Assignment

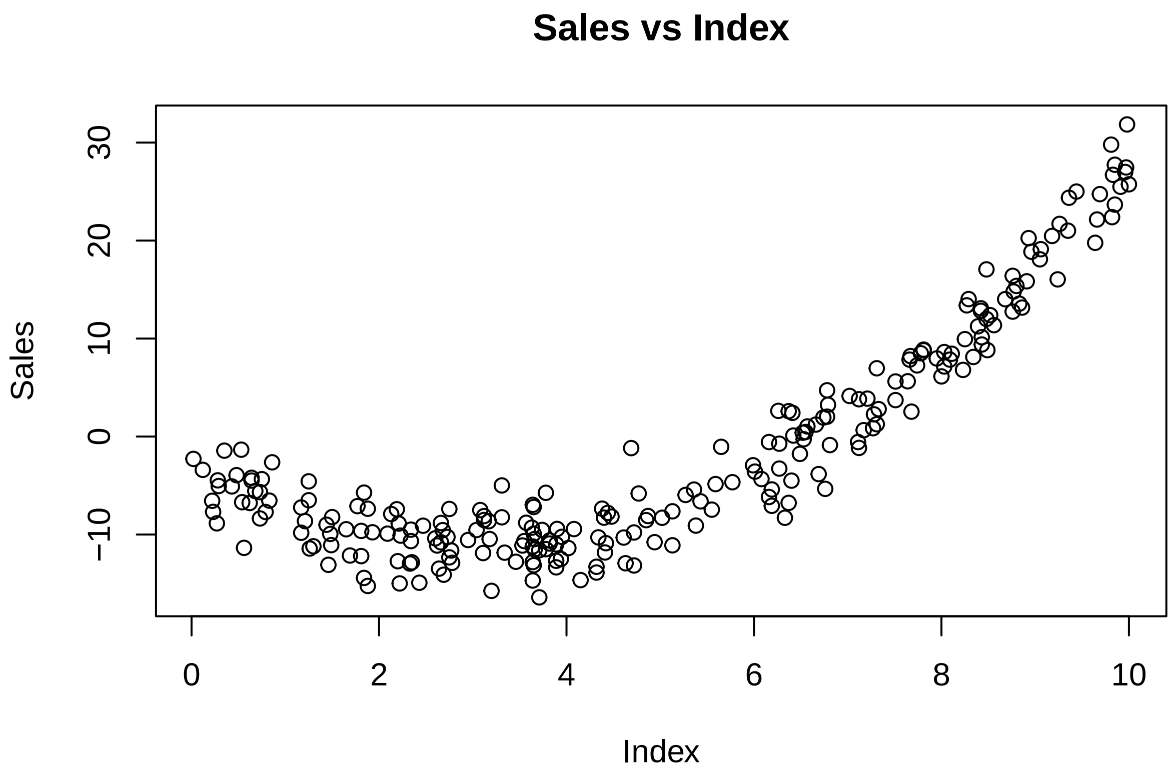
Shalin Anam

2024-05-08

Question 1

Part a

```
sales <- read.csv("sales.csv", header =TRUE)
plot(sales, main = "Sales vs Index")
```



Observing the original data in the scatter plot, it is clear that this is NOT a linear relationship. Instead, it contains a curve pattern suggesting a polynomial model might be more appropriate.

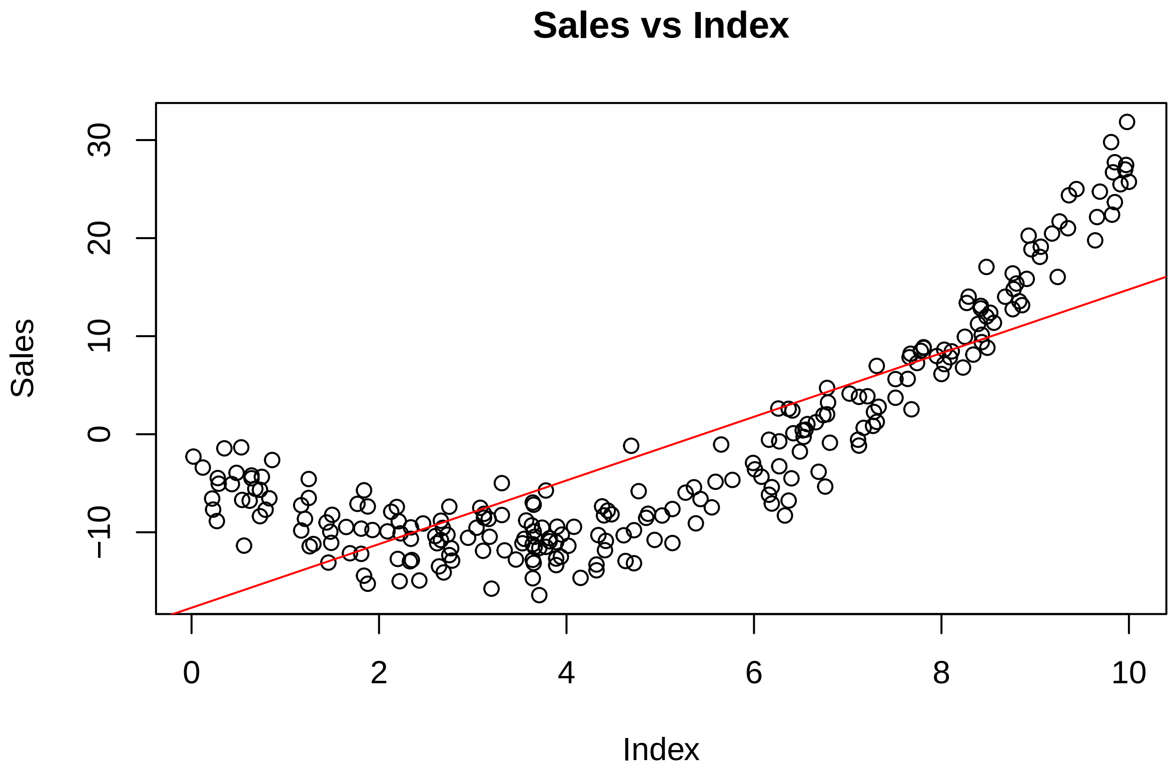
A curve in the scatter plot indicates that as Index increases, the increase or decrease in Sales does not happen at a constant rate. To prove that teh linear model is NOT the most effective, we can conduct a simple linear regression model as shown below:

Part b

```
M1 = lm(Sales ~ Index, data = sales)
summary(M1)
```

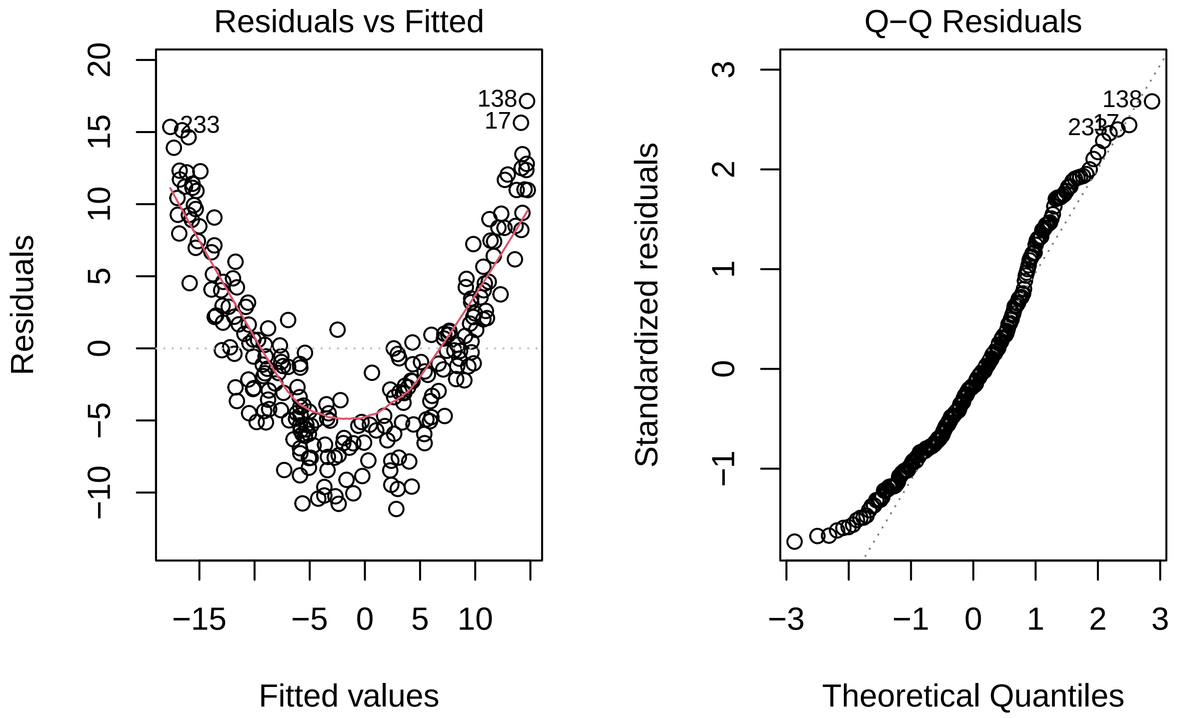
```
##
## Call:
## lm(formula = Sales ~ Index, data = sales)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.139  -4.988  -1.086   4.028  17.152
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.7007     0.8278  -21.38   <2e-16 ***
## Index         3.2464     0.1444   22.48   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 6.45 on 241 degrees of freedom
## Multiple R-squared:  0.6772, Adjusted R-squared:  0.6759
## F-statistic: 505.6 on 1 and 241 DF,  p-value: < 2.2e-16
```

```
plot(sales, main = "Sales vs Index")
abline(M1, col = "red")
```



The overlaid red line from the linear regression model is a clear indicator that this model does NOT accurately capture the underlying relationship between these variables. We can verify this with diagnostic tests:

```r
par(mfrow = c(1, 2))
plot(M1, which = 1:2)
```



The diagnostic plots reveals a concern about the effectiveness of the simple linear regression model. Firstly, the quantile plot (QQ plot) of the residuals shows a slight curve, which implies that the residuals do not perfectly follow the normal distribution assumption. However, more significantly, the Residual vs Fitted plot shows a clear concave up quadratic trend, suggesting a quadratic or a higher order fit. Hence, these patterns in the residual versus fitted plot indicate that the assumption of equal variance is violated.

This is strong evidence to support the need for a more complex model such a polynomial regression which can better represent the curve structure in our data. A standard linear regression does not meet the assumptions and a visual representation of the original plot suggests a curve structure.

Part c

```
M2 = lm(Sales ~ Index + I(Index^2), data = sales)
M3 = lm(Sales ~ Index + I(Index^2) + I(Index^3), data = sales)
summary(M2)
```

```
##
## Call:
## lm(formula = Sales ~ Index + I(Index^2), data = sales)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.755 -1.967  0.037  1.749  7.827
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.50608    0.50308  -6.969 3.06e-11 ***
## Index       -4.96591    0.23046 -21.548  < 2e-16 ***
## I(Index^2)   0.80875    0.02201  36.744  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.511 on 240 degrees of freedom
## Multiple R-squared:  0.9513, Adjusted R-squared:  0.9509
## F-statistic:  2343 on 2 and 240 DF,  p-value: < 2.2e-16
```

```r
summary(M3)
```

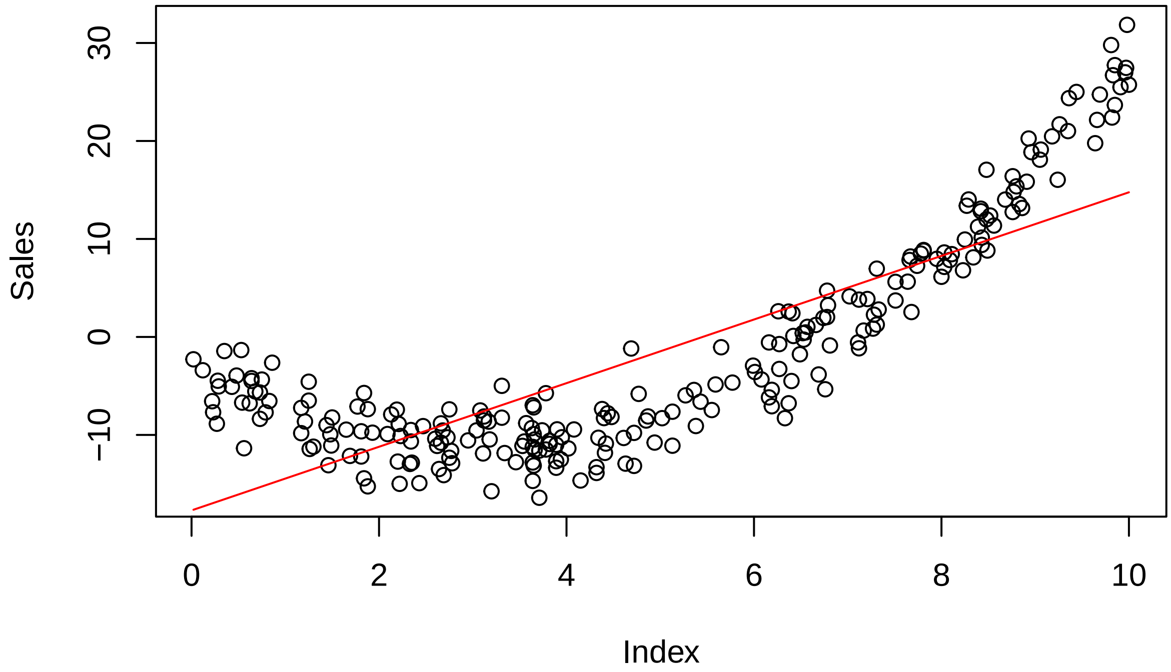
```
##
## Call:
## lm(formula = Sales ~ Index + I(Index^2) + I(Index^3), data = sales)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7850 -1.9384  0.0545  1.7424  7.8321
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.421148   0.668122  -5.121 6.27e-07 ***
## Index       -5.062632   0.550206  -9.201  < 2e-16 ***
## I(Index^2)   0.832770   0.125982   6.610 2.48e-10 ***
## I(Index^3)  -0.001599   0.008255  -0.194    0.847
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.516 on 239 degrees of freedom
## Multiple R-squared:  0.9513, Adjusted R-squared:  0.9507
## F-statistic:  1556 on 3 and 239 DF,  p-value: < 2.2e-16
```

M2 represents a quadratic model (second - level polynomial). M3 represents a cubic model (third - level polynomial). Considering the summary of each one, we can see that cubic term is insignificant. We can also observe the Goodness of Fit. The M2 Model has a slightly higher Adjusted R Squared compared to M3 which shows that it better explains the variance. Hence, M2 appears to be a better choice.

Part d

Predicting M1

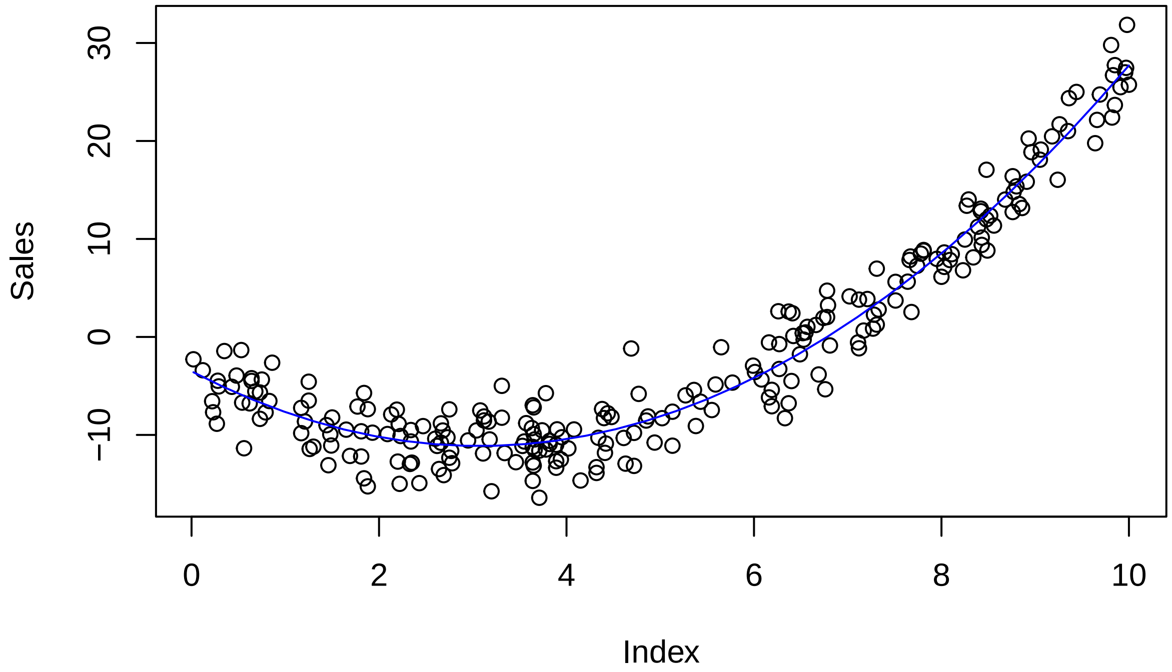
```r
plot(sales)
x = seq(from = min(sales$Index), to = max(sales$Index), length = 32)
Indexdat = data.frame(Index = x)
Saleshat = predict(M1, newdata = Indexdat)
lines(x, Saleshat, col = "red")
```



This the simple linear model (M1). Observing this graph above, we can see that this model does NOT accurately represent the curved relationship between sales and index.

Predicting M2

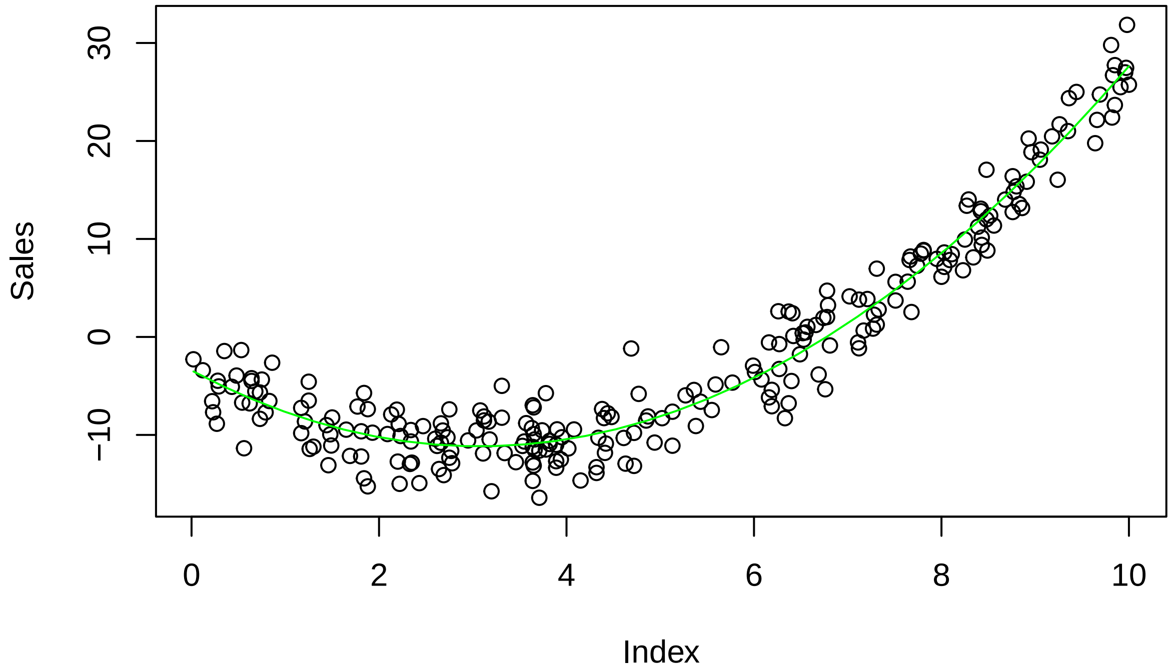
```
plot(sales)
x = seq(from = min(sales$Index), to = max(sales$Index), length = 32)
Indexdat = data.frame(Index = x)
Saleshat = predict(M2, newdata = Indexdat)
lines(x, Saleshat, col = "Blue")
```



M2 shows a quadratic polynomial. Observing this graph above, we can see that the line explains a greater portion of the variance and hence captures the curve structure of the data.

Predicting M3

```
plot(sales)
x = seq(from = min(sales$Index), to = max(sales$Index), length = 32)
Indexdat = data.frame(Index = x)
Saleshat = predict(M3, newdata = Indexdat)
lines(x, Saleshat, col = "Green")
```



M3 represents the cubic. Observing this graph above, we can see that it is similar to the quadratic model, the cubic model also captures the curve. Hence, we need to choose which model out of the quadratic and cubic is the best.

Part e

```
anova(M3)
```

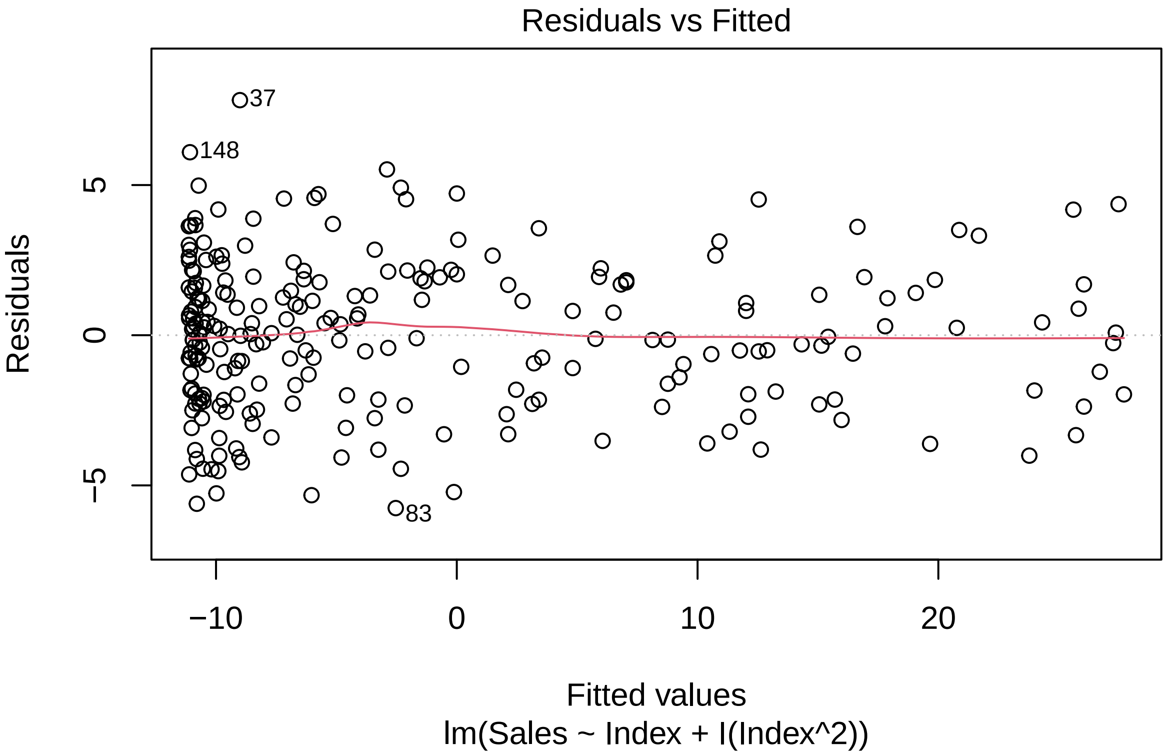
```
## Analysis of Variance Table
##
## Response: Sales
##              Df  Sum Sq Mean Sq   F value Pr(>F)
## Index         1 21036.0 21036.0 3322.5228 <2e-16 ***
## I(Index^2)    1  8513.8  8513.8 1344.7051 <2e-16 ***
## I(Index^3)    1     0.2     0.2    0.0375 0.8466
## Residuals   239  1513.2     6.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

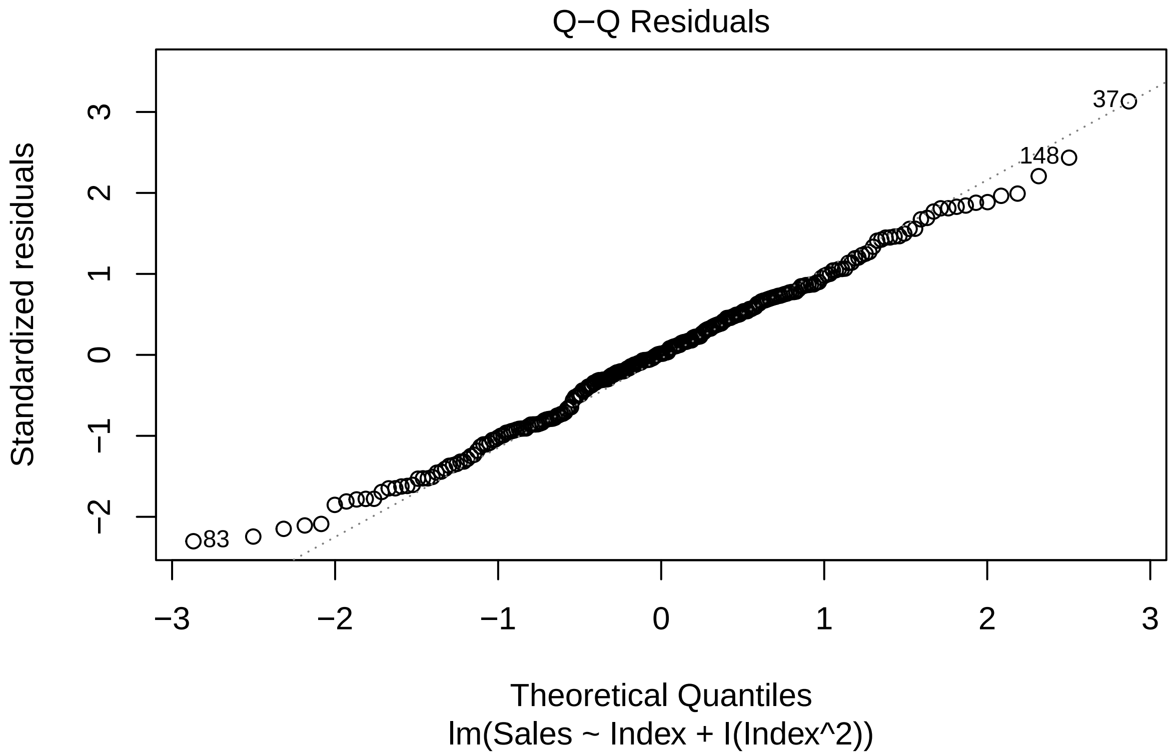
From the sequential sum of squares above we can see that the cubic term is insignificant. However, the quadratic term is significant. The cubic term contributes almost nothing to the explanation of variance of Sales with a sum square value of just 0.2. This shows it is insignificant. Furthermore, the p-value of the cubic term is 0.847 hence remove it and refit the model with the quadratic.

Part f

Comparing the p-values, we can see that the cubic has an insignificant p-value, hence remove it. Therefore, we get the quadratic which has a significant p - value. The quadratic has a significant sum square as well.

```r
plot(M2, which = 1:2)
```





Above we have plotted the Residual vs Fitted plot and the QQ plot. We can clearly see that there is equal variance and the residuals follow a normal distribution. Hence, M2 model satisfies the assumptions. Visually, M2 is the best as well since the blue line follows the trend of data points accurately.

Question 2

Part a

```
campaign <- read.csv("campaign.csv", header = TRUE)
head(campaign)
```

```
##   Score Region      Type
## 1 12.30  Rural Billboard
## 2  9.22  Rural Billboard
## 3  9.39  Rural Billboard
## 4  9.30  Rural Billboard
## 5 13.13  Rural Billboard
## 6 10.97  Rural Billboard
```

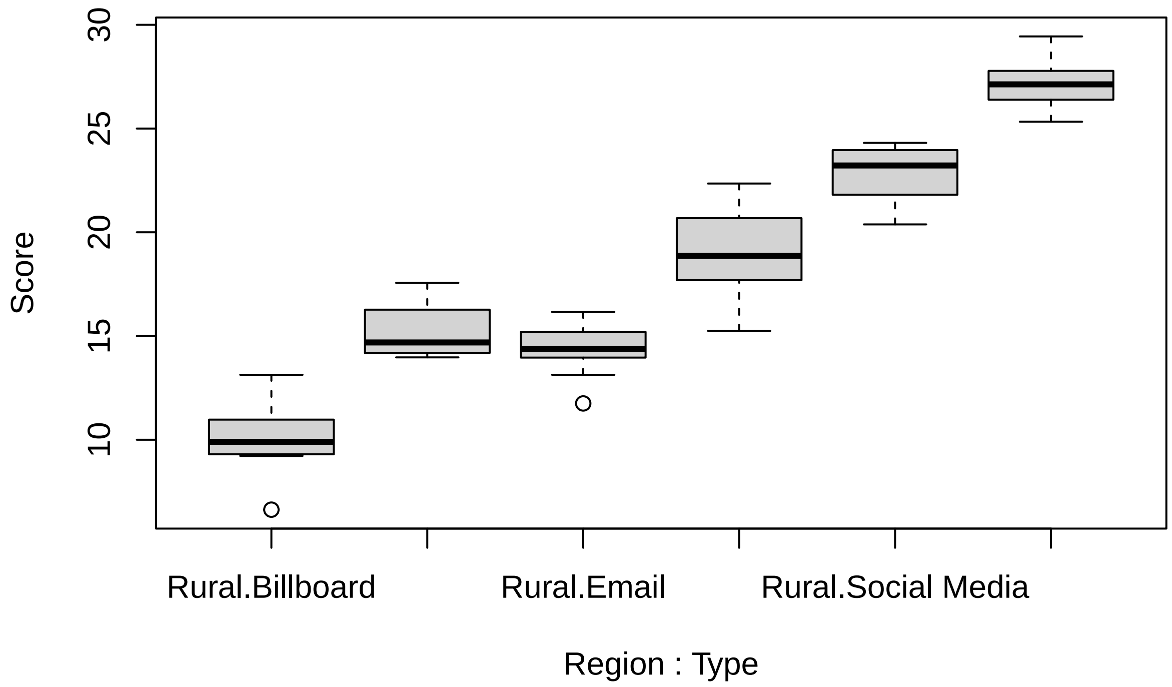
```
table(campaign [,2], campaign [,3])
```

```
##
```

```
##           Billboard Email Social Media
##     Rural        10    10           10
##     Urban        10    10           10
```

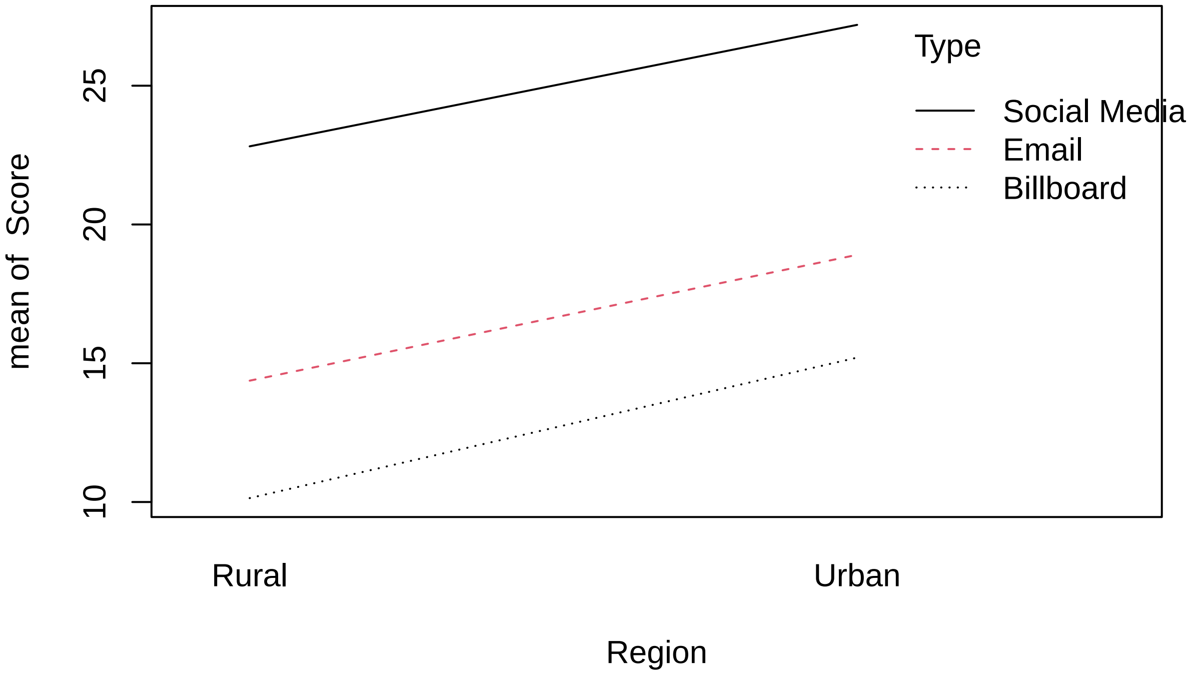
This study is balanced since there are an equal number of replicates across all the factors.

```
boxplot(Score ~ Region + Type, data = campaign)
```



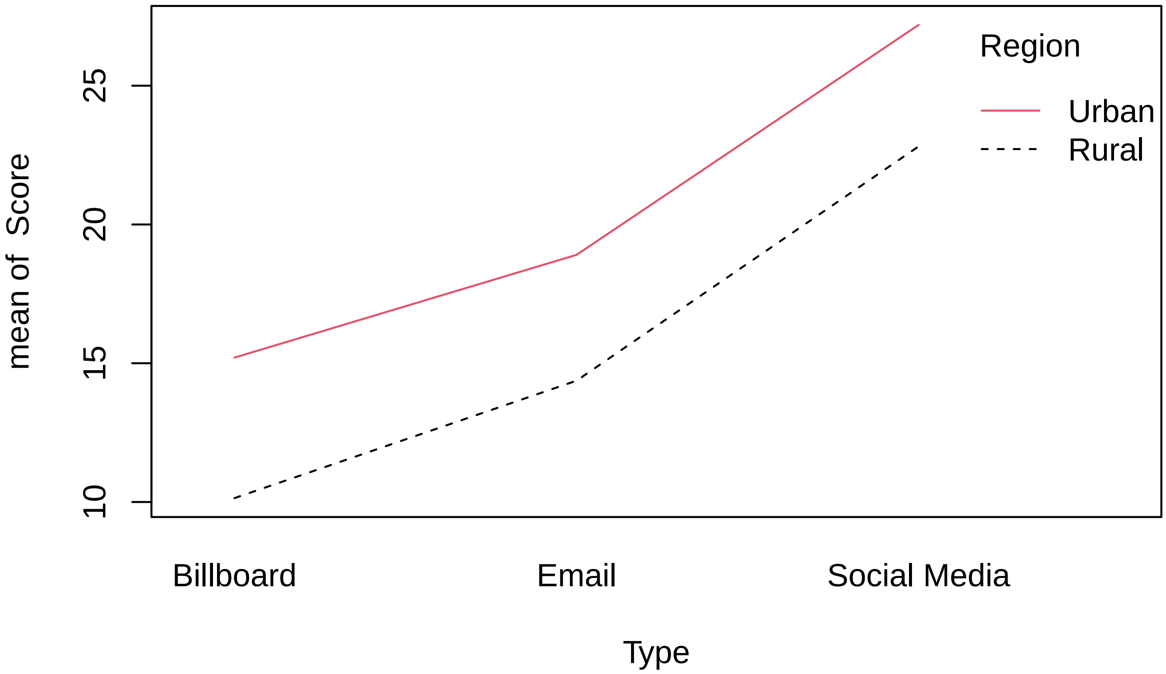
The boxplots show equal variance since the lengths are similar. Hence, the assumption of equal variance is satisfied.

```
with(campaign, interaction.plot(Region, Type, Score, col = 1:2))
```



- This first interaction plot suggests that accross all types of marketing campaigns, the engagement scores are higher in urban regions compared to rural.
- This indicates that region has a significant impact on engagement scores. This is shown by urban regions consistently outperforming rural regions.
- As shown in the interaction plot, the lines for social media, email, and billboard are relatively parallel. This indicated that an increase in engagement scores from rural to urban is consistent across all catagories and different types of campaigns
- The parallel lines suggest that there is significant interaction effect between region and type.

```r
with(campaign, interaction.plot(Type, Region, Score, col = 1:2))
```



- From this interaction plot, we can see that billboards have the least effect on engagement scores across both rural and urban. Its clear that social media has the highest engagement score in both regions.
- The interaction plot shows that scoial mendia is the most effective in increasing customer engagement.
- While some interaction is present in the plot, it clearly demonstrates that the Type of the campaign and Region independantly effect the customer engagement score.

Part b

The full Two-Way ANOVA interaction model is:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk},$$

where the parameters are:

- Y_{ijk}: Increase in engagement score;
- α_i: the Type main effect, there are three levels - Billboard, Email, Social Media;
- β_j: the Region main effect, there are two levels - Rural and Urban;
- γ_{ij}: interaction effect between Type and Region;
- $\epsilon_{ijk} \sim N(0, \sigma^2)$: is the unexplained variation.

Part c

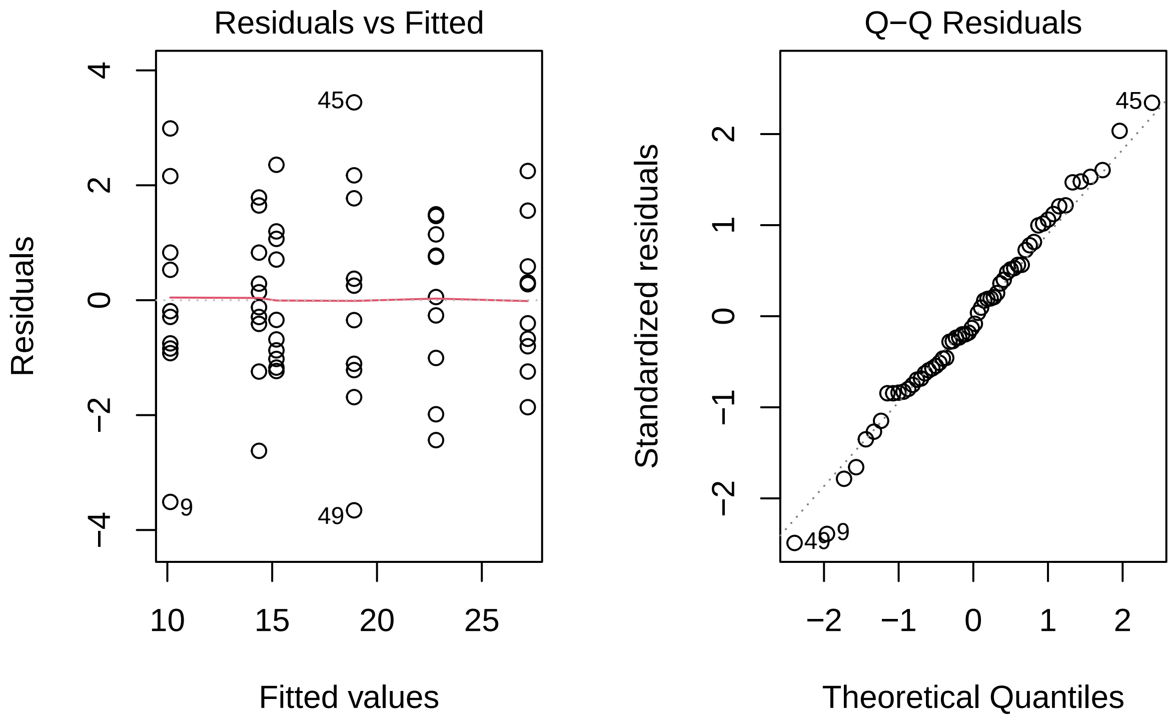
We test the null hypothesis:

$H_0 : \gamma_{ij} = 0$ for all i, j against the alternative hypothesis H_1 : at least one $\gamma_{ij} \neq 0$.

```
model = lm(Score ~ Region * Type, data = campaign)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: Score
##             Df  Sum Sq Mean Sq  F value    Pr(>F)
## Region       1  325.45  325.45 135.7281 2.336e-16 ***
## Type         2 1585.09  792.54 330.5242 < 2.2e-16 ***
## Region:Type  2    1.29    0.64   0.2683    0.7657
## Residuals   54  129.48    2.40
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
par(mfrow = c(1, 2))
plot(model, which = 1:2)
```



- The equal variance assumption is met adequately as shown in the residual vs fitted plot where the residuals show an equal distribution around the fitted value.

- The quantile plot (QQ plots) follow a linear line, hence indicating that residuals follow a formal distribution.
- From the anova model, the F-test of the interaction term contains a P-vale of 0.7657. This indicates that the interaction term is insignificant and hence needs to be removed from the model.

Part d

Currently in our study, we have a balanced study with an insignificant interaction term. Therefore we must refit the model:

Hypotheses:

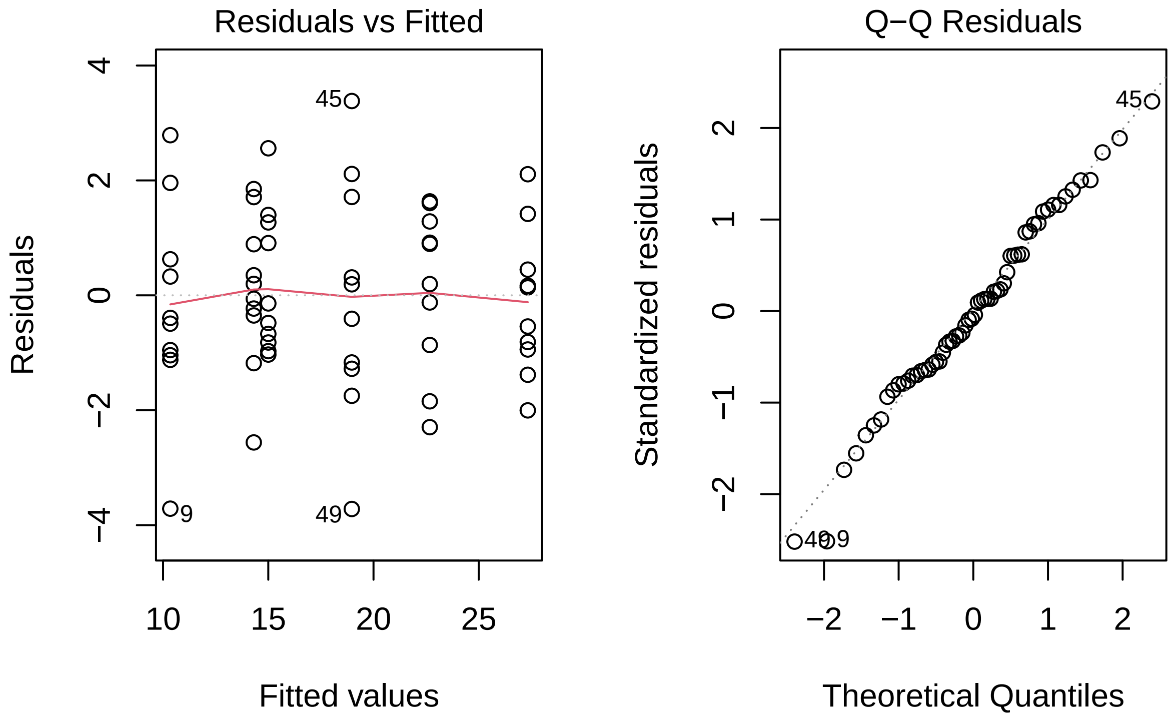
- For the **Type** of campaign: (H_0): $\alpha_i = 0$ for all i against (H_1): At least one $\alpha_i \neq 0$
- For the **Region**: (H_0): $\beta_j = 0$ for all j against (H_1): At least one $\beta_j \neq 0$

Using the regression method, we have:

```
model_main = lm(Score ~ Type + Region, data = campaign)
summary(model_main)
```

```
##
## Call:
## lm(formula = Score ~ Type + Region, data = campaign)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7185 -0.9575 -0.0925  1.0039  3.3815
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        10.3430     0.3946   26.21  < 2e-16 ***
## TypeEmail           3.9675     0.4832    8.21 3.49e-11 ***
## TypeSocial Media   12.3315     0.4832   25.52  < 2e-16 ***
## RegionUrban         4.6580     0.3946   11.81  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.528 on 56 degrees of freedom
## Multiple R-squared:  0.9359, Adjusted R-squared:  0.9325
## F-statistic: 272.7 on 3 and 56 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(1, 2))
plot(model_main, which = 1:2)
```



The effect of changing one factor is considered while maintaining all other factors or predictors at their baseline. Each predictor factor is conditional to all other factors used in the model. In our study, the baseline comparison is type billboard and region rural. The anova will check the significance of the Type and Region:

```
anova(model_main)
```

```
## Analysis of Variance Table
##
## Response: Score
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Type       2 1585.09  792.54  339.39 < 2.2e-16 ***
## Region     1  325.45  325.45  139.37 < 2.2e-16 ***
## Residuals 56  130.77    2.34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The residule vs fitted plot shown validates the equal vairance assumption since there is no pattern on that plot.
- The residuals follows a normal distribution as shoown in the QQ plot which has a linear structure.
- From the Anova table, both Type and Region are significant at 5% significance level since both p values are less then 0.05. Hence, cannot be removed.
- Therefore we have accomplished our final model.

Part e

```
table(campaign[, c("Type", "Region")])
```

```
##                Region
## Type           Rural Urban
##   Billboard       10    10
##   Email           10    10
##   Social Media    10    10
```

Hence the study is balanced since since there is an equal number of obervations. Therefore, we can conduct the TukeyHSD Test.

```
TukeyHSD(aov(model_main))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = model_main)
##
## $Type
##                            diff       lwr       upr p adj
## Email-Billboard         3.9675  2.804077  5.130923     0
## Social Media-Billboard 12.3315 11.168077 13.494923     0
## Social Media-Email      8.3640  7.200577  9.527423     0
##
## $Region
##              diff      lwr      upr p adj
## Urban-Rural 4.658 3.867599 5.448401     0
```

The balanced design of this study allows for us to compare using Tukey's HSD test more effectively. The results clearly show that social media campaigns have the highest impact on the customer engagement scores. This is followed by email campaigns and lastly billboard with the least effectiveness. Furthermore, the campigns that are conducted in the urban regions significantly outperform those in rural regions.