Stat2170 Assignment

Rashrav Shrestha (46295194)

19/05/2022

Contents

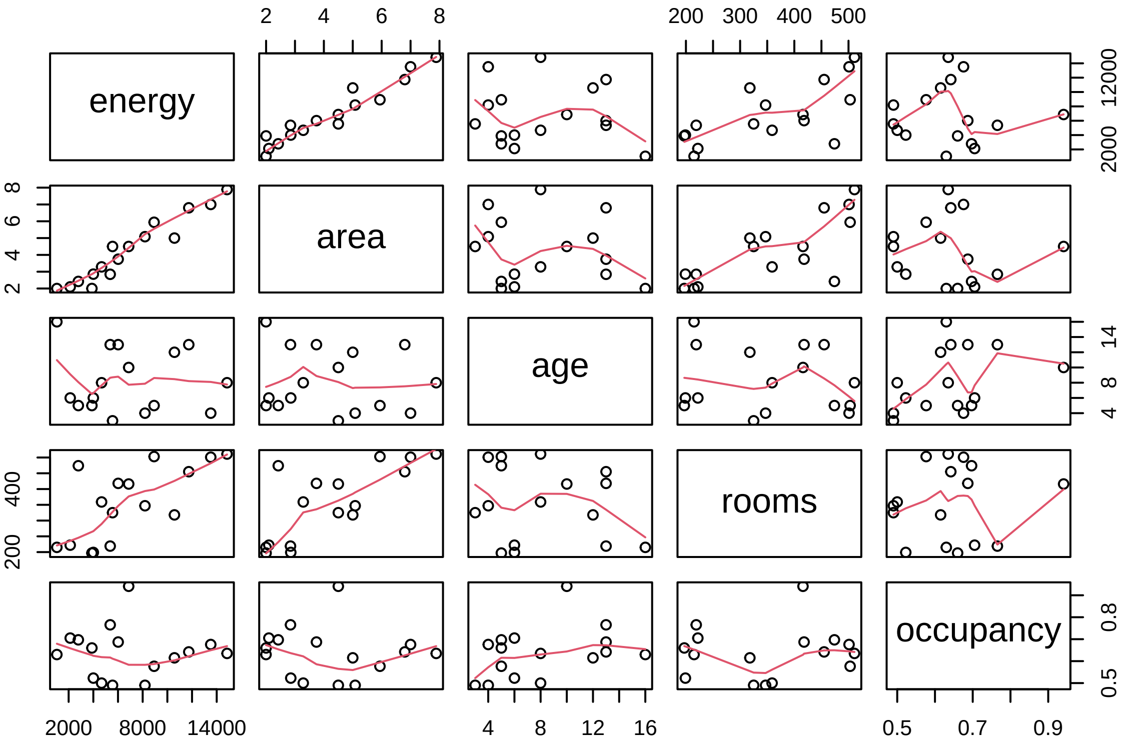
Question 1 [45 Marks]

Loading the data set

```
hotel = read.table(here::here("data", "hotel2022.dat"), header=TRUE)
```

a. Plotting and producing a correlation matrix of the data.

```
pairs(hotel, panel = panel.smooth)
```



```
round(cor(hotel), 2)
```

```
##           energy  area   age rooms occupancy
## energy      1.00  0.96 -0.06  0.68     -0.02
## area        0.96  1.00 -0.11  0.76     -0.09
## age        -0.06 -0.11  1.00 -0.16      0.36
## rooms       0.68  0.76 -0.16  1.00      0.10
## occupancy  -0.02 -0.09  0.36  0.10      1.00
```

Commenting on the possible relationship

Energy has a **Strong Positive Correlation** with Area. There also appears to be a **Moderate Positive Correlation** between energy and rooms. Looking at the correlation matrix, age and occupancy seem to have a **weak negative correlation** against Energy. Judging from this, Area and Room would be a great predictor of Energy Consumption as they have high coefficient values.

b. Fitting a model using all the features (area, age, rooms, occupancy) to explain the response variable (energy)

```r
hotel_full_model = lm(energy ~ area + age + rooms + occupancy, data = hotel)
summary(hotel_full_model)
```

```
##
## Call:
## lm(formula = energy ~ area + age + rooms + occupancy, 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
```

Producing a 95% confidence interval that measures the change in the consumption of energy for an increase in each unit of the hotel’s area.

Looking at the coefficients of the full model

```r
summary(hotel_full_model)$coefficients
```

```
##                 Estimate  Std. Error     t value     Pr(>|t|)
## (Intercept) -3197.278911 1871.533234 -1.70837410 1.155950e-01
## area         2331.116239  250.918960  9.29031525 1.534817e-06
## age             2.357756   80.841496  0.02916517 9.772553e-01
## rooms          -5.383194    4.168325 -1.29145243 2.230244e-01
## occupancy    3234.553418 2928.605342  1.10446886 2.929634e-01
```

The terms of interests:

- $\beta_{area} = 2331.116$
- $s.e.(\beta_{area}) = 250.919$
- $\alpha = 0.05$
- $df = 11$
- $t_{df,1-\alpha/2} = 2.200985$

Quantile Calculation

```
qt(1-0.05/2,11)
```

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

Calculating Confidence Interval

$\beta_{area} \pm t_{11,1-0.05/2} s.e.(\beta_{area}) = 2331.116 \pm 2.200985 \times 250.919 = (\mathbf{1778.847,\ 2883.385})$

Comment: For a unit increase in area of the hotel, we are 95% confident to expect a change in energy consumption between 1778.85 and 2883.39.

c. Conducting an F-test for the full regression model and examining the relationship between predictors and response.

Full Mathematical Multiple Regression Model

- $\hat{energy} = \beta_0 + \beta_1$ Area + β_2 Age β_3 Rooms β_4 Occupancy
- $\hat{energy} = -3197.28 + 2331.12$ area $+2.36$ age -5.38 rooms $+3234.56$ occupancy

```
coefficients(hotel_full_model)
```

```
##  (Intercept)         area          age        rooms    occupancy 
## -3197.278911  2331.116239     2.357756    -5.383194  3234.553418
```

Hypothesis for the Overall Anova test of multiple regression

The **Null Hypothesis** suggests that area, age, occupancy and room doesn't have an impact in the energy consumption of the hotel. This tells us that there is no linear relationship.
- $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$

The **Alternative Hypothesis** suggests that atleast one of the predictor variables impact the hotel's energy consumption.
- H_1 : at least one $\beta_i \neq 0$

ANOVA Table for Full Multiple Regression Model

```r
anova(hotel_full_model)
```

```
## 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
```

Calculating the Full Regression S.S. and M.S.

Full Reg S.S. $= RegSS_{area} + RegSS_{age|area} + RegSS_{rooms|area,age} + RegSS_{occupancy|area,age,rooms}$

Full Reg S.S. $= 232665641 + 556602 + 1293045 + 1625643 = 236140931$

Reg M.S.$=\frac{Reg.S.S}{k} = \frac{236140931}{4} = 59035232.75$

Computing the F-statistic

F-stat $= \frac{RegM.S}{ResM.S} = \frac{59035232.75}{1332656} = 44.30$

Null distribution for the test statistic

$f_{obs} \sim f_{4,11}$

Computing the P-Value

```r
pf(44.298928418136413, 4, 11, lower.tail = FALSE)
```

```
## [1] 1.01904e-06
```

$P(f_{4,11} \geq 44.30) = 1.01904e - 06$

Conclusion

At the 5% level of significance, the Null hypothesis is rejected because the P-value is less than 0.05. Thus, we can conclude that there is a significant relationship between energy response and at least one of the four predictor variables.

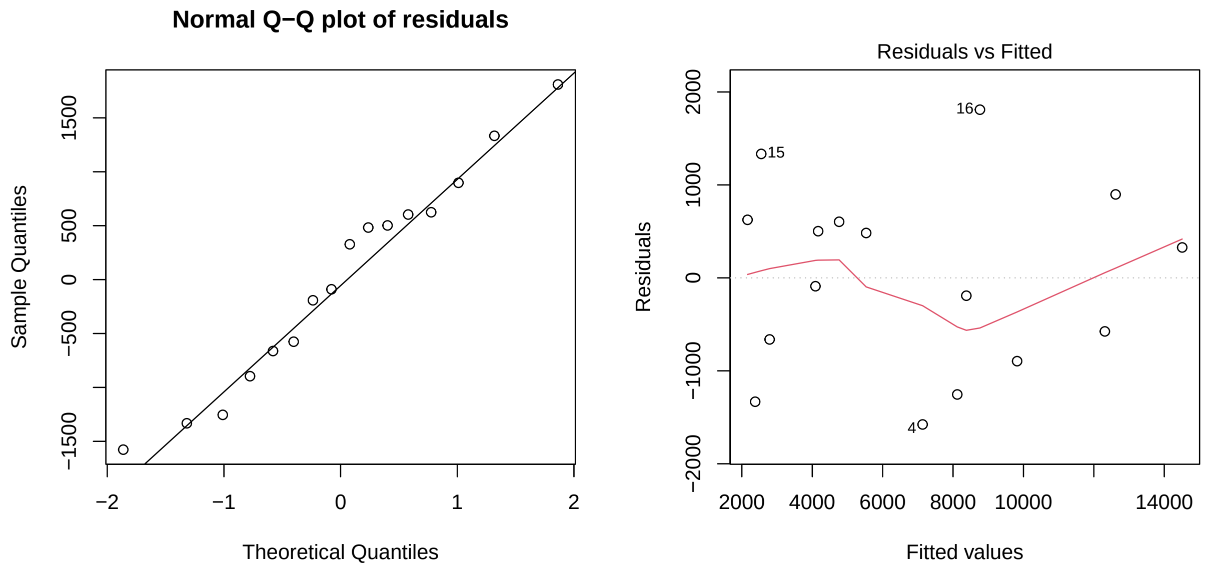
d. Validating the Full Model

Diagnostic Checking

- Regression Equation: $Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon$ where $N(0, \sigma^2)$
- Residuals Q-Q plot checks if the residuals are normally distributed: if $\epsilon \sim Normal$
- fitted vs residuals values plot checks: if the variance of the residuals changes along $\hat{Y}$
- Residuals vs predictor plot, for each i, checks: if variance of residuals changes along X_i

Plotting the figures

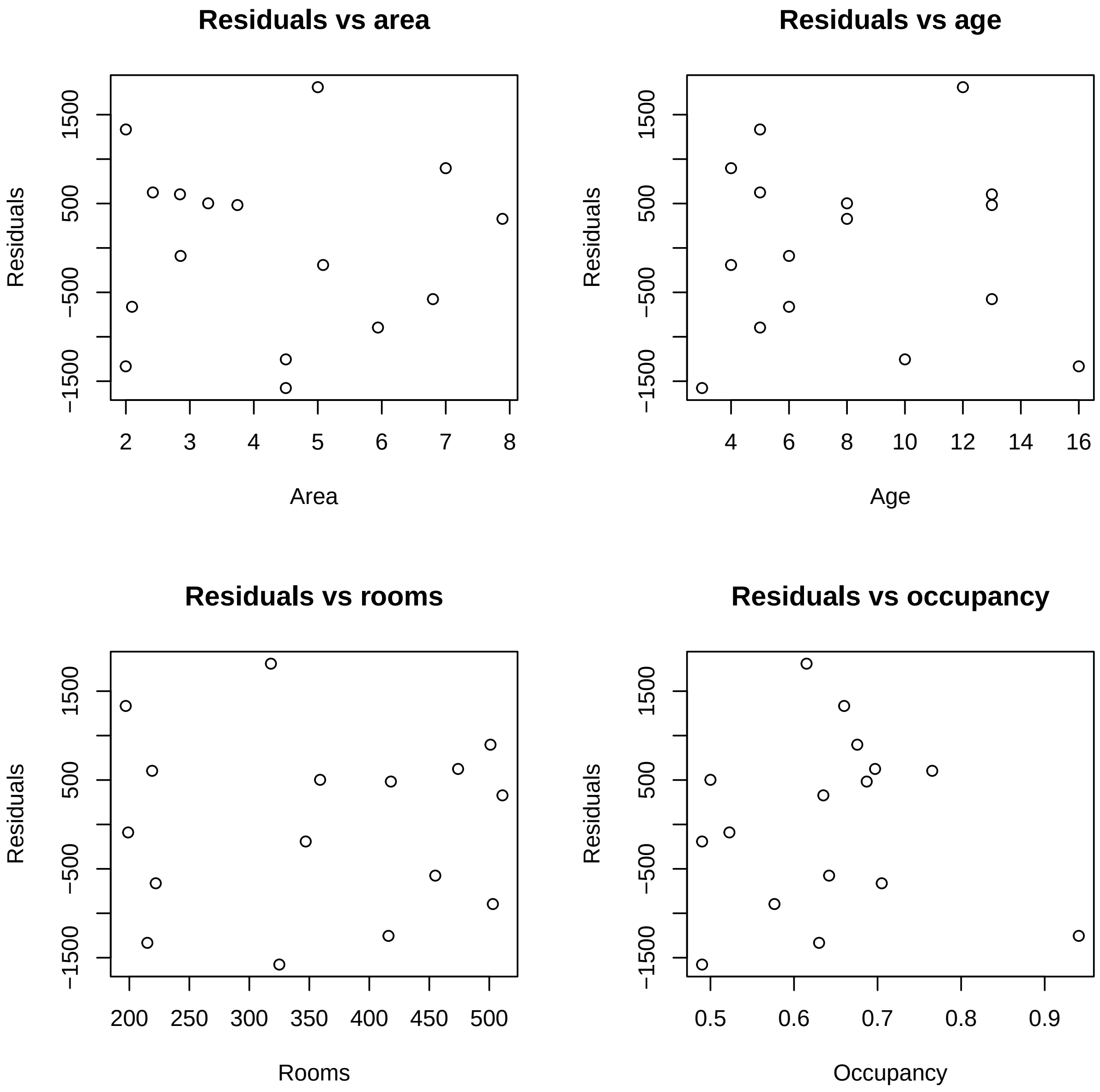
```
par(mfrow = c(1, 2))
qqnorm(hotel_full_model$residuals, main = "Normal Q-Q plot of residuals")
qqline(hotel_full_model$residuals)
plot(hotel_full_model,which=1:1)
```



In a Normal Q-Q plot, the points will roughly fall on the diagonal line if the errors and residuals are normally distributed. We can see that the bulk of the data seems to fit the diagonal line pretty well, therefore, the assumption that residuals are normally distributed is met.

The Residuals vs Fitted plot basically shows you if the residuals have non-linear patterns. Some of the data seems to be roughly scattered in the downward portion of the plot but isn't enough to suggest that it doesn't meet the linearity assumption.

```
par(mfrow = c(2, 2))
plot(hotel$area, hotel_full_model$residuals, main = "Residuals vs area",
xlab = "Area", ylab = "Residuals")
plot(hotel$age, hotel_full_model$residuals, main = "Residuals vs age",
xlab = "Age", ylab = "Residuals")
plot(hotel$rooms, hotel_full_model$residuals, main = "Residuals vs rooms",
xlab = "Rooms", ylab = "Residuals")
plot(hotel$occupancy, hotel_full_model$residuals, main = "Residuals vs occupancy",
xlab = "Occupancy", ylab = "Residuals")
```



In the residuals vs predictor plot, we can see the points are randomly distributed across the horizontal axis throughout all the graphs. This tells us that the regression model is suitable for the data. However, there is less scatter on the residuals vs occupancy plot towards the right side. This may suggest that occupancy doesn't have any relationship with energy and it is better if the feature is excluded from the model.

All in all, the full model seems to be appropriate as it meets all the assumptions. However, the model can be fine-tuned by removing some insignificant features which only increases the model complexity and doesn't impact the response variable much. In this way, the model can be more parsimonious.

e. Finding the R^2 value of the full model

```
summary(hotel_full_model)$r.squared
```

```
## [1] 0.9415502
```

Comment: This means that 94% variance of the response variable (energy) is explained by the predictor variables in the full regression model (area, age, rooms, occupancy).

f. Determining the best Regression Model which explains the data.

Stepwise backward selection

The following steps are followed to prepare the final model
- Step 1: Include all the predictor variables in the model
- Step 2: remove the predictor that is insignificant and has the highest P-Value
- Step 3: Fit the model with the reduced features
- Step 4: Repeat Step 2 and 3 until all predicting variables are not insignificant

```
summary(hotel_full_model)
```

```
##
## Call:
## lm(formula = energy ~ area + age + rooms + occupancy, 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
```

- Only the area variable seems to be significant.
- Age has the highest P-Value (0.977), Thus, Age explains the least variation when fitted to the model.
- As the variable with the highest P-value is dropped, the new model will be regressed with the age variable.

```
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
```

- Both Rooms and Occupancy variables are insignificant.
- Occupancy has the largest P-Value (0.23), therefore, a new model with Occupancy is regressed

```
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
```

- The P-value of room is insignificant at the 5% significance level as its P-value (0.300) is higher than 0.05.
- To make the model parsimonious, the model is regressed with only age as the predictor variable.

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

```
##
## 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
```

Final Fitted Regression Model

```
  coefficients(hotel_final_model)
```

```
## (Intercept)        area
##   -1874.601    2061.408
```

$\widehat{energy} = -1874.60 + 2061.41$ area

g. Comparing the value of R^2 and adjusted R^2 between the full and the final model.

The R^2 value in the full model decreased from **0.9416** in the full model to **0.9277** in the final model. The difference is only under **2%** which indicates that relevant variables are not removed. The value of R^2 increases when more predictors are added to the model. In our full model, all the features are used to predict the energy which leads to better goodness of fit. Hence, focusing solely on high R^2 leads to infinitely many predictors which may be irrelevant and does not add much to the model. If a feature does not increase the R^2 value, there is no need to add a lot of complexity to the model. The R squared doesn’t take into account how significant each one of these features truly is. On the other hand, the adjusted r squared is a better indicator of whether or not the model is getting better with the increasing number of predictors. The Adjusted R^2 attempts to balance R^2 with the number of predictors. It penalizes the number of parameters to offset the increase in R^2. The adjusted R^2 value increased from **0.9203** in the full model to **0.9225** in the final one. This tells us that the new model is parsimonious and also has a high R^2 value.

Question 2 [29 marks]

Loading the dataset

```
movie = read.table(here::here("data", "movie.dat"), header=TRUE)
head(movie)
```

```
##   Gender  Genre Score
## 1      F Action     1
## 2      F Action     1
## 3      F Action     1
## 4      F Action     1
## 5      F Action     2
## 6      F Action     2
```

Column 1 and 2 contains the two categorical predictors

a. Checking if the design is unbalanced

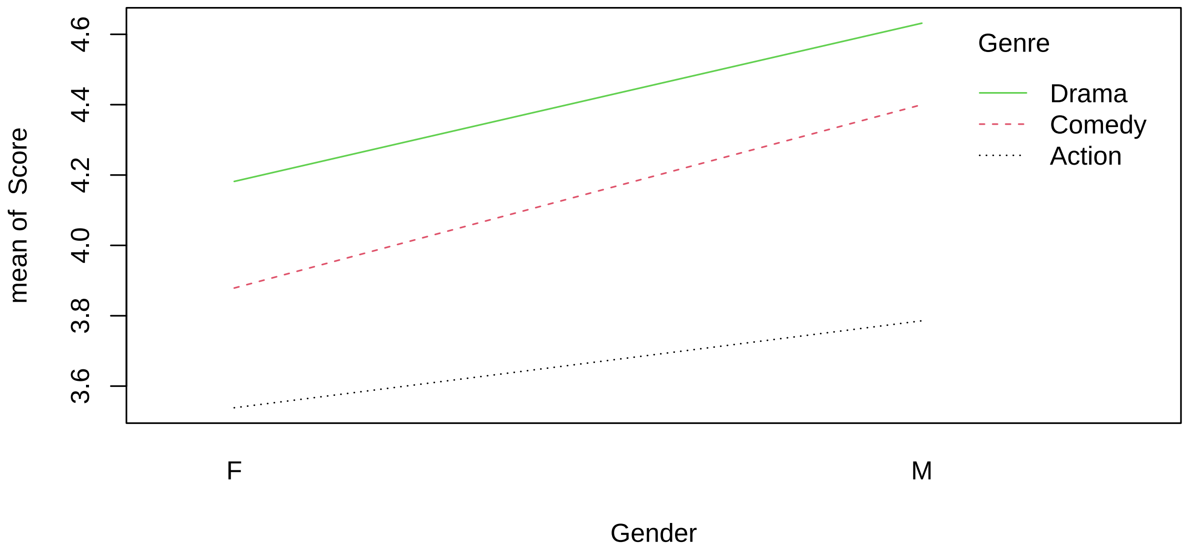
```
table(movie[, 1:2])
```

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

Here we can observe that the count of replicates differ for different levels of each factor combination. Therefore, the design is unbalanced.

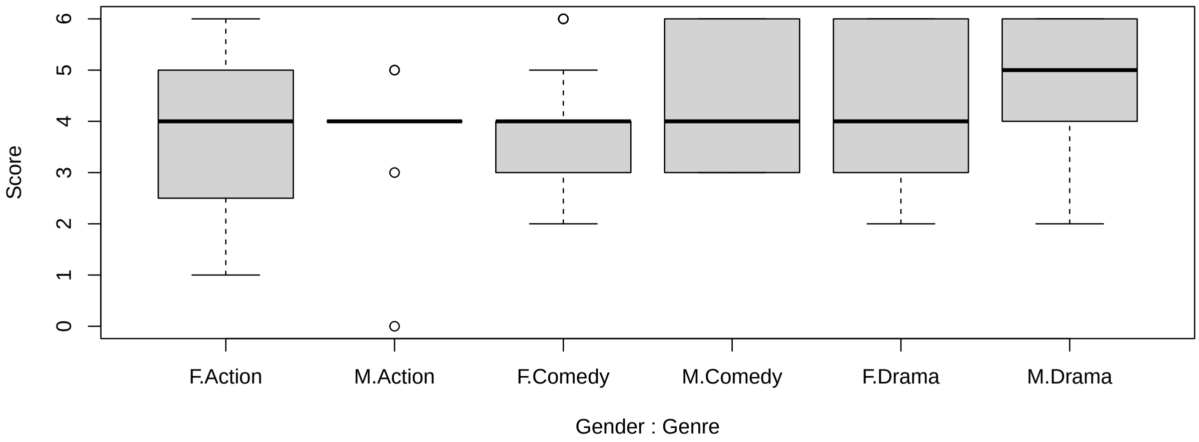
b. Constructing preliminary graphs that investigates various characteristics of the data

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



Comment: The plot is showing signs of possible interaction between the response and the factors due to the fact that the lines are not parallel. Since the sample sizes for some factor combinations are small, it is difficult to make valid conclusions from the graphs.

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



Comment: The second group, third and last group seems to have an unusual spread. This may be due to low replicates. The spread of the first, fourth and the fifth group is identical. The range of the second group seems to be not existent, this is because there is not a lot of variation of Score data in that category.

c. Stating the full mathematical model and defining all the parameters

$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \epsilon_{ijj} \sim N(0, \sigma^2)$

- Response: $Y_{ijk} = k^{th}$ replicate of the treatment at i^{th} level in Gender and j^{th} level in Genre.
- Y_{ijk} = Score response
- μ = Overall population mean
- α_i= Gender effect
- β_j= Genre effect
- γ_{ij}= Interaction effect between Gender and Genre
- ϵ_{ijk}= is the unexplained variation for each replicated observation
- $\epsilon_{ijk} \sim N(0, \sigma^2)$

Final Mathematical Model

```
movie.model = lm(Score ~ Gender + Genre, data = movie)
coefficients(movie.model)
```

```
## (Intercept)     GenderM GenreComedy  GenreDrama
##   3.4993887   0.3951712   0.4087110   0.7077271
```

$Sc\hat{o}re = 3.50 + 0.40$ GenderM + 0.41 GenreComedy + 0.71 GenreDrama

d. Analyzing the effect of Gender and Genre on Score

Null and Alternate Hypothesis

- H_0: no interaction | $\gamma_{ij} = 0$ for all i, j.
- The effect of Gender is the same whatever the type of Genre is and vice versa.
- H_1: there is interaction | not all $\gamma_{ij} = 0$
- The effect of Gender will change depending on the level of Genre and vice versa.

Fitting the model: Regression approach

```
movie.interaction = lm(Score ~ Gender * Genre, data = movie)
anova(movie.interaction)
```

```
## 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
```

- The Interaction term is not significant, therefore, we fail to reject the null hypothesis. From this, we have no evidence to suggest that Gender and Age are dependent.

From the above anova table, we know that the interaction term for gender and genre is insignificant and can be dropped. We choose a model that will examine the impact of genre after accounting gender.

```
movie.final = lm(Score ~ Gender + Genre, data = movie)
anova(movie.final)
```

```
## 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
```

We can clearly observe that both factors are significant. This means that both Genre and Gender have a significant effect on the Movie Score.

Does order matters when the design is unbalanced?

- Reversing the order of the Two-way Analysis of Variance

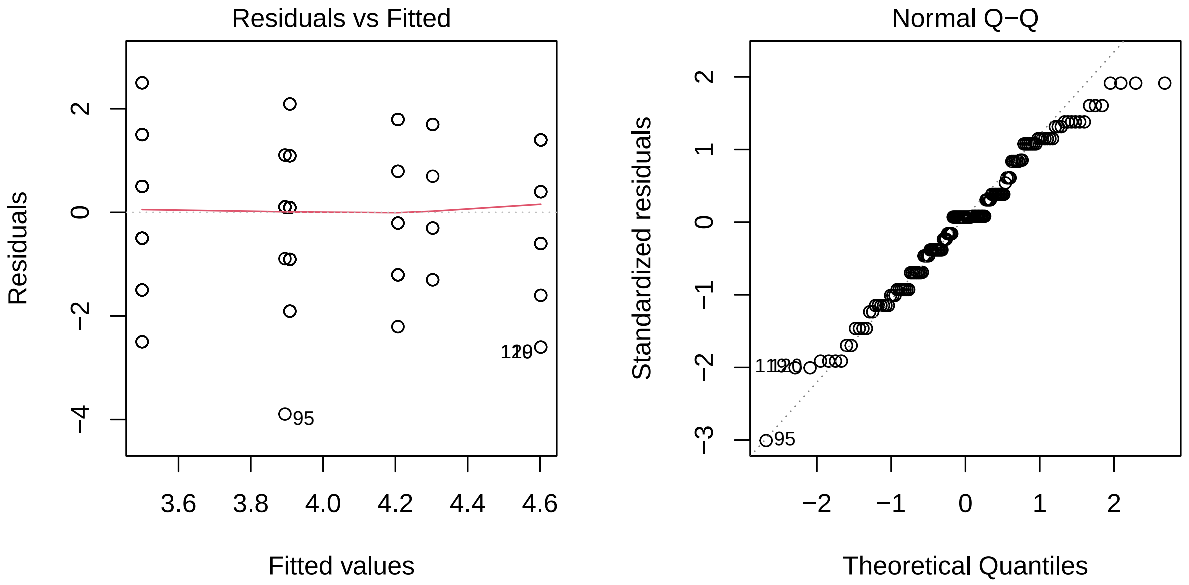
```
movie.final.rev = lm(Score ~ Genre + Gender, data = movie)
anova(movie.final.rev)
```

```
## Analysis of Variance Table
##
## Response: Score
##            Df  Sum Sq Mean Sq F value  Pr(>F)
## Genre       2  14.382  7.1911  4.1218 0.01833 *
## Gender      1   4.399  4.3993  2.5216 0.11467
## Residuals 133 232.036  1.7446
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

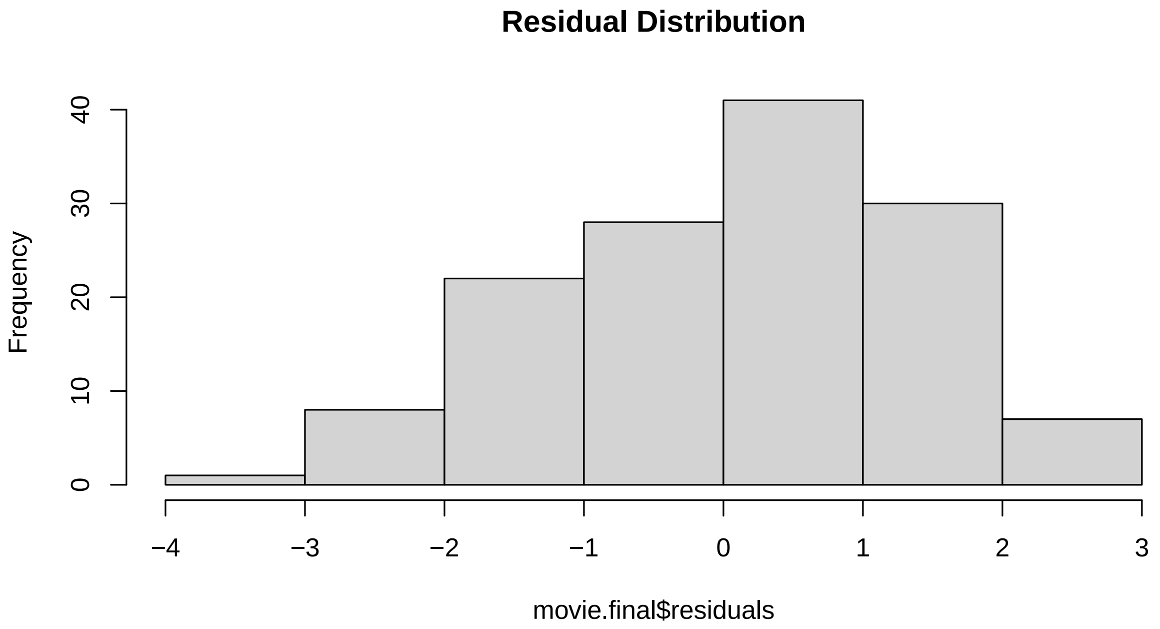
When we reverse the model, We notice that the SS values and the P-Values becomes slightly different when the order is reversed. In the reversed model, the P-value of Gender is greater than the alpha value of 0.05, thus, is insignificant. Therefore, in unbalanced design order does matter.

Checking assumptions for the final model

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



```
hist(movie.final$residuals, main = "Residual Distribution")
```



Observation: The diagnostic plots presented above appears to validate the final model. There doesn't seem to be any trend or pattern in the residual or the fitted values, the variability between effects appears to be constant. The Normal QQ shows slight curvature on the tails but the majority of the points are closer to the diagonal line suggesting that the residuals are close to normally distributed. Furthermore, the histogram of the residual seems to be skewed to the left but not enough to prove that it isn't normally distributed.

Conclusion: The design seems to be unbalanced as the count of replicates is different for each factor combination. The preliminary graph showed signs of interaction but after fitting the interaction model, we found out that the value of the interaction term is insignificant, therefore, we fail to reject the null hypothesis. After that, we proceeded to fit the model with just the main effects (Gender and Genre) and both of their P-values were significant. However, upon reversing the model, the SS and P-value changed. This suggests that order does matter when the design is unbalanced. The assumptions of the final model were finally checked using normal QQ plot, residual vs fitted plot and the histogram of the residuals. Upon checking the assumptions, the model appears to be suitable to describe the data.