# Modeling and prediction for movies

## Setup

### Load packages

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.3
```

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.3
```

```r
library(statsr)
```

```
## Warning: package 'statsr' was built under R version 4.0.3
```

```r
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.0.3
```

```r
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.0.3
```

### Load data

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

## Part 1: Data

The movies data-set gathers data on movies and a few attributes about its structure, like movie length in minutes, release day, month and year, MPAA rating, and so on. Besides these variables, there are a few columns of interest regarding the movie evaluation, such as the Internet Movie Database - IMDB rating, critics rating and audience ratings. The goal of this project is to bring a perspective in one of these classifications, the IMDB ratings, and its relationship with the other variables in the data-set, excluding the other variables of interest.

For this project, it is essential to state the following points:-

i. The original data-set about each movie's score was taken from both the IMDb and Rotten Tomatoes database. Characteristics of the site user community influence trends in specific movie scores, thus this results can be generalized to the whole population.
ii. This article was an observational study. There is no causation can be established because random assignment is not used in this study. In this study, I will assume some voting results about a film on website A could be used to predict the same movie's score on website B. However, in the real world the score for the same movie occurs at the same time on two sites A and B.

## Part 2: Research question

A few features regarding the movie can influence how people perceive the movie. These features not only include genre and run-time, but also consider facts like whether the movie was in the Top 200 Box Office list on BoxOfficeMojo.

What are the factors/variables that influence the IMDb ratings received by a movie from the audience? To what extent are these factors significant in predicting the IMDb score of a movie?

This question is important for us because it is important for online streaming services to predict ratings based on these factors. So, they can accordingly invest in contents meeting certain conditions.

## Part 3: Exploratory data analysis

Let us consider the following variables for constructing a multiple linear regression (MLR) model,

genre: Genre of the movie

runtime: Runtime of movie (in movies)

mpaa_rating: MPAA rating of the movie (G, PG, PG-13, R)

critics_score: Critics score on Rotten Tomatoes

audience_score: Audience score on Rotten Tomatoes

best_pic_nom: Whether or not the movie was nominated for a best picture

best_pic_win: Whether or not the movie won a best picture Oscar

top200_box: Whether or not the movie is in the Top 200 Box Office list on BoxOfficeMojo

best_actor_win: Whether or not one of the main actors in the movie ever won an Oscar

best_actress_win: Whether or not one of the main actresses in the movie ever won an Oscar

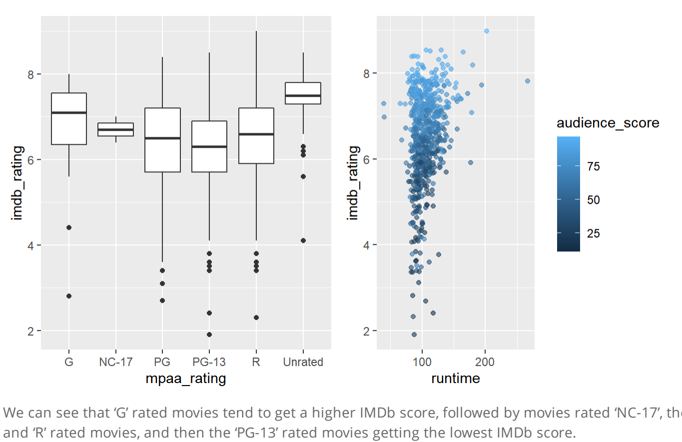
We will use the above variables to try and predict the variable imdb_rating using Multiple Linear Regression model.

First let us clean the model of observations having missing values

```r
stats <- movies %>% filter(!is.na(genre) & !is.na(runtime) & !is.na(mpaa_rating) & !is.na(critics_score) & !is.na(au
```

Note that there is a single observation which has a missing value of one of the variables of our interest. Now, we move on to visualizing the data of IMDb scores based on these factors.

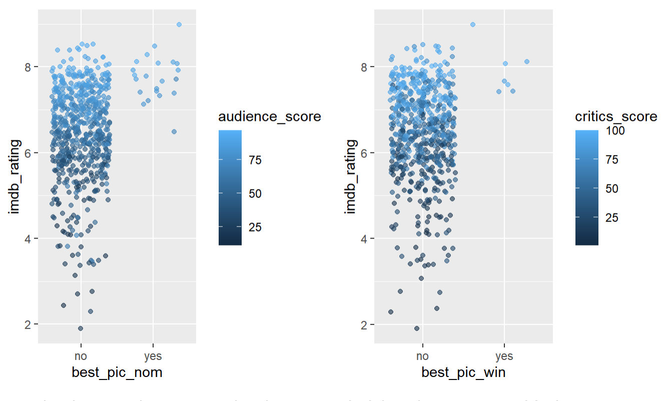
```r
p1 <- ggplot(data = stats, aes(x=mpaa_rating, y=imdb_rating))+geom_boxplot()
p2 <-  ggplot(data=stats, aes(x=runtime, y=imdb_rating, colour=  audience_score))+geom_jitter(aes(alpha=I(0.6)))
grid.arrange(p1,p2, ncol=2)
```




We can see that 'G' rated movies tend to get a higher IMDb score, followed by movies rated 'NC-17', then with a close tie between 'PG' and 'R' rated movies, and then the 'PG-13' rated movies getting the lowest IMDb score.

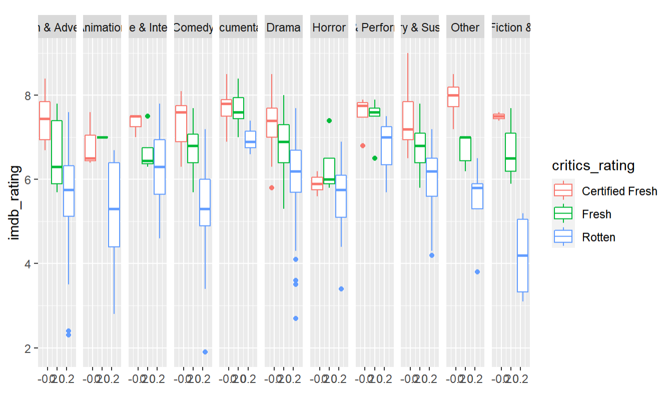
Also, it is to be noted that there is no clear relation between runtime and IMDb score. Although it can be said that audience score is always reliable because there are considerable number of observations with High Audience Score but low IMDb score and vice versa.

```r
p1<-ggplot(data=stats, aes(x=best_pic_nom, y=imdb_rating))+geom_jitter(aes(colour=audience_score, alpha=I(0.6)))
p2<- ggplot(data=stats, aes(x=best_pic_win, y=imdb_rating))+geom_jitter(aes(colour= critics_score, alpha=I(0.6)))
grid.arrange(p1,p2, ncol=2)
```



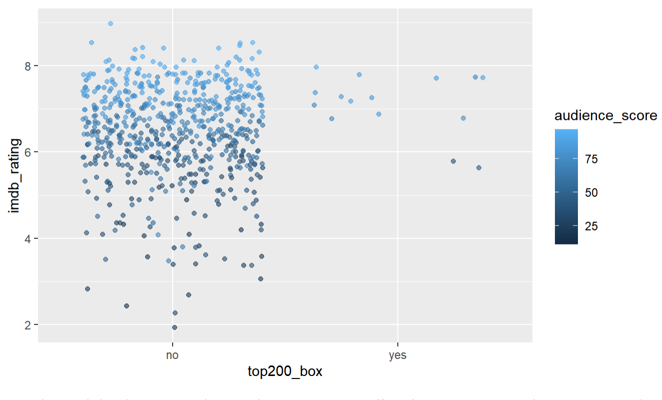

From the above graph we can say that the movies which have been nominated for best picture or won the award for best picture have moderate to high IMDb score. In fact, while comparing the two graphs it can be observed that the movies which were nominated and received a moderate IMDb score did not win the prize at the end.

```r
ggplot(data=stats, aes(y=imdb_rating, colour=critics_rating))+geom_boxplot()+facet_grid(.~genre, scale="free")
```



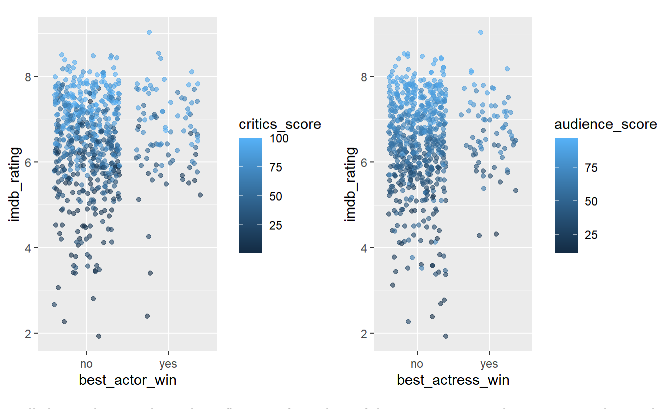

It can be concluded from the above facet that the Horror genre gets the lowest IMDb rating on an average. So that genre is pretty risky for production houses to invest into. While genres like Drama, Documentary, Comedy and Action and Adventure are good sectors for investing.

```r
ggplot(data=stats, aes(x=top200_box, y=imdb_rating))+geom_jitter(aes(colour=audience_score, alpha=I(0.6)))
```




It can be said that bagging a place in the Top 200 Box Office does not necessarily ensure a good IMDb rating. In some observations we even see that the Audience have given a less score to the movies in the Top 200 Box Office list.

```r
p1<- ggplot(data=stats, aes(x=best_actor_win, y=imdb_rating))+geom_jitter(aes(colour=critics_score, alpha=I(0.6)))
p2<- ggplot(data=stats, aes(x=best_actress_win, y=imdb_rating))+geom_jitter(aes(colour=audience_score, alpha=I(0.6)
grid.arrange(p1,p2, ncol=2)
```




Finally let us discuss about the influence of stardom of the main actors and actresses on the IMDb score of the movie they were featured in. It can be observed that featuring an Oscar winning actor doesn't save a bad movie from receiving a low IMDb score. So, is the case for featuring an Oscar winning actress, although the lowest score is comparatively greater then the former category.

## Part 4: Modeling

Now let us form the multiple linear regression model using all the variables we have analyzed in the Exploratory Data Analysis. We use the method of backward elimination based on p-values.

```r
model <- lm(imdb_rating ~ genre+runtime+mpaa_rating+critics_score+audience_score+best_pic_nom+best_pic_win+top200_bo
summary(model)
```

```
##
## Call:
## lm(formula = imdb_rating ~ genre + runtime + mpaa_rating + critics_score +
##     audience_score + best_pic_nom + best_pic_win + top200_box +
##     best_actor_win + best_actress_win, data = stats)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.35800 -0.19831  0.03439  0.28111  1.19759
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    3.2939966  0.1712484  19.235  < 2e-16 ***
## genreAnimation                -0.4301985  0.1833393  -2.346  0.01926 *
## genreArt House & International  0.2113828  0.1415625   1.493  0.13589
## genreComedy                   -0.1485145  0.0783843  -1.895  0.05859 .
## genreDocumentary               0.3112779  0.1072217   2.903  0.00382 **
## genreDrama                     0.0393476  0.0684997   0.574  0.56589
## genreHorror                    0.0828272  0.1171954   0.707  0.47999
## genreMusical & Performing Arts  0.0323248  0.1512875   0.214  0.83088
## genreMystery & Suspense        0.2242696  0.0880947   2.546  0.01114 *
## genreOther                    -0.0466777  0.1327961  -0.351  0.72533
## genreScience Fiction & Fantasy -0.1943448  0.1667645  -1.165  0.24431
## runtime                        0.0050309  0.0011058   4.550 6.46e-06 ***
## mpaa_ratingNC-17              -0.1712827  0.3551033  -0.482  0.62973
## mpaa_ratingPG                 -0.1564002  0.1293084  -1.210  0.22692
## mpaa_ratingPG-13              -0.1041214  0.1332650  -0.781  0.43491
## mpaa_ratingR                  -0.0716802  0.1286623  -0.557  0.57765
## mpaa_ratingUnrated            -0.1952550  0.1468137  -1.330  0.18402
## critics_score                  0.0104261  0.0009597  10.864  < 2e-16 ***
## audience_score                 0.0338682  0.0013394  25.286  < 2e-16 ***
## best_pic_nomyes               -0.0306353  0.1218988  -0.251  0.80165
## best_pic_winyes                0.1080759  0.2044895   0.529  0.59733
## top200_boxyes                 -0.0132653  0.1273853  -0.104  0.91710
## best_actor_winyes              0.0377107  0.0556715   0.677  0.49842
## best_actress_winyes            0.0615546  0.0617252   0.997  0.31904
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.467 on 626 degrees of freedom
## Multiple R-squared:  0.8213, Adjusted R-squared:  0.8147
## F-statistic: 125.1 on 23 and 626 DF,  p-value: < 2.2e-16
```

The variable top200_box has the highest p-value and thus it bears the least statistical significance. So, we refit the MLR model without that variable.

```r
model <- lm(imdb_rating ~ genre+runtime+mpaa_rating+critics_score+audience_score+best_pic_nom+best_pic_win+best_acto
summary(model)
```

```
##
## Call:
## lm(formula = imdb_rating ~ genre + runtime + mpaa_rating + critics_score +
##     audience_score + best_pic_nom + best_pic_win + best_actor_win +
##     best_actress_win, data = stats)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.35791 -0.19841  0.03438  0.28130  1.19760
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    3.2934919  0.1710447  19.255  < 2e-16 ***
## genreAnimation                -0.4283592  0.1823424  -2.349  0.01912 *
## genreArt House & International  0.2122249  0.1412199   1.503  0.13339
## genreComedy                   -0.1478041  0.0780253  -1.894  0.05864 .
## genreDocumentary               0.3124765  0.1065180   2.934  0.00347 **
## genreDrama                     0.0402380  0.0679103   0.593  0.55372
## genreHorror                    0.0833656  0.1169889   0.713  0.47636
## genreMusical & Performing Arts  0.0336115  0.1506631   0.223  0.82354
## genreMystery & Suspense        0.2250678  0.0876914   2.567  0.01050 *
## genreOther                    -0.0462421  0.1326255  -0.349  0.72746
## genreScience Fiction & Fantasy -0.1947910  0.1665779  -1.169  0.24270
## runtime                        0.0050197  0.0010997   4.565 6.02e-06 ***
## mpaa_ratingNC-17              -0.1696438  0.3544744  -0.479  0.63241
## mpaa_ratingPG                 -0.1553253  0.1287940  -1.206  0.22827
## mpaa_ratingPG-13              -0.1028108  0.1325647  -0.776  0.43831
## mpaa_ratingR                  -0.0701105  0.1276755  -0.549  0.58311
## mpaa_ratingUnrated            -0.1936864  0.1459237  -1.327  0.18489
## critics_score                  0.0104205  0.0009575  10.883  < 2e-16 ***
## audience_score                 0.0338619  0.0013370  25.327  < 2e-16 ***
## best_pic_nomyes               -0.0306327  0.1218026  -0.251  0.80151
## best_pic_winyes                0.1073023  0.2041933   0.525  0.59943
## best_actor_winyes              0.0376123  0.0556196   0.676  0.49914
## best_actress_winyes            0.0612119  0.0615888   0.994  0.32066
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4666 on 627 degrees of freedom
## Multiple R-squared:  0.8213, Adjusted R-squared:  0.815
## F-statistic:   131 on 22 and 627 DF,  p-value: < 2.2e-16
```

We notice a slight increase in the adjusted R-square after removing the previous insignificant variable. Now, we remove the next least significant variables 'best_pic_nom', 'best_pic_win', 'best_actor_win', 'best_actress_win', 'mpaa_rating', 'top200_box' with pretty high p-value.

```r
model <- lm(imdb_rating ~ genre+ runtime+critics_score+audience_score,data=stats)
summary(model)
```

```
##
## Call:
## lm(formula = imdb_rating ~ genre + runtime + critics_score +
##     audience_score, data = stats)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34430 -0.20090  0.03524  0.27085  1.17364
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    3.1675348  0.1251241  25.315  < 2e-16 ***
## genreAnimation                -0.3681453  0.1668808  -2.206   0.0277 *
## genreArt House & International  0.1997289  0.1376430   1.451   0.1473
## genreComedy                   -0.1410076  0.0766630  -1.839   0.0663 .
## genreDocumentary               0.2611971  0.0945446   2.763   0.0059 **
## genreDrama                     0.0573713  0.0655556   0.875   0.3818
## genreHorror                    0.0953283  0.1141619   0.835   0.4040
## genreMusical & Performing Arts  0.0156689  0.1491699   0.105   0.9164
## genreMystery & Suspense        0.2613679  0.0846405   3.088   0.0021 **
## genreOther                    -0.0599035  0.1311583  -0.457   0.6480
## genreScience Fiction & Fantasy -0.1913924  0.1660092  -1.153   0.2494
## runtime                        0.0052878  0.0010182   5.193 2.78e-07 ***
## critics_score                  0.0104037  0.0009376  11.096  < 2e-16 ***
## audience_score                 0.0339006  0.0013210  25.663  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4657 on 636 degrees of freedom
## Multiple R-squared:  0.8194, Adjusted R-squared:  0.8157
## F-statistic:   222 on 13 and 636 DF,  p-value: < 2.2e-16
```

In the above model we see that some of the categories of the factor 'Genre' are pretty insignificant, but we retain the factor in our model because it has some levels which are statistically very significant with low p-values. This is our final MLR model.

Note that, if the Genre falls into the categories 'Animation', 'Comedy', 'Science Fiction and Fantasy' and 'Other', the predicted IMDb score tends to decrease (when all other factors are held constant), as the estimated slope coefficient of these factors are negative.

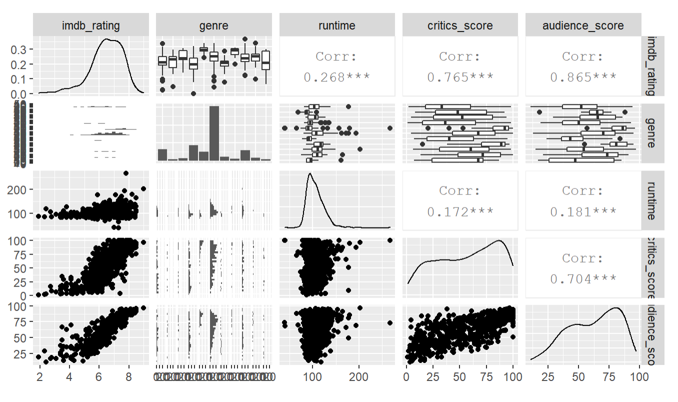
Also, note that the coefficient of determination r-square is pretty high (also the Adjusted r-square), so 81.94% of the variability can be explained.

### Checking conditions and Model Diagnostics

Let us first check for collinearity between the explanatory variables using pairwise plot.

```r
ggpairs(data=stats, columns = c(13,3,4,16,18))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




We do not see any collinearity between the explanatory variables because we have already removed the factors which are statistically insignificant and might show collinearity with the significant factors. Also, we had concluded that runtime has low correlation with IMDb score (as is evident from the graph above). But, we have kept the factor in the model because from the Regression table we obtained a pretty low p-value and thus conclude that it is statistically significant in the prediction of IMDb score.

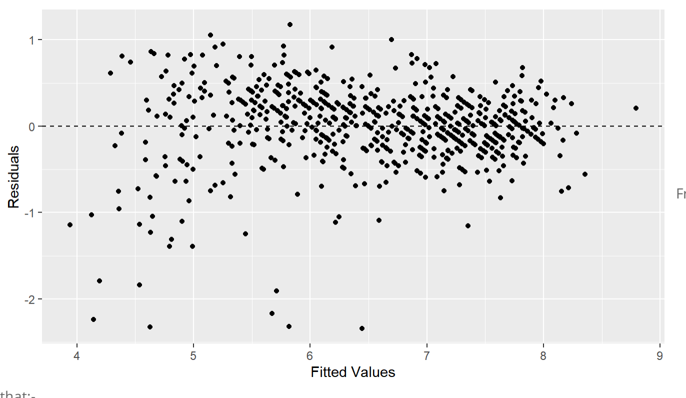
So, our final linear regression model is given by:-

$$\begin{aligned} IMDb\_score = 3.17 - 0.37 genre:Ani + 0.2 genre:Art - 0.14 genre:Com \\ + 0.26 genre:Doc + 0.06 genre:Dra + 0.1 genre:Horr + 0.02 genre:Mus \\ + 0.26 genre:Mys + -0.06 genre:Oth - .019 genre:SciFi + 0.005 runtime \\ + 0.01 crit + 0.03 aud \end{aligned}$$

Note: The model coefficients have already been interpreted.

### Residuals vs Predicted

```r
ggplot(data=model, aes(x=.fitted, y=.resid))+geom_point()+geom_hline(yintercept = 0, linetype="dashed")+labs(x="Fit
```

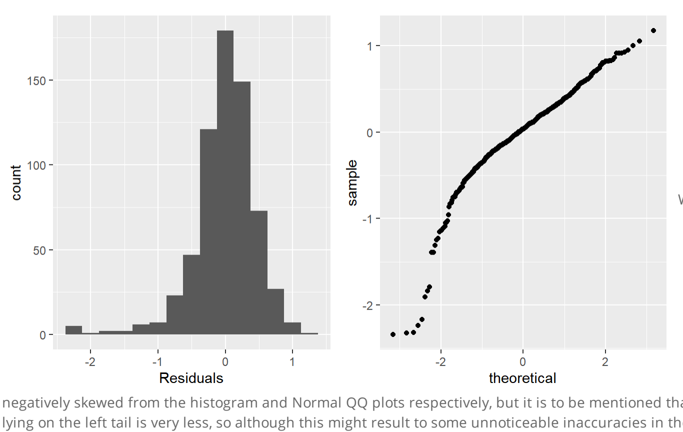



From the above plot we conclude that:-

i. Points are randomly scattered around the x-axis showing no trends, indicating that the residuals are randomly distributed around 0.
ii. The points show constant variability indicating the residuals are almost homoskedastically distributed.

### Histogram and QQ Plot of residuals

```r
p1<- ggplot(data=model, aes(x=.resid))+geom_histogram(binwidth = 0.25)+xlab("Residuals")
p2 <- ggplot(data=model, aes(sample=.resid))+stat_qq()
grid.arrange(p1,p2, ncol=2)
```




We see that the residuals are a bit negatively skewed from the histogram and Normal QQ plots respectively, but it is to be mentioned that the number of observations lying on the left tail is very less, so although this might result to some unnoticeable inaccuracies in the prediction, it shouldn't largely affect the predictive accuracy of our model.

## Part 5: Prediction

For prediction, let us consider the movie "Silence" by Martin Scorsese, released in 2016.

For our model, the required data are (source: Google) :-

i. Genre: Drama
ii. Runtime: 161 minutes
iii. Critics Score: 83
iv. Audience Score: 69

Let us create a data frame for predicting the IMDb score of this movie.

```r
Silence <- data.frame(genre="Drama", runtime=161, critics_score=83, audience_score=69)
```

Now, we predict the IMDb score of this movie and also the interval of precision based on the available data.

```r
predict(model, Silence, interval = "prediction", level=0.95)
```

```
##        fit      lwr    upr
## 1 7.278885 6.357171 8.2006
```

Note that, the observed IMDb score is 7.2 which is almost equal to the predicted value of 7.27. Thus, the fit is extremely good. The 95% confidence interval for this estimate is (6.36, 8.20).

So, we can say that the probabiliy that the above interval contains the IMDb score of a movie under the "Drama" genre, running for 161 minutes, with a Rotten Tomatoes Critic Score of 83% and an Audience Score of 69%, is 0.95

## Part 6: Conclusion

We have analyzed different factors so far to predict the IMDb scores and later ended up only 4 of them, because the others were statistically insignificant (judging from the p-values). We have used the backward reduction based on p-values. In some cases, it may be observed that in this process, the coefficient of determination (r-square) has decreased after removing the insignificant factors from the model, but in our model, the adjusted r-square was initially very high and then went up by a small amount after reduction, indicating that the factors removed could have caused collinearity.

During the model selection we saw that the score has a low linear association with the run-time, but we still keep the factor in our model as it is statistically significant and we also see that the final model yields a pretty accurate prediction of the IMDb score, indicating that runtime is an important factor of our model.

Lastly, we can conclude by saying that there is a lot of research that can be done and the one explained in this project shows that we can accurately predict how people will rate (like) the movies from its general characteristics.