Assignment Part II Shrutee Kadage 48126268

QUESTION 1

The association between the change in retail sales in 243 different cities and the Consumer Confidence Index is being studied by economists. The following variables' information can be found in the sales.csv dataset:

Index: Consumer Confidence Index Sales: Change in retail sales (in percentage)

a)Load the data and create a scatter plot

```
# Load necessary packages
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.3.3
```

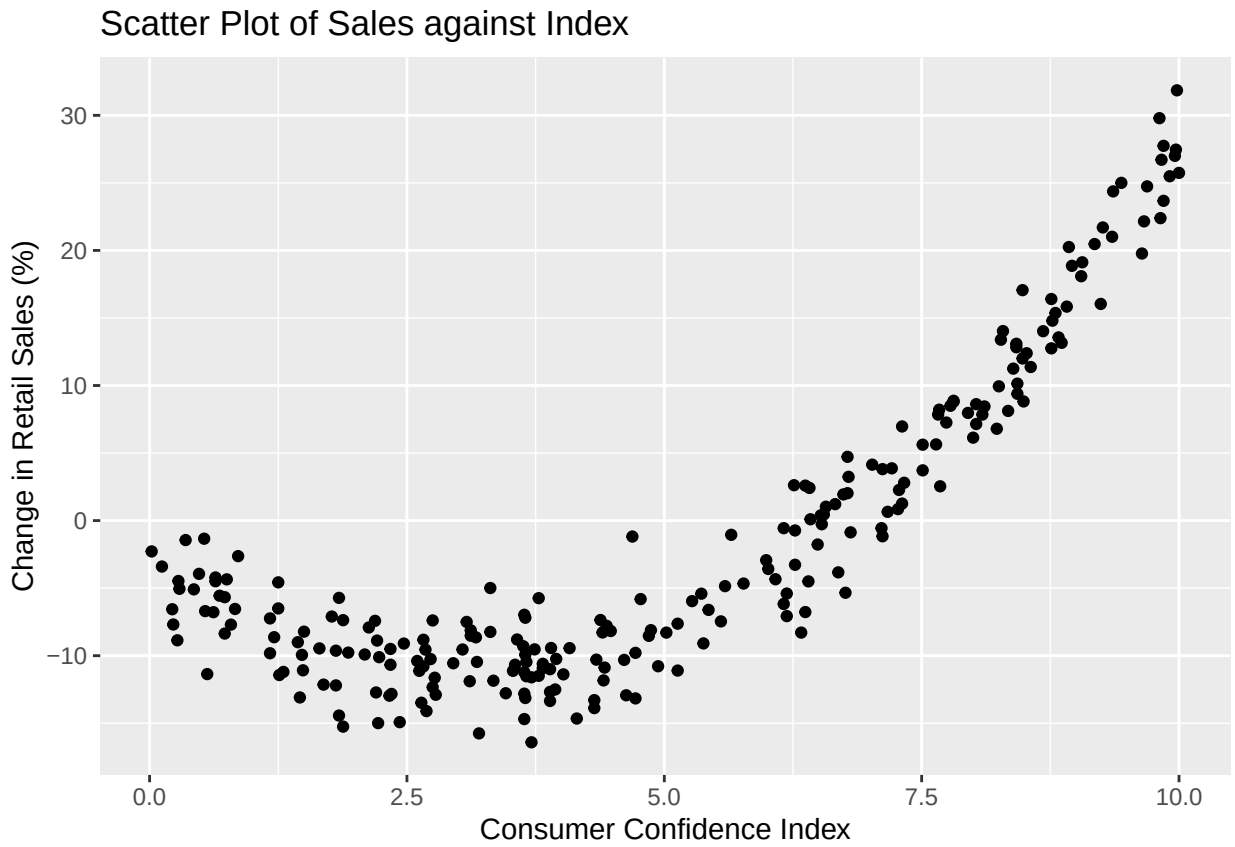
```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
# Load the data from the data folder
sales <- read_csv("data/sales.csv")
```

```
## Rows: 243 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (2): Index, Sales
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Create a scatter plot
ggplot(sales, aes(x = Index, y = Sales)) +
  geom_point() +
  labs(x = "Consumer Confidence Index", y = "Change in Retail Sales (%)") +
  ggtitle("Scatter Plot of Sales against Index")
```



Comment: A positive linear relationship between the change in retail sales and the consumer confidence index appears to exist, according to the scatter plot. Retail sales change typically rises in lockstep with the Consumer Confidence Index.The two have a linear correlation with one another.Both are linearly correlated with each other.

b)Fit a simple linear regression model

```
# Fit simple linear regression model
M1 <- lm(Sales ~ Index, data = sales)

# Diagnostic checks
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
```

Comment:

Retail sales are highly impacted by the Consumer Confidence Index ($p < 0.001$), as confirmed by the model. Sales are boosted by 3.2464 percentage points for every increase in the index unit. Error margin residual: 6.45. The model accounts for 67.72% of the variability in sales. Index and sales have a substantial positive correlation, as indicated by the F-statistic of 505.6 and $p <$ 2.2e-16, which support the model significance.

c)Fit polynomial models of order 2 (M2) and order 3 (M3)

```r
# Fit quadratic model (M2)
M2 <- lm(Sales ~ poly(Index, 2), data = sales)
summary(M2)
```

```
##
## Call:
## lm(formula = Sales ~ poly(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)      -1.5799     0.1611  -9.807   <2e-16 ***
## poly(Index, 2)1 145.0378     2.5112  57.757   <2e-16 ***
## poly(Index, 2)2  92.2701     2.5112  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
```

Comment:

Overtaking the linear model is the polynomial model. -1.5799 is the intercept. High significance ($p <$ 2e-16) is found for both polynomial terms. Compared to the linear model (6.45), the residual standard error (2.511) indicates a superior match. Significant improvement is seen by the multiple R-squared of 0.9513, which accounts for 95.13% of sales fluctuation. Strong significance of the model: $p <$ 2.2e-16, F-statistic: 2343.

```r
# Fit cubic model (M3)
M3 <- lm(Sales ~ poly(Index, 3), data = sales)
summary(M3)
```

```
##
## Call:
## lm(formula = Sales ~ poly(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)      -1.5799     0.1614  -9.788   <2e-16 ***
## poly(Index, 3)1 145.0378     2.5162  57.641   <2e-16 ***
## poly(Index, 3)2  92.2701     2.5162  36.670   <2e-16 ***
## poly(Index, 3)3  -0.4874     2.5162  -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
```

Comment:

With an intercept of -1.5799, the quadratic model is still strong. The initial two polynomial terms exhibit strong significance ($p < 2e\text{-}16$), however the cubic component is not significant ($p = 0.847$). In comparison to the quadratic model, the residual standard error remains almost constant at 2.516. Multiple R-squared stays at 0.9513; no additional explanatory power is shown with the cubic term. With an F-statistic of 1556, the model is still considered highly significant overall ($p < 2.2e\text{-}16$).

Comparison :

With reduced residual standard error and better R-squared values, the polynomial models (M2 and M3) significantly outperform the linear model (M1). But in contrast to the statistically better quadratic model (M2), the cubic factor in M3 is not significant and provides no additional explanatory power.

d)Plot the data and add predicted lines from M1, M2, and M3

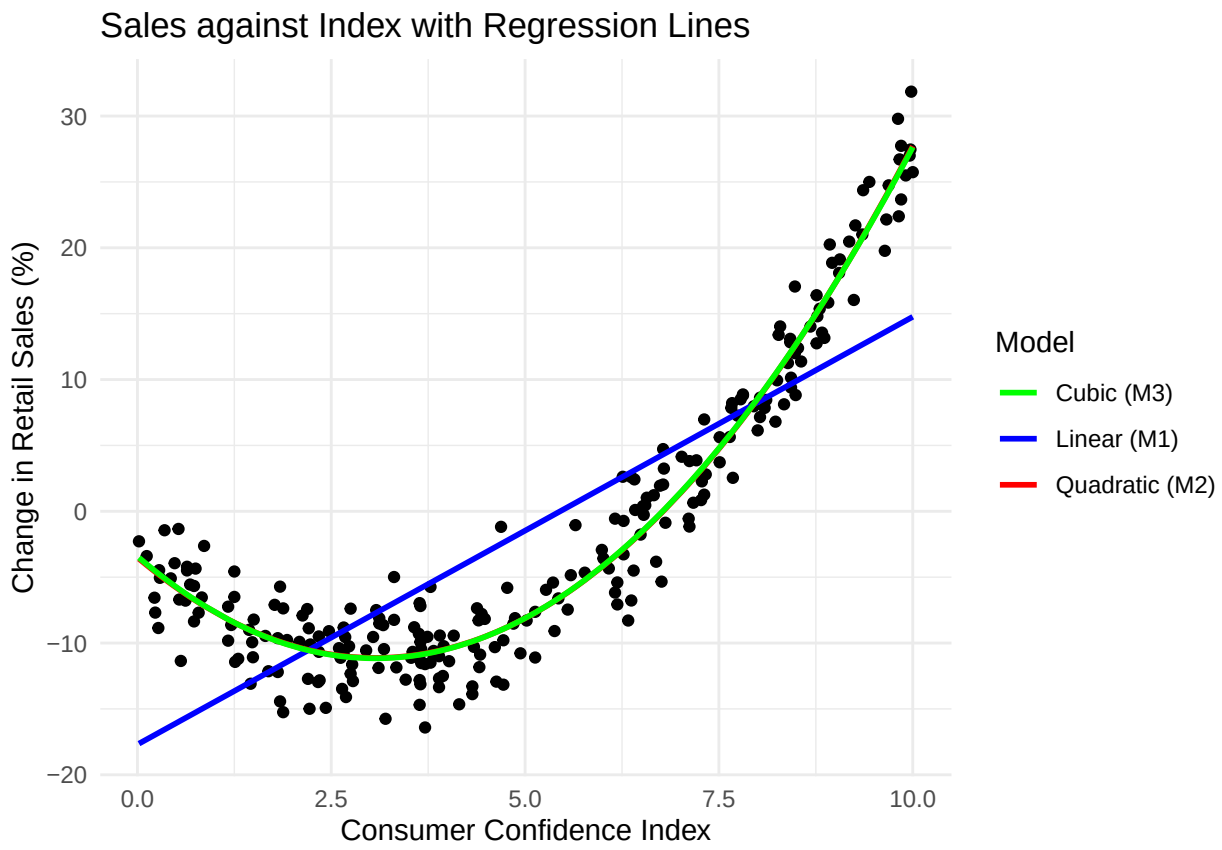
```r
# Create scatter plot
# Load necessary libraries
library(ggplot2)
library(readr)

# Load the data
data <- read_csv("data/sales.csv")
```

```
## Rows: 243 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (2): Index, Sales
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# Fit models
M1 <- lm(Sales ~ Index, data = data)
M2 <- lm(Sales ~ poly(Index, 2), data = data)
M3 <- lm(Sales ~ poly(Index, 3), data = data)

# Generate predictions
```

```r
data$pred_M1 <- predict(M1, newdata = data)
data$pred_M2 <- predict(M2, newdata = data)
data$pred_M3 <- predict(M3, newdata = data)

# Plot the data and predicted lines
ggplot(data, aes(x = Index, y = Sales)) +
  geom_point(color = "black") +
  geom_line(aes(y = pred_M1, color = "Linear (M1)"), size = 1) +
  geom_line(aes(y = pred_M2, color = "Quadratic (M2)"), size = 1) +
  geom_line(aes(y = pred_M3, color = "Cubic (M3)"), size = 1) +
  labs(title = "Sales against Index with Regression Lines",
       x = "Consumer Confidence Index",
       y = "Change in Retail Sales (%)") +
  scale_color_manual(name = "Model",
                     values = c("Linear (M1)" = "blue",
                                "Quadratic (M2)" = "red",
                                "Cubic (M3)" = "green")) +
  theme_minimal()
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```



Comment:

1) Blue Line(Linear Model M1) - It represents the linear relationship between the Consumer Confidence Index and the Change in Retail Sales.

2)Red Line (Quadratic Model M2) - The red line and green line are overlapping.

It means that the predictions from M3 are very close to those from M2.When the cubic component does not significantly increase the explanatory power over the quadratic term, this can occur.In the range of the data, it implies that the quadratic term could not be significantly improving upon the linear model.

3)Green Line (Cubic Model M3) - An even more flexible relationship is made possible by this dotted line. The red dashed line is closely followed by this line because the cubic term does not significantly increase the explanatory power.

e)Assess the significance of terms in M3 using Sequential Sum of Squares

```
# Sequential ANOVA
anova(M3)
```

```
## Analysis of Variance Table
##
## Response: Sales
##                  Df  Sum Sq Mean Sq F value    Pr(>F)
## poly(Index, 3)    3 29550.0  9850.0  1555.8 < 2.2e-16 ***
## Residuals       239  1513.2     6.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comment:

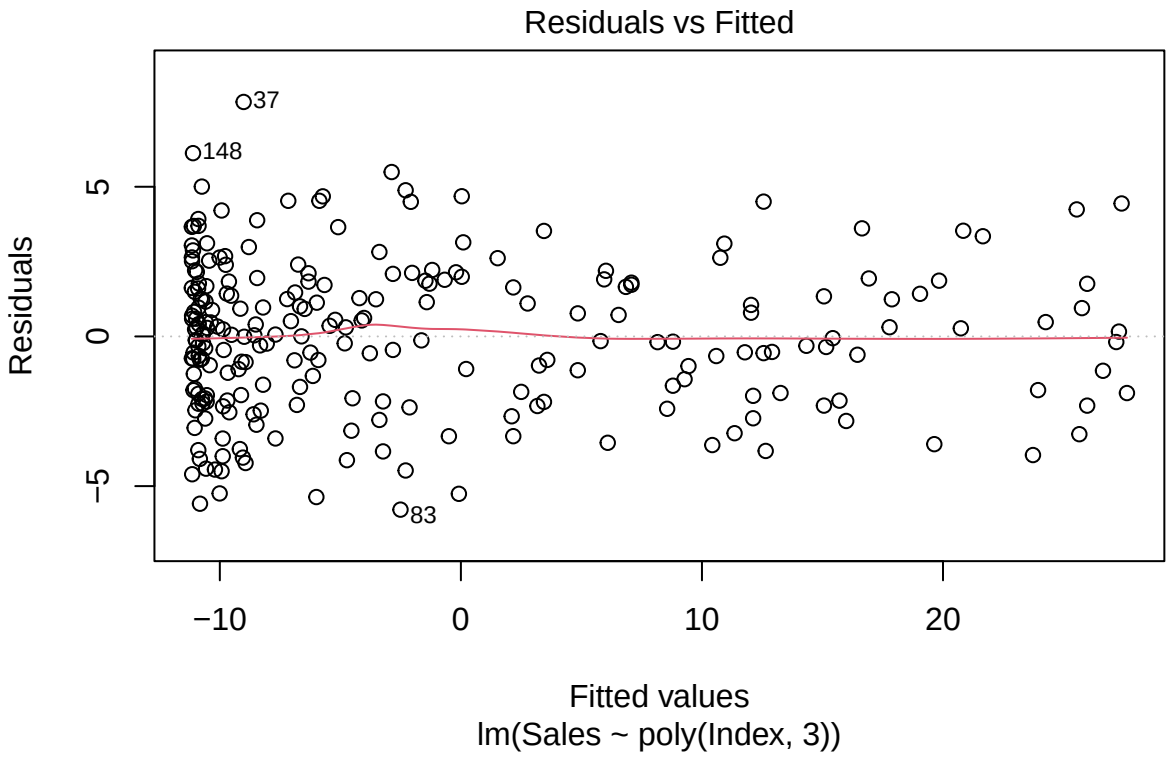
Given that the cubic components substantially contribute to the model, as demonstrated by the sequential ANOVA, the cubic polynomial model (M3) is a good choice to explain the relationship between the Change in Retail Sales and the Consumer Confidence Index. The ideal balance between model complexity and explanatory power must be ensured, notwithstanding the high relevance, by additional model validation and comparison with simpler models (such as M1 and M2).

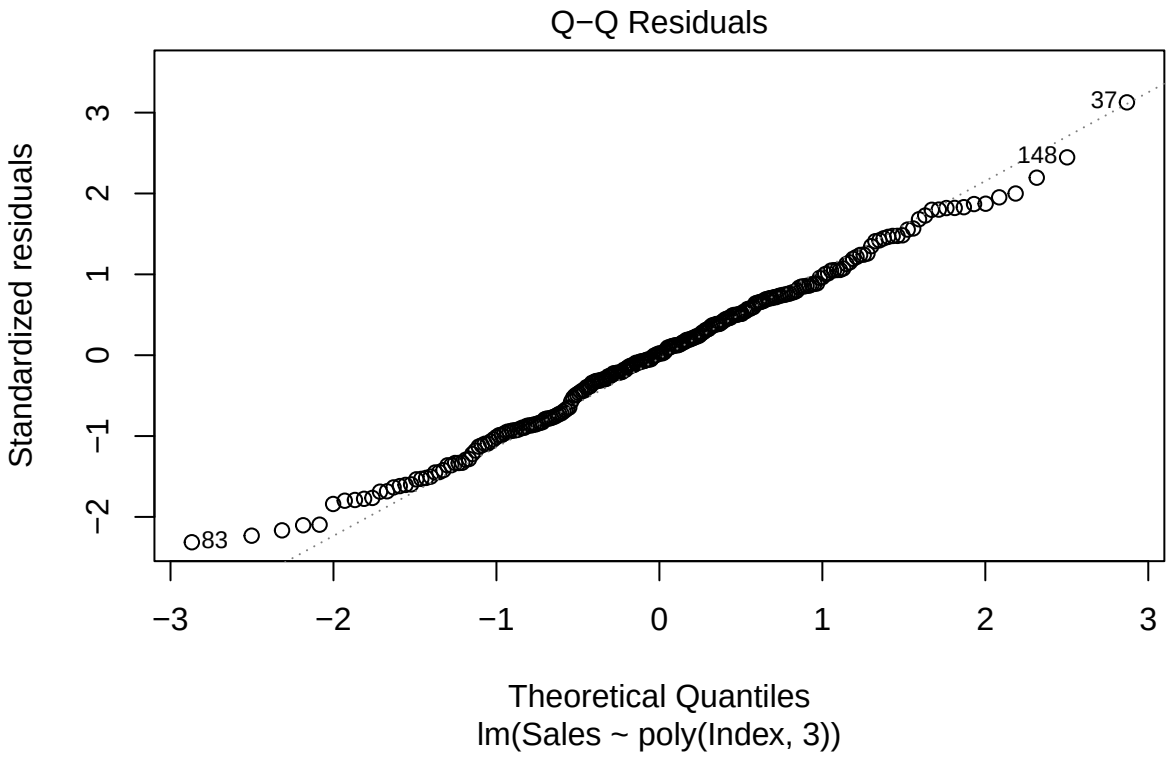
f)Choose the best model among M1, M2, and M3 and validate it

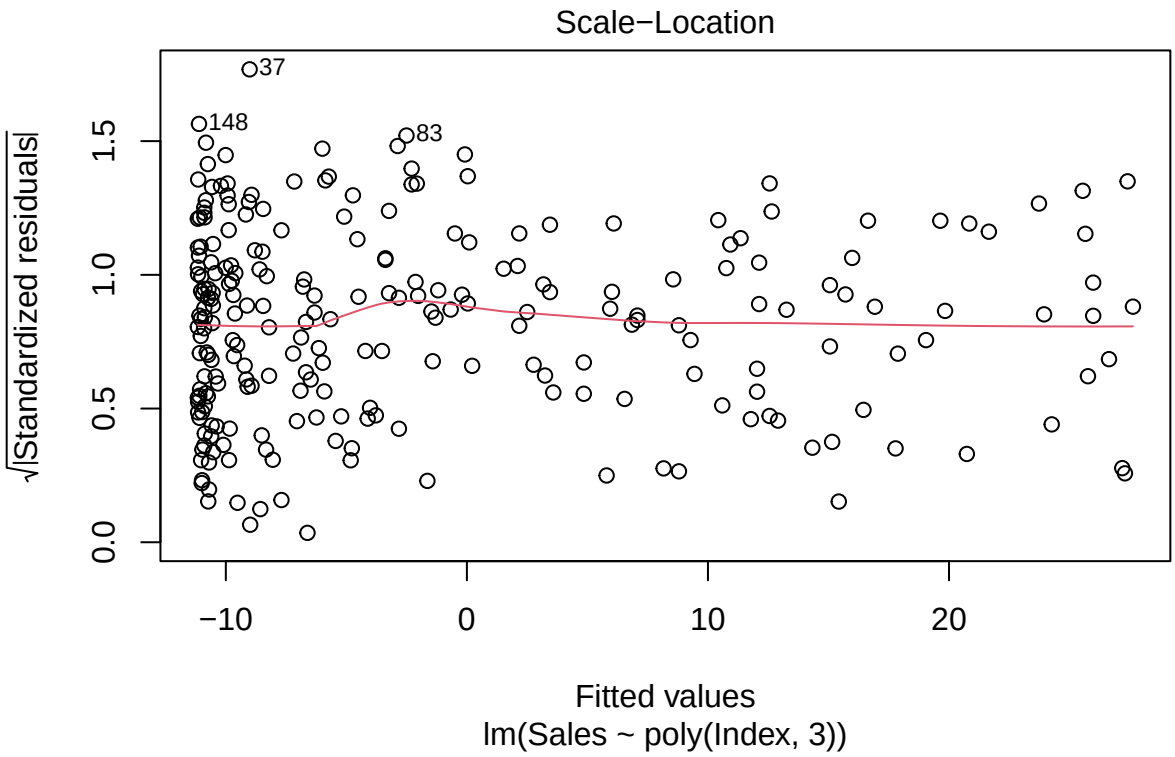
```
# Model validation
# We choose the model with the highest R-squared value as the best model

# R-squared values
rsquared <- c(summary(M1)$r.squared, summary(M2)$r.squared, summary(M3)$r.squared)

# Best model
best_model <- c("M1", "M2", "M3")[which.max(rsquared)]

# Print best model
print(paste("Best model:", best_model))
```

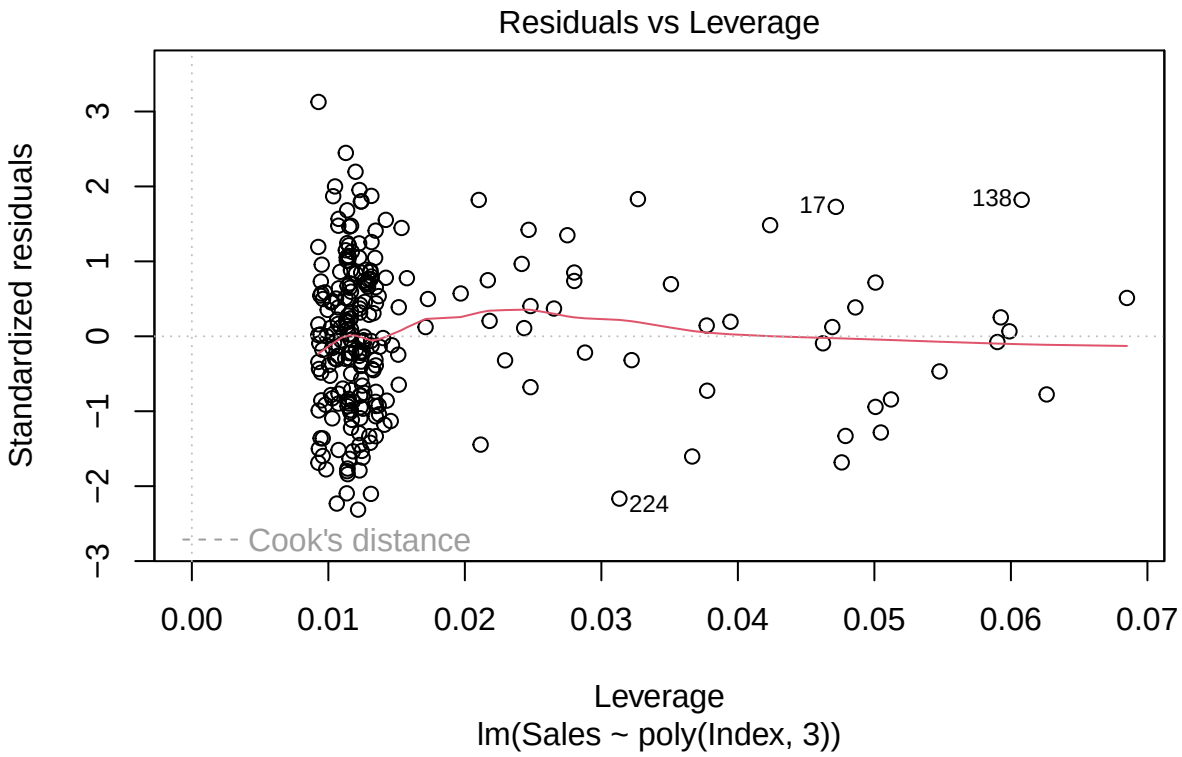
```
## [1] "Best model: M3"
```

```
# Validate best model (e.g., diagnostic plots)
plot(M3)
```


Residuals vs Fitted
37
148
83
Residuals
5
0
−5
−10
0
10
20
Fitted values
lm(Sales ~ poly(Index, 3))


Q−Q Residuals
Standardized residuals
−2
−1
0
1
2
3
83
148
37
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(Sales ~ poly(Index, 3))


Scale−Location
37
148
83
√|Standardized residuals|
0.0
0.5
1.0
1.5
−10
0
10
20
Fitted values
lm(Sales ~ poly(Index, 3))



Comment:

Based on the R-squared values of the three models (M1, M2, and M3), the code compares them to get the best fit. With the highest R-squared value, the model known as M3 is the best. For M3, diagnostic plots are produced in order to verify the fit and look for any possible problems, including heteroscedasticity, non-linearity, or outliers.

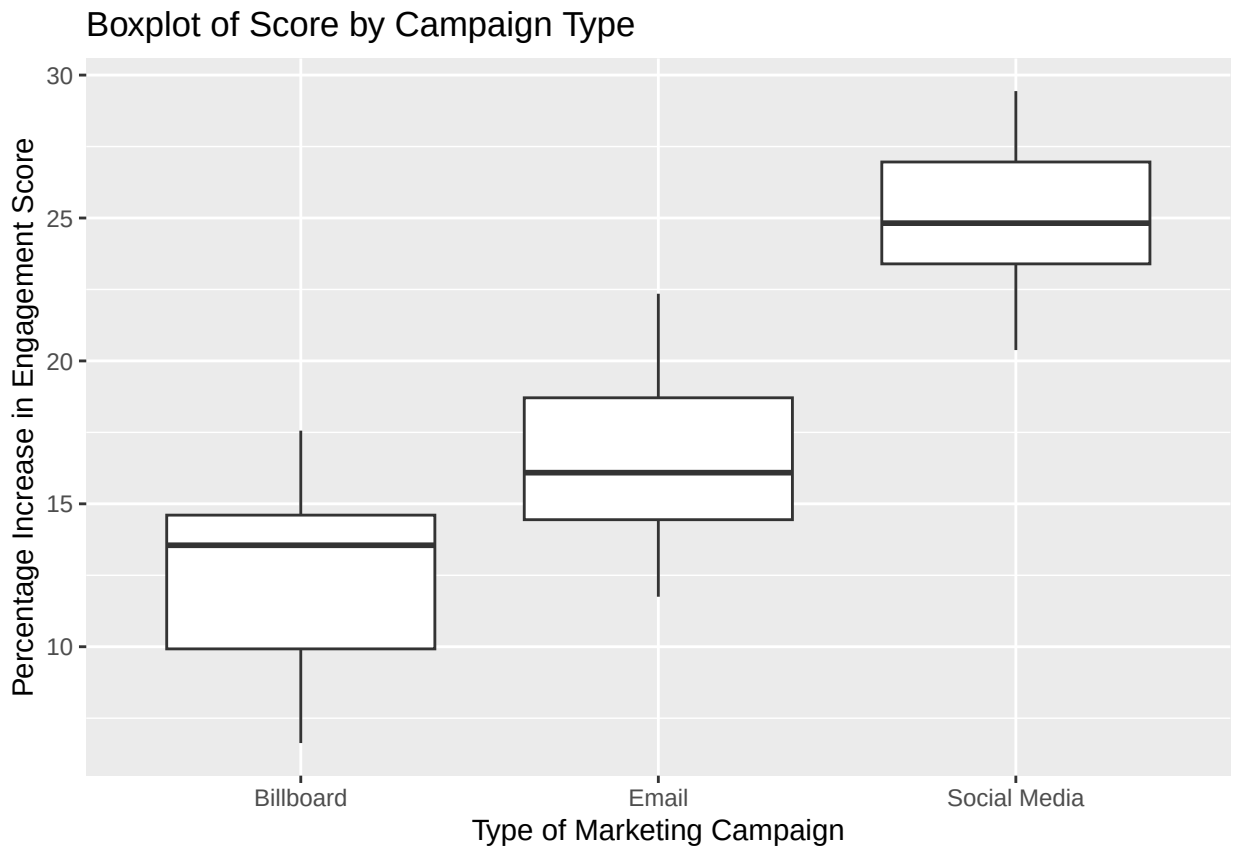
QUESTION 2

a)Construct two different preliminary graphs

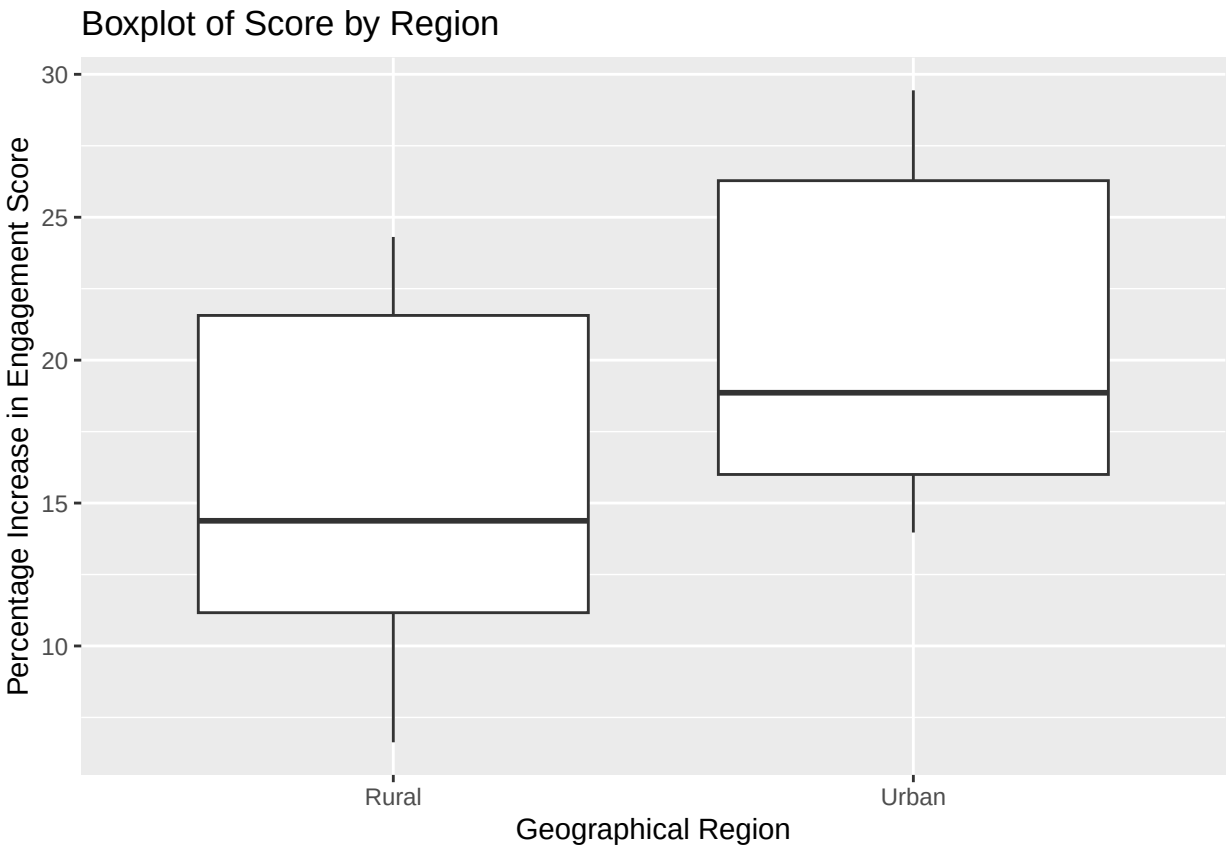
```
# Load necessary packages
library(readr)
library(ggplot2)

# Load the data
campaign <- read_csv("data/campaign.csv")
```

```
## Rows: 60 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): Region, Type
## dbl (1): Score
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# Preliminary graph 1: Boxplot of Score by Type
ggplot(campaign, aes(x = Type, y = Score)) +
  geom_boxplot() +
  labs(x = "Type of Marketing Campaign", y = "Percentage Increase in Engagement Score") +
  ggtitle("Boxplot of Score by Campaign Type")
```




```r
# Preliminary graph 2: Boxplot of Score by Region
ggplot(campaign, aes(x = Region, y = Score)) +
  geom_boxplot() +
  labs(x = "Geographical Region", y = "Percentage Increase in Engagement Score") +
  ggtitle("Boxplot of Score by Region")
```



Comment:

The first graph compares the engagement score distribution across different types of marketing campaigns, while the second graph looks into the engagement rating distribution across different geographic locations. These figures provide some initial understanding of how campaign kind and location impact engagement levels.

b) Full interaction model

The full interaction model:

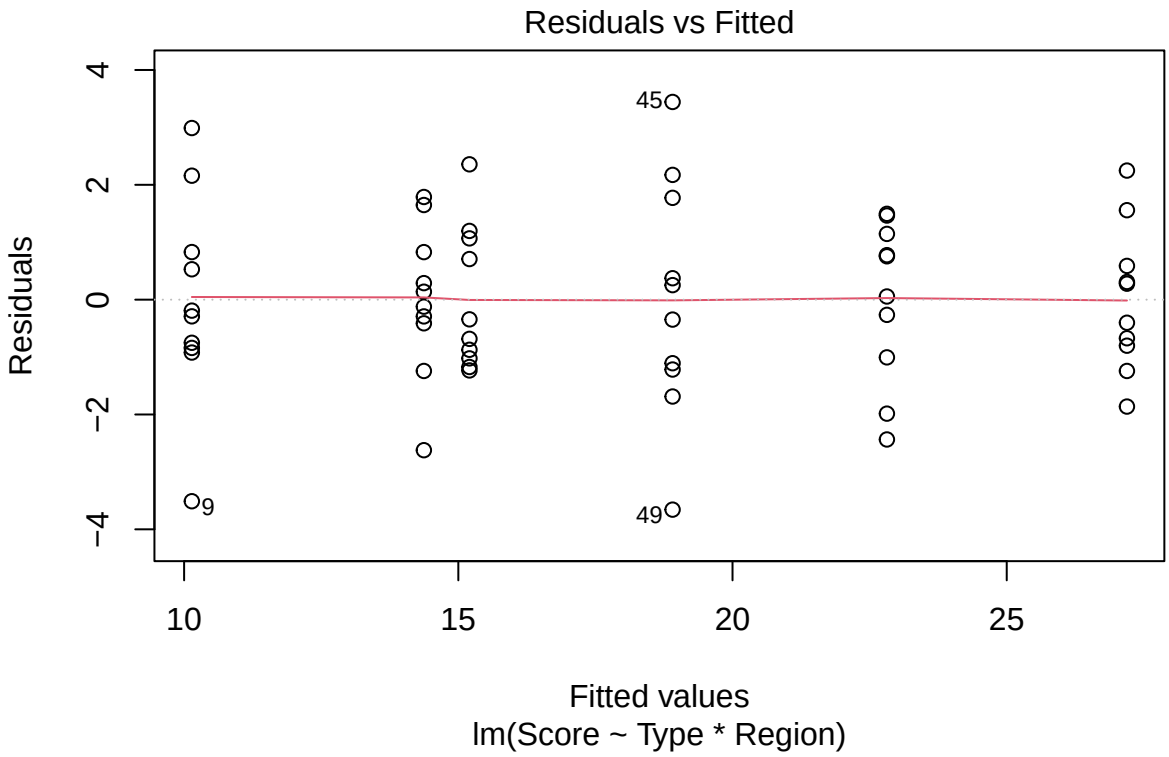
$$Score = \beta_0 + \beta_1 \cdot \text{Type} + \beta_2 \cdot \text{Region} + \beta_3 \cdot \text{Type} \cdot \text{Region} + \epsilon$$

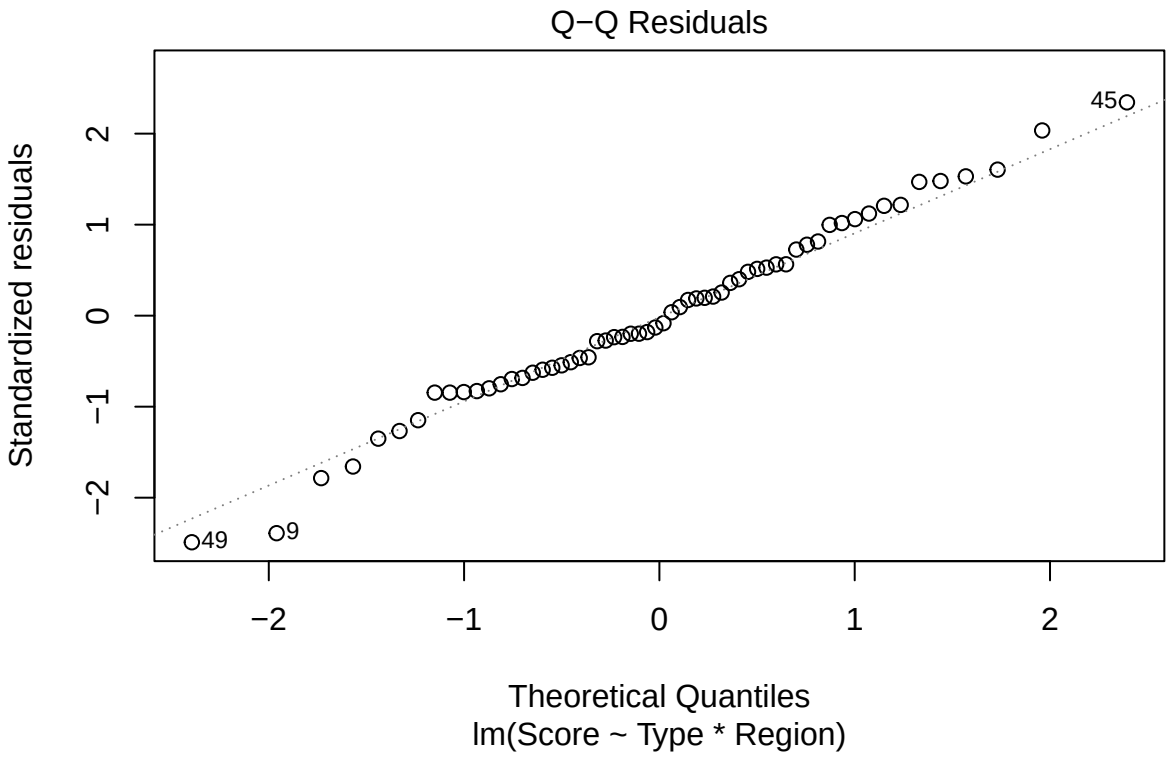
Where:

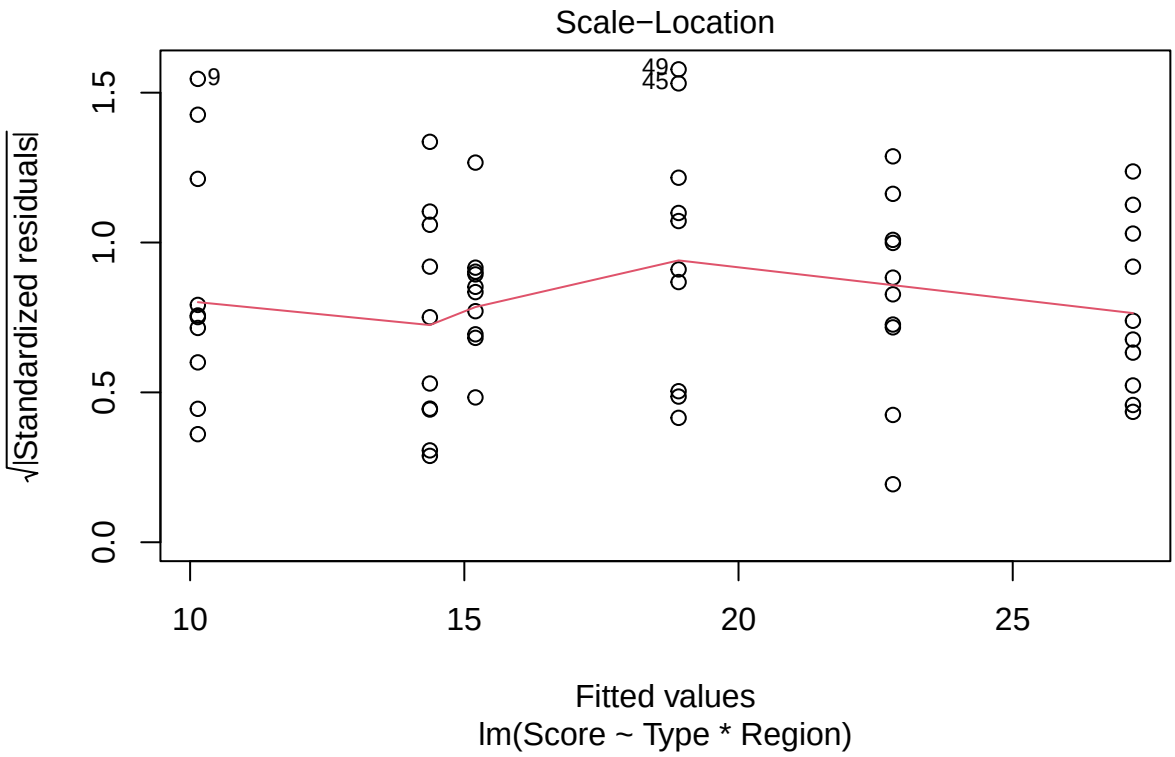
- β_0 is the intercept term,
- β_1 is the coefficient for the effect of campaign Type,
- β_2 is the coefficient for the effect of Region,
- β_3 is the coefficient for the interaction effect between Type and Region,
- ϵ is the error term.

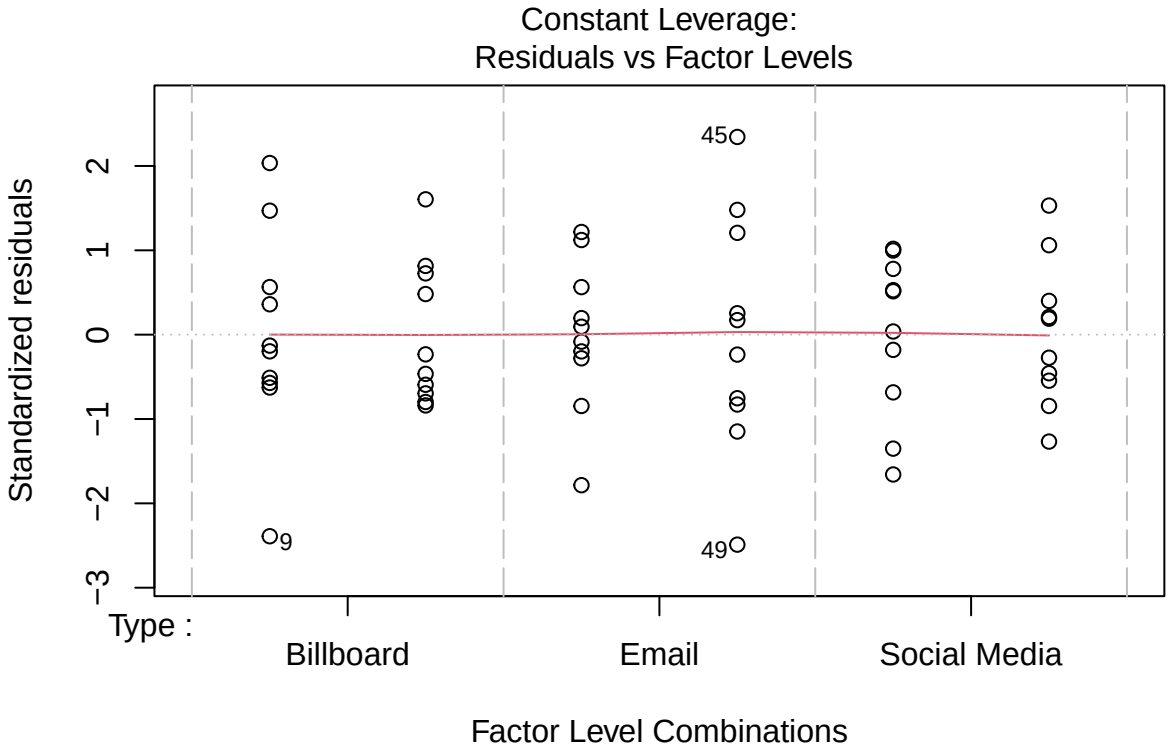
c)Analyze the data for the effect of Type and Region on Score

```
# Fit interaction model
interaction_model <- lm(Score ~ Type * Region, data = campaign)

# Check assumptions (model diagnostics)
plot(interaction_model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Score ~ Type * Region)
45
9
49
10
15
20
25
−4
−2
0
2
4


Q−Q Residuals
Standardized residuals
−2
−1
0
1
2
45
49
9
−2
−1
0
1
2
Theoretical Quantiles
lm(Score ~ Type * Region)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
9
49
45
10
15
20
25
Fitted values
lm(Score ~ Type * Region)




```
# Interpret the results
summary(interaction_model)
```

```
##
## Call:
## lm(formula = Score ~ Type * Region, data = campaign)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6570 -0.9420 -0.1565  0.8885  3.4430
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   10.1410     0.4897  20.710  < 2e-16 ***
## TypeEmail                      4.2310     0.6925   6.110 1.14e-07 ***
## TypeSocial Media              12.6740     0.6925  18.302  < 2e-16 ***
## RegionUrban                    5.0620     0.6925   7.310 1.29e-09 ***
## TypeEmail:RegionUrban         -0.5270     0.9794  -0.538    0.593
## TypeSocial Media:RegionUrban  -0.6850     0.9794  -0.699    0.487
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.548 on 54 degrees of freedom
## Multiple R-squared:  0.9366, Adjusted R-squared:  0.9307
## F-statistic: 159.5 on 5 and 54 DF,  p-value: < 2.2e-16
```

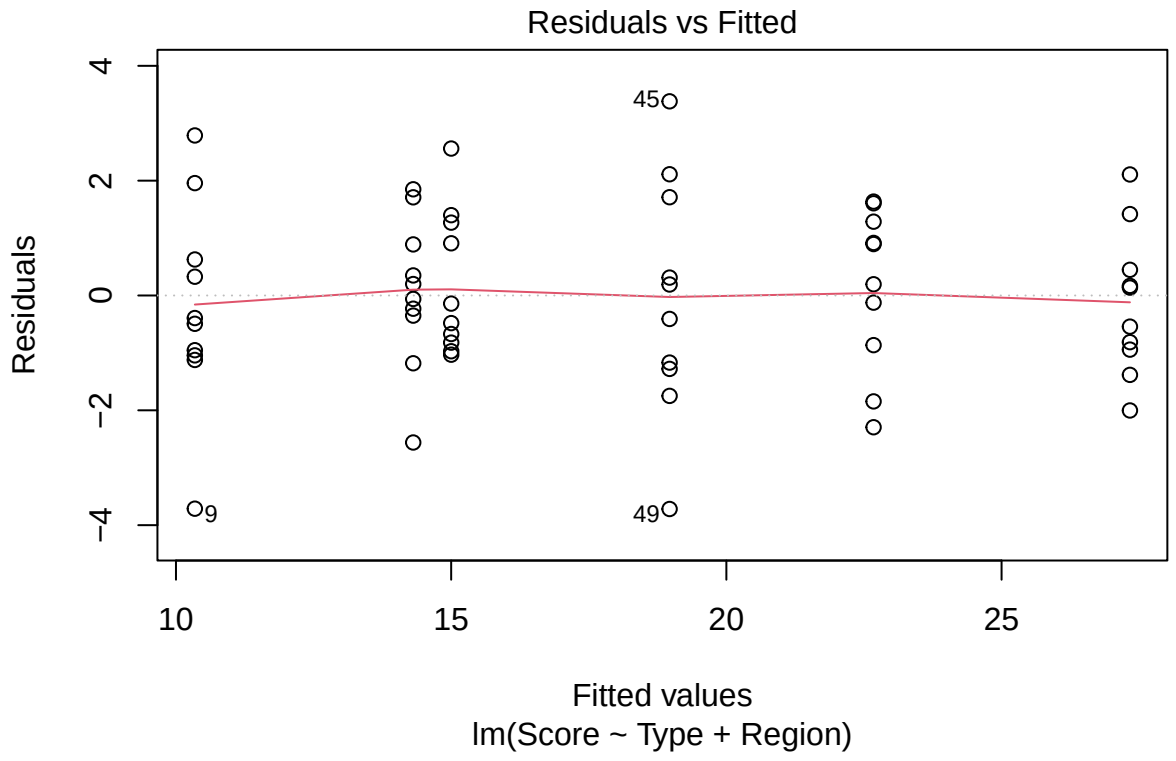
Comment:

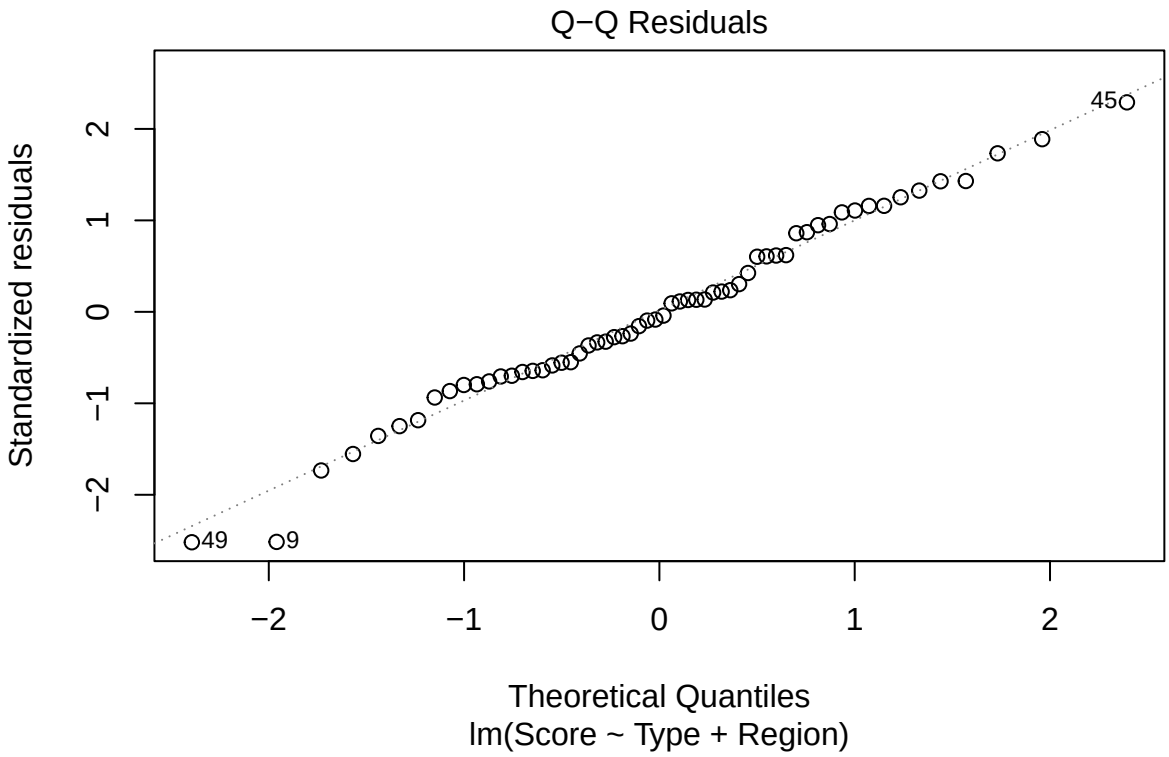
With notable effects for Type (email, social media) and Region (urban), the interaction model predicts Score based on Type and Region. Terms that interact are negligible. With a well-fitting Multiple R-squared of 0.9366, the model accounts for 93.07% of Score variability. There is a reasonable distribution of residuals (Residual Standard Error: 1.548). The model is statistically significant overall ($p < 2.2e\text{-}16$, F-statistic: 159.5).

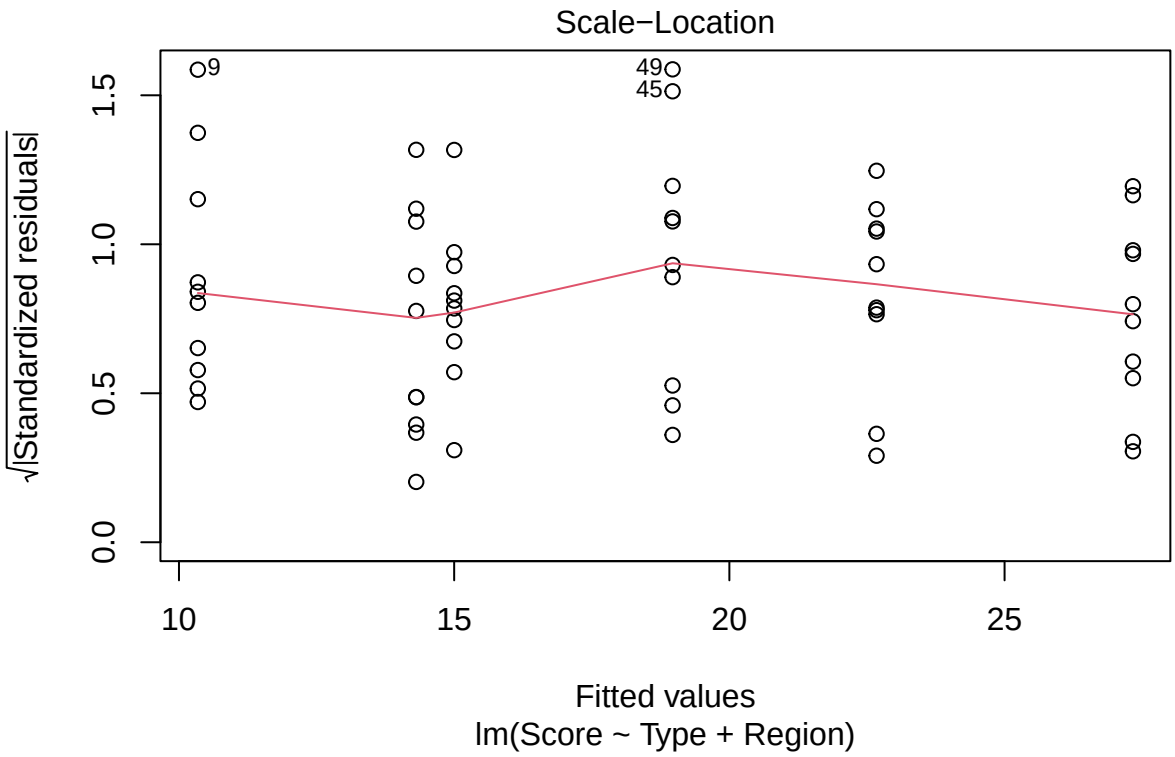
Null Hypotheses: • H0: There is no significant interaction effect between Type and Region on Score. • H0: There is no significant effect of Type on Score. • H0: There is no significant effect of Region on Score. Alternative Hypotheses: • H1: There is a significant interaction effect between Type and Region on Score. • H1: There is a significant effect of Type on Score. • H1: There is a significant effect of Region on Score.

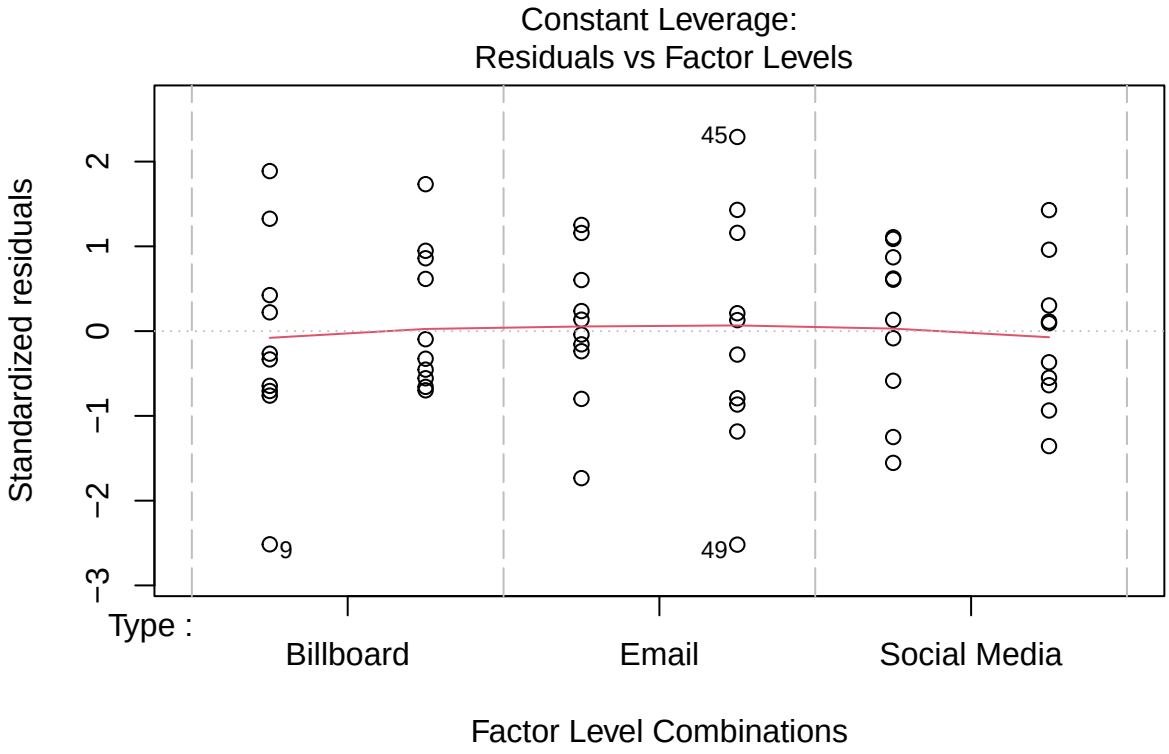
d)Repeat the analysis for main effects

```r
# Fit main effects model
main_effects_model <- lm(Score ~ Type + Region, data = campaign)

# Check assumptions (model diagnostics)
plot(main_effects_model)
```




Q−Q Residuals
Standardized residuals
−2
−1
0
1
2
45
49
9
−2
−1
0
1
2
Theoretical Quantiles
lm(Score ~ Type + Region)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
9
49
45
10
15
20
25
Fitted values
lm(Score ~ Type + Region)




```
# Interpret the results
summary(main_effects_model)
```

```
##
## 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
```

Comment:

Significant effects are found for Type (Email, Social Media) and Region (Urban) in the main effects model, which looks at the individual influences of Type and Region on Score. With a Multiple R-squared of 0.9359, the model accounts for 93.25% of the variability in scores. With a residual standard error of 1.528, residuals seem to be fairly dispersed. The model is statistically significant overall ($p < 2.2e\text{-}16$, F-statistic: 272.7).

Null Hypotheses: • H0: There is no significant effect of Type on Score. • H0: There is no significant effect of Region on Score. Alternative Hypotheses: • H1: There is a significant effect of Type on Score. • H1: There is a significant effect of Region on Score.

e)Multiple comparisons using TukeyHSD

```
# Check if the design is balanced
table(campaign$Type, campaign$Region)
```

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

```
# Perform Tukey's HSD test
tukey_result_type <- TukeyHSD(aov(Score ~ Type, data = campaign))
print(tukey_result_type)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Score ~ Type, data = campaign)
##
## $Type
##                             diff       lwr       upr     p adj
## Email-Billboard         3.9675  1.814604  6.120396 0.0001247
## Social Media-Billboard 12.3315 10.178604 14.484396 0.0000000
## Social Media-Email      8.3640  6.211104 10.516896 0.0000000
```

```
tukey_result_region <- TukeyHSD(aov(Score ~ Region, data = campaign))
print(tukey_result_region)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Score ~ Region, data = campaign)
##
## $Region
##              diff     lwr     upr     p adj
## Urban-Rural 4.658 1.84685 7.46915 0.0015755
```

Comment:

The campaign Type and Region levels that substantially differ in their influence on engagement scores are identified by the Tukey's HSD test results. This lets us assess how well various campaign kinds work together and how different geographic areas affect participation levels. To ensure that the Tukey's HSD test is conducted properly, make sure the design is balanced.