Modeling and prediction for movies

Therese Smit

Setup

Load packages

```r
library(ggplot2)
library(dplyr)
library(statsr)
library(DAAG)
library(knitr)
library(DT)
library(xtable)
```

Load data

```r
load("movies.Rdata")
```

Part 1: Data

This is an observational study, where the data collected is a random sample of movies collected from the IMDB website. This is a case of simple random sampling.

There was no random assignment in this study, therefore no causality. This, and the random selection of the movies, makes the study generalisable.

Part 2: Research question

Which of the following variables, critics_score, audience_score, and top200_box, have an association with a movies popularity?

A multiple linear regression model will be used to fit the above variables to imdb_rating, to determine the best fit.
This is of interest because it will provide valuable insight into the success/failure of a movie, and the overall rating compared to that of critics and/or audiences.

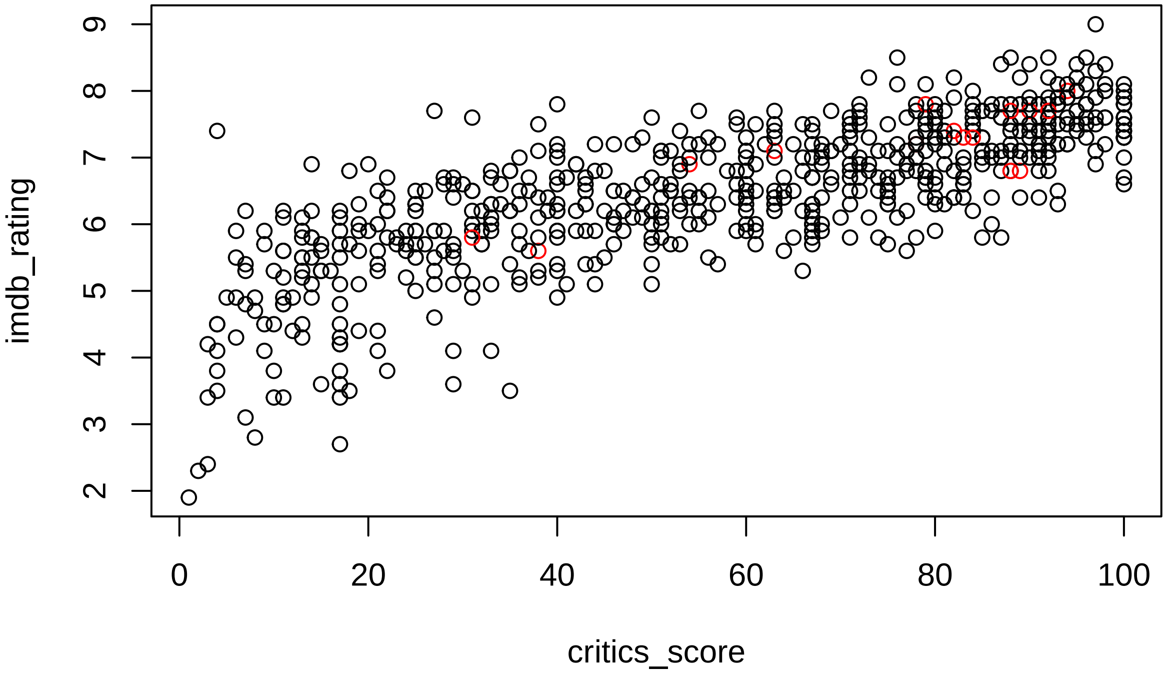
Part 3: Exploratory data analysis

In the figure below, there is a scatterplot comparing the variables imdb_rating and critics_score. The categorical variable, top200_box, is represented by the colour red.

```
movies_1 <- movies[,c("title", "imdb_rating", "critics_score", "audience_score", "top200_box")]

plot1 <- with(movies_1, plot(critics_score, imdb_rating, col = top200_box,
         main = "Figure 1: Scatter plot to represent IMDB rating vs critics score"))
```

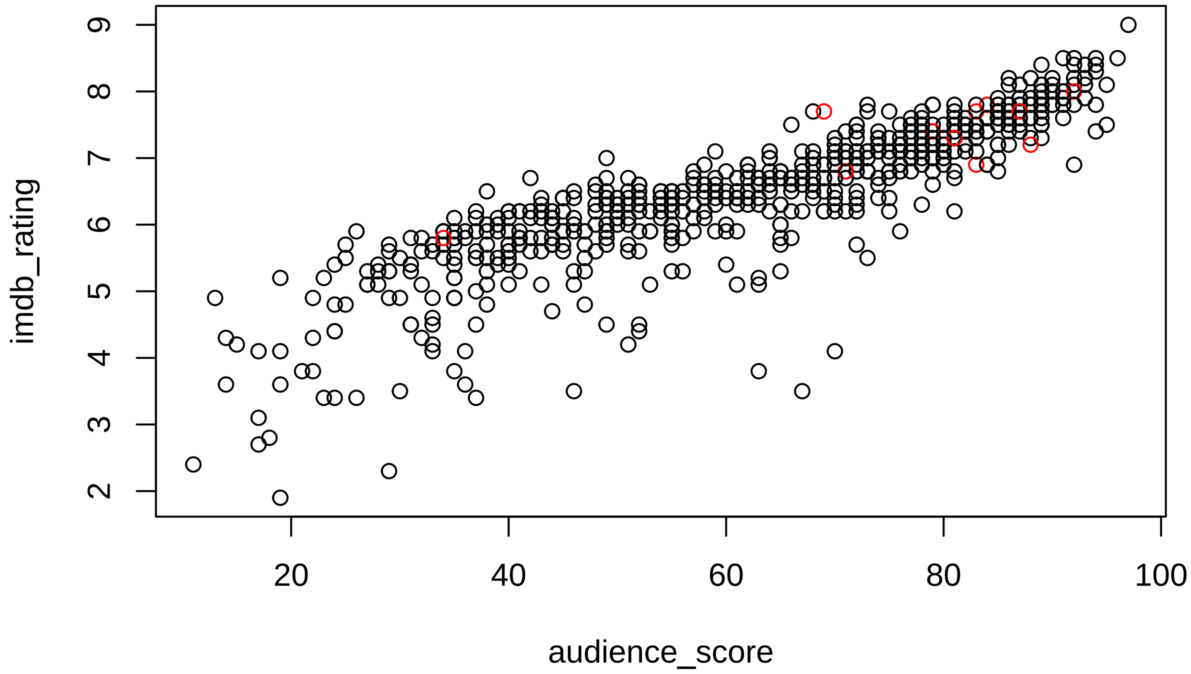
Figure 1: Scatter plot to represent IMDB rating vs critics score



In the figure below, there is a scatterplot comparing the variables imdb_rating and audience_score. The categorical variable, top200_box, is represented by the colour red.

```
plot2 <- with(movies_1, plot(audience_score, imdb_rating, col = top200_box,
         main = "Figure 2: Scatter plot to represent IMDB rating vs audience score"))
```

Figure 2: Scatter plot to represent IMDB rating vs audience score



By comparing the figures above, it can be seen that there is a stronger relationship between the variables imdb_rating and audience_score (Figure 2) than between the variables imdb_rating and critics_score (Figure 1).

In both of the figures above, it can be seen that there are more movies in the top200_box category when the imdb_rating, critics_score, and audience_score is higher.

By looking at the figures above, only an estimation of the relationship between the variables can be given. A more accurate result would be to find the correlation coefficient.

```
movies_1 %>%
  summarise(cor(imdb_rating, critics_score))
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.1
```

```
## # A tibble: 1 x 1
##   `cor(imdb_rating, critics_score)`
##                               <dbl>
## 1                         0.7650355
```

The correlation coefficient between imdb_rating and critics_score is 0.7650355.

```
movies_1 %>%
  summarise(cor(imdb_rating, audience_score))
```

```
## # A tibble: 1 x 1
##   `cor(imdb_rating, audience_score)`
##                                  <dbl>
## 1                            0.8648652
```

The correlation coefficient between imdb_rating and audience_score is 0.0.8648652.

These results confirm the observation made earlier. The relationship between the variables imdb_rating and audience_score is stronger than between the variables imdb_rating and critics_score.

Part 4: Modeling

The following variables will be considered in the full multiple linear regression model: imdb_rating, critics_score, audience_score, top200_box.

```r
rating <- lm(imdb_rating ~ critics_score + audience_score + top200_box, data = movies)
summary(rating)
```

```
## 
## Call:
## lm(formula = imdb_rating ~ critics_score + audience_score + top200_box, 
##     data = movies)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.51954 -0.19711  0.02982  0.30703  1.22706 
## 
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    3.6475021  0.0625645   58.30   <2e-16 ***
## critics_score  0.0118114  0.0009556   12.36   <2e-16 ***
## audience_score 0.0346984  0.0013417   25.86   <2e-16 ***
## top200_boxyes  0.0141542  0.1288856    0.11    0.913    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.4908 on 647 degrees of freedom
## Multiple R-squared:  0.7962, Adjusted R-squared:  0.7953 
## F-statistic: 842.8 on 3 and 647 DF,  p-value: < 2.2e-16
```

The model selection that shall be used is Forward Selection - Adjusted R. This method has been chosen because the Adjusted R yields more reliable predictions.

The adjusted R of each explanatory variable vs. response variable was first calculated.

```r
rating_1 <- lm(imdb_rating ~ critics_score, data = movies)
summary(rating_1)$adj.r.squared
```

```
## [1] 0.5846403
```

```
rating_2 <- lm(imdb_rating ~ audience_score, data = movies)
summary(rating_2)$adj.r.squared
```

```
## [1] 0.7476034
```

```
rating_3 <- lm(imdb_rating ~ top200_box, data = movies)
summary(rating_3)$adj.r.squared
```

```
## [1] 0.006873213
```

The model with the highest adjusted R was selected, and then the remaining variables were added one at a time to this model, with the adjusted R calculated for each one.

```
rating_A <- lm(imdb_rating ~ audience_score + critics_score, data = movies)
summary(rating_A)$adj.r.squared
```

```
## [1] 0.7956067
```

```
rating_B<- lm(imdb_rating ~ audience_score + top200_box, data = movies)
summary(rating_B)$adj.r.squared
```

```
## [1] 0.7473515
```

The above step was then repeated until the addition of any of the remaing variables did not result in a higher adjusted R.

```
rating_C <- lm(imdb_rating ~ audience_score + critics_score + top200_box, data = movies)
summary(rating_C)$adj.r.squared
```

```
## [1] 0.7952946
```

See below the tables containing all the calculated adjusted R's.

```
Variables_Included <- c('imdb_rating ~ critics_score', 'imdb_rating ~ audience_score', 'imdb_rating ~ to
Adjusted_R <- c(0.5846403, 0.7476034, 0.006873213, 0.7956067, 0.7473515, 0.7952946)
df <- data.frame(Variables_Included, Adjusted_R)

kable(df)
```

Variables_Included	Adjusted_R
imdb_rating ~ critics_score	0.5846403
imdb_rating ~ audience_score	0.7476034
imdb_rating ~ top200_box	0.0068732
imdb_rating ~ audience_score + critics_score	0.7956067
imdb_rating ~ audience_score + top200_box	0.7473515
imdb_rating ~ audience_score + critics_score + top200_box	0.7952946

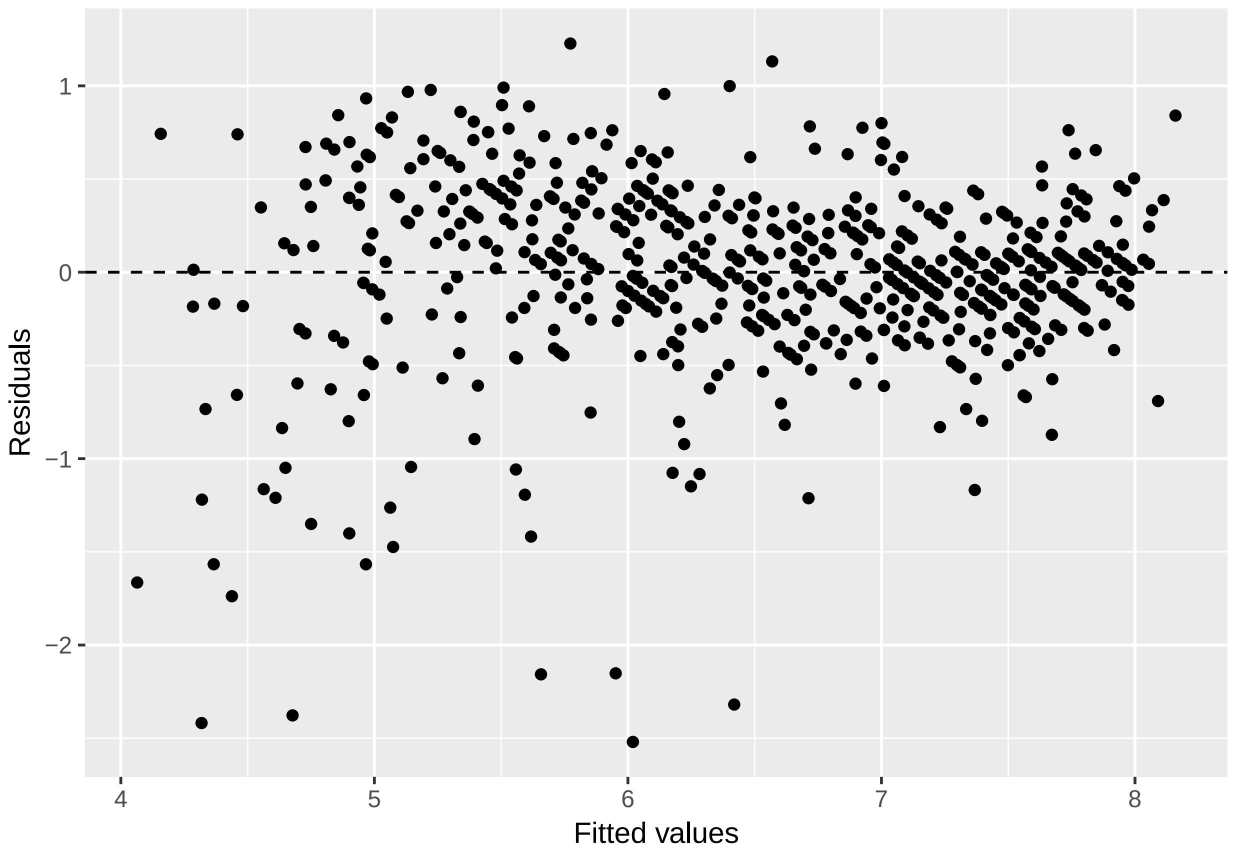
The adjusted R value for the model containing the variables audience_score and critics_score is 0.7956067.
The adjusted R value for the model containing the variables audience_score, critics_score, and top200_box is 0.7952946.
Therefore, the multiple linear regression model chosen is with the variables audience_score and critics_score.

```
final_rating <- lm(imdb_rating ~ critics_score + audience_score, data = movies_1)
summary(final_rating)
```

```
##
## Call:
## lm(formula = imdb_rating ~ critics_score + audience_score, data = movies_1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.51964 -0.19767  0.03466  0.30671  1.22691
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.647241   0.062471   58.38   <2e-16 ***
## critics_score  0.011816   0.000954   12.39   <2e-16 ***
## audience_score 0.034703   0.001340   25.90   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4904 on 648 degrees of freedom
## Multiple R-squared:  0.7962, Adjusted R-squared:  0.7956
## F-statistic:  1266 on 2 and 648 DF,  p-value: < 2.2e-16
```

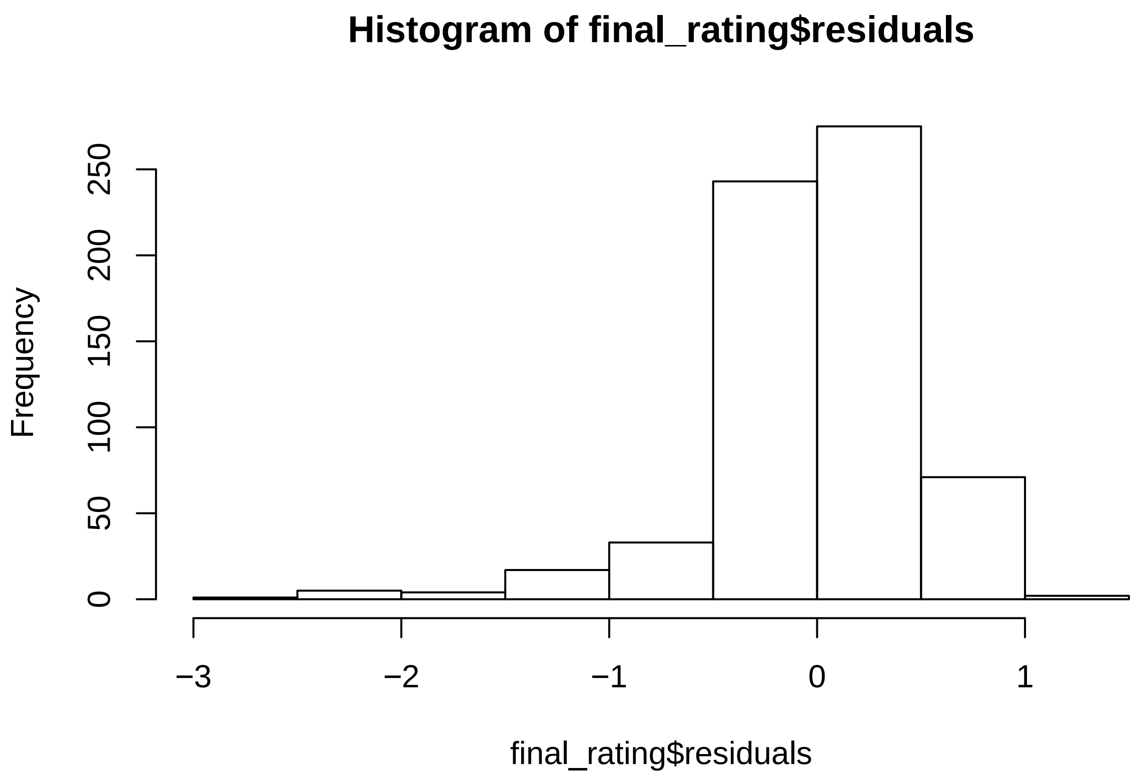
The diagnostics for the multiple linear regression model will be carried out using the models below.
In the residual plot below it can be seen that there is random scatter around 0. This confirms a linear relationship between the x and y variables.

```
ggplot(data = final_rating, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
    xlab("Fitted values") +
  ylab("Residuals")
```

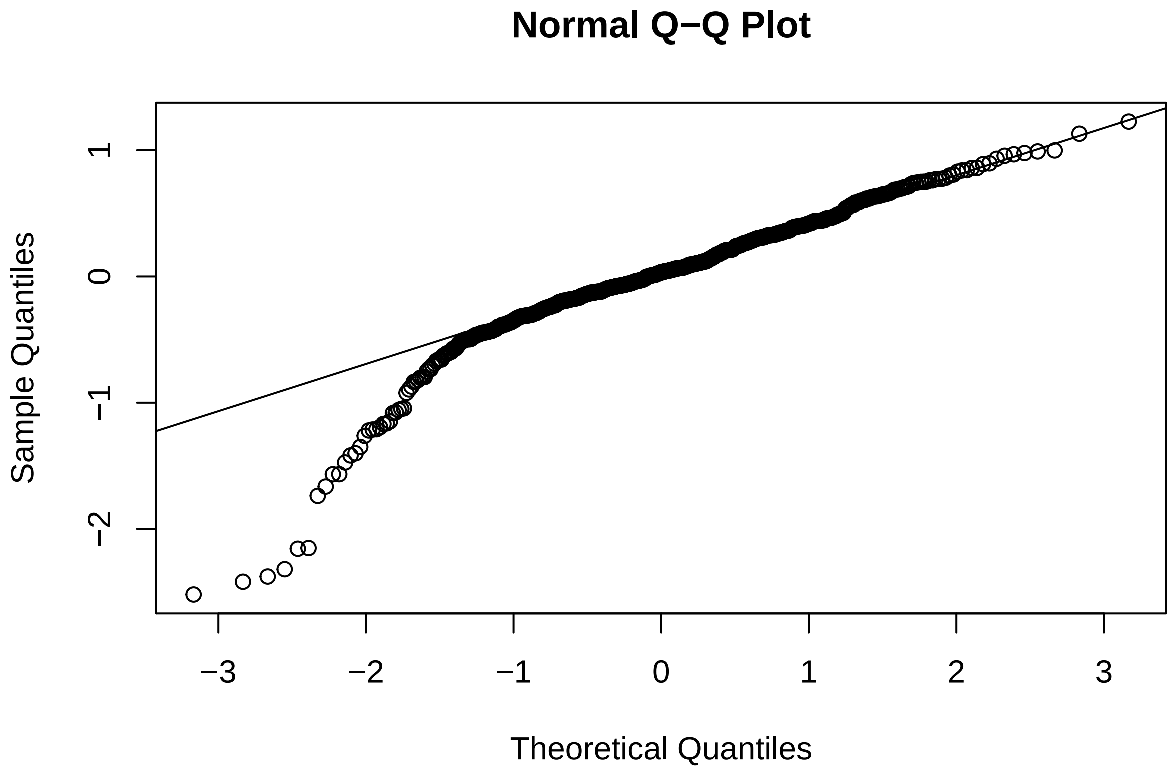


In the histogram and normal Q-Q plot below, it can be seen that the residuals are centered around 0. Therefore the residuals are nearly normal with mean 0.

```
hist(final_rating$residuals)
```

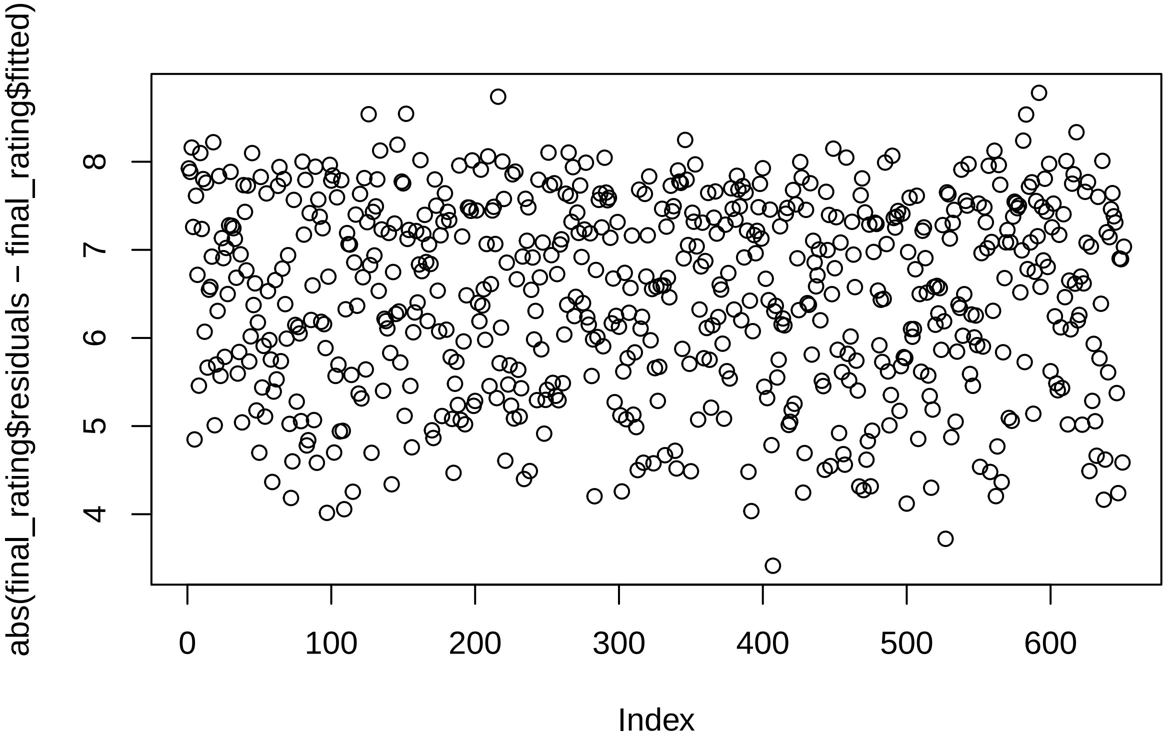


```
qqnorm(final_rating$residuals)
qqline(final_rating$residuals)
```



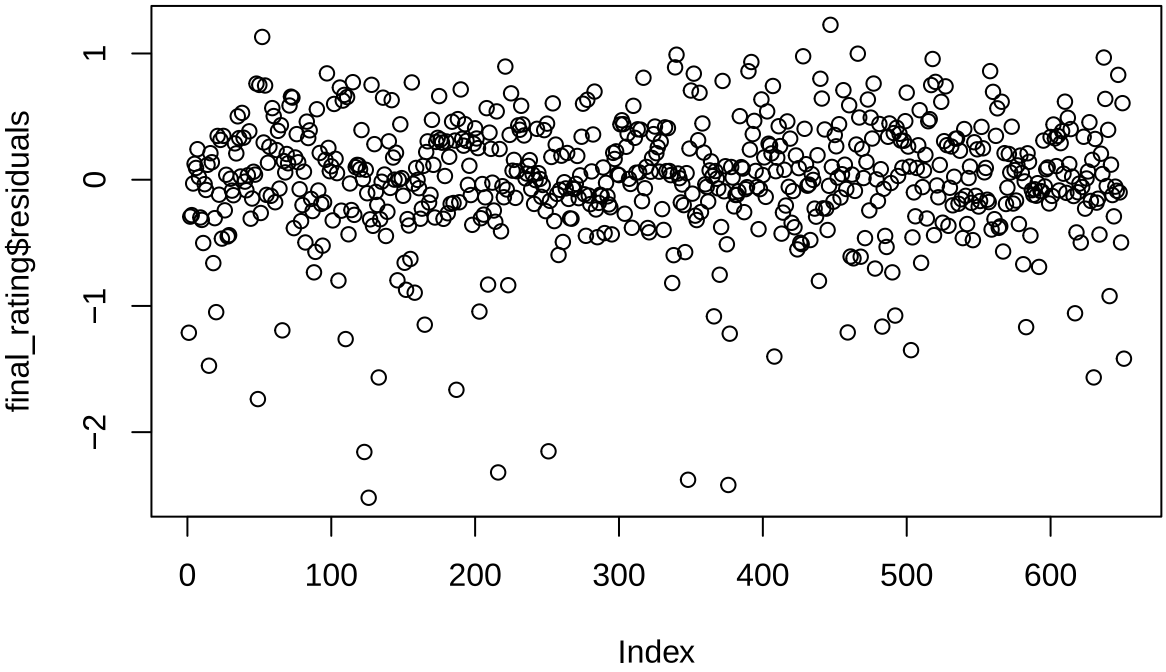
In the residual plot below, where the residuals are plotted against the predicted values, it can be seen that the residuals are not randomly scattered in a band with a constant width around 0. Therefore there is not a constant variability of the residuals.

```
plot(abs(final_rating$residuals - final_rating$fitted))
```



As mentioned previously, in the residual plot below it can be seen that there is random scatter around 0. Therefore the residuals are independent.

```
plot(final_rating$residuals)
```



Part 5: Prediction

The multiple linear regression model above will now be used to predict a movie from 2016. The chosen movie is La La Land, which received an IMDB rating of 8.2/10. This movie is not in the dataset provided for this project.
The parameters we will be including in the prediction is the critics score (92%) and the audience score (81%). This information was found on the IMDB website and the Rotten Tomatoes website.

```r
new_df <- data.frame(title = "La La Land", critics_score = 92,
          audience_score = 81, top200_box = "yes")

predict(final_rating, newdata = new_df, interval = 'prediction')
```

```
##        fit      lwr      upr
## 1 7.545305 6.580488 8.510121
```

As can be seen, the prediction given for the IMDB rating of La La Land is 7.5/10, whereas the real rating is 8.2/10.
The prediction interval calculated is (6.58, 8.51), which tells us that 95% of movies with a critics score of 92% and an audience score of 81% have an IMDB rating somewhere between 6.58 and 8.51.
Since 8.2 is within that interval, the multiple linear regression model created is a good prediction tool for movies.

Part 6: Conclusion

Out of the following vriables critics_score, audience_score, and top200_box, it can be said that critics_score and audience_score had a stronger association with a movies popularity than top200_box did, based on the findings above.
The prediction done on the multiple linear regression model created accurately placed the real IMDB rating in the prediction interval, however the direct prediction on 7.5 based on the model was not accurate enough. Improvements can be made on the model, such as including more variables to increase the accuracy of the prediction.