STAT6180: Assignment I

Name: Do Nam Phong Phung
ID: 47828013

Question 1

a. The relationship between Consumer confidence index and Sales change:

Load the dataset

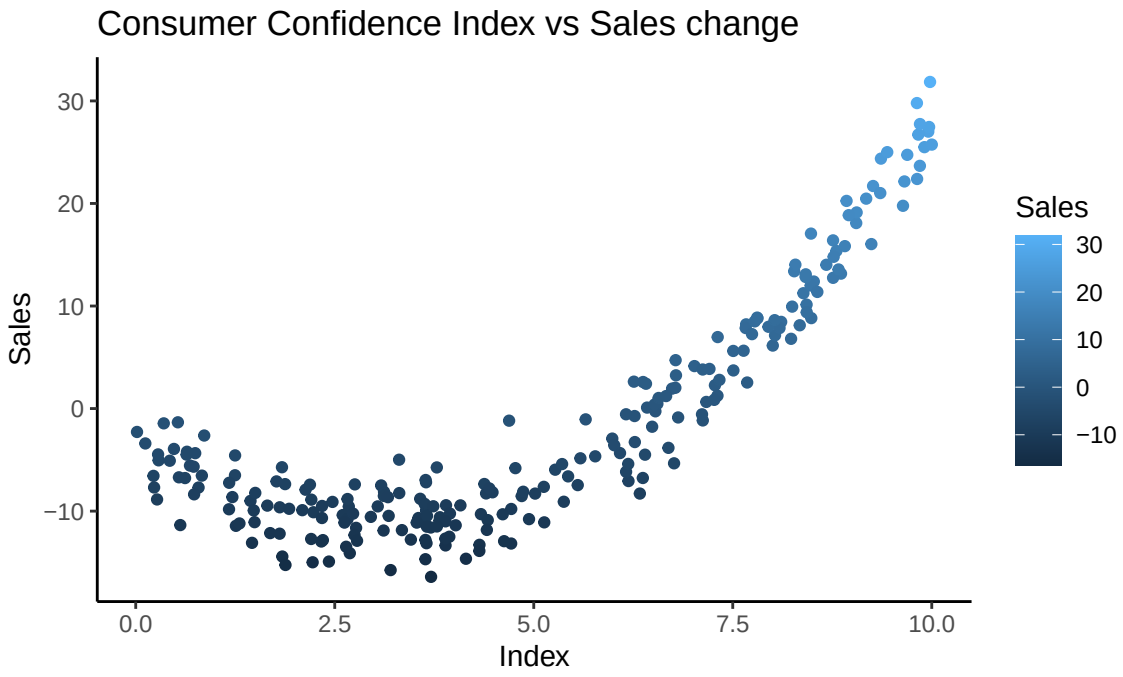
```
# Import datasets
sales <- read.csv('data/sales.csv', header = TRUE)
head(sales)
```

```
  Index  Sales
1  3.89 -13.35
2  8.27  13.39
3  6.49  -1.77
4  7.12  -1.16
5  3.55 -10.67
6  8.68  14.02
```

Create a scatter plot of `Sales` against `Index`.

```
# use `ggplot2` to draw the scatter plot
sales %>%
    ggplot(
        aes(
            x = Index,
```

```
            y = Sales,
            color = Sales
        )
    )+
    geom_jitter(
    )+
    labs(
        title = 'Consumer Confidence Index vs Sales change',
        x = 'Index',
        y = 'Sales'
    )+
    theme_classic()
```



Comment on the relationship between `Sales` and `Index`:

- With index in the range ~ 0-5:
  - The sales change is always negative.
  - With the increase of index, the sales change variable gradually decreases, which means that the retail sales drop bigger. The climax is at index about 4.

- From the index range ~ 5-10:
  - There is a positive change in the number of sales change.
  - With the increase of index, the sales change, starts at negative, then rapidly increases, top at sales change of more than 30.
  - We can clearly see that the sales change increase at a much higher rate in this half compared to the decrease rate of the first half.

```
The plot suggests:

- Starts from a high enough confidence the higher the confidence, the more customers tend to purchase

- There is a non-linear relationship between two variables.
```

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

Fit a simple linear regression model M1 to predict `Sales` using `Index`

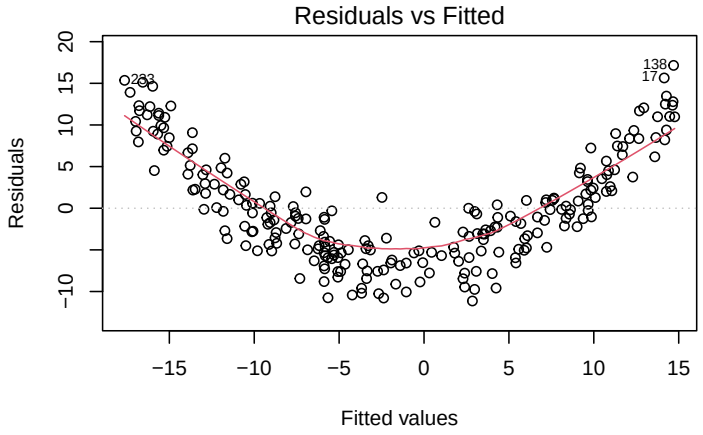
```
set.seed(5) # For constant output
M1 <- lm(Sales ~ Index, data = sales)
summary(M1)
```

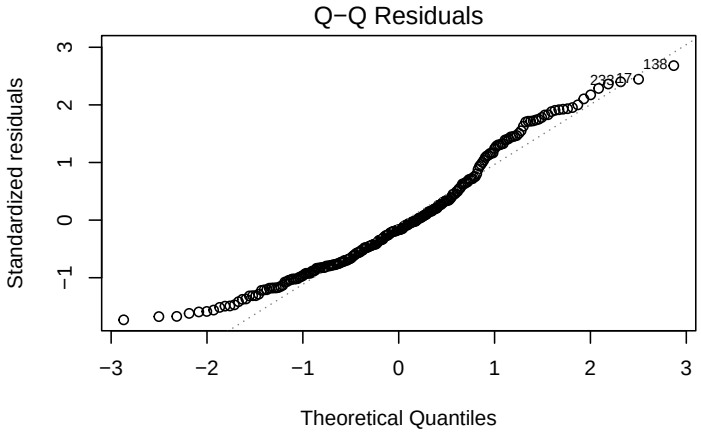
```
Call:
lm(formula = Sales ~ Index, data = sales)

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
```

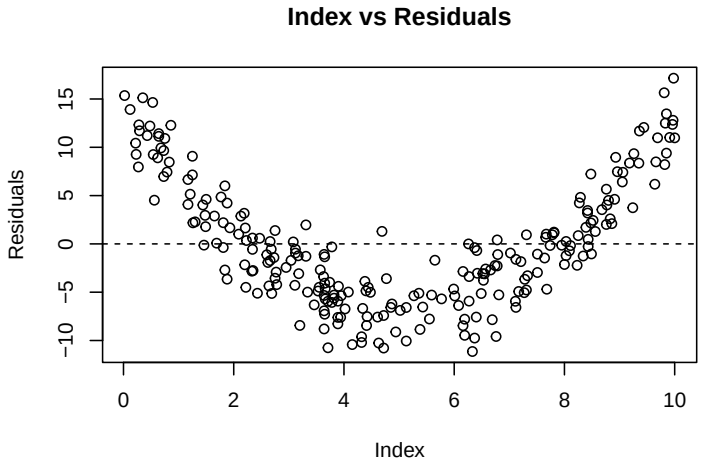
```
Residual standard error: 6.45 on 241 degrees of freedom
Multiple R-squared:  0.6772,    Adjusted R-squared:  0.6759
F-statistic: 505.6 on 1 and 241 DF,  p-value: < 2.2e-16
```

Validate the model through diagnostic plots checks & comments

```r
par(mfrow = c(2,2))
# Plot M1 residuals vs fitted and QQ model
plot(M1, which = 1:2)

# Index vs residual plot
plot(
    resid(M1) ~ Index,
    data = sales,
    main = 'Index vs Residuals',
    xlab = 'Index',
    ylab = 'Residuals'
)
abline(
    h = 0,
    lty = 2
)
```







Comments

- The pattern of the residuals in the Q-Q plot does not strictly follow a straight line. This suggests a potentially not normal distribution and may violates the normality assumption.
- The residuals vs fitted plot suggests unconstant variance. The pattern shows a hyperbolic/curvative shaped-spread of the residuals, which violates the constant variance assumption (pattern should be in a horizontal distribution).
- The residuals plot of `Index` also suggests unconstant variance. Similar to what we observed from the Residuals vs Fitted plot.
- The output suggest a non-linear relationship in the data, therefore it is appropriate to try fitting a polynomial model.

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. There is no need to validate the polynomial models at this stage. Compare and comment.

Fit a quadratic polynomial model

```
M2 = lm(Sales ~ Index + I(Index^2), data = sales)
summary(M2)$coefficients # print only the coefficients for comparison
```

```
             Estimate Std. Error    t value      Pr(>|t|)
(Intercept) -3.5060769  0.5030819  -6.969197  3.055813e-11
Index       -4.9659062  0.2304601 -21.547796  4.952328e-58
I(Index^2)   0.8087458  0.0220103  36.743973 1.586234e-100
```

```
[1] "R-squared: 0.951279"
```

Fit a cubic polynomial model

```
M3 = lm(Sales ~ Index + I(Index^2) + I(Index^3), data = sales)
summary(M3)$coefficients # print only the coefficients for comparison
```

```
               Estimate  Std. Error   t value     Pr(>|t|)
(Intercept) -3.421147623 0.668122407 -5.120540 6.270087e-07
Index       -5.062632109 0.550206234 -9.201335 1.822691e-17
I(Index^2)   0.832769781 0.125982447  6.610205 2.479951e-10
I(Index^3)  -0.001598833 0.008254854 -0.193684 8.465878e-01
```

```
[1] "R-squared: 0.951287"
```

Compare and comment

We will compare by looking at 2 aspects: Coefficients significance and Model's fit. Note that confidence level is default at 95%.

Coefficient significance

- M2 model: All coefficients are significant (super low p-value).
- M3 model:

- Overally, the p-values of all M3 coeffcients are higher than M2's.
- The coefficients of `Intercept`, `Index`, `Index^2` are significant (low p-value).
- The coefficient of `Index^3` is insignificant (p=0.846>0.05).

Model's fit

- R-squared of M2 and M3 model are extremeyly high (0.95). This shows the how fit the model is to the predicted data.
- Both models have almost identical R-squared value.

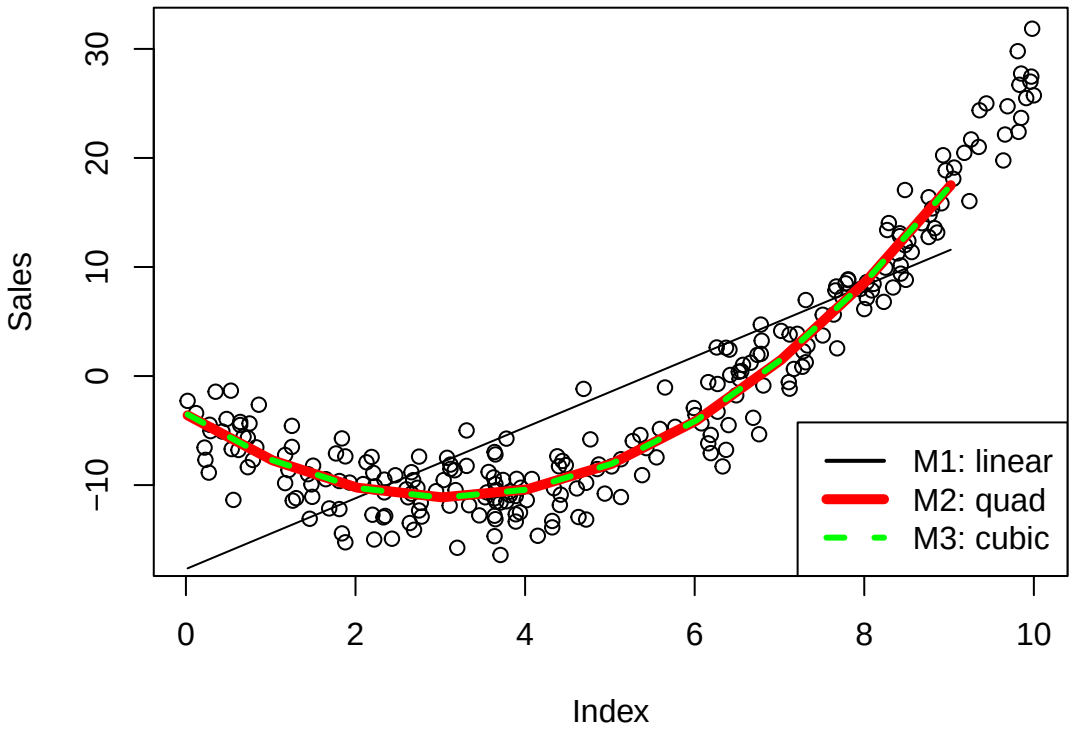
All of the information suggests that M2 is a better model because it is more simple (lower term), possesses significant coefficients and a great (higher) R-squared.

d. Plot the data and add the three predicted lines from models M1, M2, and M3 to your plot. Comment on the fit of the models.

Plot data and predicted y lines

```
# Change the plot format back to usual
par(mfrow = c(1,1))

# Convert x into a separate df
x = seq(from = min(sales$Index), to = max(sales$Index))
df = data.frame(Index = x)

# Determine the fitted y value
yhat_M1 = predict(M1, df)
yhat_M2 = predict(M2, df)
yhat_M3 = predict(M3, df)

# Plot the relationship `Sales` ~ `Index`
plot(Sales ~ Index, data = sales)

# Plot yhat distribution lines
lines(x, yhat_M1)
lines(x, yhat_M2, col = 'red', lwd = 5, lty = 1)
lines(x, yhat_M3, col = 'green', lwd = 3, lty = 2)

# Create a legend box
```

```r
legend(
    "bottomright",
    legend = c("M1: linear", "M2: quad", "M3: cubic"),
    col = c("black", "red", "green"),
    lty = c(1, 1, 2), lwd = c(2, 5, 3), cex = 1 # text size
)
```



Comment on the fit of the models

- The linear line represents M1 model seems to not fit the data distribution.
- It is visible that M2 and M3 predicted line follows the distribution of the data.
- The high accuracy from M2 and M3 lines, which are almost identical, are similar to what we observed from the model summary table.

e. Assess the significance of the linear, quadratic and cubic terms in M3 using a Sequential Sum of Squares (Hint: A sequential ANOVA is necessary). Comment on the results.

Use ANOVA to assess the significance of each term in M3

```r
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
```

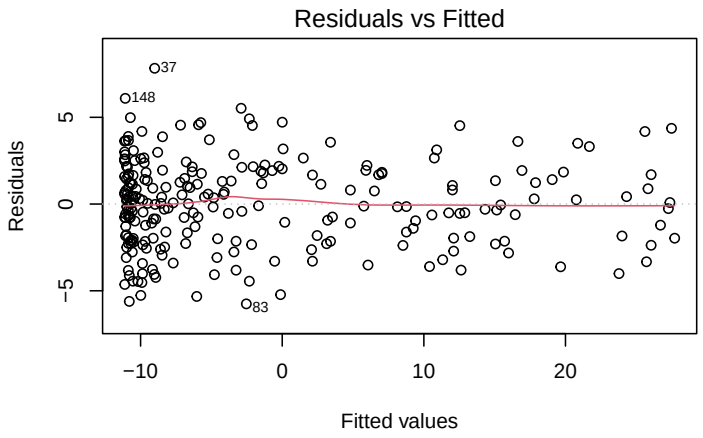
Comment on the results

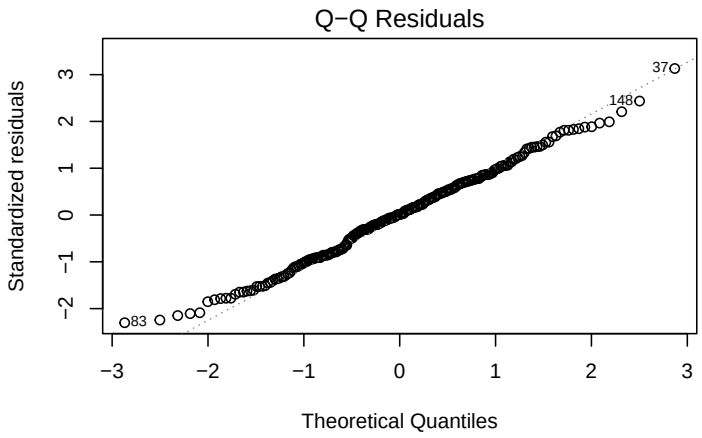
- The linear and quadratic terms are significant with small p-value (<2e-16).
- The cubic term is insignificant (p = 0.846). Which means it has low influence on represent the relationship between `Index` and `Sales`.
- The linear and quadratic model explain more variation than the cubic model as they are significant.

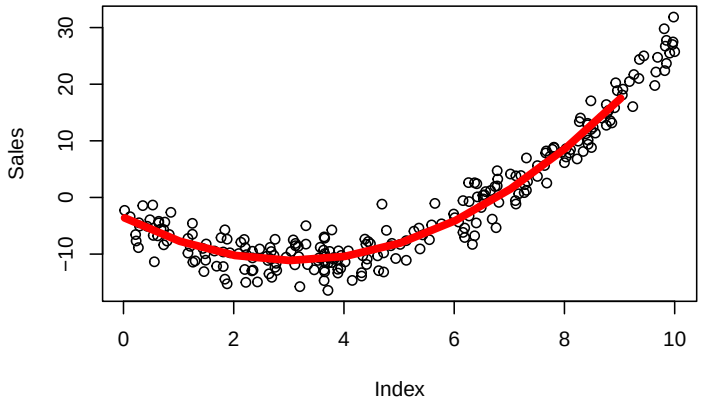
f. Choose the best model among M1, M2 and M3 and validate it. Provide reasoning and comment on the model fit.

Gained insights suggest that M2 should be the most appropriate model. Let's validate it by creating the Residuals vs Fitted and the QQ plot.

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

# Replot the relationship `Sales` ~ `Index` with M2 line
plot(Sales ~ Index, data = sales)
lines(x, yhat_M2, col = 'red', lwd = 5, lty = 1)
```







Reasoning and comments on the model fit

Reasons for choosing M2 as the best model:

- M3 is not suitable: The cubic model shows insignificance with a high p-value. This shows that the cubic term does not provide any meaningful improvement to the model even though the predicted line suggests that the cubic follows the distribution of the data.
- M1 is not suitable: Based on the predicted line of M1, the linear model does not fit the data distribution. Therefore, it is inappropriate to use the M1 model.
- M2: From ANOVA, the quadratic model is significant with a very small p-value. Looking at the above output, it seems that the quadratic model is not only able to represent the variance of the model but also is most efficient since it is more simple than the cubic model.

Comment on the validation plots:

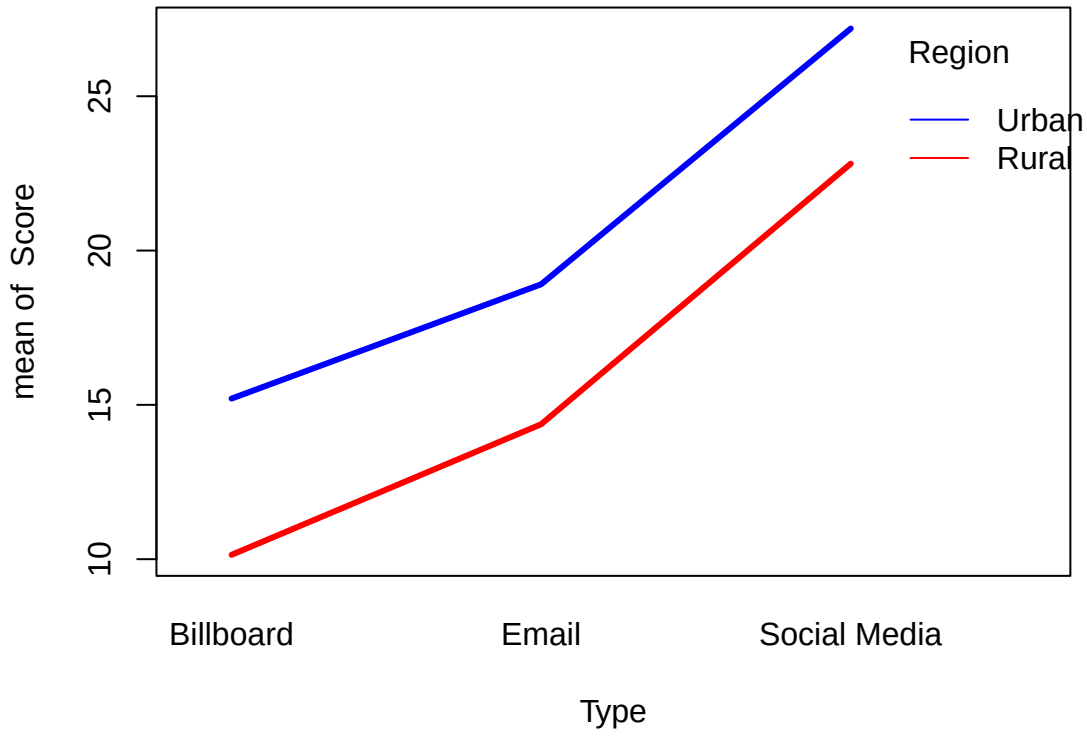
- Observing the residuals vs fitter plot, we can see the distribution of the values are distributed horizontally. This suggest constant std, suitable with the assumption.
- The quantile plot of reisiduals is close to linear suggests that the errors are close to a normal distribution.
- The predicted line follows the distribution of the data, suggests that the model fits the data.

Question 2

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

Create an interaction plot to determine the relationship between Type of marketing campaign and Region

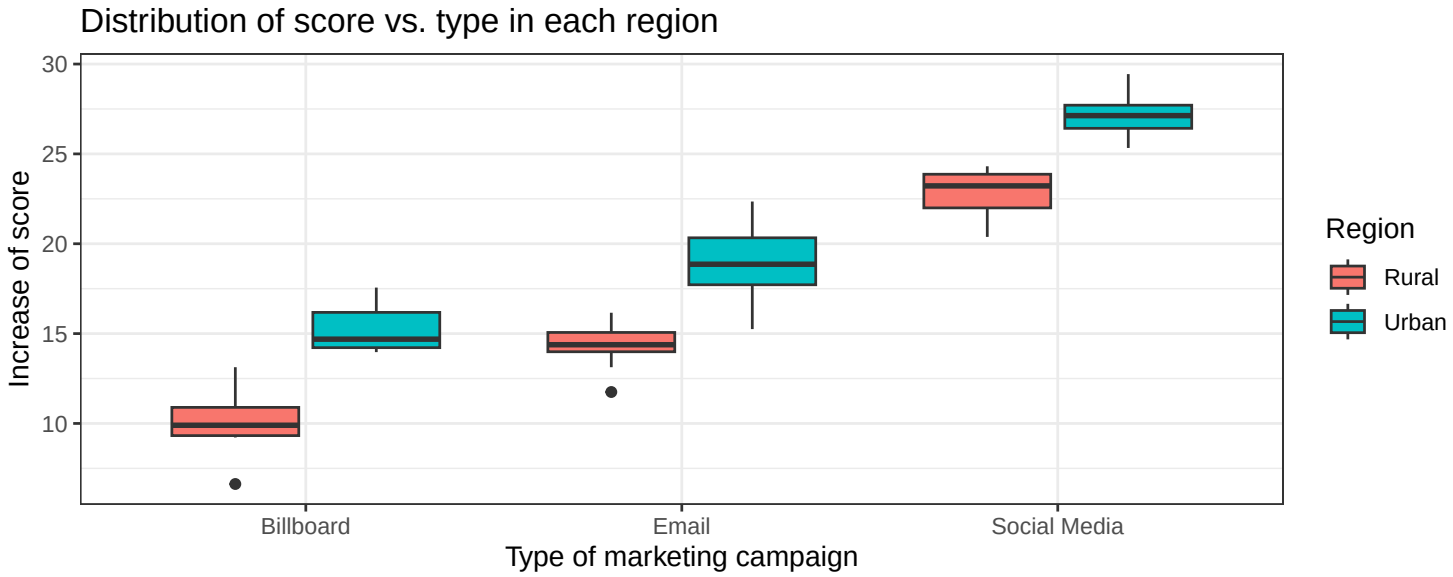
```
par(mfrow = c(1,2))
with(
    campaign,
    interaction.plot(
        x.factor = Type,
        fun = mean,
        trace.factor = Region,
        response = Score,
        col = c("red", "blue"), lty = 1, lwd = 3
    )
)
```



The two lines on the interaction plot are parallel, means that there is no interaction effect between the type of marketing campaign and region. Region has a constant effect on `Score` irrespective of `Type`.

Create an box plot to observe the variance of the data

```
campaign %>%
    ggplot(
        aes(
            x = Type, y = Score, fill = Region
        )
    )+
    geom_boxplot()+
    labs(
        title = "Distribution of score vs. type in each region",
        x = 'Type of marketing campaign',
        y = 'Increase of score'
    )+
    theme_bw()
```



- `Social Media` in general gets the highest engagment score increase, while `Billboard` has the lowests. When considering the region, `Urban` seems to have bigger but also more vary scores (bigger spread) compared to `Rural`.

- We can see the similarity in sizes of the boxes, means that assumption of equal variance among each type/region is approximately valid.
- Let's check the std of each group to make sure it is valid.

```
# A tibble: 6 x 3
  Type         Region   std
  <chr>        <chr>  <dbl>
1 Billboard    Rural   1.80
2 Billboard    Urban   1.24
3 Email        Rural   1.31
4 Email        Urban   2.08
5 Social Media Rural   1.41
6 Social Media Urban   1.26
```

- The stds are close to each other, we can assure the assumption of equal variance.

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

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

Where:

- Y_{ijk}: Increase in engagement score response.
- μ: The overall population mean.
- α_i: The `Type` effect, three levels: `Billboard`, `Email`, `Social Media`.
- β_j: The `Region` effect, two levels: `Urban`, `Rural`.
- γ_{ij}: The interaction effect between `Type` and `Region`.
- $\epsilon_{ijk} \sim N(0, \sigma^2)$: The unexplained variation.

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

Let's state the null and alternative hypothesis

- The null hypothesis says that there is no interaction between `Type` and `Region`

$$H_0 : \gamma_{ij} = 0 \text{ for all } i, j$$

- The alternative hypothesis says that there is an interaction between `Type` and `Region`

$$H_1 : \text{At least one } \gamma_{ij} \neq 0$$

Fit the interaction model

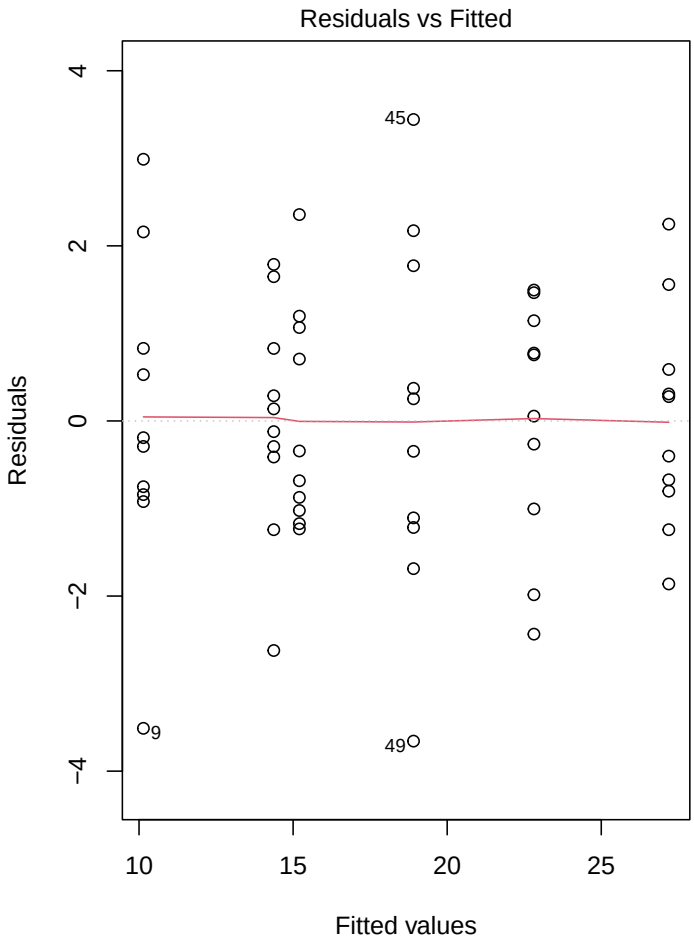
```
campaign_aov <- lm(Score ~ Type * Region, data = campaign)
anova(campaign_aov)
```

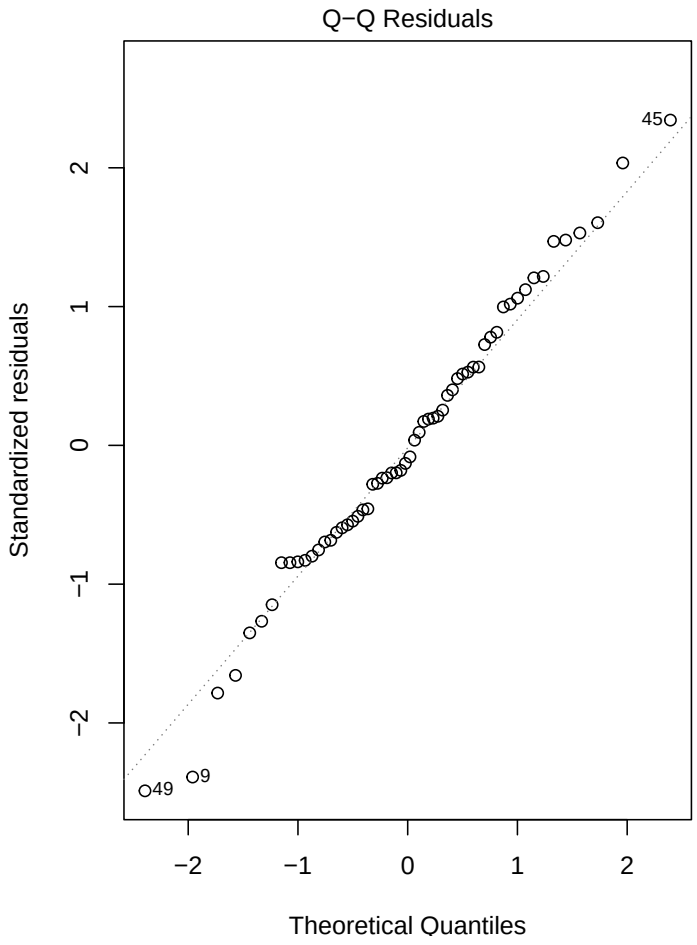
```
Analysis of Variance Table

Response: Score
            Df  Sum Sq Mean Sq  F value    Pr(>F)
Type         2 1585.09  792.54 330.5242 < 2.2e-16 ***
Region       1  325.45  325.45 135.7281 2.336e-16 ***
Type: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
```

- It is visible that the p-value represents the interaction of `Type:Region` is large (p=0.765>0.05), which means that the interaction between the two variable is insignificant. The effect of `Type` stays the same whatever the level of `Region` is (and vice versa).
- The interaction can be removed from the model.

Do diagnostics for model validation

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





- The residual vs fitted plot shows a horizontal distribution. This is a sign of equal variance and matches the requirement for the constant variance assumption.
- The residuals in the QQ plot are close to linear. This proves the validity of normality assumption.

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

For both main effects, the model changes to

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

- With γ_{ij} is removed as we no longer consider the interaction term between `Type` and `Region`.

- The null hypothesis changes to

$$H_0 : \alpha_i(\text{ or } \beta_j) = 0$$

- While the alternative hypothesis is

$$H_1 : \text{At least one } \alpha_i(\text{ or } \beta_j) \neq 0$$

Let's fit the model without the interaction effect

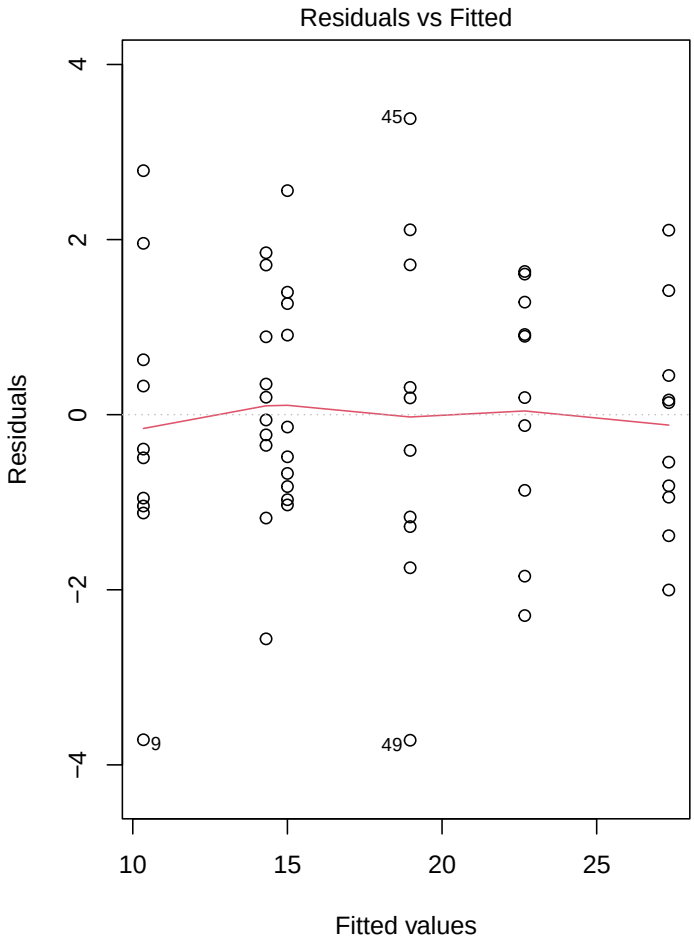
```
# Remove the interaction from the initial model
campaign_aov_noint = update(campaign_aov, . ~ . - Type:Region)
anova(campaign_aov_noint)
```

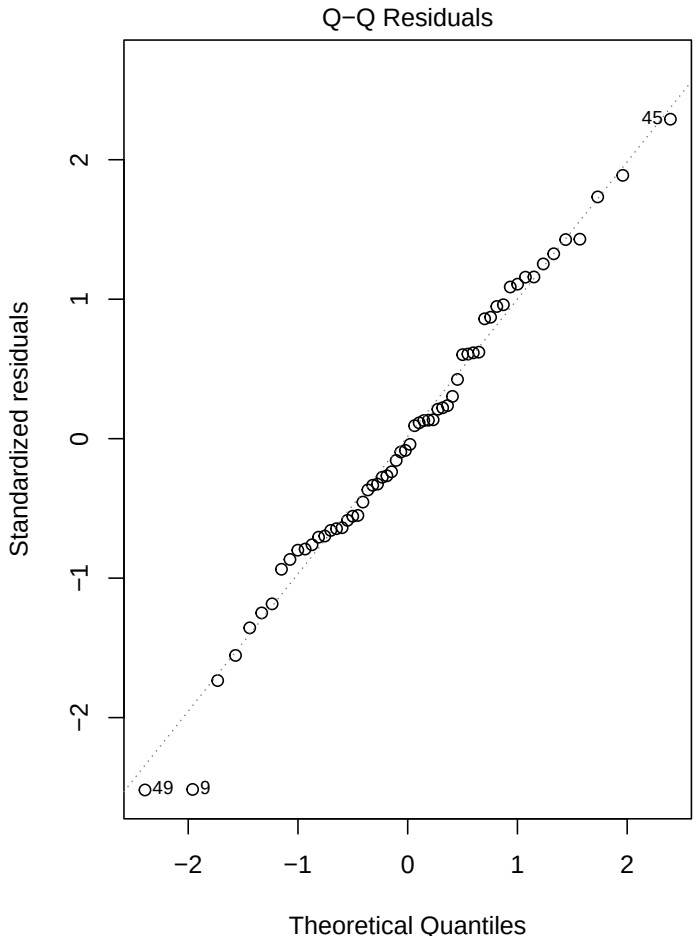
```
Analysis of Variance Table

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
```

- It is visible that both main effects `Type` and `Region` are significant with $p < 0.05$.

Validate the model

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





- The residual vs fitted plot shows a horizontal distribution despite that it is not as straight as the model with the interaction. It seems that the variability near 15 and 23 spike up a bit. However, this is still a sign of equal variance and matches the requirement for the constant variance assumption.
- The residuals in the QQ plot are closer to linear compared to the previous model. This proves the validity of normality assumption.

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.)

Check for balance design

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

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

Equal number of replicates in each group proves balance design.

Compare between each level of `Type` and `Region` using `TukeyHSD`

```r
TukeyHSD(aov(Score ~ Type + Region, data = campaign))
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Score ~ Type + Region, data = campaign)

$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
```

- All comparisons have $p<0.05$, indicates that each difference is significant.
- On marketing type, `Social Media` has the biggest difference with the other two types on the increase of score. Moreover, these differences are relatively high. This means that `Social Media` has the best performance among all marketing campaign types and outperforms the other two types.
- `Email` has higher increase in score than `Billboard` by 3.967. This shows that `Billboard` has the lowest increase in engagment score.
- `Urban` has significantly higher increase (4.658) in engagement score compared to `Rural`.
- Note that these insights match what we have observed in preliminary analysis.

Compiled with LATEX