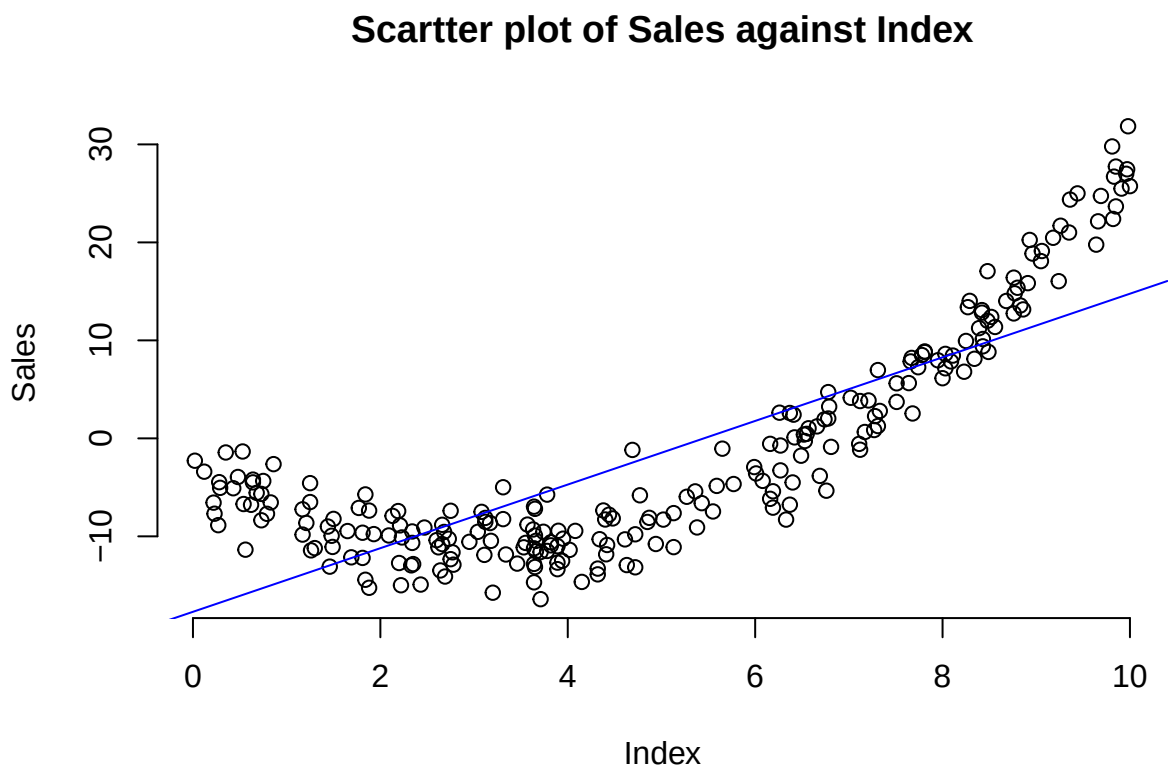
Sales Analysis

2024-05-15

Part 1

a. Sccater Plot

```
sales <- read.csv("/Users/nguyenphucnhuhai/Downloads/sales.csv")
plot(sales$Index, sales$Sales, main = "Scartter plot of Sales against Index",
     xlab = "Index", ylab = "Sales",
    frame = FALSE)
abline(lm(Sales ~ Index, data = sales), col = "blue")
```

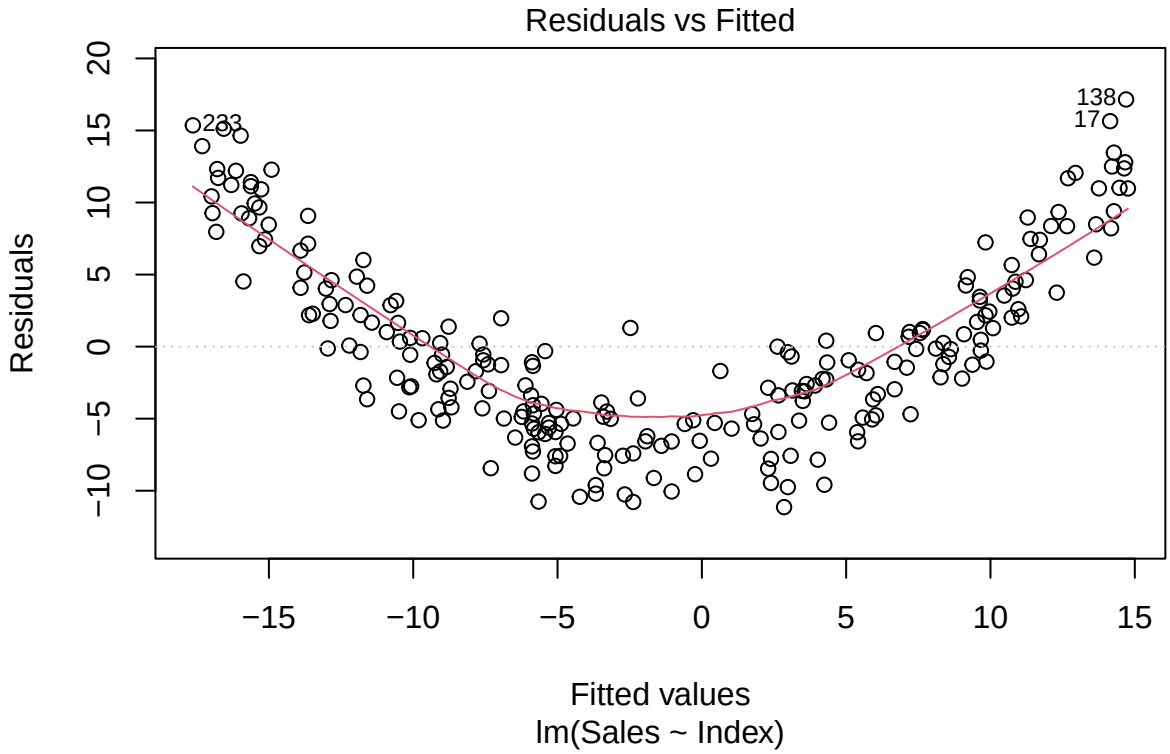


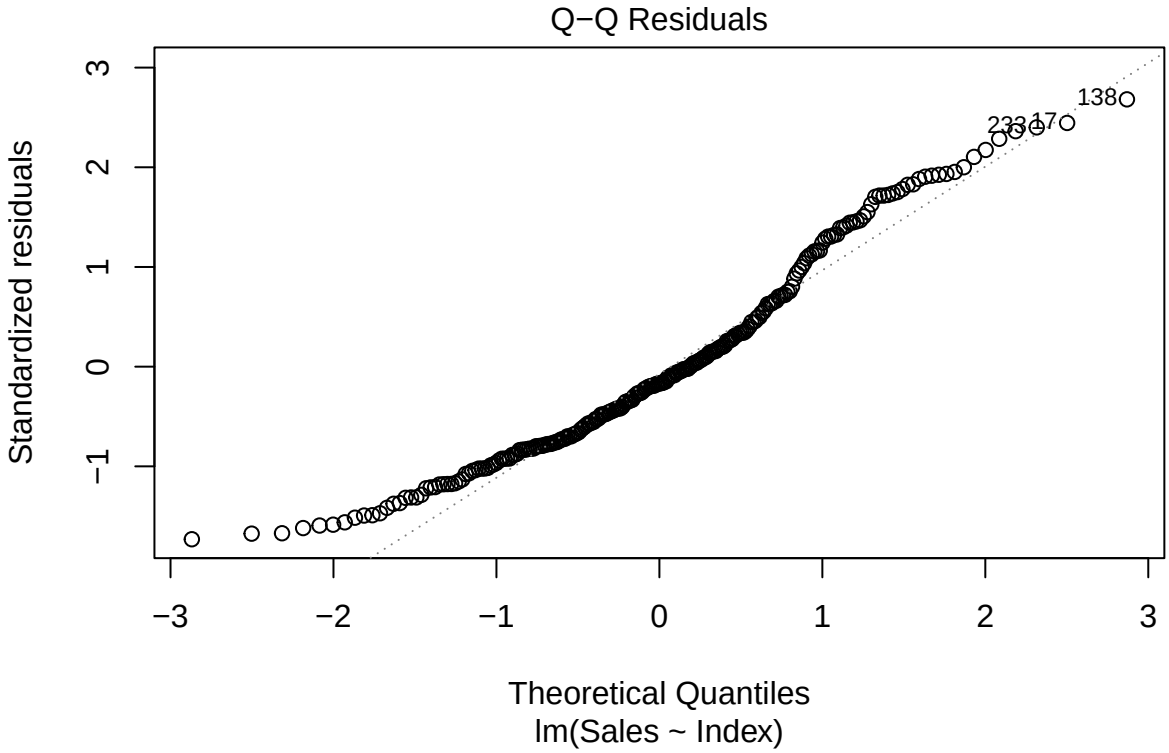
The sccater plot and the blue line show that there is a strong, positive, linear regression between **Sales** and **Index** variables as there is no outlier, and the blue line is slightly goes up following the plots.

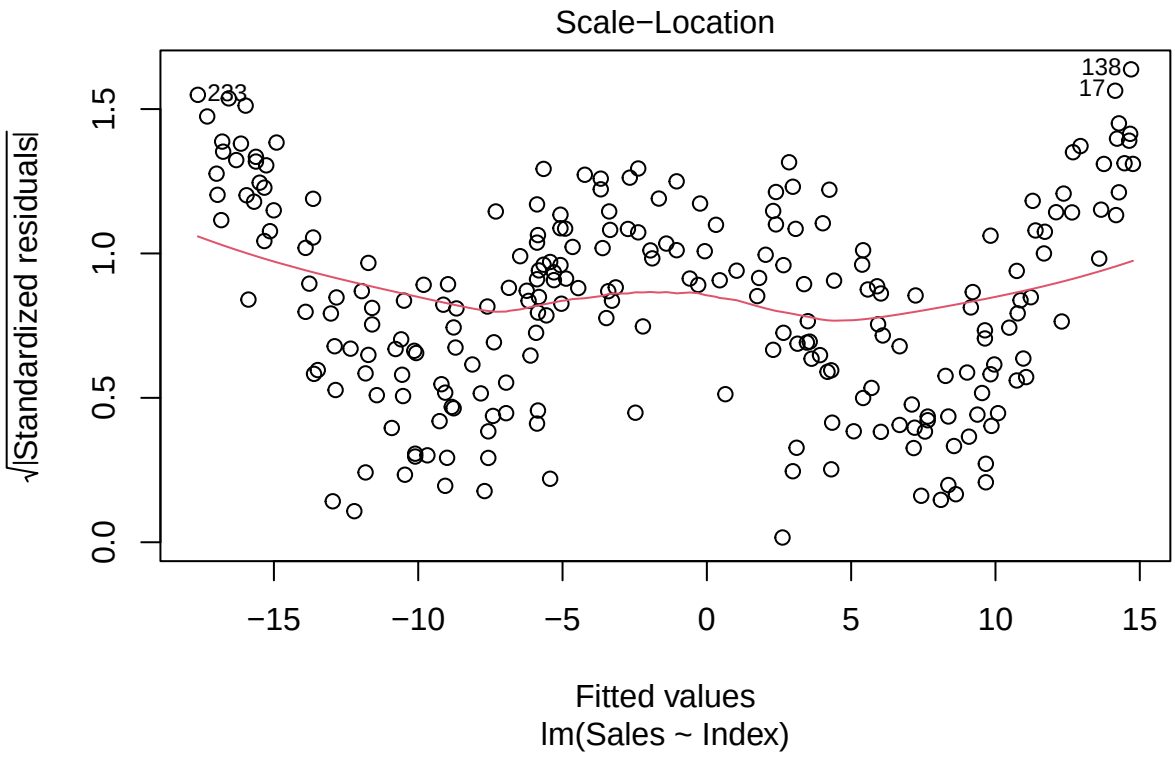
b. Fit the model diagnostic regression

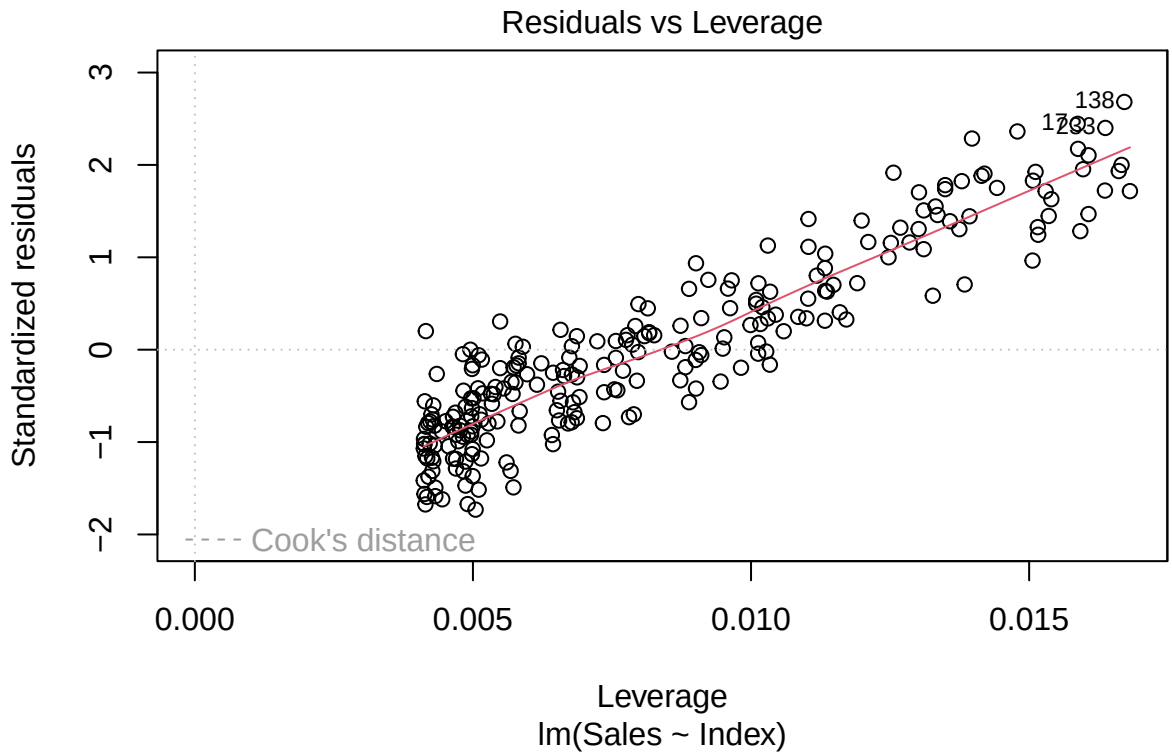
$$y = \beta_0 + \beta_1 X + \epsilon$$

```r
M1 <- lm(Sales ~ Index, data = sales)
plot(M1)
```




Q−Q Residuals
Standardized residuals
3
2
1
0
−1
138
233
17
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(Sales ~ Index)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
233
138
17
−15
−10
−5
0
5
10
15
Fitted values
lm(Sales ~ Index)



Comment: According to the the residual and fitted model, it can be clearly seen that there is a U-shaped parabola form, it suggests that there might be a non-linear relationship between the independent and dependent variables (which is sales and index) The quantile plot of residuals look approximately linear, suggesting the normality assumption for residuals is appropriate

c. Quadratic Model

$$y(x) = ax^2 + bx + c + \epsilon_i (i = 1, 2, .., n), a \neq 0$$

Cubic Model

$$y(x) = ax^2 + bx + cx + d + \epsilon_i (i = 1, 2, .., n), a \neq 0$$

```
M2 <- lm(Sales ~ Index + I(Index^2), 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
```

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

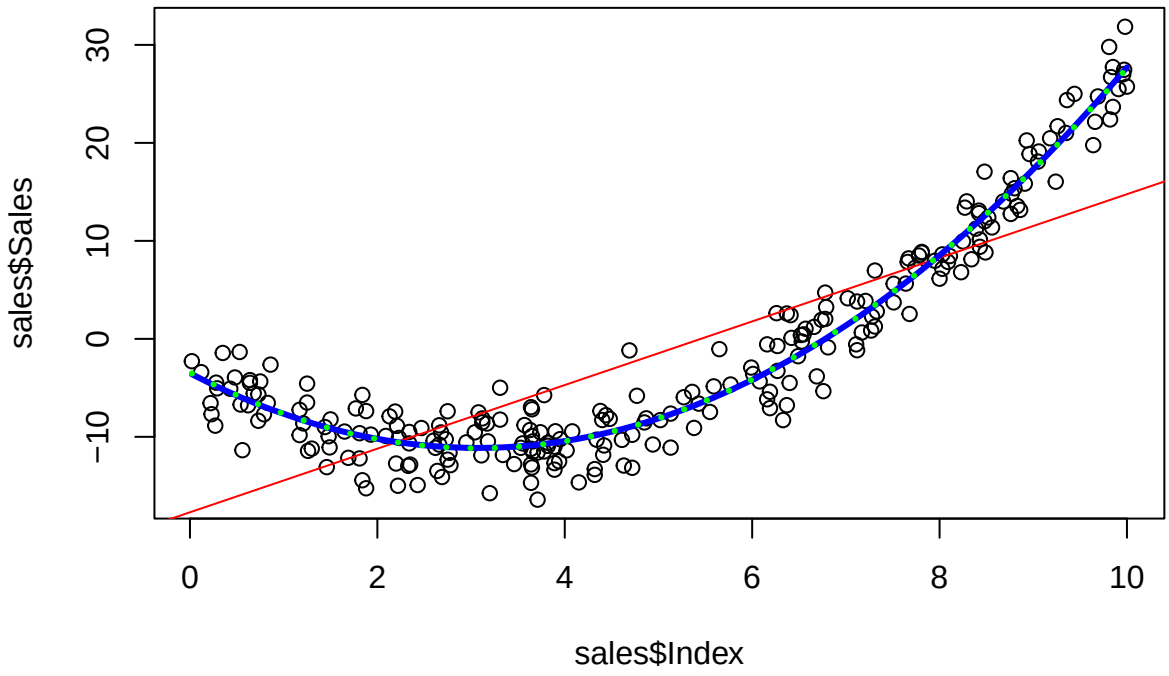
Compare and Comment

d.

```
sales.linear = lm(Sales ~ Index, data = sales)
anova(sales.linear,M1, M2, M3)
```

```
## Analysis of Variance Table
##
## Model 1: Sales ~ Index
## Model 2: Sales ~ Index
## Model 3: Sales ~ Index + I(Index^2)
## Model 4: Sales ~ Index + I(Index^2) + I(Index^3)
##   Res.Df     RSS Df Sum of Sq         F Pr(>F)
## 1    241 10027.2
## 2    241 10027.2  0       0.0
## 3    240  1513.4  1    8513.8 1344.7051 <2e-16 ***
## 4    239  1513.2  1       0.2    0.0375 0.8466
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
plot(sales$Sales ~ sales$Index)
abline(M1,col="red")
lines(smooth.spline(sales$Index, predict(M2)), col = "blue", lwd=3)
lines(smooth.spline(sales$Index, predict(M3)), col = "green", lwd = 3, lty =3)
```



Comment:

```r
M1.lm <-lm(Sales ~ Index, data = sales)
anova(M1.lm)
```

```
## Analysis of Variance Table
##
## Response: Sales
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Index       1  21036 21036.0  505.59 < 2.2e-16 ***
## Residuals 241  10027    41.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
M2.lm <-lm(Sales ~ Index  + I(Index^2), data = sales)
anova(M2.lm)
```

```
## Analysis of Variance Table
##
```

```
## Response: Sales
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## Index        1 21036.0 21036.0  3335.9 < 2.2e-16 ***
## I(Index^2)   1  8513.8  8513.8  1350.1 < 2.2e-16 ***
## Residuals  240  1513.4     6.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
M3.lm <-lm(Sales ~ Index + I(Index^2) + I(Index^3), data = sales)
anova(M3.lm)
```

```
## 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
```

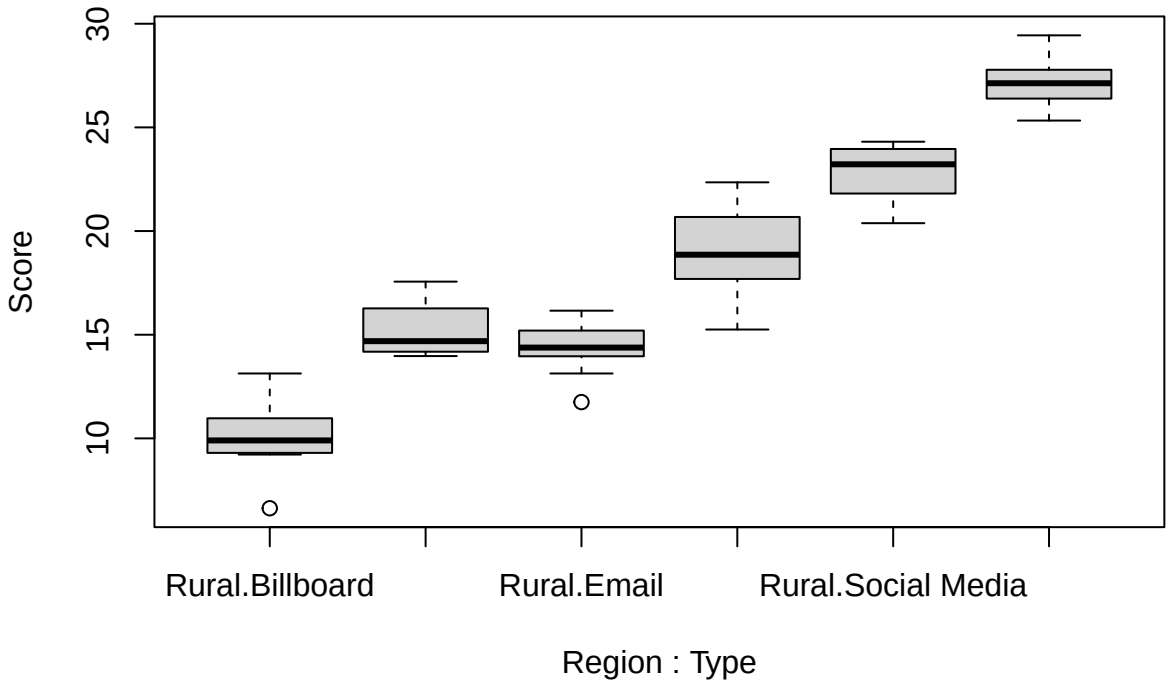
Comment: We can see that the interaction terms are significant since the F-test of the interaction term has a P-value of $0.8466 > 0.05$, they are considered insignificant, indicating that not adding them into the model fit.

f. M2, which is the quadratic polynomial model, is the best fit for this analyis. First, as this analysis is non-linear so we will use polynomial regression. They can be used to model two-dimensional objects to allow us to identify and explore the result of intereaction between these two dimensions. Following what we have analyzed earlier, the model 3 is considered insignificant, so they are not fit in this analysis.

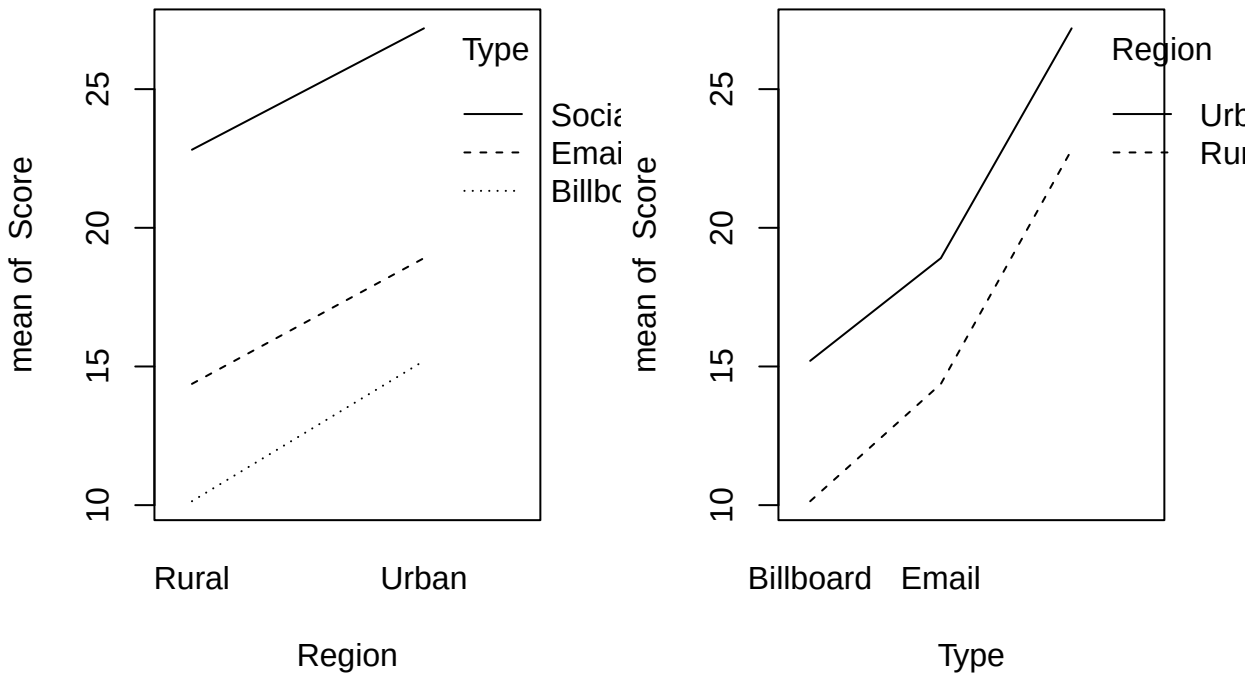
Part 2

a. Constructing the preliminary plots

```
campaign <- read.csv("/Users/nguyenphucnhuhai/Downloads/campaign.csv")
boxplot(Score ~ Region + Type, data = campaign)
```




```
par(mfrow = c(1, 2))
with(campaign, interaction.plot(Region, Type, Score))
with(campaign, interaction.plot(Type, Region, Score))
```



- From the boxplot, we can see that the assumption of equal variance among levels seems approximately valid due to the similar box sizes.
- From both interaction plots we can see parallel lines for the means of each group at different levels of the independent variables, this indicates a insignificant interaction effect between the two independent variables.

b. The full Two-Way ANOVA model with interaction is:

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

where the parameters are :

- Y_{ijk}: the campaign recall **Score** response
- μ: overall mean
- α_i: the **Region** effect, there are two levels - Urban and Rural
- β_j: the **Type** effect, there are three levels - Billboard, Email, Social Media
- γ_{ij}: interaction effect between **Region** and **Type**.
- $\epsilon_{ijk} \sim N(0, \sigma^2)$ is the unexplained variation.

c. Null and alternative hypothesis:

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

Fitting this interaction model

```r
campaign.int <- lm(Score ~ Region * Type, data = campaign)
anova(campaign.int)
```

```
## 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
```

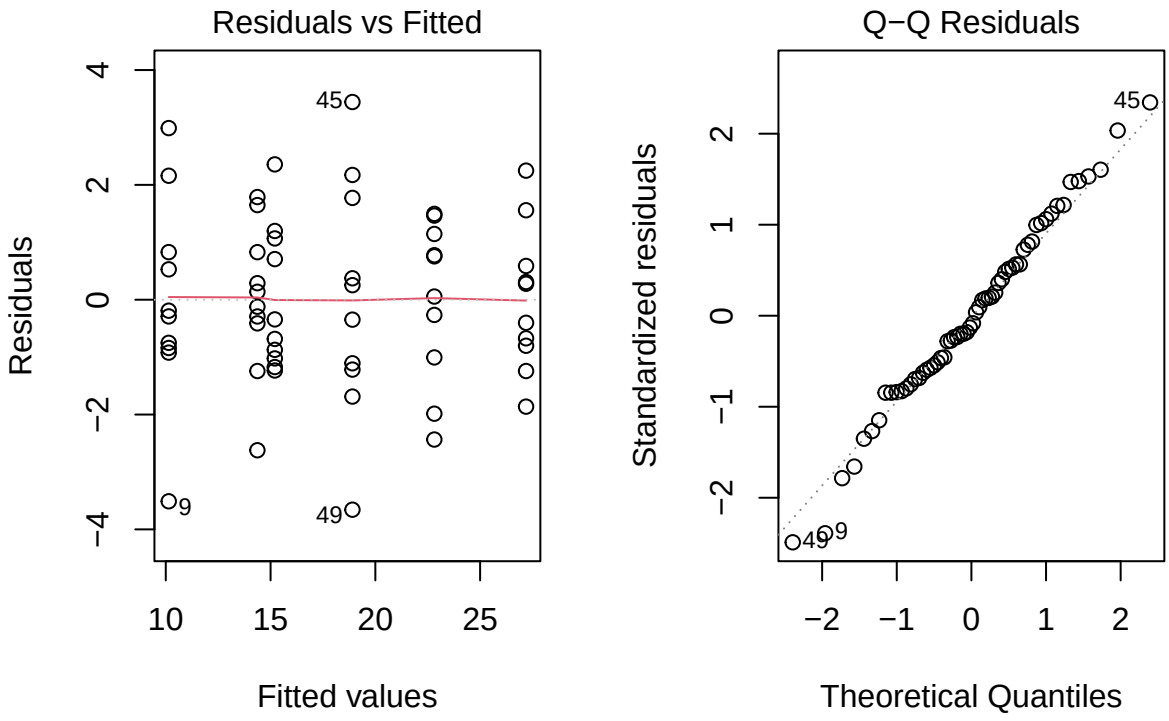
As interaction term has the P-value = 0.7657 > 0.05,

(Statistical) There is not enough evidence to reject H0.

(Contextual) We can conclude that the interaction between the **Region** effect and **Type** effect is insignificant and do not have significant effects on the response **Score**.

We should check assumptions with the plots

```r
par(mfrow = c(1, 2))
plot(campaign.int, which = 1:2)
```



- The residual plot seems to show equal spread around the fitted values and so the constant variance assumption is also appropriate.
- The residuals are close to linear in the QQ-plot, and so the normal assumption should be valid.

e. Null and alternative hypothesis:

- For the **Region** effect: $H_0 : \alpha_i = 0$ for all i, j against H_1: at least one $\alpha_i \neq 0$
- For the **Type** effect: $H_0 : \beta_j = 0$ for all i, j against H_1: at least one $\beta_j \neq 0$

Fitting main effects model

```
campaign.lm = lm(Score ~ Region + Type, data = campaign)
anova(campaign.lm)
```

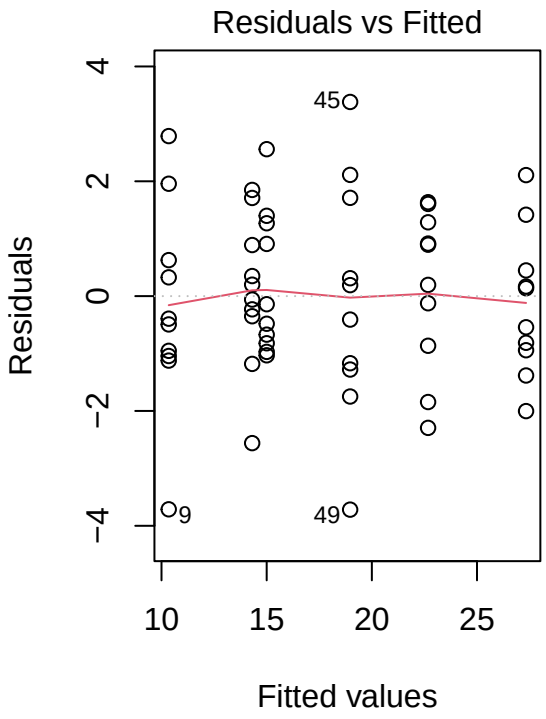
```
## Analysis of Variance Table
##
## Response: Score
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Region     1  325.45  325.45  139.37 < 2.2e-16 ***
## Type       2 1585.09  792.54  339.39 < 2.2e-16 ***
## Residuals 56  130.77    2.34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05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

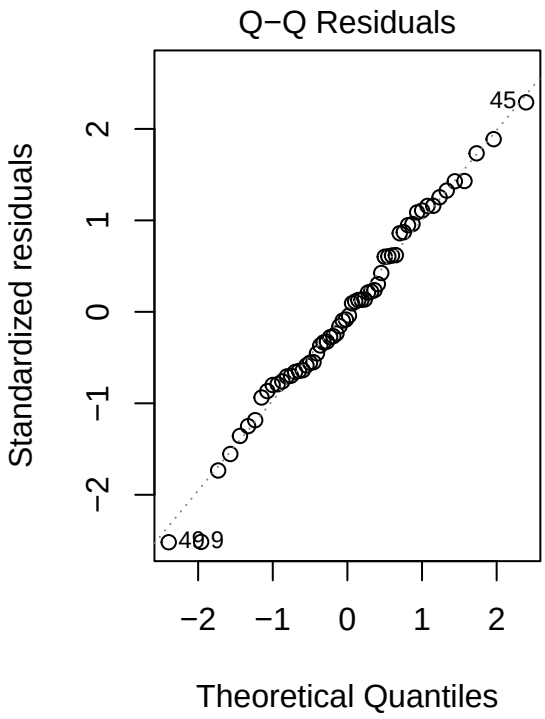
As the P-value for the **Region** effect = 2.2e-16 $<$ 0.05, for the **Type** effect = 2.2e-16 $<$ 0.05

(Statistical) There is enough evidence to reject H0.

(Contextual) We can conclude that the **Type** effect and **Region** effect have significant effects on the response **Score**.

```
par(mfrow = c(1, 2))
plot(campaign.lm, which = 1:2)
```





- The residual plot seems to show equal spread around the fitted values and so the constant variance assumption is also appropriate.
- The residuals are close to linear in the QQ-plot, and so the normal assumption should be valid.

f. Repeat the above test analysis for the main effects. Overall, the **Region** which is the surveyed area and the **Type** that was used for Billboard, Email and Social media types have a significant effect on the **Score**. The Social media has a outstanding effect on score espacially in Urban area, while Billboard and Email only have a minor contribution on Score. These findings emphasize the importance of focusing in the specific platform and area, enabling businesses to increase the effective of marketing campaign, advertisements which create desirable, consistent products that resonate with consumer preferences and improve market competitiveness.

g. Tukey HSD

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

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

From the above we can see that the design is balanced with an equal number of replicates for each combination of levels of the two factors.

```
TukeyHSD(aov(campaign.lm))$Type
```

```
##                             diff       lwr       upr        p adj
## Email-Billboard          3.9675  2.804077  5.130923 1.118377e-10
## Social Media-Billboard 12.3315 11.168077 13.494923 7.305045e-12
## Social Media-Email       8.3640  7.200577  9.527423 7.305045e-12
```

Comment: Email-Billboard is significant Social Media-Billboard is significant Social Media-Email is significant

```
TukeyHSD(aov(campaign.lm))$Region
```

```
##              diff      lwr      upr        p adj
## Urban-Rural 4.658 3.867599 5.448401 7.310597e-12
```

Comment: Urban- Rural is significant

The Social media has a outstanding effect on score espacially in Urban area, while Billboard (have a least effect) and Email only have a minor contribution on Score, specifically in rural area.