Assignment Part 2

Arian Kalantari (48549819)

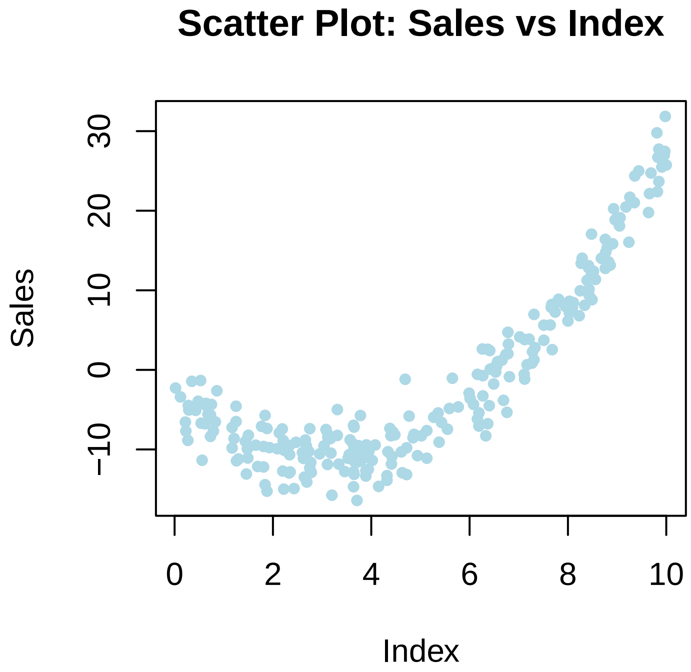
2024-05-17

Question 1

Economists are exploring the relationship between Consumer Confidence Index and the change in retail sales in 243 different cities. The dataset saldsdses.csv contains information about the following variables.

1.a) Load the data from the file sales.csv and create a scatter plot of Sales against Index.

```r
sales <- read.csv('data/sales.csv', header=TRUE)
plot(Sales~Index, data=sales, pch = 20, col = 'light blue',
     main='Scatter Plot: Sales vs Index')
```

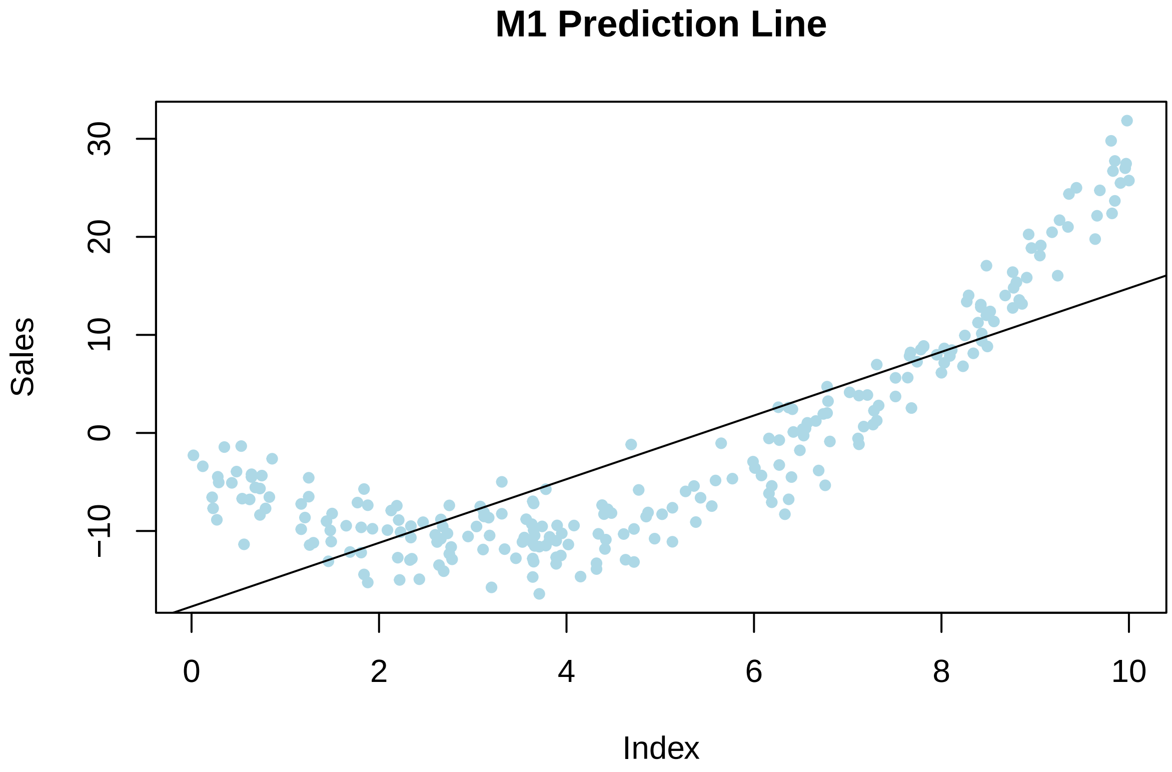


Comment: We can not see linear relationship between 2 x variables. However there seems to be a non-linear relationship that is quadratic or higher order.

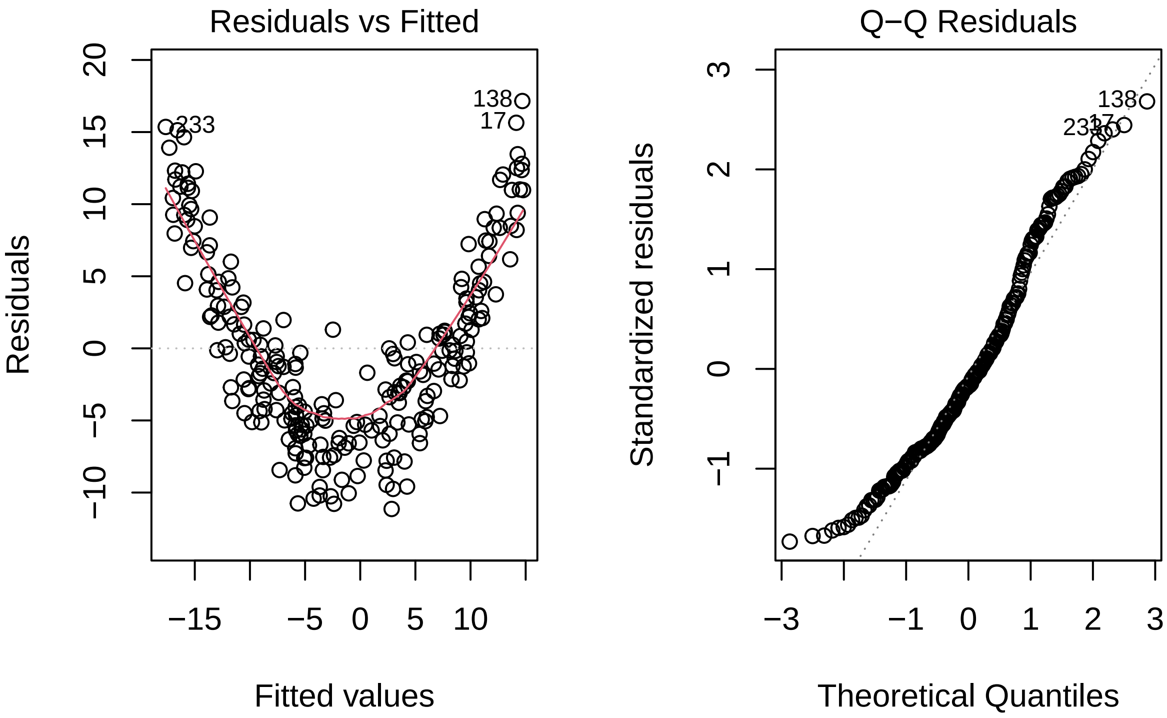
1.b) Fit a simple linear regression model and name it as M1 to predict Sales using Index. Validate the model through diagnostic checks and comment.

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

plot(Sales~Index, data=sales, pch = 20, col = 'light blue', main='M1 Prediction Line')
abline(M1)
```



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



```r
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
```

M1 Diagonostics

- QQ plot of the residuals look slightly off from normality at the beginning and the end of the plot, however we can say this is close to normality, suggesting the **normality assumption for residuals is appropriate**.

- The main problem is the residual vs fitted plot. We can see a quadratic trend in residuals, suggesting **quadratic or higher order fits better**.
- Having plotted the fitted line on the scatter plot, we can see that the **simple linear regression is far accurate from the observed sales**.
- We can see from the summary where the **R-squared:0.6772** as well as **Adjusted R-squared:0.6759**, suggesting that the predicted values from simple linear regression are far from the observed values. Moreover, this means that nearly 67.5% of the variation in Sales is explained by this simple linear regression model.

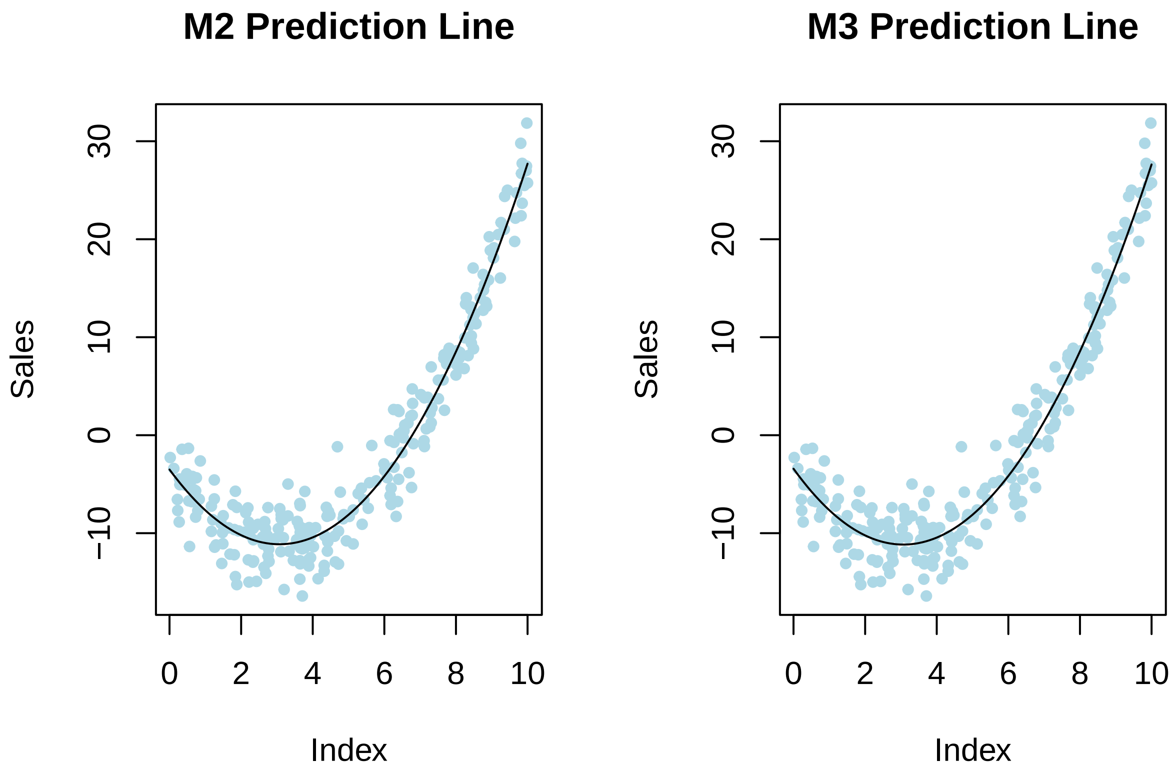
1.c) Fit two polynomial models of order 2 and order 3 to predict Sales using Index. Name the quadratic model as M2 and the cubic model as M3.

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

# M2 Predictions
grid <- seq(from = 0, to = 10, by = 0.1)
m2_dat <- data.frame(Index = grid)
pred_m2 <- predict.lm(M2, newdata = m2_dat, interval = 'prediction')
```

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

# M3 Predictions, we can re-use the grid.
m3_dat <- data.frame(Index = grid)
pred_m3 <- predict.lm(M3, newdata = m3_dat, interval = 'prediction')
```



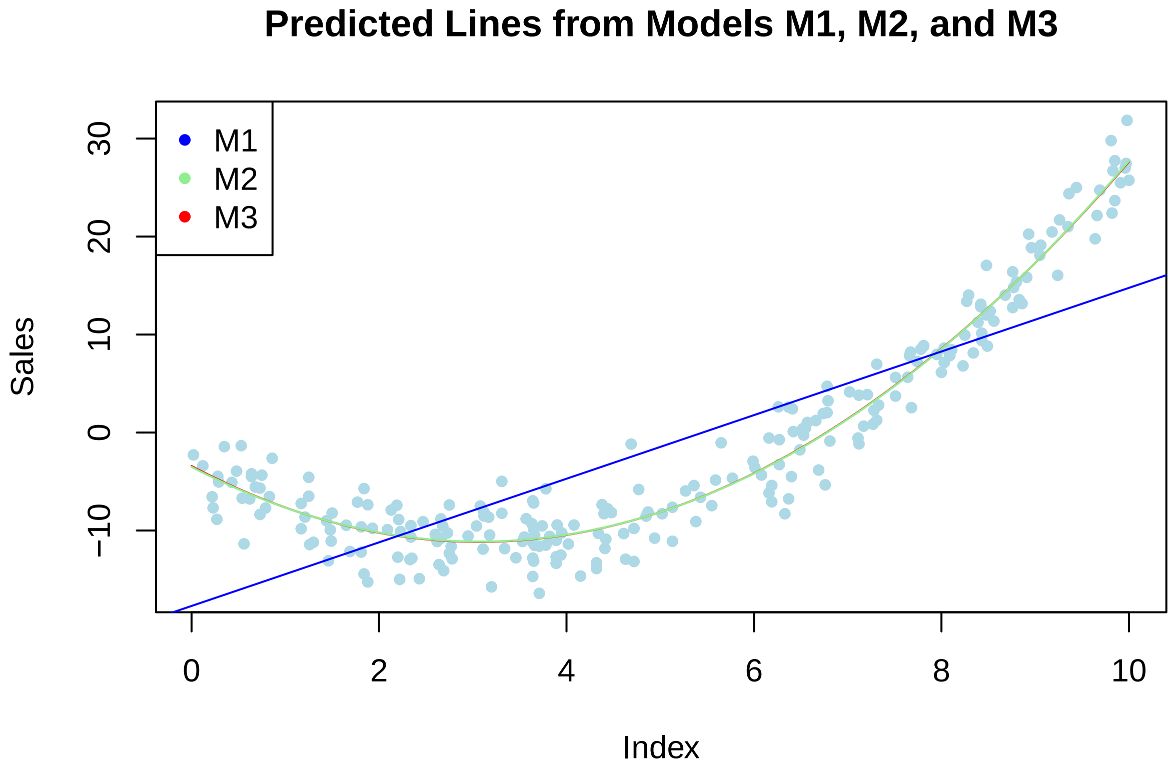
Comment: M2 and M3 resemble each other, perhaps there is no much difference between 2 x models. Their prediction lines fit to the observed values (data points) very well, much better than simple linear model (M1).

We will validate the fit of the models in next section.

1.d) Plot the data and add the three predicted lines from models M1, M2, and M3 to your plot.

```
plot(Sales~Index, data=sales, pch = 20, col = 'light blue',
     main='Predicted Lines from Models M1, M2, and M3')
lines(grid, pred_m3[,1], lty = 1, col = 'red')
lines(grid, pred_m2[,1], lty = 1, col = 'light green')
abline(M1, col='blue')

legend("topleft", legend = c("M1","M2", "M3"), col = c('blue', 'light green', 'red'), pch= 20)
```



By simply looking at the predicted lines, we can see that M1 has the least fit. M2 and M3 are nearly identical. We will check the model summaries.

```
#anova(M1) #As I want to also comment on R-Squared Score, for M1 I am just using summary(M1)
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
```

```
#anova(M2)
summary(M2)
```

```
##
## Call:
## lm(formula = Sales ~ I(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 ***
## I(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
```

```
summary(M3)
```

```
##
## Call:
## lm(formula = Sales ~ I(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 ***
## I(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
```

Findings from the sequential sum of squares:

- **M1:**
  - R-squared: 0.6772
  - Adjusted R-squared: 0.6759
  - Linear term (Index) is statistically significant.
  - Residual standard error is bigger compared to other models (6.45)
- **M2:**
  - R-squared: 0.9513
  - Adjusted R-squared: 0.9509
  - Linear term (Index) and quadratic term (Index^2) are both statistically significant.
  - Residual standard error is much lower than M1 (2.511)
- **M3:**
  - R-squared: 0.9513
  - Adjusted R-squared: 0.9507
  - Cubic term (Index^3) is ***not*** statistically significant.
  - Residual standard error is close to M2 (2.516)

Summary Comment:

M2 and M3 have almost same goodness of fit, but M2 is slightly better fit as its adjusted R-squared is greater than M3's. Nearly 95.1% of the variation in Sales is explained by M2.

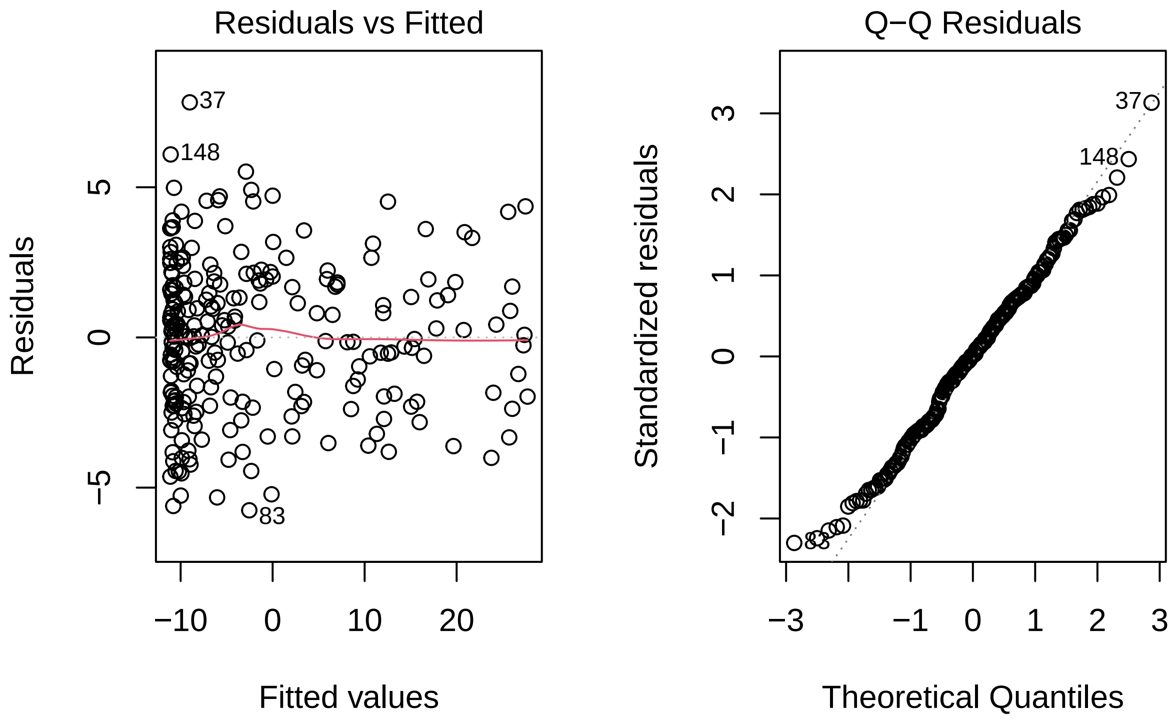
1.f) Choose the best model among M1, M2, and M3 and validate it.

```
anova(M3)
```

```
## Analysis of Variance Table
##
## Response: Sales
##             Df  Sum Sq Mean Sq   F value Pr(>F)
## I(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
```

As it's stated in the previous section, cubic term (Index^3) does not seem significant. We can see on the ANOVA table above that the F-Value for the cubic term is very low at 0.0375, and P-Value: 0.8466 > 0.05, suggesting that the cubic term is not necessary for model build. Therefore, we can check if M2 is a good model or not next.

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



Comment:

QQ-plot of the residuals is not far from normality, suggesting the normality assumption is met. Residuals vs Fitted plot does not show curvature pattern like M1, moreover the variance seems close to constant between -6 and 6, excluding the outlier. Therefore, we can conclude that the M2 model is the most approprate model out of three. This is clear when we commented on the goodness of fit in early section.

Question 2

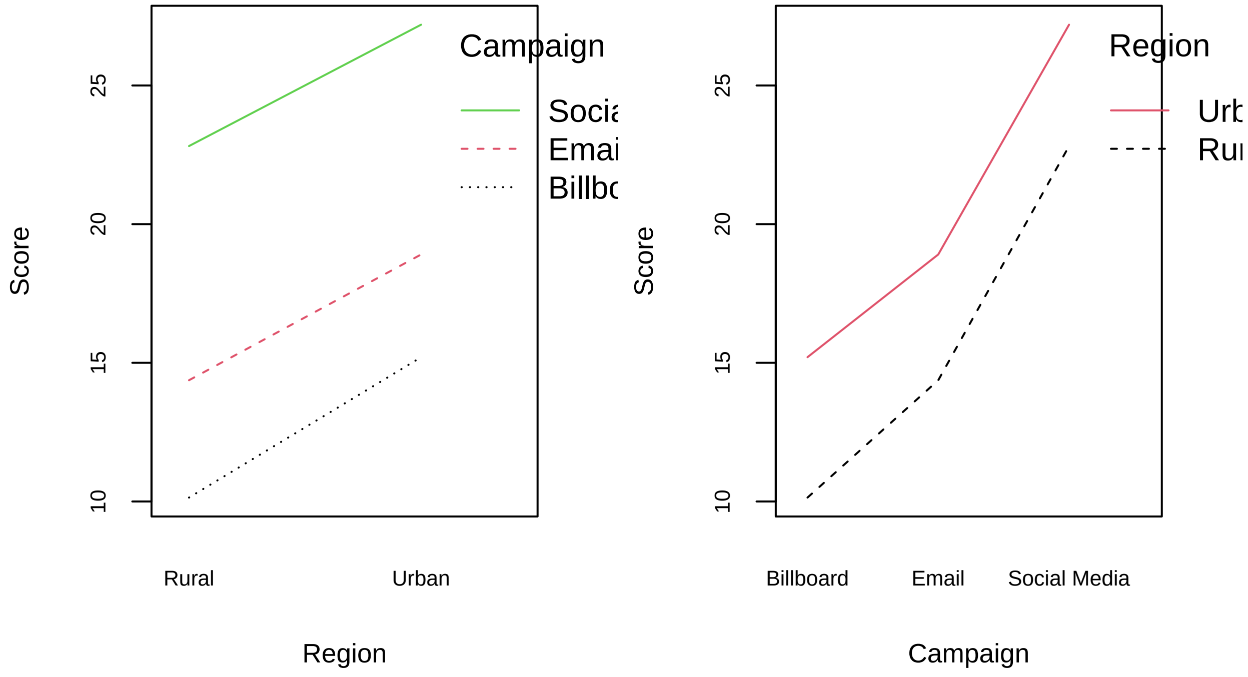
In the rapidly evolving landscape of marketing, companies are constantly seeking strategies to capture consumer attention and foster brand loyalty. The effectiveness of these marketing campaigns can be influenced by a myriad of factors, including the medium of the campaign and the geographical region in which it is deployed. Your task is to determine the impact of different marketing campaign types on customer engagement across various regions. The dataset campaign.csv contains information about the following variables.

```r
campaign <- read.csv('data/campaign.csv', header=TRUE)
head(campaign)
```

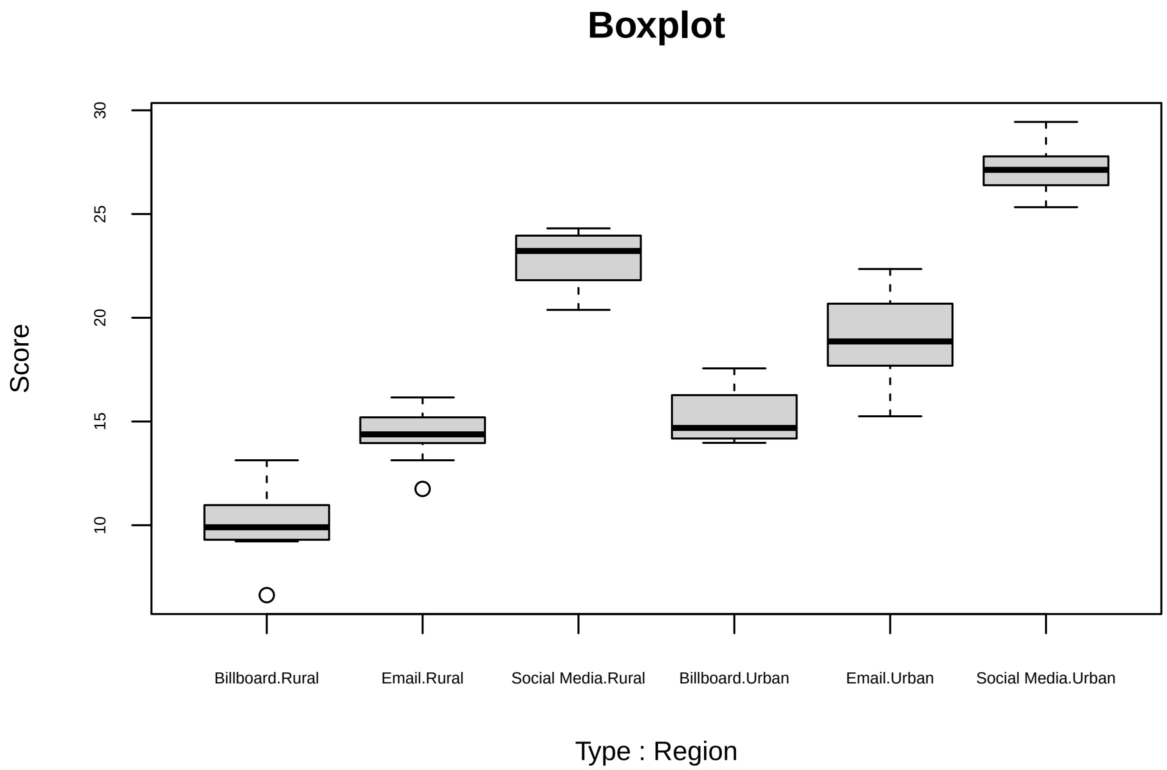
```
##   Score Region      Type
## 1 12.30  Rural Billboard
## 2  9.22  Rural Billboard
## 3  9.39  Rural Billboard
## 4  9.30  Rural Billboard
## 5 13.13  Rural Billboard
## 6 10.97  Rural Billboard
```

2.a) Construct two different preliminary graphs that investigate different features of the data and comment.

```
par(mfrow=c(1,2), cex.lab = 0.8, cex.axis = 0.7)
with(campaign, interaction.plot(Region, Type, Score,
trace.label = "Campaign", xlab = "Region", ylab = "Score", col = 1:3))
with(campaign, interaction.plot(Type, Region, Score,
trace.label = "Region", xlab = "Campaign", ylab = "Score", col = 1:3))
```



```
par(mfrow=c(1,1))
boxplot(Score ~ Type + Region, data = campaign, cex.axis = 0.53, main="Boxplot")
```



From a glance of the 3 x graphs we can see below:

- Parallel lines on both interaction plots suggests that there is no interaction effect between Region and Campaign Type. (We will look into this in later question.)
- From the boxplot, we can see that the assumption of equal variance among levels seems approximately valid due to the similar box sizes. (Standard deviation of the each group is calculated below)
- Social Media.Urban has the highest group mean of about 27.5.
- Billboard.Rural has the lowest group mean of about 10.

```
## [1] "Urban.Email"      "2.08084945261411"

## [1] "Urban.Social Media" "1.25895724047059"

## [1] "Urban.Billboard"  "1.24345441769648"

## [1] "Rural.Email"      "1.30757451455399"

## [1] "Rural.Social Media" "1.40695297244317"

## [1] "Rural.Billboard" "1.7990827910046"
```

2.b) Write down the full interaction model for this situation, defining all appropriate parameters.

$$Y_{i,j,k} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},\ \varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$$

- $Y_{i,j,k}$ = the customer engagement score
- µ = overall population mean
- α_i = the Type effect, there are three levels - Email, Social Media and Billboard
- β_j = the Region effect, there are three levels - Urban and Rural
- γ_{ij} = interaction effect between Type and Region
- ε_{ijk} = unexplained variation

```
# Fitting the model with interaction in mind.
campaign.int = lm(Score ~ factor(Type) * factor(Region), data = campaign)
```

2.c) Analyse the data to study the effect of Type and Region on the percentage increase in engagement Score at 5% significance level.

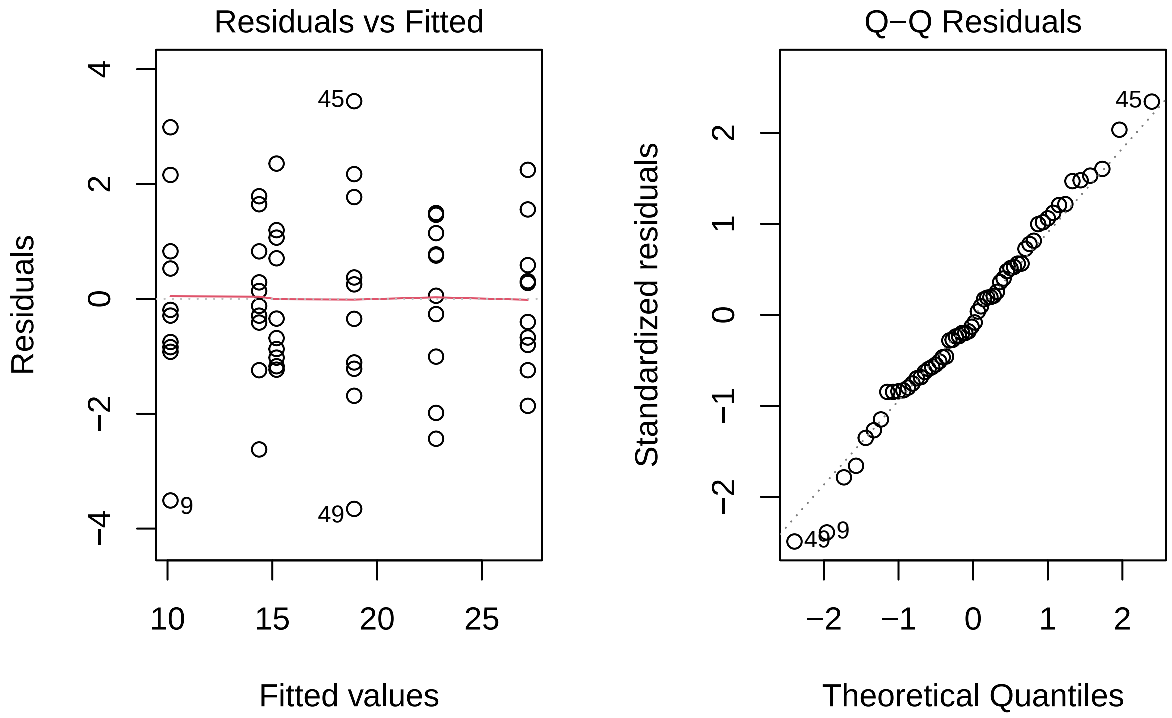
We first want to see if there is an interaction, hence:

H_0: $\gamma_{ij} = 0$ for all ij | H_1: at least one $\gamma_{ij} = 0$

```
anova(campaign.int)
```

```
## Analysis of Variance Table
##
## Response: Score
##                              Df  Sum Sq Mean Sq  F value    Pr(>F)
## factor(Type)                  2 1585.09  792.54 330.5242 < 2.2e-16 ***
## factor(Region)                1  325.45  325.45 135.7281 2.336e-16 ***
## factor(Type):factor(Region)  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(campaign.int, which = 1:2)
```



We can see that the interaction terms are not significant since the F -test of the interaction term has a P-value of 0.7657 > 0.05, suggesting that there is no interactions. The interaction can be removed from the model.

Other than that, the QQ-plot of the residuals show patterns close to normality. And the Residuals vs Fitted plot show constant variance. Both assumptions are appropriate, but the interaction is the only concern.

We may need to build a model without interaction. See below ANOVA for reference, interactions are not significant (P-Value > 0.05)

```
summary(campaign.int)
```

```
##
## Call:
## lm(formula = Score ~ factor(Type) * factor(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
## (Intercept)                               10.1410     0.4897  20.710
## factor(Type)Email                          4.2310     0.6925   6.110
## factor(Type)Social Media                  12.6740     0.6925  18.302
## factor(Region)Urban                        5.0620     0.6925   7.310
## factor(Type)Email:factor(Region)Urban     -0.5270     0.9794  -0.538
## factor(Type)Social Media:factor(Region)Urban  -0.6850     0.9794  -0.699
##                                               Pr(>|t|)    
## (Intercept)                                    < 2e-16 ***
## factor(Type)Email                             1.14e-07 ***
## factor(Type)Social Media                       < 2e-16 ***
## factor(Region)Urban                           1.29e-09 ***
## factor(Type)Email:factor(Region)Urban            0.593    
## factor(Type)Social Media:factor(Region)Urban     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
```

```
confint(campaign.int)
```

```
##                                                   2.5 %    97.5 %
## (Intercept)                                    9.159255 11.122745
## factor(Type)Email                              2.842603  5.619397
## factor(Type)Social Media                      11.285603 14.062397
## factor(Region)Urban                            3.673603  6.450397
## factor(Type)Email:factor(Region)Urban         -2.490489  1.436489
## factor(Type)Social Media:factor(Region)Urban -2.648489  1.278489
```

Here, we build a model without interaction. This way we can interprete the results.

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

```
##                               2.5 %    97.5 %
## (Intercept)                9.552599 11.133401
## factor(Type)Email          2.999460  4.935540
## factor(Type)Social Media 11.363460 13.299540
## factor(Region)Urban        3.867599  5.448401
```

Comment:

From the result, with the 95% confidence interval, we can see that:

- customer engagement score increases by (2.999460, 4.935540) if Email marketing is chosen.
- customer engagement score increases by (11.363460, 13.299540) if Social Media marketing is chosen.
- customer engagement score increases by (3.867599, 5.448401) if the region is Urban.
- Reference point is (9.552599, 11.133401) when it's Billboard and performed in Rural area.

2.d) Repeat the above test analysis for the main effects.

Main effects: Mean differences among the levels of each factor Interaction effects.

2.e) Using TukeyHSD produce multiple comparisons between each level for both Type and Region. Comment on the effectiveness of the marketing campaign type on customer engagement scores and also the impact of region on customer engagement scores. (Hint: Confirm the design is balanced before proceeding with the TukeyHSD test.)

```r
table(campaign$Region, campaign$Type)
```

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

```r
TukeyHSD(aov(campaign.lm))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = campaign.lm)
##
## $‘factor(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
##
## $‘factor(Region)‘
##              diff      lwr      upr p adj
## Urban-Rural 4.658 3.867599 5.448401     0
```

Comment:

Based on the table above, this experiment is balanced. Each combination has 10 x observations. After conducting TukeyHSD, we can see that the Social Media is the most effective marketing strategy and the worst was the Billboard strategy. And the location to perform the campaign also contributes to the engagement score. In Urban area, marketing team can expect minimum 3.8 points more to the engagement score.