STAT2170 Assignment 1

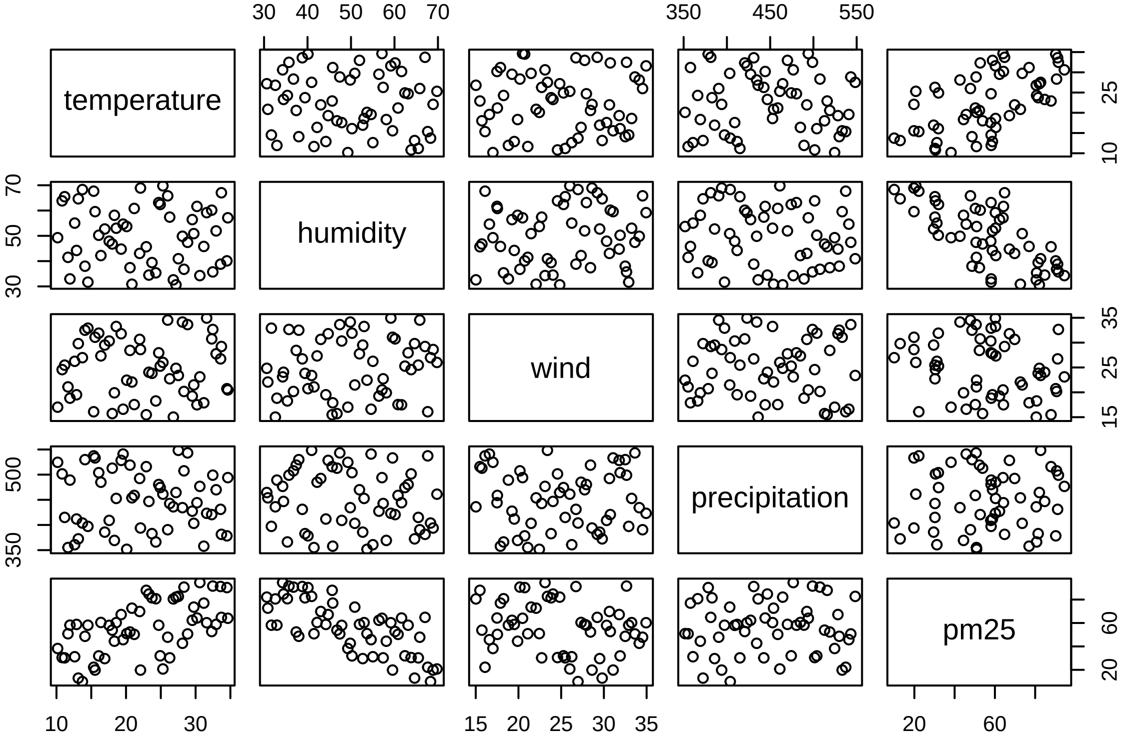
Elijah Sua-Huirua 47398167

2023-05-19

```
## [1] "temperature"   "humidity"      "wind"          "precipitation"
## [5] "pm25"
```

Question 1

a) [7 marks] 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.



A Scatterplot will assist in finding linear relationships among the variables. The results indicate that there are no distinct linear relationships among the variables. Some linear relationships can be observed between pm25 and humidity, pm25 and wind, and pm25 and precipitation. The remaining variables show little apparent relationship, appearing to exhibit a relatively random pattern.

```
##               temperature   humidity        wind precipitation        pm25
## temperature    1.00000000 -0.07264891  0.02861166   -0.05050014  0.57191961
## humidity      -0.07264891  1.00000000  0.12406351   -0.13550607 -0.71965591
## wind           0.02861166  0.12406351  1.00000000   -0.01525977 -0.21866823
## precipitation -0.05050014 -0.13550607 -0.01525977    1.00000000  0.03759033
## pm25           0.57191961 -0.71965591 -0.21866823    0.03759033  1.00000000
```

After examining the correlation matrix, we can determine the presence and strength of linear relationships between variables based on the sign and magnitude of the correlation coefficients. The strength of the correlation proportional to how close the value is to 1. From our correlation matrix output, we can see the correlation between pm25 and humidity, pm25 and wind, and pm25 and precipitation. Specifically a positive linear relationships between pm25 and temperate (0.57), and two negative linear relationships with humidity (-0.72), and wind (-0.22).

b) [6 marks] Fit a model using all the predictors to explain the pm25 response. Using the full model, estimate the impact of humidity on PM2.5 concentration. Do this by producing a 95% confidence interval that quantifies the change in PM2.5 concentration for each extra percentage of relative humidity and comment.

```
##
## Call:
## lm(formula = pm25 ~ temperature + humidity + wind + precipitation,
##     data = pm25dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.759  -6.804  -1.649   6.857  20.975
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   102.72259   14.71953   6.979 5.88e-09 ***
## temperature     1.62142    0.18762   8.642 1.46e-11 ***
## humidity       -1.27742    0.11854 -10.776 9.49e-15 ***
## wind           -0.58016    0.23405  -2.479   0.0165 *
## precipitation  -0.01091    0.02350  -0.464   0.6444
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.06 on 51 degrees of freedom
## Multiple R-squared:  0.8127, Adjusted R-squared:  0.7981
## F-statistic: 55.34 on 4 and 51 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = pm25 ~ temperature + humidity + wind + precipitation,
##     data = pm25dat)
##
## Coefficients:
##   (Intercept)    temperature       humidity           wind  precipitation
##     102.72259        1.62142       -1.27742       -0.58016       -0.01091
```

```
# Estimate the impact of humidity
humidity_coef <- coef(pm25model)["humidity"];humidity_coef
```

```
##  humidity
## -1.277423
```

```
humidity_se <- sqrt(vcov(pm25model)["humidity", "humidity"]);humidity_se
```

```
## [1] 0.1185437
```

```
humidity_ci <- humidity_coef + c(-1, 1) * qt(0.975, df = pm25model$df.residual) * humidity_se ;humidity_
```

```
## [1] -1.515409 -1.039436
```

```
## Impact of humidity on PM2.5 concentration:
```

```
## Coefficient Estimate: -1.277423
```

```
## Standard Error: 0.1185437
```

```
## 95% Confidence Interval: -1.515409 -1.039436
```

The results displayed using the cat() function show the impact of humidity on PM2.5 concentration by printing the coefficient estimate, standard error, and 95% confidence interval.

c) [14 marks] Conduct an F-test for the overall regression i.e. is there any relationship between the response and the predictors.

$$PM2.5_i = \beta_0 + temperature\beta_1 + humidity\beta_2 + wind\beta_3 + precipitation\beta_4 + \epsilon_i$$

The multiple regression model can be used to estimate the relationship between the response variable (PM2.5) and the predictor variables (temperature, humidity, wind, precipitation).

Hypotheses for overall ANOVA test:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4$$

$$H_1 : \beta_i \neq 0$$

```
## Analysis of Variance Table
##
## Response: pm25
##               Df  Sum Sq Mean Sq  F value    Pr(>F)    
## temperature    1  9014.4  9014.4  89.0853 8.908e-13 ***
## humidity       1 12739.7 12739.7 125.9013 2.200e-15 ***
## wind           1   622.6   622.6   6.1533   0.01646 *  
## precipitation  1    21.8    21.8   0.2156   0.64440    
## Residuals     51  5160.6   101.2                       
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Full regression SS = 9014.4 + 12739.7 + 622.6 + 21.8 = 22,398.5 Regression MS = 22,398.5/4 = 5,599.63 Test Stat: F observed = Regression MS / Residual MS = 5,599.625/101.2 = 55.34

```r
# Null Distribution
p_value <- pf(55.34, 4, 51, lower.tail = FALSE); p_value
```

```
## [1] 6.079192e-18
```

Null Distribution: 6.079192e-18

P-Value: P(4,51>55.33) = 0.01 < 2.2e-16

Conclusion: Given alpha = 0.05

In the provided ANOVA table, the p-values for the predictor variables and the overall regression are as follows:

temperature: 8.908e-13 (p < 0.05) humidity: 2.200e-15 (p < 0.05) wind: 0.01646 (p < 0.05) precipitation: 0.64440 (p > 0.05)

Since the p-values for temperature, humidity, and wind are all less than the significance level of 0.05, we have sufficient evidence to reject the null hypothesis for these variables. This means that these variables have a statistically significant relationship with the response variable (pm25).

On the other hand, the p-value for precipitation is greater than the significance level of 0.05. Therefore, we do not have enough evidence to reject the null hypothesis for precipitation, indicating that it does not have a statistically significant relationship with pm25.

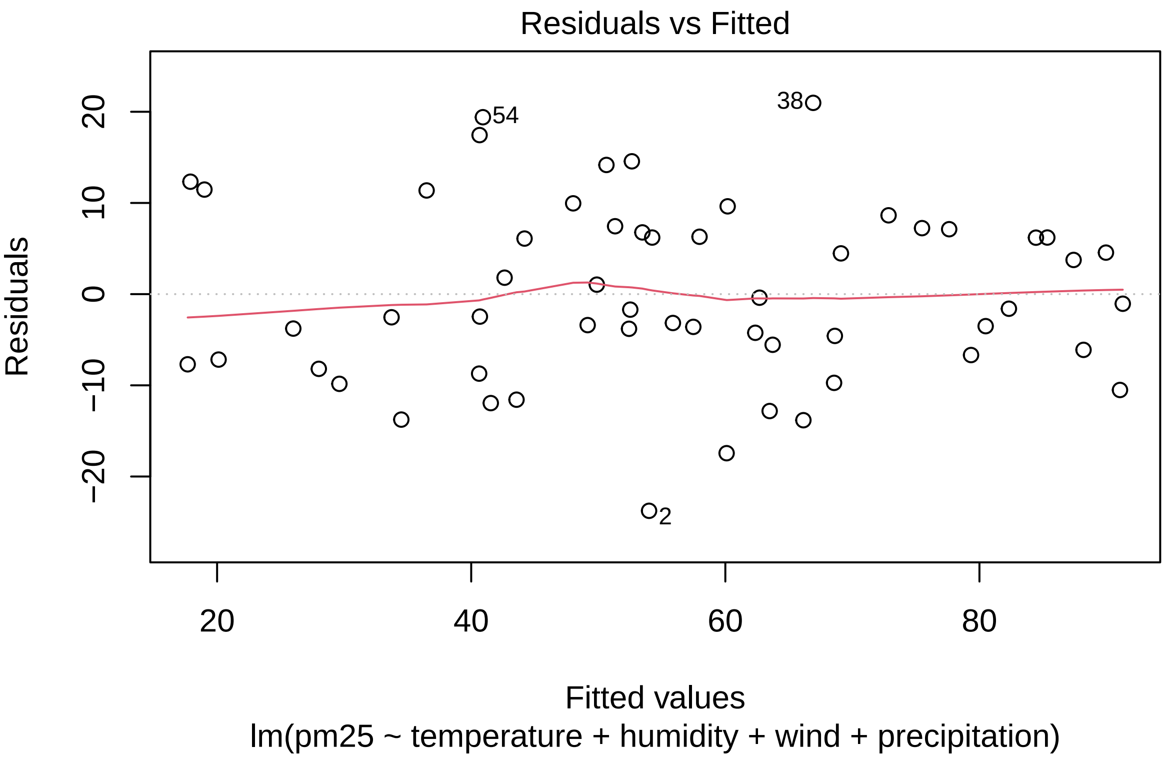
In summary, we reject the null hypothesis for temperature, humidity, and wind, but we fail to reject the null hypothesis for precipitation.

```
full_model <- lm(pm25 ~ temperature + humidity + wind + precipitation, data = pm25dat)
summary(full_model)
```

d) Validate the full model and comment on whether the full regression model is appropriate to explain the PM2.5 concentration at various test locations.

```
##
## Call:
## lm(formula = pm25 ~ temperature + humidity + wind + precipitation,
##     data = pm25dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.759  -6.804  -1.649   6.857  20.975
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   102.72259   14.71953   6.979 5.88e-09 ***
## temperature     1.62142    0.18762   8.642 1.46e-11 ***
## humidity       -1.27742    0.11854 -10.776 9.49e-15 ***
## wind           -0.58016    0.23405  -2.479   0.0165 *
## precipitation  -0.01091    0.02350  -0.464   0.6444
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.06 on 51 degrees of freedom
## Multiple R-squared:  0.8127, Adjusted R-squared:  0.7981
## F-statistic: 55.34 on 4 and 51 DF,  p-value: < 2.2e-16
```

```
plot(full_model, which = 1)
```



The full regression model provides an appropriate explanation for the PM2.5 concentration at various test locations. Key observations from the analysis include:

Significant predictors: The coefficients of the predictors (temperature, humidity, wind, and precipitation) are statistically significant, indicating that these variables have a significant impact on the PM2.5 concentration.

Good model fit: The adjusted R-squared value of 0.7981 suggests that approximately 79.81% of the variability in the PM2.5 concentration is accounted for by the predictors in the model. This indicates a reasonably good fit of the model to the data.

Overall model significance: The F-statistic of 55.34 with a p-value of < 2.2e-16 indicates that the overall model is statistically significant. This implies that the combined effect of the predictors significantly influences the PM2.5 concentration.

Residual standard error: The residual standard error of 10.06 represents the average difference between the observed PM2.5 concentrations and the model's predicted values. A lower value indicates a better fit of the model to the data.

e) [2 marks] Find the R2 and comment on what it means in the context of this dataset. In the context of this dataset, an R-squared value of 0.8127 means that approximately 81.27% of the variability in the PM2.5 concentration can be accounted for by the variation in the predictor variables. This indicates a relatively high degree of predictability of the PM2.5 concentration based on the selected predictors.

f) [3 marks] Using model selection procedures discussed in the course, find the best multiple regression model that explains the data. State the final fitted regression model. To identify the most appropriate multiple regression model that explains the data, we can employ stepwise backward estimation:

Step 1: Initially, regress the model using all predictor variables. Step 2: Identify the variable with the highest p-value in the t-test and remove it from the model. Step 3: Re-estimate the regression model using the reduced set of variables. Step 4: Iterate steps 2 and 3 until all variables in the model are statistically significant.

By iteratively eliminating non-significant variables, this approach aims to refine the regression model and identify the subset of predictors that have a significant impact on the log.survival variable.

```
##
## Call:
## lm(formula = pm25 ~ temperature + humidity + wind + precipitation,
##     data = pm25dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.759  -6.804  -1.649   6.857  20.975
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   102.72259   14.71953   6.979 5.88e-09 ***
## temperature     1.62142    0.18762   8.642 1.46e-11 ***
## humidity       -1.27742    0.11854 -10.776 9.49e-15 ***
## wind           -0.58016    0.23405  -2.479   0.0165 *
## precipitation  -0.01091    0.02350  -0.464   0.6444
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.06 on 51 degrees of freedom
## Multiple R-squared:  0.8127, Adjusted R-squared:  0.7981
## F-statistic: 55.34 on 4 and 51 DF,  p-value: < 2.2e-16
```

Remove "precipitation and re-run"

```
##
## Call:
## lm(formula = pm25 ~ temperature + humidity + wind, data = pm25dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.7588  -6.4368  -0.5659   6.4006  20.2813
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  97.3234     8.9561  10.867 5.45e-15 ***
## temperature   1.6267     0.1859   8.753 8.39e-12 ***
## humidity     -1.2698     0.1165 -10.899 4.89e-15 ***
## wind         -0.5806     0.2323  -2.500   0.0156 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.983 on 52 degrees of freedom
## Multiple R-squared:  0.812,  Adjusted R-squared:  0.8011
## F-statistic: 74.84 on 3 and 52 DF,  p-value: < 2.2e-16
```

Now all variable are significant as shown by the p-values, we therefore stop the backwards estimation and conclude this is our final model. The beta values are the coefficient estimates. The r-squared value is 0.812 which means 81.2% of the variance is captured within the model. Adujsted R-sqaured also changes from 0.798 to 0.8011.

```
AIC(full_model)
```

```
## [1] 424.2347
```

```
AIC(readjusted_model)
```

```
## [1] 422.4709
```

We can check how well our model fits using AIC. AIC of the full model gives 424.2347, while AIC of readjusted model gives 422.4709. The smaller the AIC value the better.

The full model has a Multiple R-squared of 0.8127, suggesting that 81.27% of the PM2.5 concentration variability is explained by the predictors. The Adjusted R-squared of 0.7981 adjusts for the number of predictors and provides a more conservative estimate. After removing the non-significant predictor, the adjusted model has a slightly lower Multiple R-squared of 0.812 but an increased Adjusted R-squared of 0.8011.

When a non-significant predictor is included in the model, it may contribute little to the variability explained by the model, leading to a relatively lower adjusted R-squared. Removing the non-significant predictor allows the remaining predictors to capture a higher proportion of the variability in the response variable, resulting in an improved adjusted R-squared.

Question 2

```
## [1] "Gender" "Genre"  "Score"
```

a) [2 marks] For this study, is the design balanced or unbalanced? Explain why. A balanced study has the same amount of replicates for each category whereas a unbalanced design with have varying numbers.

```
table(moviedat$Gender)
```

```
##
##  F  M
## 94 43
```

```
table(moviedat$Genre)
```

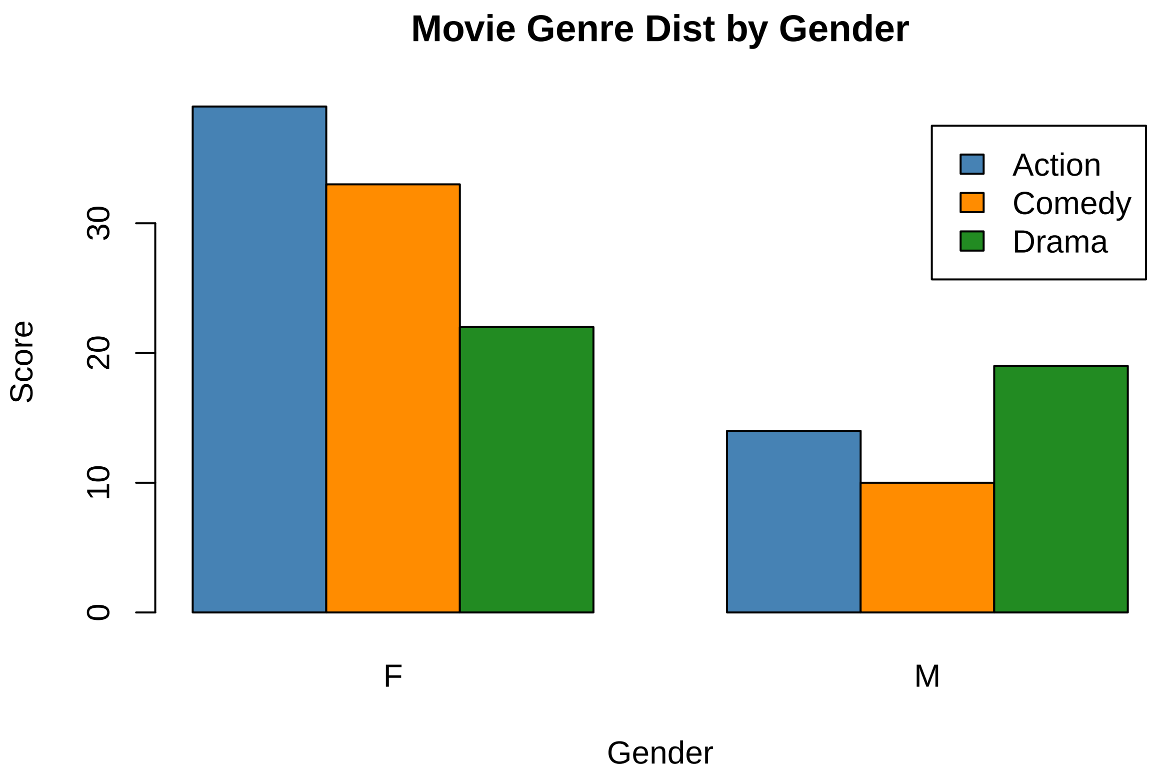
```
##
## Action Comedy  Drama
##     53     43     41
```

The study is unbalanced. The distribution of participants across the Gender variable is uneven, with 94 females and 43 males. Similarly, the distribution across the Genre variable is also uneven, with 53 observations in the Action category, 43 in Comedy, and 41 in Drama.

b) Construct two different preliminary graphs that investigate different features of the data and comment. The two preliminary graphs we will utilize to investigate different features of the data are the stacked bar plot and the interaction plot.

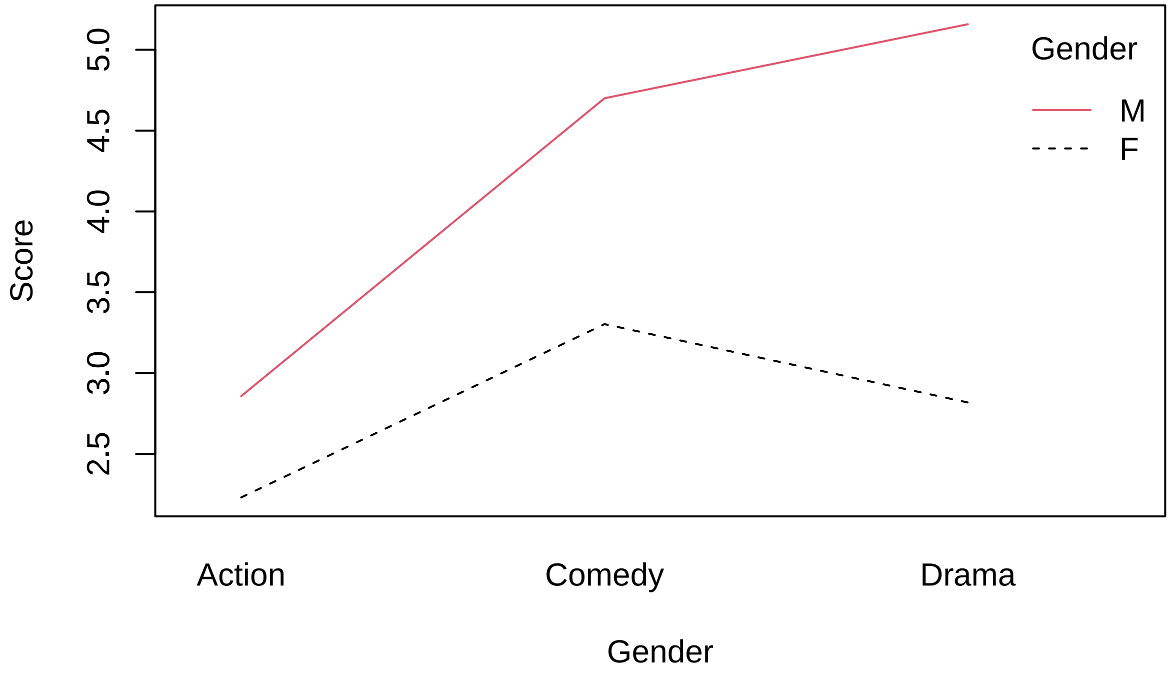
```
# Create a table of counts for the three variables
Score <- table(moviedat$Genre, moviedat$Gender)

# Create the grouped bar plot
barplot(Score, beside = TRUE,
        main = "Movie Genre Dist by Gender",
        xlab = "Gender",
        ylab = "Score",
        col = c("steelblue", "darkorange", "forestgreen"),
        legend = rownames(Score))
```



The stacked bar plot allows us to see the different categories within each variable. The unbalanced data may be having an effect with respect to male female skew.

```
with(moviedat, interaction.plot(Genre, Gender, Score,xlab = "Gender", ylab = "Score", col = 1:3))
```



Looking at the interaction plot we can see that the males gave much higher scores overall, compared to the females. They also rated drama higher with respect to comedy, as opposed to the female, who rated comedy higher. We can also see that both male and female scores increase from action to comedy, but then diverge when it comes to drama, with males favoring drama, while females favor comedy.

c) [4 marks] Write down the full mathematical model for this situation, defining all appropriate parameters. The mathematical model can be written as:

$$Score = \beta_0 + Gender\beta_1 + Genre\beta_2 + (Gender * Genre)\beta_3 + \epsilon$$

Where:

Beta_0 is the intercept, representing the baseline Score when Gender and Genre are both at their reference levels. Beta_1 represents the effect of Gender on the Score. Beta_2 represents the effect of Genre on the Score. Beta_3 represents the interaction effect between Gender and Genre on the Score. (Gender * Genre) represents the interaction term between Gender and Genre. Epsilon represents the error term, accounting for the random variability in the Score not explained by the model.

The model enables us to estimate the impact of Gender, Genre, and their interaction on the Score, considering the potential combined influence of these variables.

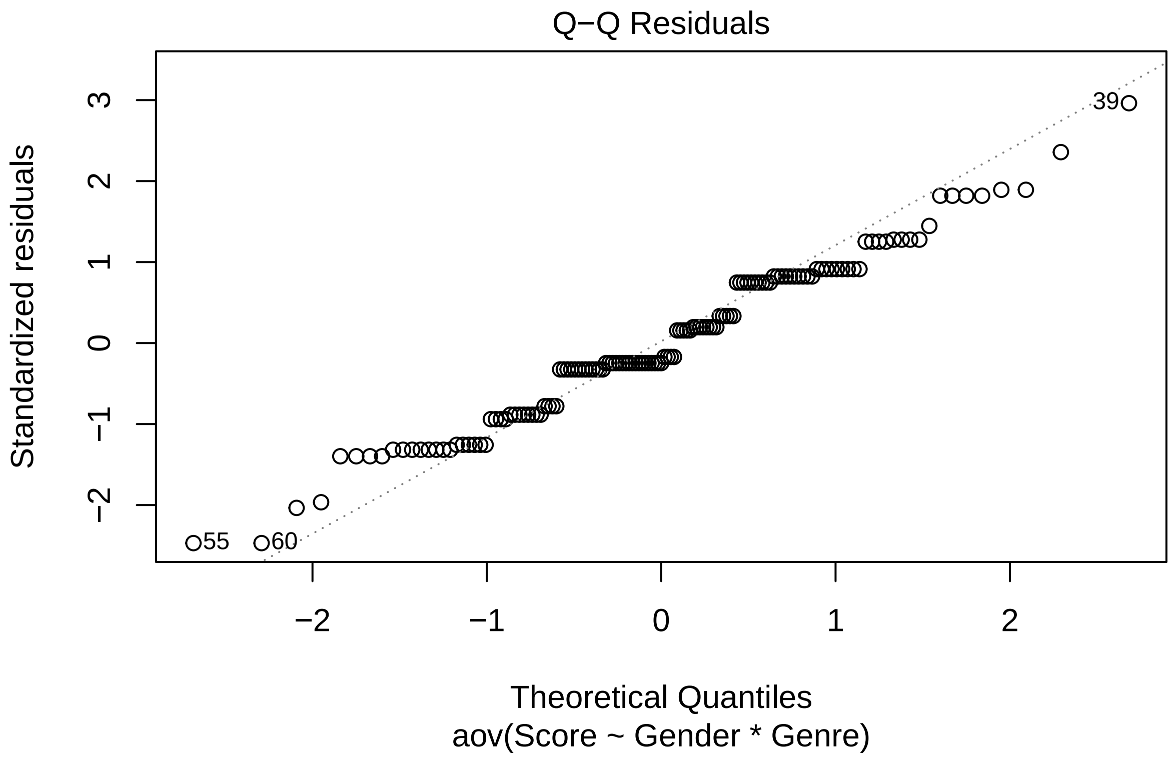
d) [9 marks] Analyse the data to study the effect of Gender and Genre on the brand recall Score. To analyze the data, three types of tests are performed: an interaction test, a main effect test for variable A, and a main effect test for variable B. The null hypothesis for each test is presented below in the respective order. Null hypothesis states True for all i,j, and alternative hypothesis states that not all i,j are true.

$$H_0 : \gamma ij = 0, H_A : \gamma ij = 0$$
$$H_0 : \alpha ij = 0, H_A : \alpha ij = 0$$
$$H_0 : \beta ij = 0, H_A : \beta ij = 0$$

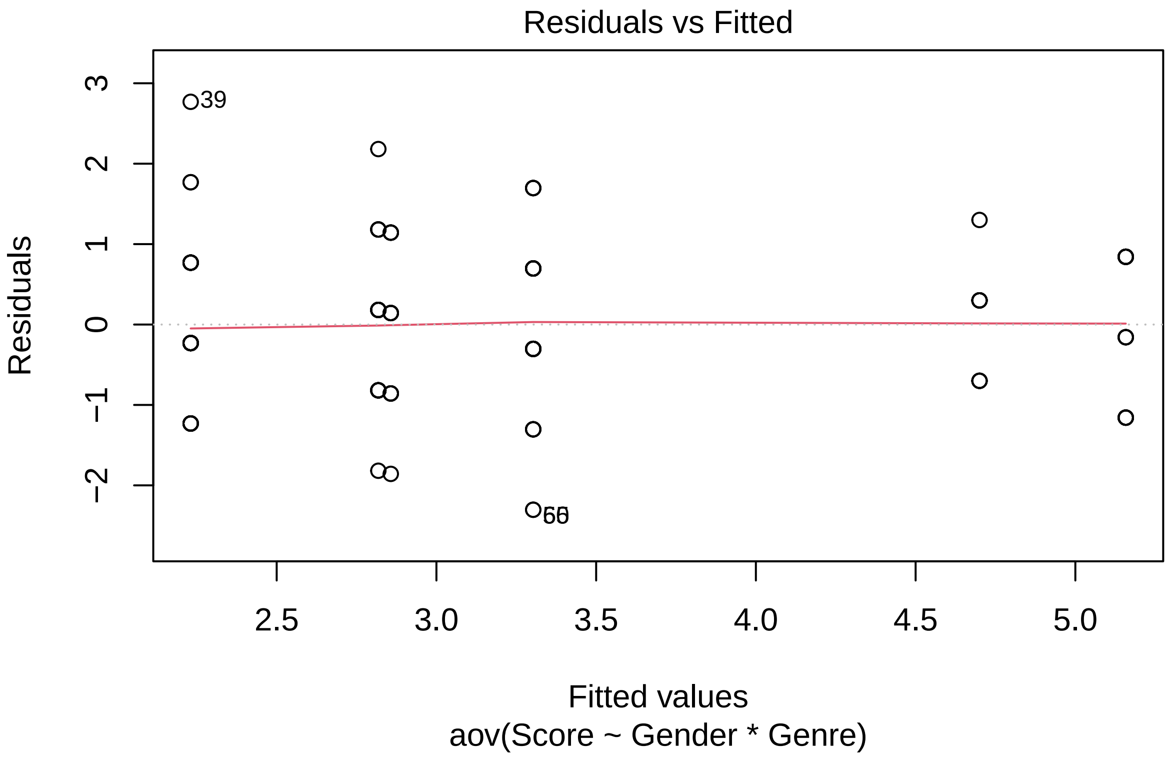
```
score.anova = aov(Score ~ Gender * Genre, data=moviedat)
summary(score.anova)
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)    
## Gender          1  71.58   71.58  79.804 3.28e-15 ***
## Genre           2  50.36   25.18  28.070 7.15e-11 ***
## Gender:Genre    2  15.08    7.54   8.405 0.000368 ***
## Residuals     131 117.51    0.90                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the ANOVA results, both Gender and Genre have independent effects on the Score, and the relationship between Gender and Score varies across different Genre levels. In an unbalanced design, the assumptions that need to be checked are similar to those in a balanced design, but with some additional considerations. We can first look at the Normality of residuals as the residuals should follow a normal distribution.



Next the variability of the residuals should be consistent across all levels of the factors.



From the graphs in b), we can check for independence. The observations should be independent of each other. This assumption is important in unbalanced designs where the number of observations may vary across groups or conditions.

e) [2 marks] Based on your results from part d), discuss the practical implications of your findings for the business that aims to maximise the brand recognition from the placement. What advice/interpretation would you provide on the effect drama genre on the brand recall Score. Based on the analysis of the data, and as we saw from the preliminary graphs in part b) we can see that the drama genre has a significant effect on the brand recall score. The F-value of 28.070 and the associated p-value of 7.15e-11, show us that the effect of the drama genre on the brand recall score is statistically significant. From this, one could suggest that incorporating the drama genre in the placement of the brand can have a positive impact on brand recognition, as audiences are more likely to recall the brand when it is associated with the drama genre. This information could be helpful for the business in terms of maximizing brand recognition and designing effective marketing strategies.

It is important to note that this conclusion is based on the data provided, and that the data is unbalanced, which may skew our analysis. Other factors such as location where data was taken, and brand postioning for example, could influence brand recall. Further data and analysis may be required to gain a more comprehensive understanding of the factors affecting brand recognition in the specific context of the business.