlet_Type_SM2 + Outlet_Outlet_Type_SM3, data = salesData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4298.7  -672.0   -78.1   567.7  7911.5
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -1835.9859    57.2816 -32.052   <2e-16 ***
## Item_MRP                 15.5626     0.1966  79.172   <2e-16 ***
## Item_Weight              -0.5848     2.8965  -0.202     0.84
## Outlet_Outlet_Type_SM1  1962.0253    37.5125  52.303   <2e-16 ***
```

```
## Outlet_Outlet_Type_SM2  1634.1243    50.5325  32.338   <2e-16 ***
## Outlet_Outlet_Type_SM3  3361.8649    50.4299  66.664   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1130 on 8517 degrees of freedom
## Multiple R-squared:  0.562,  Adjusted R-squared:  0.5618
## F-statistic:  2186 on 5 and 8517 DF,  p-value: < 2.2e-16
```

From model 3 to model 4, we can see the coefficient increases by 0.23. It is a considerable change compared to the previous models, whose coefficients change were less. Also, from model 6 onward, the models have very small changes.

Final Model Examination

Now we fit the model

sales ~Item_MRP+ Outlet_Outlet_Type_SM1 +Outlet_Outlet_Type_SM2+Outlet_Outlet_Type_SM3+Outlet_Identifier_17+Outlet_Identifier_35

as final examination model.

```
fitfin <- lm(sales ~Item_MRP+ Outlet_Location_Type_T2 +Outlet_Outlet_Type_SM1 + Outlet_Outlet_Type_SM2+Outlet_Outlet_Type_SM3+ Outlet_Identifier_17+Outlet_Identifier_35, salesData)
summary(fitfin)
```

```
##
## Call:
## lm(formula = sales ~ Item_MRP + Outlet_Location_Type_T2 + Outlet_Outlet_Type_SM1 +
##     Outlet_Outlet_Type_SM2 + Outlet_Outlet_Type_SM3 + Outlet_Identifier_17 +
##     Outlet_Identifier_35, data = salesData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4298.1  -672.6   -81.1   569.0  7911.9
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1842.9332    43.9850 -41.899  < 2e-16 ***
## Item_MRP                  15.5584     0.1963  79.245  < 2e-16 ***
## Outlet_Location_Type_T2  -111.1799    42.7446  -2.601  0.00931 **
## Outlet_Outlet_Type_SM1   1953.5367    40.4003  48.354  < 2e-16 ***
```

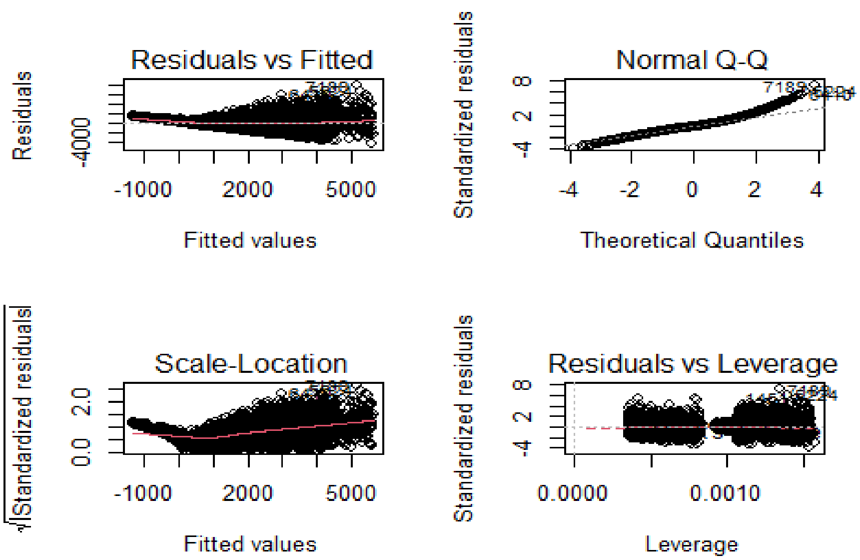
```
## Outlet_Outlet_Type_SM2   1634.1382    50.4825  32.370  < 2e-16 ***
## Outlet_Outlet_Type_SM3   3361.8787    50.3800  66.730  < 2e-16 ***
## Outlet_Identifier_17      172.0812    52.4061   3.284  0.00103 **
## Outlet_Identifier_35      212.6606    52.3505   4.062  4.9e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1129 on 8515 degrees of freedom
## Multiple R-squared:  0.563,  Adjusted R-squared:  0.5627
## F-statistic:  1567 on 7 and 8515 DF,  p-value: < 2.2e-16
```

Residual Analysis

The resulting final model examination is dependent on the ‘Item_MRP’, but also ‘SM1’, ‘SM2’, ‘SM3’,‘Outlet_Location_Type_T2’, ’ Outlet 17’ and ‘Outlet35’. All have significant p-values and the R-squared is pretty good to (0.56)

Now let’s look at the Residuals vs Fitted

```
par(mfrow=c(2,2))
plot(fitfin)
```

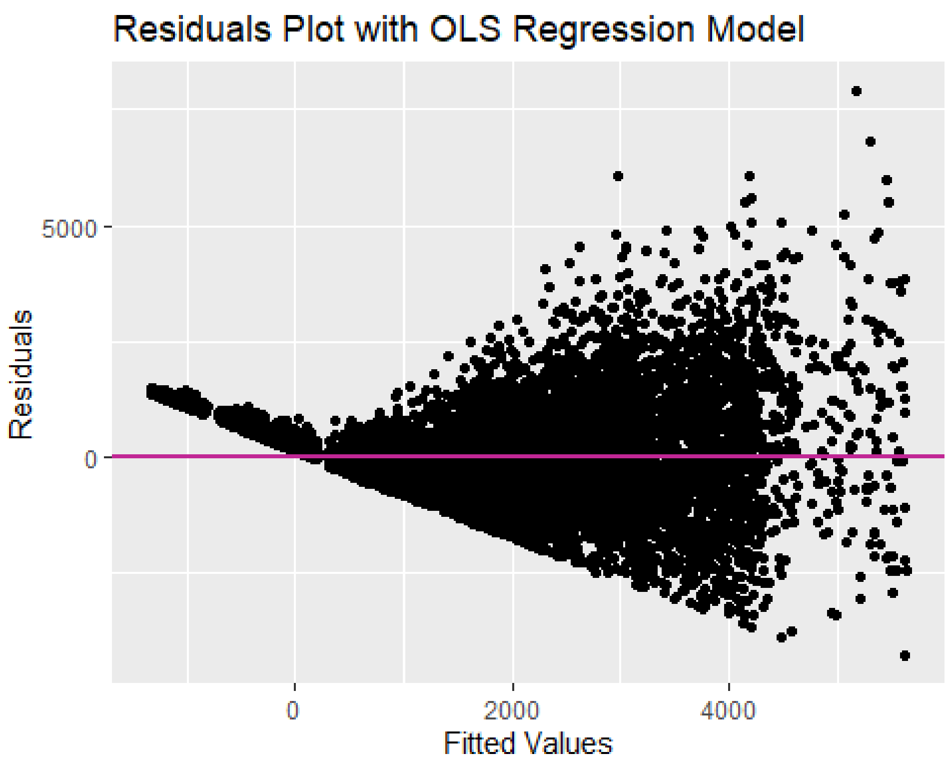


Heteroskedasticity Analysis

Heteroskedasticity occurs when the variance for all observations in a data set are not the same. In the presence of heteroskedasticity, there are two main consequences on the least squares estimators:

The least squares estimator is still a linear and unbiased estimator, but it is no longer best. That is, there is another estimator with a smaller variance. The standard errors computed for the least squares estimators are incorrect. This can affect confidence intervals and hypothesis testing that use those standard errors, which could lead to misleading conclusions.

```
ggplot(fitfin) + geom_point(aes(x=.fitted, y=.resid))+ geom_hline(yintercept=0, color = "#C12795", size = 1)+ ggtitle("Residuals Plot with OLS Regression Model") +xlab("Fitted Values") + ylab("Residuals")
```



Observing graph, we can see that the Heteroskedasticity is present but let us try a numerical methd to confirm the heteroscedasticity.

The Breusch-Pagan Test

```
bptest(fitfin)
```

```
##
## studentized Breusch-Pagan test
##
## data:  fitfin
## BP = 1203.4, df = 7, p-value < 2.2e-16
```

While it doesn't give us the critical value to compare the test statistic, but we have the p-value to determine whether or not you should reject the null. If the p-value is less than the level of significance (in this case if the p-value is less than α=0.05), then you reject the null hypothesis. Since 2.2e-16< 0.05, we can reject the null hypothesis and conclude that model have heteroscedasticity.

Resolving Heteroskedasticity - Adjusting Robust standard errors

```
coeftest(fitfin, vcov = vcovHC(fitfin, "HC1"))

##
## t test of coefficients:
##
##                           Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)            -1842.93320    40.12541 -45.9293 < 2.2e-16 ***
## Item_MRP                  15.55841     0.22698  68.5451 < 2.2e-16 ***
## Outlet_Location_Type_T2  -111.17985    41.52290  -2.6776 0.0074304 **
## Outlet_Outlet_Type_SM1   1953.53671    32.84241  59.4821 < 2.2e-16 ***
## Outlet_Outlet_Type_SM2   1634.13824    42.06774  38.8454 < 2.2e-16 ***
## Outlet_Outlet_Type_SM3   3361.87875    57.03398  58.9452 < 2.2e-16 ***
## Outlet_Identifier_17      172.08124    51.57241   3.3367 0.0008514 ***
## Outlet_Identifier_35      212.66055    51.18070   4.1551 3.283e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can clearly see that standard errors for coefficients are larger than robust standard errors. Which in turn will result into narrower confidence interval for coefficients. However, it doesn't resolve the issue of least squares estimators no longer being best.

Resolving Heteroskedasticity - Generalized Least Squares

```
salesData$resi <- fitfin$residuals


varfunc.ols <- lm(log(resi^2) ~ Item_MRP+ Outlet_Identifier_17+Outlet_Identifier_35+ Outlet_Location_Type_T2 +Outlet_Outlet_Type_SM1 + Outlet_Outlet_Type_SM2+Outlet_Outlet_Type_SM3, data = salesData)
```

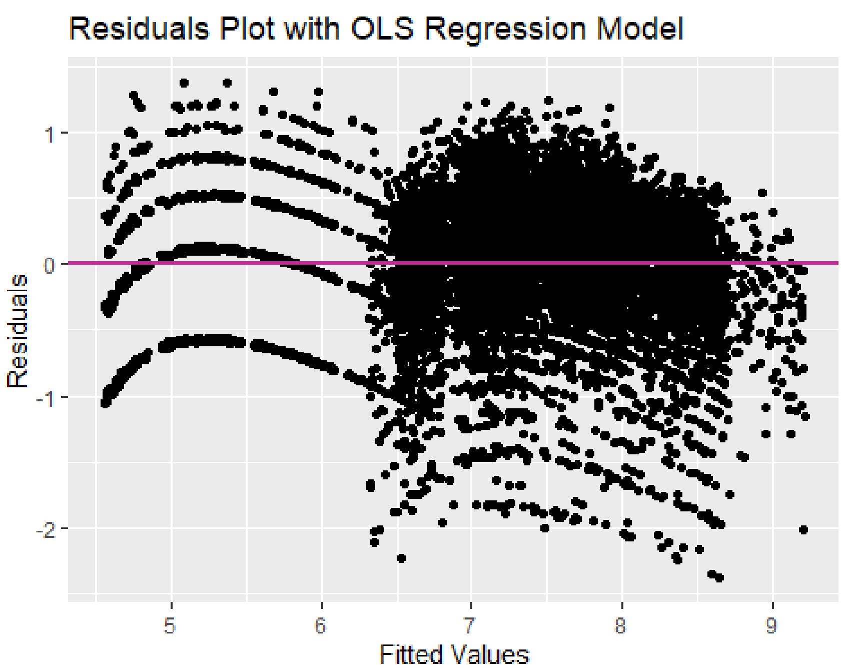
```
salesData$varfunc <- exp(varfunc.ols$fitted.values)

salesData.gls <- lm(log(sales) ~Item_MRP+ Outlet_Identifier_17+Outlet_Identifier_35+ Outlet_Location_Type_T2 +Outlet_Outlet_Type_SM1 + Outlet_Outlet_Type_SM2+Outlet_Outlet_Type_SM3,weights = 1/sqrt(varfunc), data = salesData)

bptest(salesData.gls )
```

```
##
## studentized Breusch-Pagan test
##
## data:  salesData.gls
## BP = 60.191, df = 7, p-value = 1.383e-10
```

```
ggplot(salesData.gls) + geom_point(aes(x=.fitted, y=.resid))+geom_hline(yintercept=0, color = "#C12795", size = 1)+  ggtitle("Residuals Plot with OLS Regression Model") +
  xlab("Fitted Values") + ylab("Residuals")
```



As we can see the homoskedasticity is improved but heteroskedasticity is not completely removed from model. The model shall require further transformations.

Conclusion

1. The factors affecting sales performance are MRP, Outlet Type, Outlet Location, and the outlet.
2. The sales difference between different types of stores is as follows:

a. For Supermarket 1 the sales are 1925.78 more compared to that of grocery store.

b. If the store is Supermarket 2 the sales are 1634.14 are more compared to that of grocery store.

c. For Supermarket 3 the sales are 3361.88 more compared to that of grocery store.

d. The sales at Outlet 17 and 35 exceeds the sales at output10 by 88.66 and 129.24 3

e. For every unit rise in MRP the sales is increased by 15.5

3. The company should focus on opening the supermarket 3 as they generate higher sales compared to other.
4. The company should target Tier 2 cities as the stores in these cities generated more sales compared to stores in other cities.
5. The company should target selling premium products as they generate higher sales.