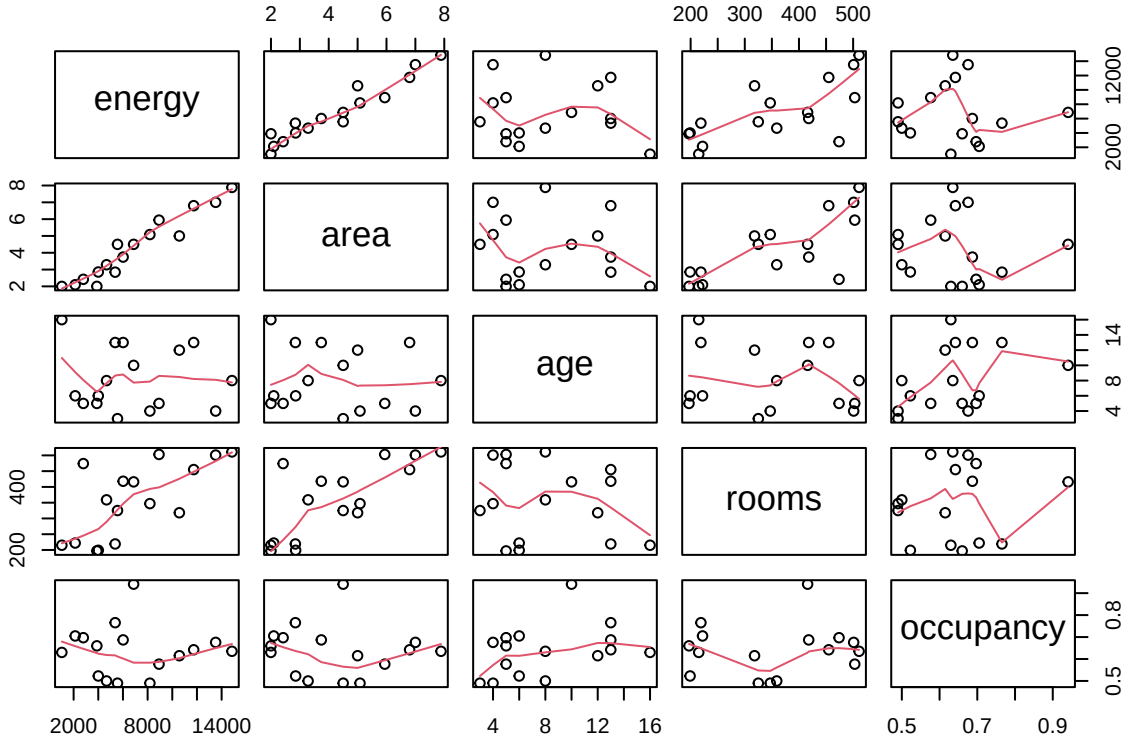
Assignment Minh_Le_Quynh_Cao 46822097

Question 1

a. Produce a plot and a correlation matrix of the data. Comment on possible relationships between the response and predictors and relationships between the predictors themselves.

```
setwd("C:/Users/cmleq/Downloads")
hotel = read.table("hotel2022.dat", header = T)
pairs(energy ~ area + age + rooms + occupancy, data = hotel, panel = panel.smooth)
```



Energy has a high correlation with area, less but still highly correlated with rooms, slight correlation with age and occupancy. There is a high correlation between area and rooms, however; the other predictors are not.

```
cor(hotel)
```

```
##                energy       area         age       rooms   occupancy
## energy     1.00000000  0.9631684 -0.05707197  0.68265406 -0.02022606
```

```
## area       0.96316841  1.0000000 -0.10787993  0.76068692 -0.08924440
## age       -0.05707197 -0.1078799  1.00000000 -0.16142400  0.36195114
## rooms      0.68265406  0.7606869 -0.16142400  1.00000000  0.09944548
## occupancy -0.02022606 -0.0892444  0.36195114  0.09944548  1.00000000
```

b. Fit a model using all the predictors to explain the energy response.

```
hotel.1 = lm(energy ~ ., data = hotel)
summary(hotel.1)
```

```
##
## Call:
## lm(formula = energy ~ ., data = hotel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1577.1  -720.8   118.7   608.6  1809.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3197.279   1871.533  -1.708    0.116
## area         2331.116    250.919   9.290 1.53e-06 ***
## age             2.358     80.841   0.029    0.977
## rooms          -5.383      4.168  -1.291    0.223
## occupancy    3234.553   2928.605   1.104    0.293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1154 on 11 degrees of freedom
## Multiple R-squared:  0.9416, Adjusted R-squared:  0.9203
## F-statistic:  44.3 on 4 and 11 DF,  p-value: 1.019e-06
```

Using the full model, estimate the impact each hectare of hotel area has on the energy efficiency of the hotel. Do this by producing a 95% confidence interval that quantifies the change in energy consumption for each extra hectare of hotel area and comment.

```
t_quant = qt(1-0.05/2, 16-5)
t_quant
```

```
## [1] 2.200985
```

$$b_1 \pm t_{n-2,1-\alpha/2} s.e.(b_1) = 2331.116 \pm 2.201 \times 250.919 = (1778.843, 2883.389)$$

Comment: We have 95% of confidence that there is an expected increase in consuming energy from 1778.843 to 2883.389 MWh for each extra hectare of hotel area.

c. Conduct an F-test for the overall regression i.e. is there any relationship between the response and the predictors. In your answer:

Write down the mathematical multiple regression model for this situation, defining all appropriate parameters.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

Y: energy

X_1, X_2, X_3, X_4: 4 different predictors (area, age, rooms, occupancy)

β_0: intercept

$\beta_1, \beta_2, \beta_3, \beta_4$: partial regression coefficients

ϵ: unexplained variation

Write down the Hypotheses for the Overall ANOVA test of multiple regression.

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

Produce an ANOVA table for the overall multiple regression model (One combined regression SS source is sufficient).

```
anova(hotel.1)
```

```
## Analysis of Variance Table
##
## Response: energy
##           Df    Sum Sq   Mean Sq  F value    Pr(>F)
## area       1 232665641 232665641 174.5879 4.297e-08 ***
## age        1    556602    556602   0.4177    0.5314
## rooms      1   1293045   1293045   0.9703    0.3458
## occupancy  1   1625643   1625643   1.2199    0.2930
## Residuals 11  14659219   1332656
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Full Model Regression SS = 232665641 + 556602 + 1293045 + 1625643 = 236140931.

Source	*df*	*SS*	*MS*	F-Value
Regression	4	236140931	59035232.75	44.2989
Residual	11	14659219	1332656.273	
Total	15	250800150		

Compute the F statistic for this test.

F-statistic has 4, 11 degrees of freedom.

$$F_{obs} = RegMS/ResMS = 59035232.75/1332656.273 = 44.2989$$

State the Null distribution for the test statistic.

If H_0 is true, there is a linear relationship between parameters.

Compute the P-Value.

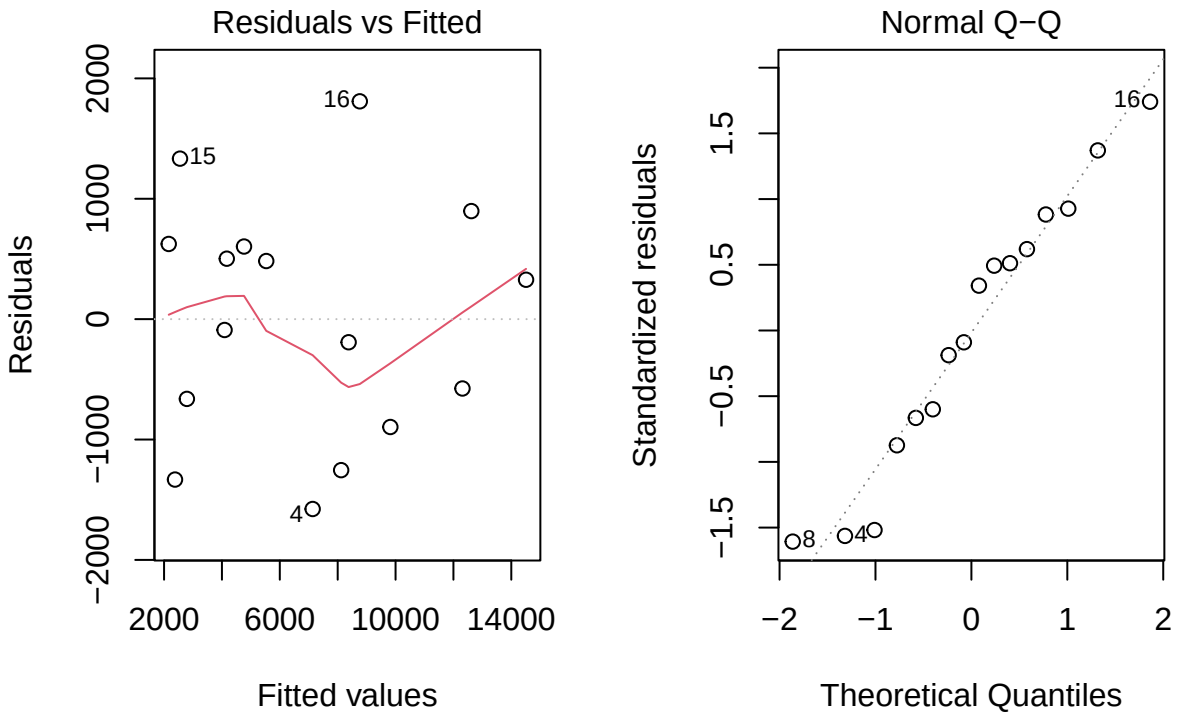
P-Value = $P(F_{4,11} >= 44.2989) = 1.019e-06$

State your conclusion (both statistical conclusion and contextual conclusion).

P-Value = 0.0025 < 0.05 Since the P-Value is smaller than 0.05, we can reject H_0 based on the evidence that the linear relationship between response and at least one of the four predictor variables is significant.

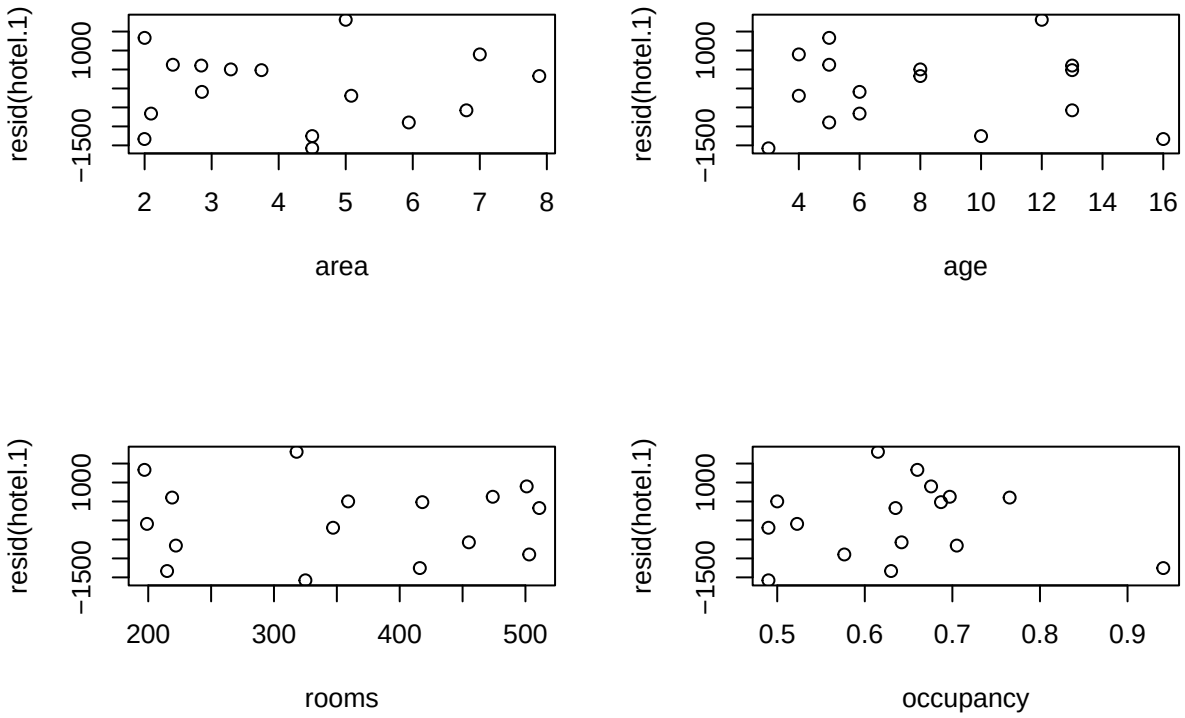
d. Validate the full model and comment on whether the full regression model is appropriate to explain the energy efficiency of the hotels.

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



The Normal Q-Q plot of residuals slightly curved but still near to linear which indicates that the residuals are close to the normal distribution. The Residuals and Fitted plot does not follow any specific trend, so the variability seems constant.

```
par(mfrow = c(2, 2))
plot(resid(hotel.1) ~ area + age + rooms + occupancy, data = hotel)
```



There is no visible pattern in the residuals and predictors plots. Thus, suggesting that the linear model seems sufficient.

e. Find the R^2 and comment on what it means in the context of this dataset.

$$R^2 = RegSS/TotalSS = 236140931/250800150 = 0.942$$

Comment: 94% of the variation in the energy efficient of the hotel is explained by the linear regression on predictor variables such as area, age, rooms, and occupancy.

f. Using model selection procedures discussed in the course, find the best multiple regression model that explains the data. State the final fitted regression model.

As can be seen, except for the area, the other predictor variables are insignificant (argument from the outcome of anova(hotel.1)). Age has the largest P-Value (0.977); therefore, there will be the least variation explained by age when it was added to the model.

```
hotel.2 = lm(energy ~ area + rooms + occupancy, data = hotel)
summary(hotel.2)
```

```
##
## Call:
## lm(formula = energy ~ area + rooms + occupancy, data = hotel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1585.3  -728.4   116.9   610.9  1817.4
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3194.017   1788.723  -1.786   0.0994 .
## area         2332.113    238.007   9.799 4.46e-07 ***
## rooms          -5.412      3.876  -1.396   0.1879
## occupancy    3269.093   2564.540   1.275   0.2265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1105 on 12 degrees of freedom
## Multiple R-squared:  0.9415, Adjusted R-squared:  0.9269
## F-statistic: 64.43 on 3 and 12 DF,  p-value: 1.14e-07
```

After dropping age, area is still significant, rooms and occupancy are insignificant.

```
hotel.3 = lm(energy ~ area + rooms, data = hotel)
summary(hotel.3)
```

```
##
## Call:
## lm(formula = energy ~ area + rooms, data = hotel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2016.2  -517.9  -180.3   655.9  1842.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1228.306    927.939  -1.324    0.208
## area         2254.660    235.587   9.570 2.99e-07 ***
## rooms          -4.133      3.833  -1.078    0.300
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1132 on 13 degrees of freedom
## Multiple R-squared:  0.9336, Adjusted R-squared:  0.9234
## F-statistic: 91.44 on 2 and 13 DF,  p-value: 2.202e-08
```

After eliminating occupancy, area variable is still significant, but rooms variable is not.

```
hotel.4 = lm(energy ~ area, data = hotel)
summary(hotel.4)
```

```
##
## Call:
## lm(formula = energy ~ area, data = hotel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1843.6  -447.1  -275.4   580.8  2141.0
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1874.6      712.5  -2.631   0.0198 *
## area          2061.4      153.8  13.402 2.24e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1138 on 14 degrees of freedom
## Multiple R-squared:  0.9277, Adjusted R-squared:  0.9225
## F-statistic: 179.6 on 1 and 14 DF,  p-value: 2.237e-09
```

The final fitted regression model: hotel.4 seems to be the final fitted regression model since there is no variables are insignificant in this model.

g. Comment on the R^2 and adjusted R^2 in the full and final model you chose in part f. In particular explain why those goodness of fitness measures change but not in the same way.

Full Model:

- $R^2 = 0.942$ (Question 1e)
- Adjusted $R^2 = 1 - (1 - R^2) \times (n - 1)/(n - p) = 1 - (1 - 0.942) \times (16 - 1)/(16 - 5) = 0.921$

Final Model:

- $R^2 = RegSS/TotalSS = 232665641/(232665641 + 18134508) = 0.928$
- Adjusted $R^2 = 1 - (1 - R^2) \times (n - 1)/(n - p) = 1 - (1 - 0.928) \times (16 - 1)/(16 - 2) = 0.923$

In general, the adjusted R squared increased in the full model and slightly maintains unchanged in the final model from the R squared. Those goodness of fitness estimates the difference but not in a similar way because of the difference in the number of predictors.

Question 2

```
setwd("C:/Users/cmleq/Downloads")
movie = read.table("movie.dat", header = T)
```

a. For this study, is the design balanced or unbalanced? Explain why.

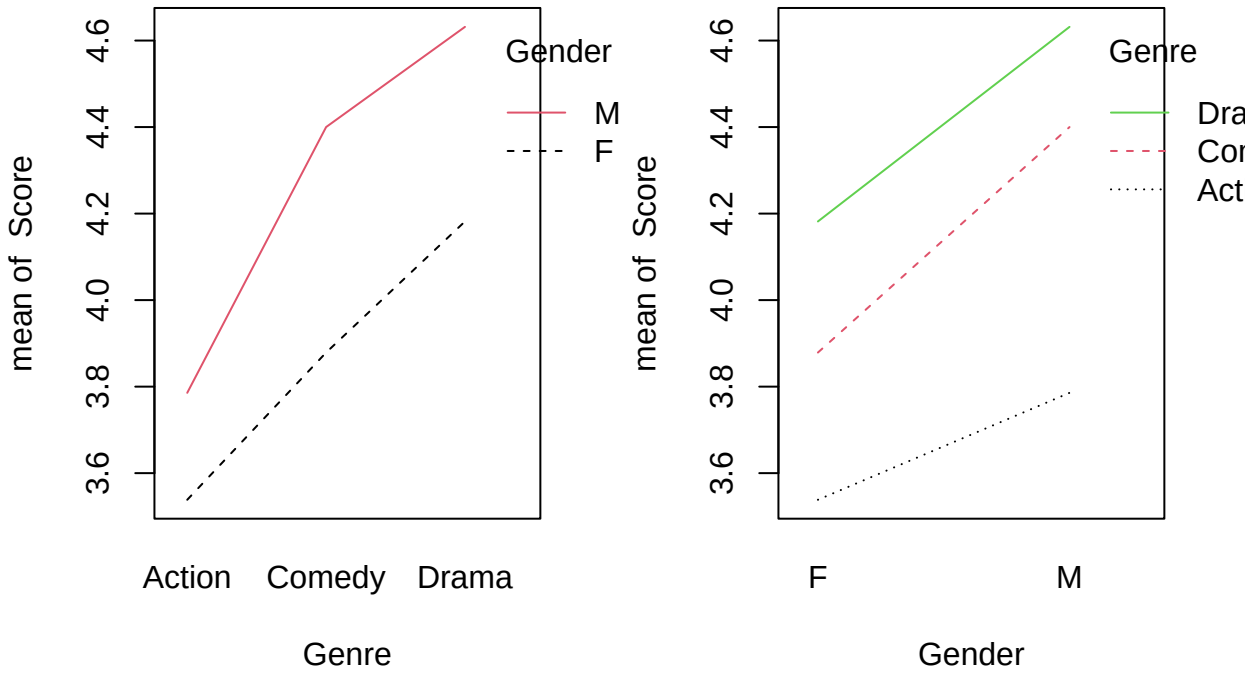
```
table(movie[, 1:2])
```

```
##        Genre
## Gender Action Comedy Drama
##      F     39     33    22
##      M     14     10    19
```

As for this study, the design is unbalanced because of the inexact replication available for all pairs of factors.

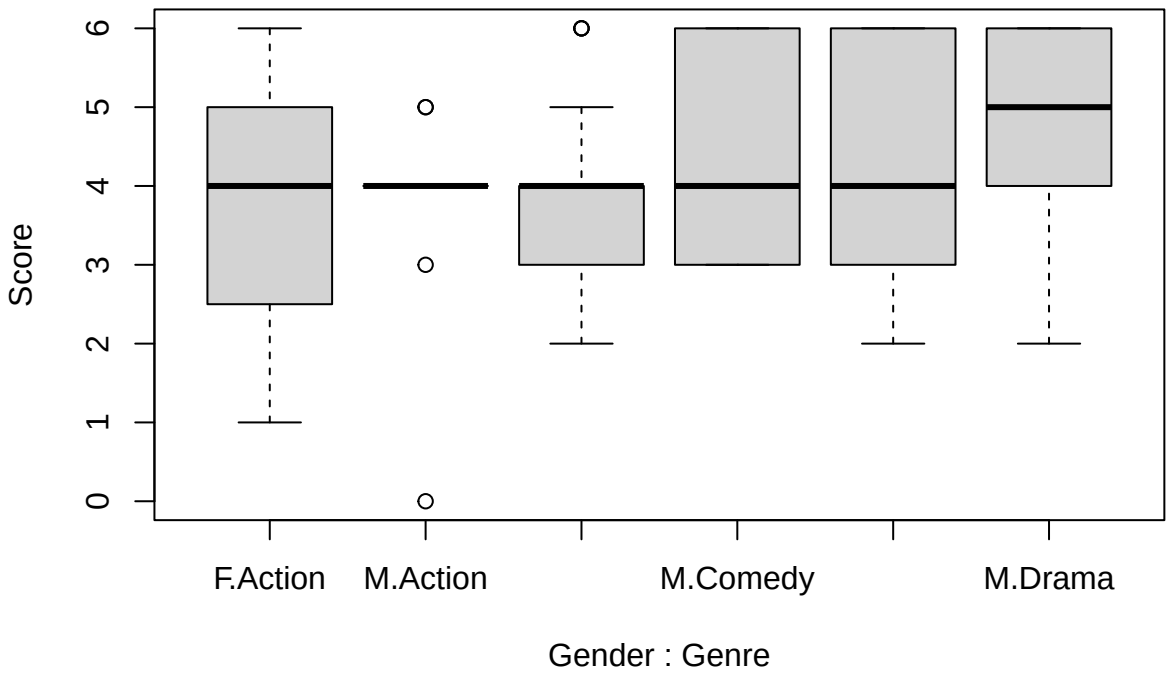
b. Construct two different preliminary graphs that investigate different features of the data and comment

```
par(mfrow = c(1, 2))
with(movie, interaction.plot(Genre, Gender, Score, col = 1:2))
with(movie, interaction.plot(Gender, Genre, Score, col = 1:3))
```



It is highly likely that there might be an interaction between factors and response, as the lines from the left plot show different trends between three levels of Genre. In general, lines seem to be parallel, which indicate that there is no interaction in the second plot. The test is hard to certainly identify because of the low sample sizes.

```
boxplot(Score ~ Gender + Genre, data = movie)
```



There is a constant variability between levels of each factor, apart from the males with action. There are also 3 outliers in males with action and an outlier in female with comedy.

c. Write down the the full mathematical model for this situation, defining all appropriate parameters.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}; \epsilon_{ijk} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$$

Y_{ijk}: k^{th} replicate of the treatment at i^{th} level in Gender and j^{th} level in Genre.

μ: overall population mean

α_i: main Gender effect for $i = 1, 2$

β_j: main Genre effect for $j = 1, 2, 3$

γ_{ij}: interaction effect between Gender and Genre

ϵ_{ijk}: unexplained variation for each replicated observation.

d. Analyse the data to study the effect of Gender and Genre on the brand recall Score. These conclusions are only required to be at the qualitative level and can be based off the outcomes of the hypothesis tests you conducted in this part and the preliminary plots in part b. You do not need to statistically examine the multiple comparisons between contrasts and interactions. Remember to state the null and alternative hypothesis for each test, and check assumptions

```
movie.1 = lm(Score ~ Gender * Genre, data = movie)
summary(movie.1)$coefficients
```

```
##                       Estimate Std. Error    t value     Pr(>|t|)
```

```
## (Intercept)         3.5384615  0.2129394 16.6172220 4.665049e-34
## GenderM             0.2472527  0.4143143  0.5967757 5.516871e-01
## GenreComedy         0.3403263  0.3145324  1.0820074 2.812370e-01
## GenreDrama          0.6433566  0.3545762  1.8144381 7.189851e-02
## GenderM:GenreComedy 0.2739594  0.6340996  0.4320447 6.664192e-01
## GenderM:GenreDrama  0.2025080  0.5874610  0.3447174 7.308596e-01
```

The interaction is not significant, so it could be removed from the model

```
movie.2 = update(movie.1,  . ~ . - Gender:Genre)
```

```
anova(movie.1)
```

```
## Analysis of Variance Table
##
## Response: Score
##               Df  Sum Sq Mean Sq F value  Pr(>F)
## Gender         1   7.195  7.1946  4.0684 0.04574 *
## Genre          2  11.587  5.7934  3.2761 0.04089 *
## Gender:Genre   2   0.378  0.1889  0.1068 0.89879
## Residuals    131 231.658  1.7684
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interaction

- Hypotheses: H_0 : γ_{ij} for all i,j; H_A : not all $\gamma_{ij} = 0$
- P-Value $= 0.89879 > 0.05$
- Conclusion: The interaction is insignificant.

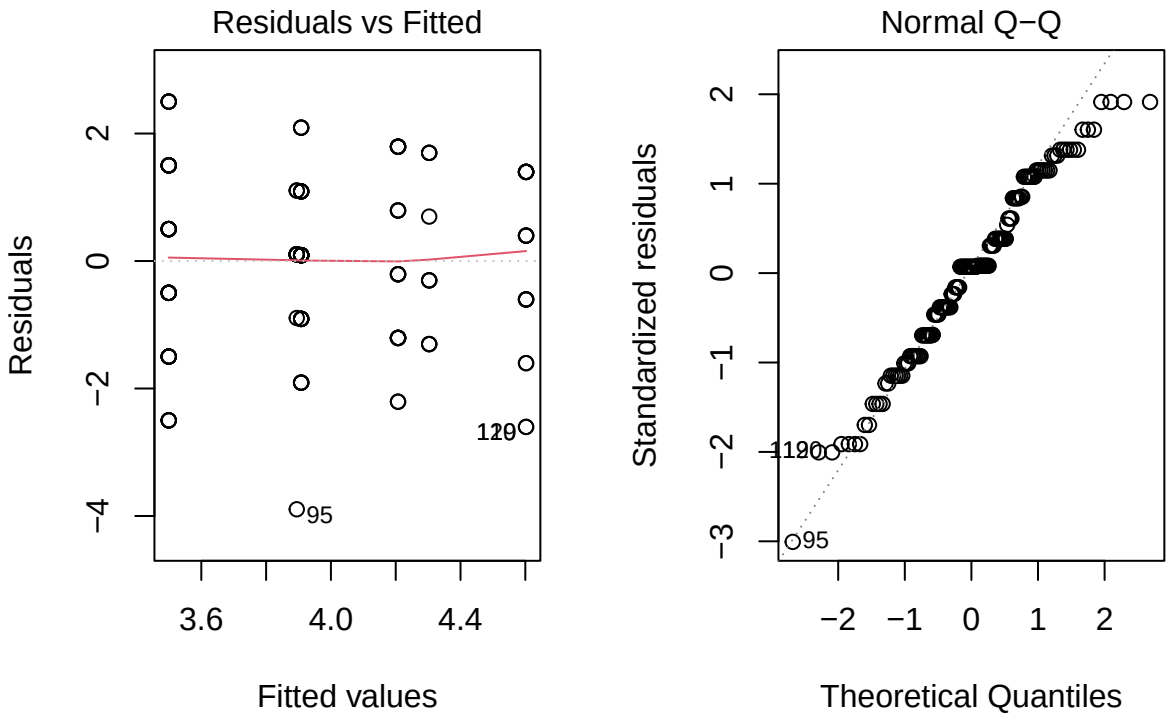
```
anova(movie.2)
```

```
## Analysis of Variance Table
##
## Response: Score
##            Df  Sum Sq Mean Sq F value  Pr(>F)
## Gender      1   7.195  7.1946  4.1238 0.04428 *
## Genre       2  11.587  5.7934  3.3207 0.03915 *
## Residuals 133 232.036  1.7446
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Main effects: Gender

- Hypotheses: H_0 : $\alpha_i = 0$ for all i; H_A : not all $\alpha_i = 0$
- P-Value $= 0.04428 < 0.05$
- Conclusion: Gender type is significant.

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

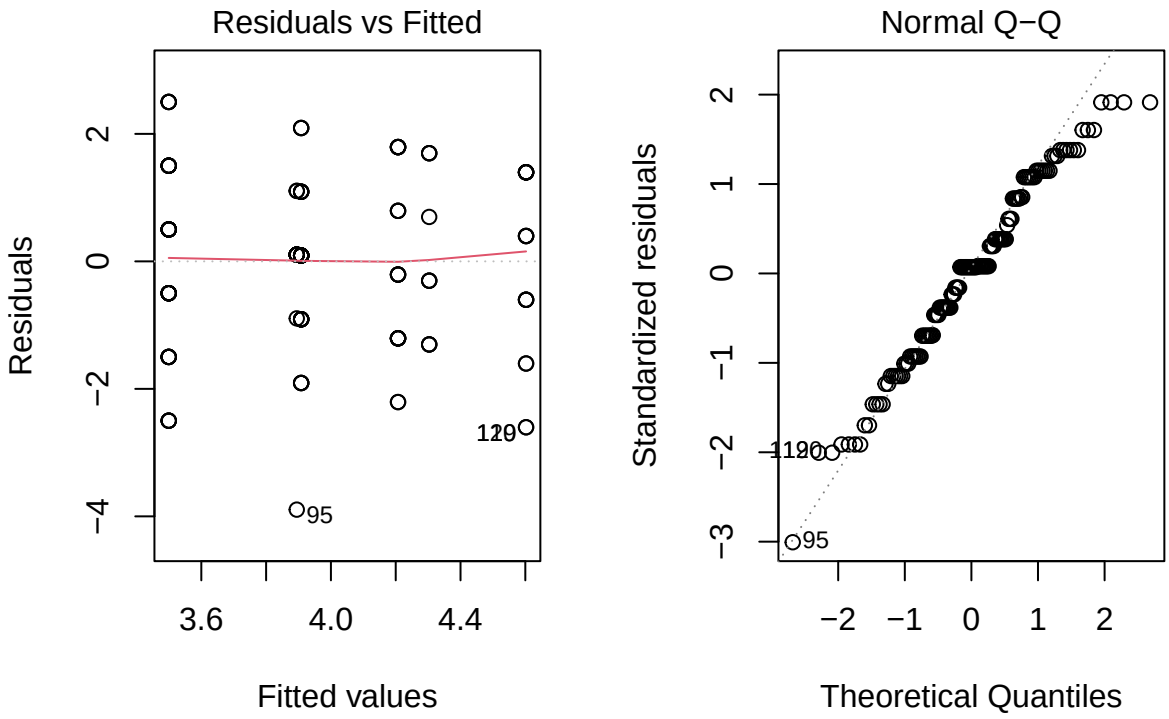


As can be seen, there is a significant effect of gender on score, since the p-value (0.04428) is lower than 0.05. Additionally, the residuals plot vs fitted plot has no trend, the variability seems to maintain the same between effects. The normal Q-Q tends to be close to linear indicating the normal distribution of residuals.

Main effect: Genre

- Hypotheses: $H_0 : \beta_j = 0$ for all j; H_A : not all $\beta_j = 0$
- P-Value $= 0.03915 < 0.05$
- Conclusion: Genre type is significant.

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



Similar to the model validation of Gender type, the residuals vs fitted plot of Genre type depicts no trend showing the likely constancy of the variability. The normal Q-Q plot is close to linear indicating that the residuals are near to normal distributed. "'