MBA 6693 Business Analytics

Assignment 1: Regression Models

Name: Vinayak B. Menon

11/07/2020

Objective:

This report aims to study the significance of location when considering the number of sales of child car seats, based on the *Carseats* data in R. Our analysis will focus on where the most car seats number are sold based on whether the store location is in the US or abroad. We will also consider a classification based on urban or rural locations. By the end of this study, we aim to establish whether or not there is any significance in store location, and if so, where it significance lies most, with respect to the number of child car seats.

We will first explore the *Carseats* dataset provided, which would include cleaning the data and establishing any patterns through graphical tools. We then proceed to create a multivariate regression model using some of the factors and study the efficiency of the model. To reduce the complexity and strengthen the model, we will perform a backward step regression. Finally we compare all considered models in terms of their explanatory power and fit, and check how much the location variables influence the number of car seats that are sold at the store.

Data Exploration:

The *Carseats* data consists of 400 observations from different stores, with coloumns representing the unit sales in thousands at each location *Sales*, the local advertising budget for the company at each location in thousands of dollars *Advertising*, price charged by the store for the car seat at each location *Price*, quality of the shelving location for the car seats at each store location *ShelvLoc*, location of store in terms of urban and rural *Urban* and the location of the store in terms of whether it is situated in the US *US*. Of these factors, *Urban*, *US* and *ShelvLoc* are categorical while the rest are numeric in nature.

We can first convert the boolean results for *Urban*, *ShelvLoc* and *US* to digits. Since *ShelvLoc* consists of three indicators, Good, Bad and Medium, we will convert them to 1,-1 and 0 respectively. *Urban* and *US* can be converted to binary forms since they present boolean results.

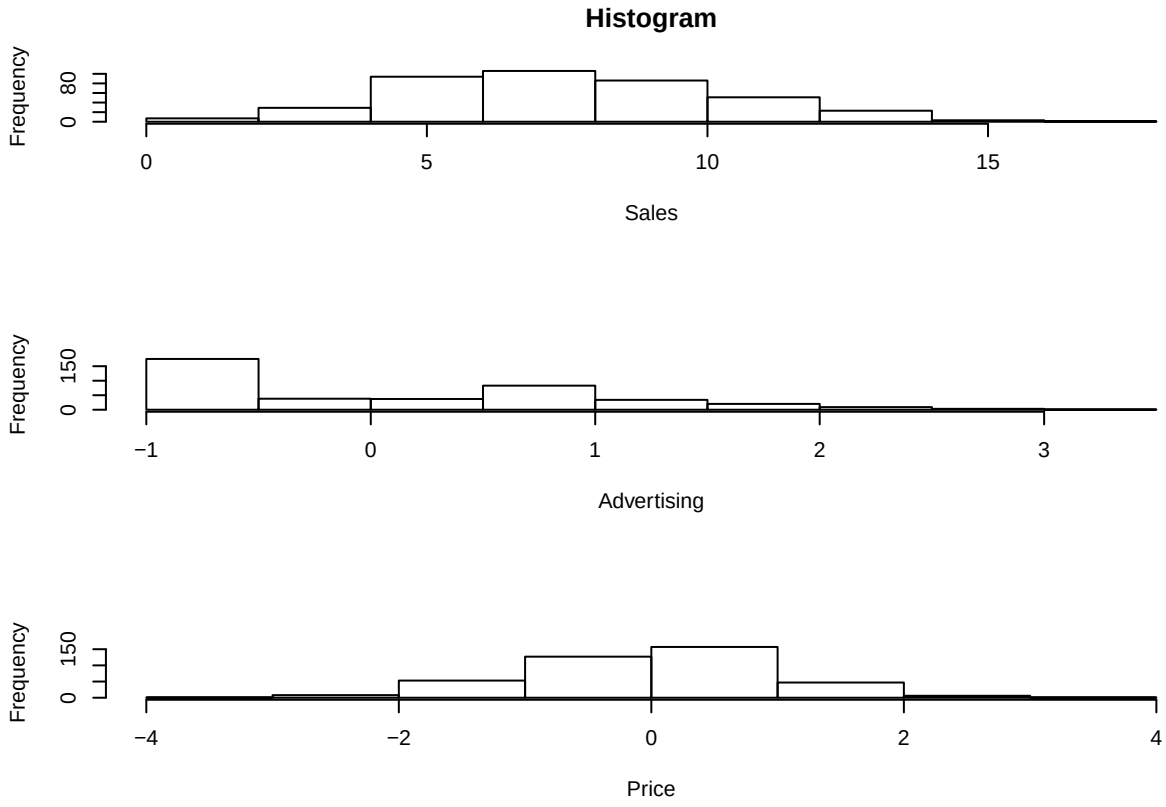
```
#converting Urban to numeric
carseats$Urban[carseats$Urban =="Yes"] <- 1
carseats$Urban[carseats$Urban =="No"] <- 0
carseats$Urban <- as.numeric(carseats$Urban)
#converting US to numeric
carseats$US[carseats$US =="Yes"] <- 1
carseats$US[carseats$US =="No"] <- 0
carseats$US <- as.numeric(carseats$US)
#converting ShelvLoc to numeric
carseats$ShelveLoc[carseats$ShelveLoc=="Good"] <-1
carseats$ShelveLoc[carseats$ShelveLoc=="Bad"] <--1
carseats$ShelveLoc[carseats$ShelveLoc=="Medium"] <-0
carseats$ShelveLoc <- as.numeric(carseats$ShelveLoc)
```

Now that we have transformed the categorical inputs to a usable format, the next step would be to normalize the numeric independent variables *Price* and *Advertising*. This is to avoid disparate coefficients from arising in the regression process. We leave the *Sales* information as it is since the intercept can adjust accordingly.

```
for (i in c(2,3)) {
  carseats[,i] <- (carseats[,i]-mean(carseats[,i]))/sd(carseats[,i])
}
```

We now plot the histograms for each of the numeric independent factors to see if there are any widely deviating values:

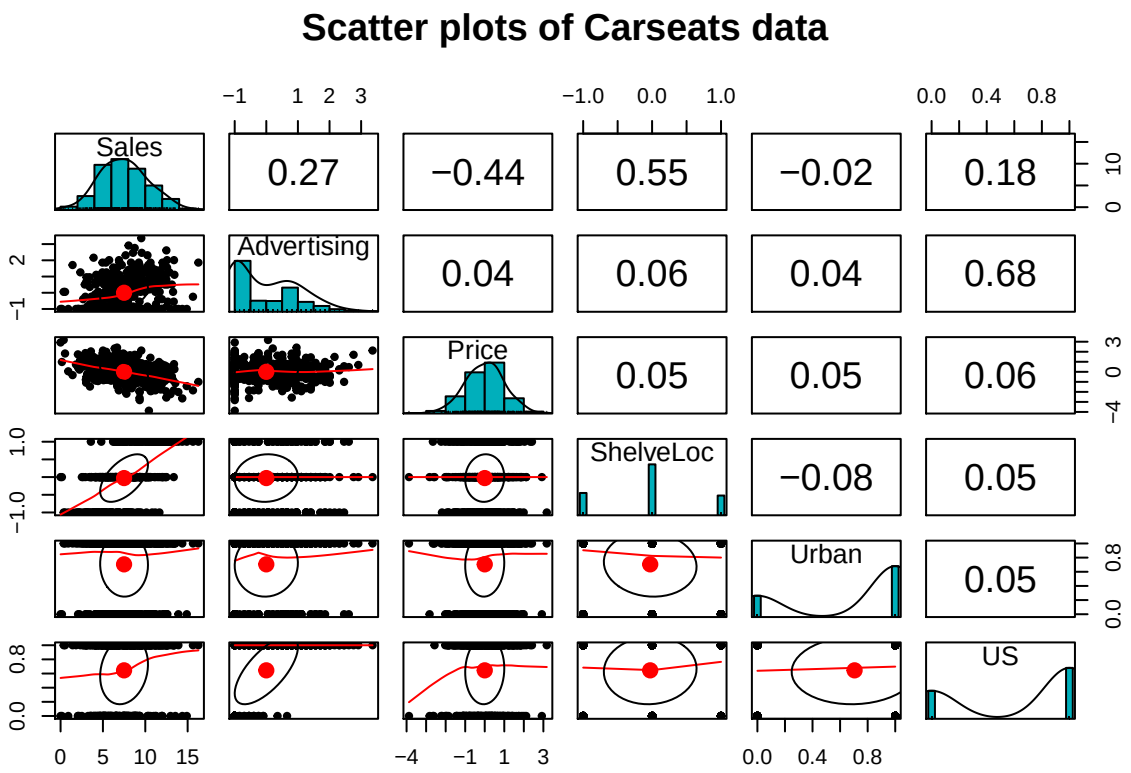
```
#histogram output
par(mfrow=c(3,1))
hist(carseats$Sales,xlab = "Sales",main = "Histogram")
hist(carseats$Advertising,xlab = "Advertising",main = "")
hist(carseats$Price,xlab = "Price",main = "")
```



The values are comparable. Now we create a scatterplot matrix embedded with correlation plots and values. This is possible using the pairs.panel() function using the *psych* package.

```
#paired plots
pairs.panels(carseats,
             method = "pearson", # correlation method
             hist.col = "#00AFBB",
```

```
            density = TRUE,  # show density plots
            main="Scatter plots of Carseats data",)
```



Most of the correlations are quite moderate or low and we can assume that these variables do not have much dependence. Note that there is a visible downward trend in the *scatterplot* between Sales and *Price*.

Modelling:

As mentioned initially, we will first create a model that encompasses all the dependent features to predict *Sales*, after which we will remove insignficant variables:

```
#Model 1,
model <- lm(Sales~ Advertising+Price+Urban+US+ShelveLoc,data = carseats)
summary(model)
```

```
##
## Call:
## lm(formula = Sales ~ Advertising + Price + Urban + US + ShelveLoc,
##     data = carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5263 -1.2049  0.0855  1.1516  4.3766
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.387489   0.232443  31.782  < 2e-16 ***
## Advertising  0.725883   0.123464   5.879 8.79e-09 ***
## Price       -1.375899   0.090313 -15.235  < 2e-16 ***
## Urban        0.244054   0.198068   1.232    0.219
## US           0.003553   0.257788   0.014    0.989
## ShelveLoc    2.382298   0.134479  17.715  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.797 on 394 degrees of freedom
## Multiple R-squared:  0.6003, Adjusted R-squared:  0.5952
## F-statistic: 118.3 on 5 and 394 DF,  p-value: < 2.2e-16
```

This initial model has a moderately good R^2 value (60.03%) suggesting that it manages to explain a considerable amount of the variance. Yet we can see that there is room for improvement as two of the variables do not have any significant effect on the sales of the child car seats, based on their high p-values. We move on to seeing if removing any one can bring a difference to the model.

```
#Proto model 1
model_11 <- lm(Sales~ Advertising+Price+US+ShelveLoc,data = carseats)
summary(model_11)
```

```
##
## Call:
## lm(formula = Sales ~ Advertising + Price + US + ShelveLoc, data = carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4724 -1.2069  0.0481  1.1887  4.4382
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.55419    0.18914  39.940  < 2e-16 ***
## Advertising  0.72830    0.12353   5.896 8.01e-09 ***
## Price       -1.37052    0.09027 -15.183  < 2e-16 ***
## US           0.01130    0.25788   0.044    0.965
## ShelveLoc    2.36894    0.13413  17.662  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.798 on 395 degrees of freedom
## Multiple R-squared:  0.5988, Adjusted R-squared:  0.5947
## F-statistic: 147.4 on 4 and 395 DF,  p-value: < 2.2e-16
```

Removing $Urban$ has slightly lowered the R^2 to 59.88%, with a marginal decrease in the residual standard error. The US variable provides no contribution to the model with a low coefficient and the significance is still not considerable. We check below if removing US while retaining $Urban$ provides any different results.

```
#Proto model 2
model_12 <- lm(Sales~ Advertising+Price+Urban+ShelveLoc,data = carseats)
summary(model_12)
```

```
##
## Call:
## lm(formula = Sales ~ Advertising + Price + Urban + ShelveLoc,
##     data = carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5270 -1.2044  0.0853  1.1533  4.3773
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.38973    0.16559  44.627  < 2e-16 ***
## Advertising  0.72704    0.09015   8.065 8.84e-15 ***
## Price       -1.37585    0.09014 -15.263  < 2e-16 ***
## Urban        0.24412    0.19776   1.234    0.218
## ShelveLoc    2.38232    0.13430  17.739  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.794 on 395 degrees of freedom
## Multiple R-squared:  0.6003, Adjusted R-squared:  0.5963
## F-statistic: 148.3 on 4 and 395 DF,  p-value: < 2.2e-16
```

As with the previous model, removing *US* does not show any significant changes. We thus proceed to excluding both of them from our upcoming models. Since the remaining variables provide significant contribution in explaining the *Sales* variable, we can retain them to create our first possible model to verify.

```
#Model 1
model_1 <- lm(Sales~ Advertising+Price+ShelveLoc,data = carseats)
summary(model_1)
```

```
##
## Call:
## lm(formula = Sales ~ Advertising + Price + ShelveLoc, data = carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4746 -1.2067  0.0527  1.1920  4.4405
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.56147    0.08986  84.149  < 2e-16 ***
## Advertising  0.73199    0.09012   8.123 5.86e-15 ***
## Price       -1.37038    0.09009 -15.211  < 2e-16 ***
## ShelveLoc    2.36900    0.13395  17.685  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.796 on 396 degrees of freedom
## Multiple R-squared:  0.5988, Adjusted R-squared:  0.5957
## F-statistic:   197 on 3 and 396 DF,  p-value: < 2.2e-16
```

The results are the same as our initial model. What we can infer from Model 1 is the relationships between *Sales* and the considered dependent variables. We see that *ShelvLoc* seems to have a significant effect

on sales of child car seats at a given store with a positive relationship. Specifically, a *Good* shelving quality can increase the number of sales by 2369 units whereas a *Bad* can reduce the number of sales by the same amount. Since *Medium* is associated with 0, the contribution would be included in the intercept.

The second biggest contribution comes from *Price*, which has a negative relationship with the no. of sales. This was visually noticeable in the scatterplot as previouslt mentioned. A unit increase in the price (i.e a 1000 dollar increase in price) of the car seat can lead to a decrease in sales by 1370.

Finally, *Advertising* shows a positive relationshio with the number of units of carseats sold. It shows that a 1000$ increase in advertising expenditure can lead to 731 units increase in sales.

We will consider the second model to be one with *Advertising* excluded. We exclude this factor since the contribution is the lowest, while the significance, although quite strong, is less as compared to the remaining two dependent variables. We will then observe if there is any improvement or negative effects on the model.

```
#Model 2
model_2 <- lm(Sales~Price+ShelveLoc,data = carseats)
summary(model_2)
```

```
##
## Call:
## lm(formula = Sales ~ Price + ShelveLoc, data = carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0900 -1.3005 -0.0662  1.3884  4.9322
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.56310    0.09693   78.02   <2e-16 ***
## Price       -1.33981    0.09710  -13.80   <2e-16 ***
## ShelveLoc    2.42820    0.14429   16.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.937 on 397 degrees of freedom
## Multiple R-squared:  0.5319, Adjusted R-squared:  0.5296
## F-statistic: 225.6 on 2 and 397 DF,  p-value: < 2.2e-16
```

The new model doesn not show any improvement. Rather there is a significant decrease in performance, as the R^2 drops from 60% to 53%, while the *RSE* increases from 1.7 to 1.9. It is safe to say then that advertising does play a good role in the number of units sold, suggesting that Model 1 is superior.

We conclude our modelling process with one more model, making use of logarithmic transformations on the continuous dependent variables, with the hope that it may provide some better results by working on tightly packed data.

```
#Model 3
model_3 <- lm(Sales~log(Advertising)+log(Price)+ShelveLoc,data = carseats)
```

```
## Warning in log(Advertising): NaNs produced
```

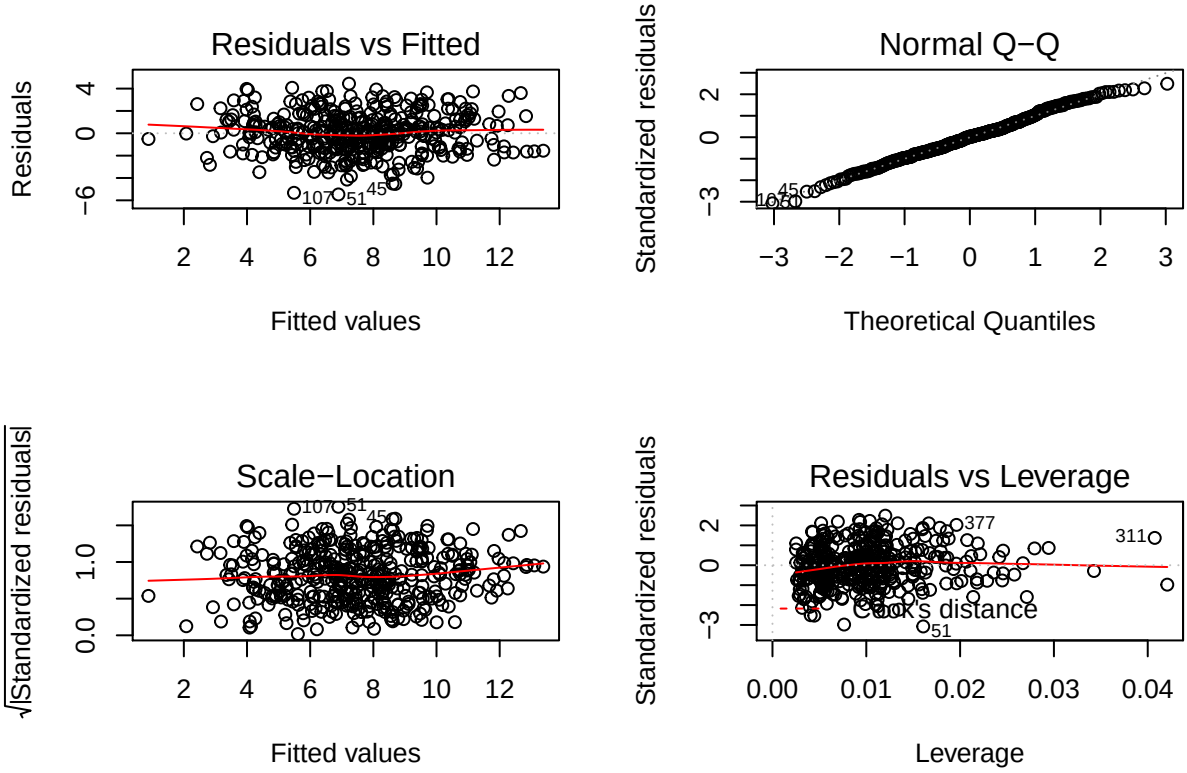
```
## Warning in log(Price): NaNs produced
```

```
summary(model_3)
```

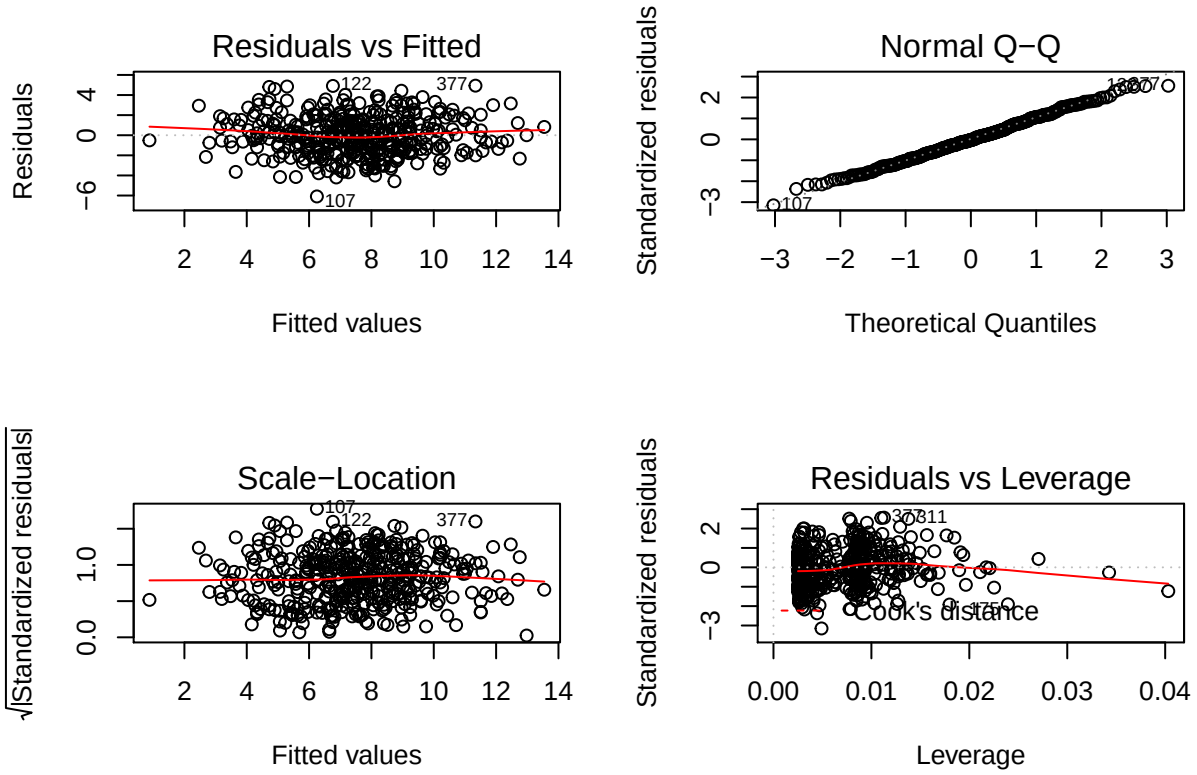
```
##
## Call:
## lm(formula = Sales ~ log(Advertising) + log(Price) + ShelveLoc,
##     data = carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7069 -1.0834  0.0412  1.0800  3.8862
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        6.9963     0.2214  31.596  < 2e-16 ***
## log(Advertising)   0.5460     0.1827   2.988  0.00361 **
## log(Price)        -0.7675     0.1691  -4.538 1.73e-05 ***
## ShelveLoc          2.4790     0.2628   9.433 3.91e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.741 on 91 degrees of freedom
##   (305 observations deleted due to missingness)
## Multiple R-squared:  0.5944, Adjusted R-squared:  0.581
## F-statistic: 44.45 on 3 and 91 DF,  p-value: < 2.2e-16
```

The R^2 is slightly lowered as compared to Model 1, though the RSE lowers from 1.79 to 1.74 as well. There is no particularly drastic change upon log transformation as well. All that is left is to check the residual plots for any additional information.

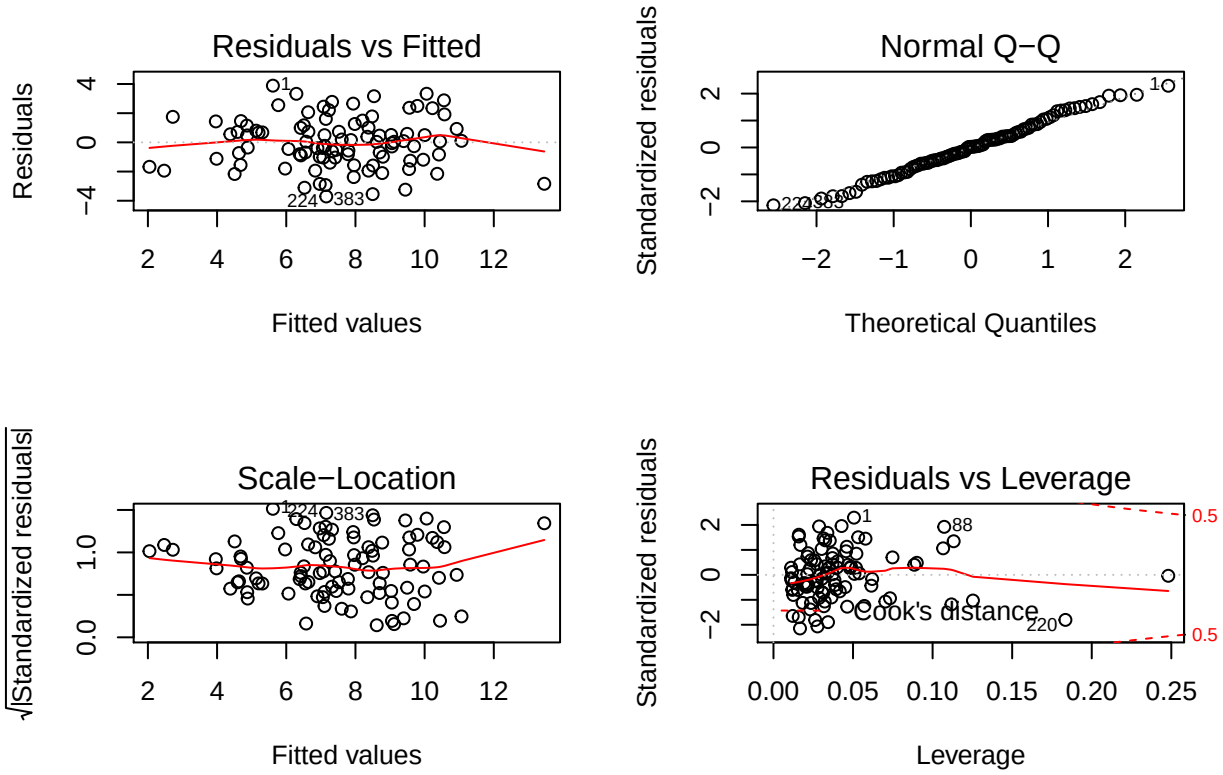
```
par(mfrow=c(2,2))
plot(model_1)
```


Residuals vs Fitted
Residuals
Fitted values
107
51
45
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
377
311
51

```
plot(model_2)
```


Residuals vs Fitted
Residuals
Fitted values
122
377
107
Normal Q−Q
Standardized residuals
Theoretical Quantiles
107
Scale−Location
√|Standardized residuals|
Fitted values
107
122
377
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

```
plot(model_3)
```

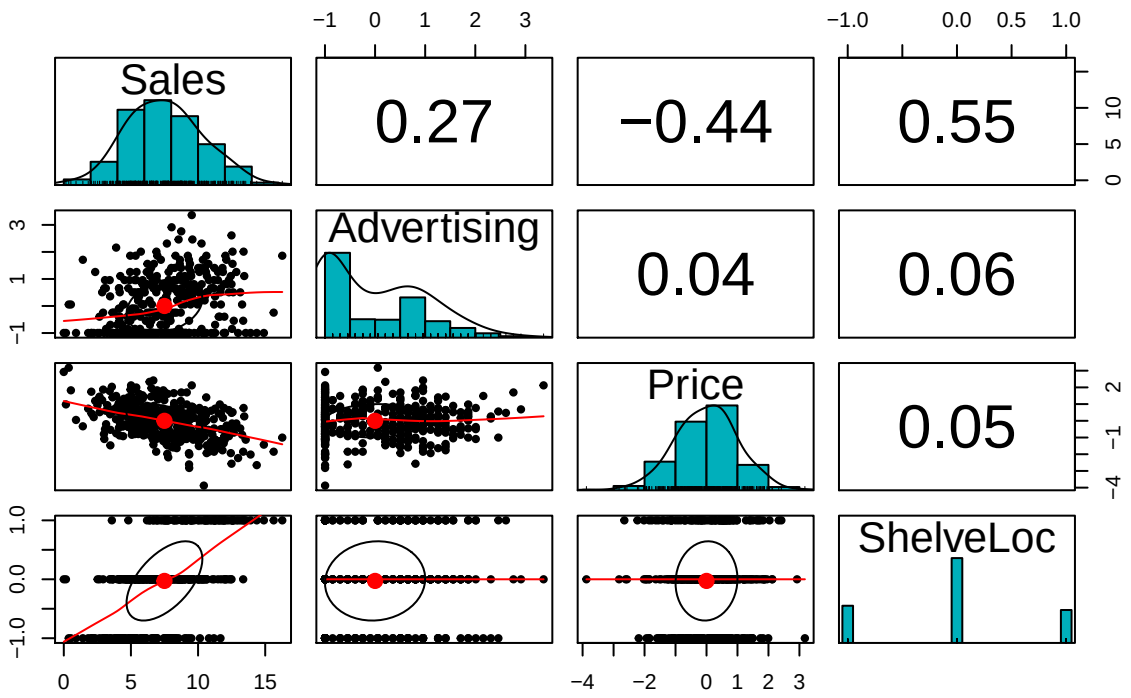


Based on the QQ plots, the residuals from all 3 models seem to closely fit the normal distribution as they lie on the straight line. Thus there is no difference in terms of residual distributionfit.

If we look at the residuals vs fitted graphs as well, there is no major differences among the models as all three are close to the central line. Infact all the residual plots seem to have very similar characteristics, showcasing that not much disparities can be gleamed from them to eliminate any of the three models.

Thus we will choose the ideal model based on the R^2 which would be Model 1 with 59% i.e Model 1 explains 59% of the variance in the data considered. We present the scatterplot of the considered factors in the chosen model below as a summary of its characteristics.

```
pairs.panels(carseats[,c(-5,-6)],
             method = "pearson", # correlation method
             hist.col = "#00AFBB",
             density = TRUE,  # show density plots
             main="Scatter plots of Carseats data")
```



Conclusion

We see that the location of the child car seat stores have very little impact on the number of car seats sold. Rather, the price of the car seats, the advertising budget and shelving quality of the stores seems to have the most considerable impact on the number of units sold. Our model consisting of these three variables seem to provide the best possible model in comparison, with a moderately good explanatory power in predicting sales.