Linear Regression and Modeling Project

Date: 4/20/2020

Contents

Part 1: Data

In this project, we study **what attributes make a movie popular**.

The data set, `movies`, is comprised of 651 randomly sampled movies produced and released before 2016. It has 32 variables. According to the codebook, we may assume random sampling is used. Additionally, given that it is an observational study, we are not able to draw conclusions regarding causality.

We first load the `movies` data set as well as the `dplyr` and `ggplot2` packages.

```r
# Load data
load("C:/Other/eLearning/Coursera/Linear Regression and Modeling/Week 4/movies.Rdata")

# Load packages
lapply(c('ggplot2','dplyr','stargazer','lmtest'),require,character.only=TRUE)
```

Part 2: Research Question

Our research question is **what attributes are associated with a higher IMDB rating for a movie.**

We think the following attributes affect the popularity of a movie:

Type of movie, **runtime of movie**, **critics rating on Rotten Tomatoes**, **critics score on Rotten Tomatoes**, **audience rating on Rotten Tomatoes**, **audience score on Rotten Tomatoes**, and **whether or not the movie has won a best picture Oscar**.

The reasoning is as followed. First, there is no doubt that both the critics and the audience's evaluations heavily affect the final score of the movie. Second, whether the movie has won a title indicates if the movie is a success. Third, people may dislike movies that are too lengthy, so runtime of movie is a factor worths consideration. Finally, type of movie is included as a control variable.

Part 3: Exploratory Data Analysis

Accordingly, our dependent variable is *imdb_rating* (rating on IMDB).

The independent variables are:

title_type: type of movie.

runtime: runtime of movie.

critics_rating: critics rating on Rotten Tomatoes.

critics_score: critics score on Rotten Tomatoes.

audience_rating: audience rating on Rotten Tomatoes.

audience_score: audience score on Rotten Tomatoes.

best_pic_win: whether or not the movie has won a best picture Oscar.

Specifically, *title_type*, *critics_rating*, *audience_rating* and *best_pic_win* are categorical variables, while *imdb_rating*, *runtime*, *critics_score* and *audience_score* are numeric ones. The descriptive statistics of four numeric varibales are illustrated below.

```
summary(movies %>%
  select(imdb_rating, runtime, critics_score, audience_score))
```

```
##   imdb_rating       runtime      critics_score    audience_score
##  Min.   :1.900   Min.   : 39.0   Min.   :  1.00   Min.   :11.00
##  1st Qu.:5.900   1st Qu.: 92.0   1st Qu.: 33.00   1st Qu.:46.00
##  Median :6.600   Median :103.0   Median : 61.00   Median :65.00
##  Mean   :6.493   Mean   :105.8   Mean   : 57.69   Mean   :62.36
##  3rd Qu.:7.300   3rd Qu.:115.8   3rd Qu.: 83.00   3rd Qu.:80.00
##  Max.   :9.000   Max.   :267.0   Max.   :100.00   Max.   :97.00
##                  NA's   :1
```

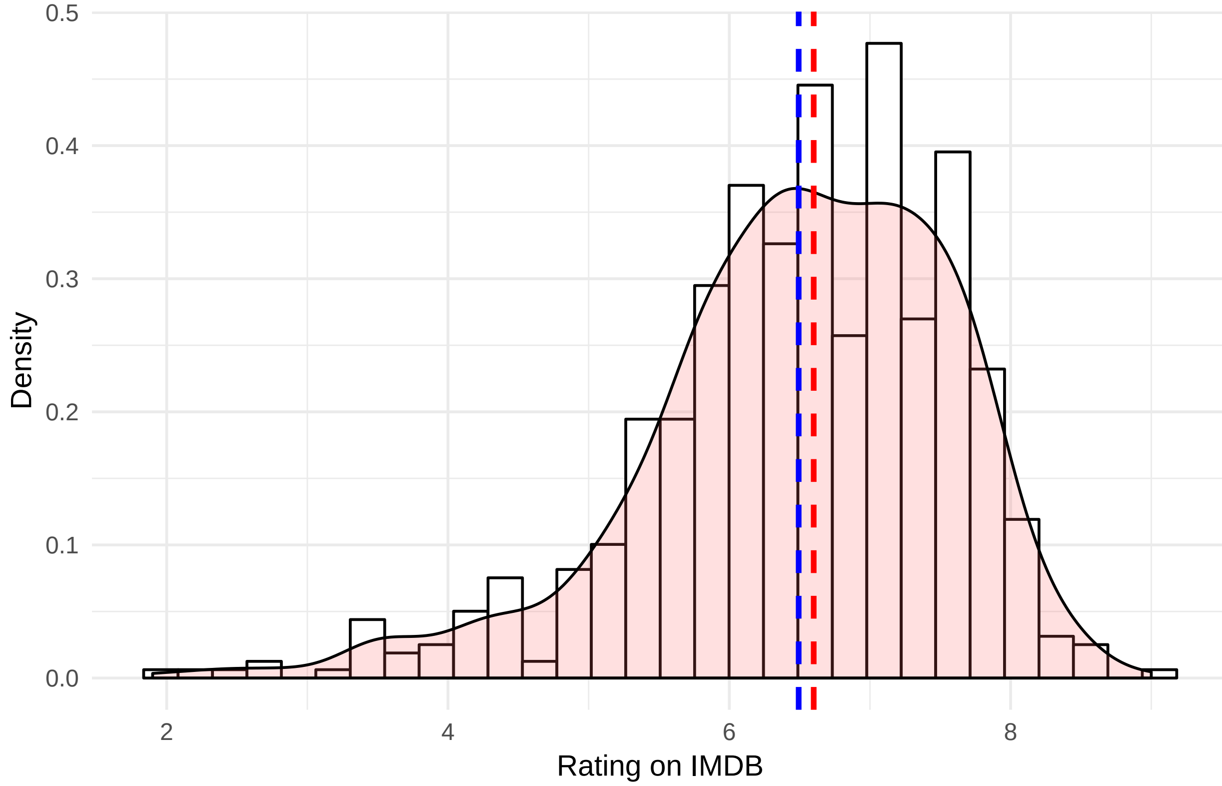
The level and absolute frequency of each level for four categorical variables are listed below.

```
summary(movies %>%
  select(title_type, critics_rating, audience_rating, best_pic_win))
```

```
##         title_type          critics_rating audience_rating best_pic_win
##  Documentary : 55   Certified Fresh:135    Spilled:275     no :644
##  Feature Film:591   Fresh          :209    Upright:376     yes:  7
##  TV Movie    :  5   Rotten         :307
```

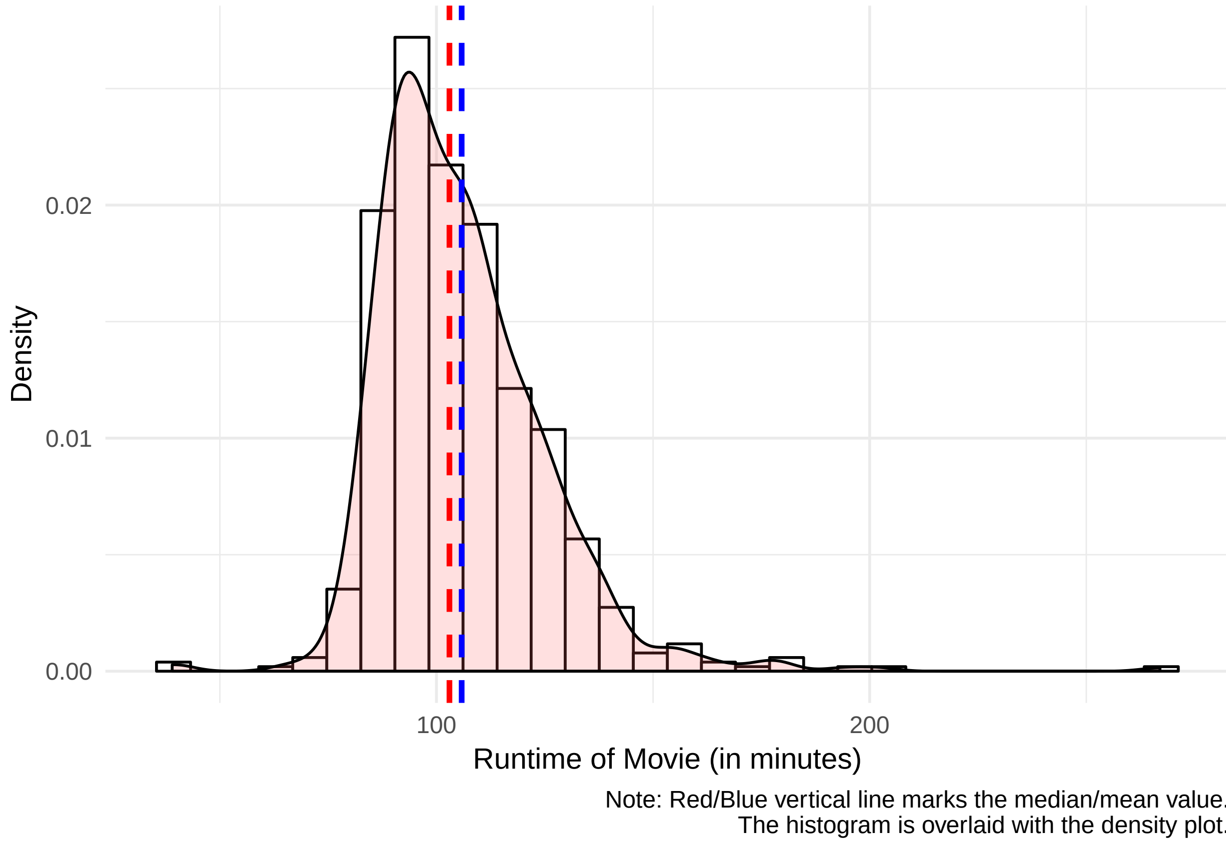
Next, we visualize the distribution of four numeric variables. We first plot the distribution of *imdb_rating*. It is left skewed with mean of 6.5 and median of 6.6.

```
# Define a "histogram" function
distribution<-function(var){

  ggplot(movies, aes(x=var)) +
  geom_histogram(aes(y=..density..), colour="black", fill="white")+
  geom_density(alpha=.2, fill="#FF6666")+
  geom_vline(aes(xintercept=mean(var)),
             color="blue", linetype="dashed", size=1)+
  geom_vline(aes(xintercept=median(var)),
             color="red", linetype="dashed", size=1)+
  ylab("Density")+
  labs(caption="Note: Red/Blue vertical line marks the median/mean value.
       The histogram is overlaid with the density plot.")+
  theme_minimal()
}

# Draw imdb_rating distribution
distribution(movies$imdb_rating)+xlab('Rating on IMDB')
```



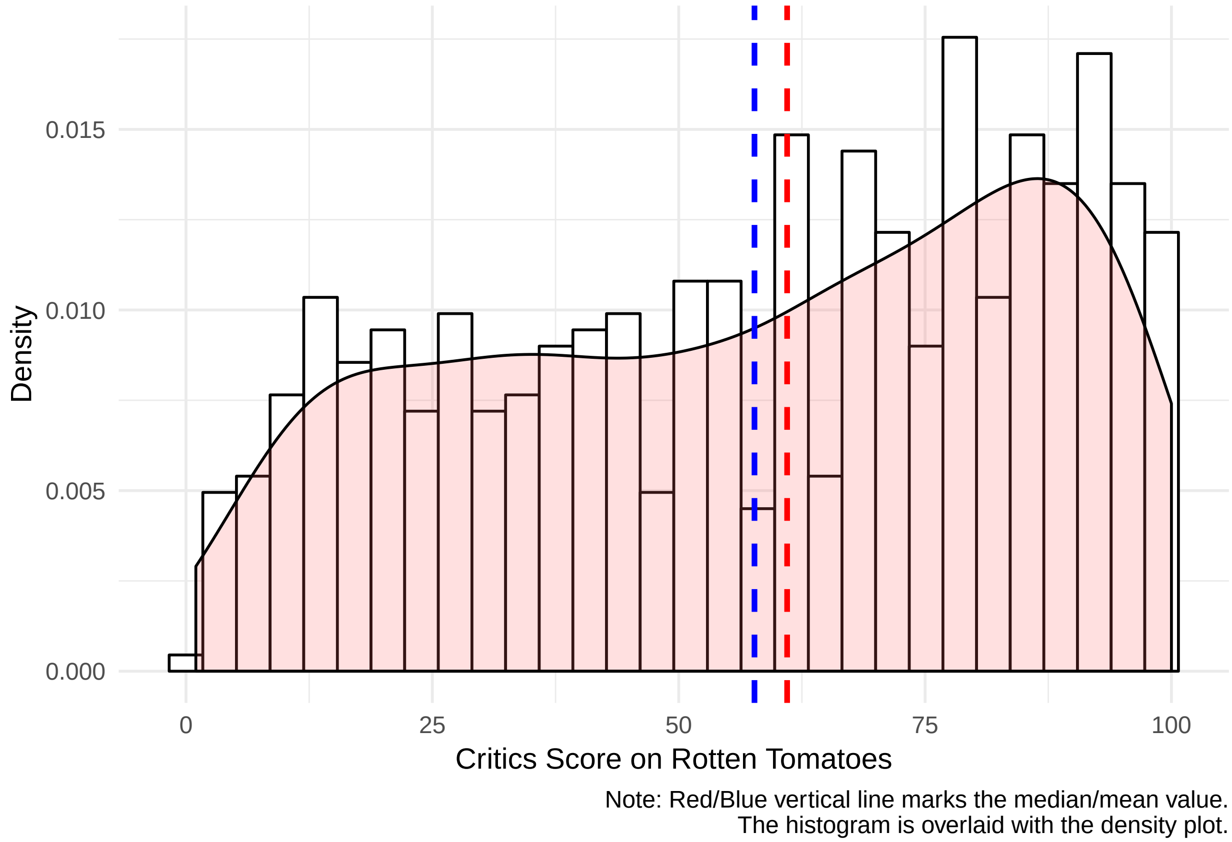
Note: Red/Blue vertical line marks the median/mean value.
The histogram is overlaid with the density plot.

The distribution of *runtime* is right skewed with mean of 105.8 minutes and median of 103 minutes.



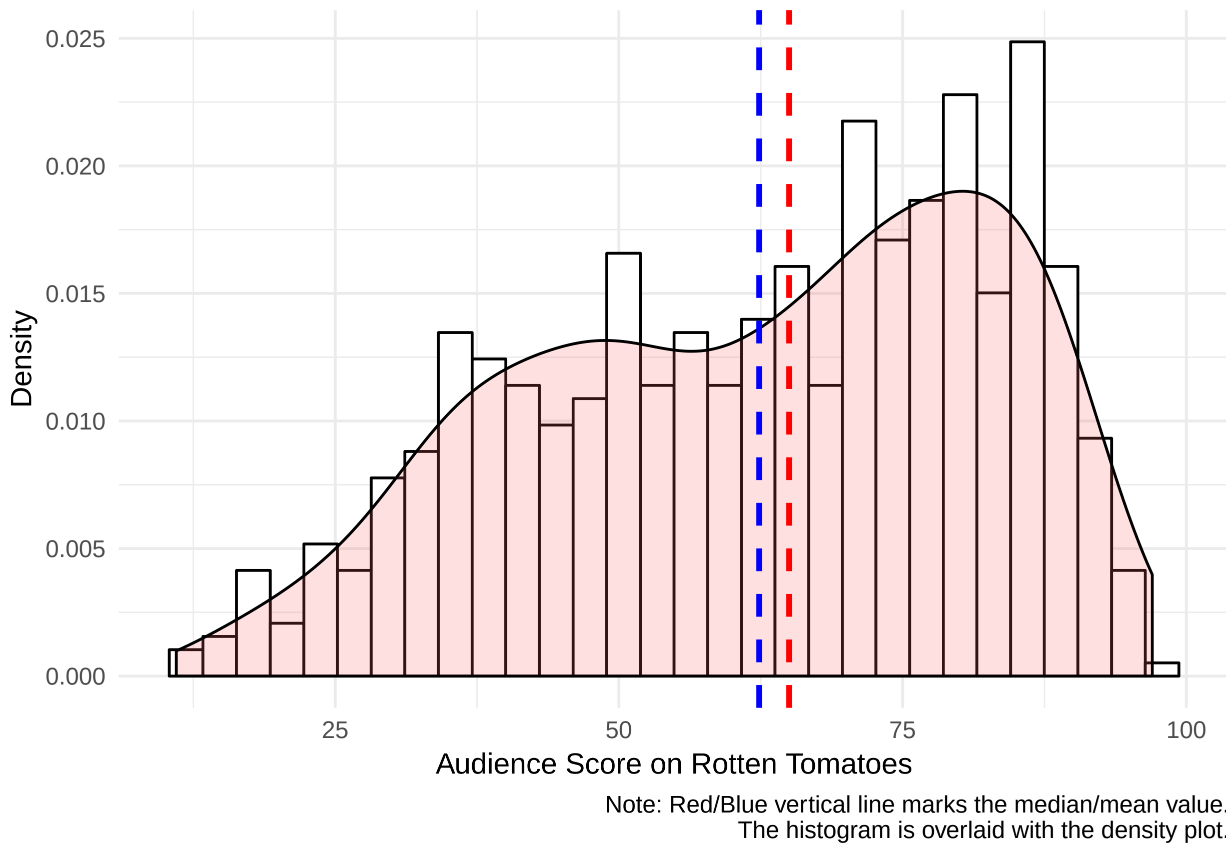
The distribution of *critics_score* is fairly left skewed with mean of 58 and median of 61.

```
distribution(movies$critics_score)+xlab('Critics Score on Rotten Tomatoes')
```



Finally, the distribution of *audience_score* is left skewed with mean of 62 and median of 65.

```
distribution(movies$audience_score)+xlab('Audience Score on Rotten Tomatoes')
```



Part 4: Modeling

We fit a multiple linear regression model with *imdb_rating* as the response variable, and *title_type*, *runtime*, *critics_rating*, *critics_score*, *audience_rating*, *audience_score* and *best_pic_win* as predictors. Regression result is shown in the first column in the following table.

The coefficient of *best_pic_win* is not significant, therefore we exclude this variable based on **backwards elimination (p-value)**. The new regression result is in the second column. The coefficients of the remaining variables as well as the model's adjusted R-squared do not change too much.

```
OLS1<-lm(imdb_rating ~ title_type+runtime+critics_rating+
          critics_score+audience_rating+audience_score+best_pic_win,data = movies)
OLS2<-lm(imdb_rating ~ title_type+runtime+critics_rating+
          critics_score+audience_rating+audience_score,data = movies)
stargazer(OLS1,OLS2,type='text')
```

```
##
## ========================================================================
##                                         Dependent variable:
##                               ------------------------------------------------
##                                             imdb_rating
##                                     (1)                       (2)
## ------------------------------------------------------------------------
## title_typeFeature Film           -0.166**                  -0.165**
##                                  (0.073)                   (0.072)
##
```

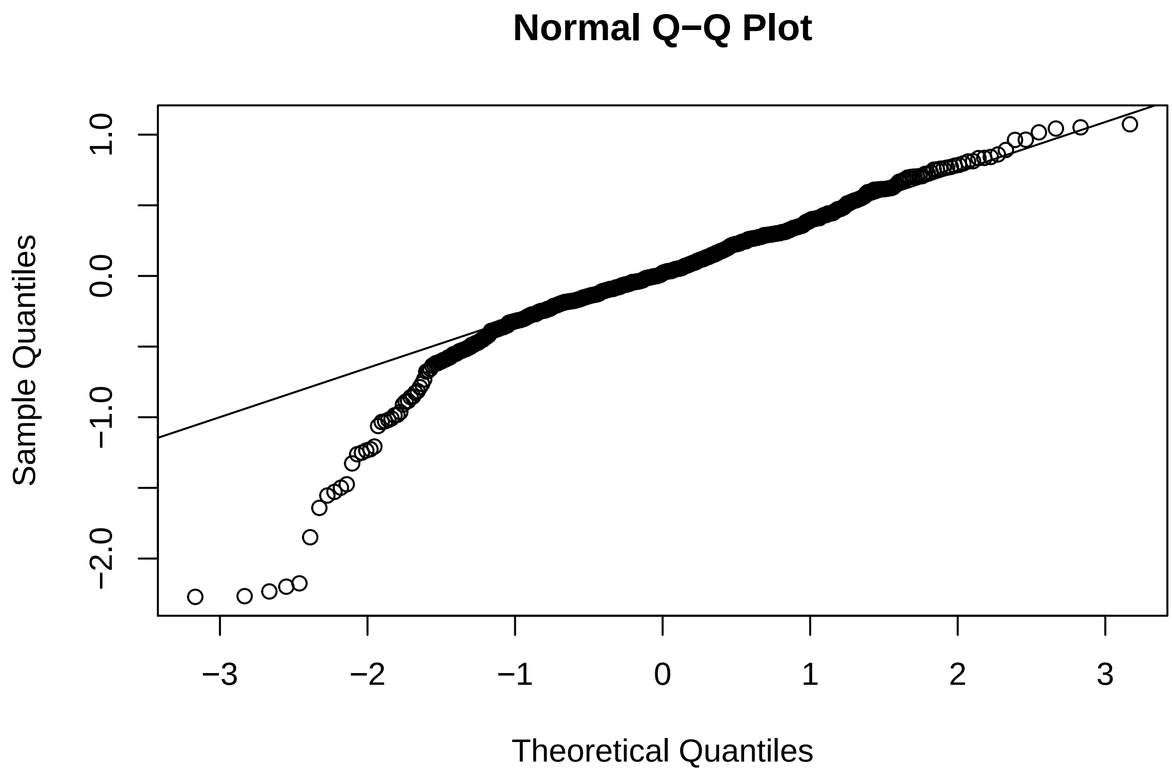
```
## title_typeTV Movie             -0.369*                 -0.368*
##                                (0.219)                 (0.219)
##
## runtime                         0.006***                0.006***
##                                (0.001)                 (0.001)
##
## critics_ratingFresh             0.028                   0.024
##                                (0.054)                 (0.054)
##
## critics_ratingRotten            0.273***                0.271***
##                                (0.090)                 (0.090)
##
## critics_score                   0.015***                0.015***
##                                (0.002)                 (0.002)
##
## audience_ratingUpright        -0.366***               -0.367***
##                                (0.074)                 (0.074)
##
## audience_score                  0.041***                0.041***
##                                (0.002)                 (0.002)
##
## best_pic_winyes                 0.071
##                                (0.183)
##
## Constant                        2.668***                2.663***
##                                (0.182)                 (0.181)
##
## -----------------------------------------------------------------------
## Observations                      650                     650
## R2                               0.819                   0.819
## Adjusted R2                      0.817                   0.817
## Residual Std. Error        0.464 (df = 640)        0.464 (df = 641)
## F Statistic            322.669*** (df = 9; 640) 363.465*** (df = 8; 641)
## =======================================================================
## Note:                                       *p<0.1; **p<0.05; ***p<0.01
```

We move forward to check the assumptions of the model.

1. Normally distributed error terms

The QQ plot shows that the error terms are almost normally distributed.

```r
qqnorm(OLS2$residuals)
qqline(OLS2$residuals)
```



2. Homoskedasticity

A particularly small p-value as the result of the `Breush-Pagan test` implies that the null hypothesis of homoskedasiticity is rejected. There is, indeed, heteroskedasticity in our model.

```
lmtest::bptest(OLS2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  OLS2
## BP = 92.225, df = 8, p-value < 2.2e-16
```

3. No autocorrelation

Finally, a large p-value of the `Durbin-Watson test` indicates that there is no autocorrelation.

```
lmtest::dwtest(OLS2)
```

```
##
##  Durbin-Watson test
##
## data:  OLS2
## DW = 1.9483, p-value = 0.2551
## alternative hypothesis: true autocorrelation is greater than 0
```

We hereby interpret the coefficient of one categorical variable and one numeric variable. First, all else held equal, compared to a documentary, a feature film scores 0.165 lower in IMDB rating. Second, all else held equal, 1 point increase in audience rating leads to 0.041 points increase in IMDB rating.

Part 5: Prediction

We randomly pick a movie, i.e. *The Lighthouse*, from the Rotten Tomatoes to make prediction.

Because of lack of time to get all the data of the movie, we arbitrarily set values for some explanatory variables. *The Lighthouse* is (presumably) a feature film, with a run time of 110 minutes, "certified fresh" critics rating, 90 critics score, (presumably) "spilled" audience rating, and 72 audience score.

Our model predicts that *The Lighthouse* has an IMDB rating of about 7.5. In addition, with 95% confidence, this score is within the interval of 6.5 to 8.4.

```r
# A new data frame for The Lighthouse
lighthouse<-data.frame(title_type="Feature Film", runtime=110,
                       critics_rating="Certified Fresh",critics_score=90,
                       audience_rating="Spilled",audience_score=72)

# Prediction
predicted<-predict(OLS2,lighthouse)

# Prediction with interval
interval<-predict(OLS2,lighthouse,interval = "prediction", level = 0.95)

print(paste("The predicted IMDB rating for The Lighthouse is", predicted))
```

```
## [1] "The predicted IMDB rating for The Lighthouse is 7.46186851689173"
```

```r
print(paste("The lower bound of the 95% prediction interval is",interval[2]))
```

```
## [1] "The lower bound of the 95% prediction interval is 6.53980349970546"
```

```r
print(paste("The upper bound of the 95% prediction interval is",interval[3]))
```

```
## [1] "The upper bound of the 95% prediction interval is 8.38393353407801"
```

Part 6: Conclusion

In this project, we explore the factors that affect the IMDB rating of a movie. By using the multiple linear regression, we find that runtime, critics score and audience score are positively correlated with a higher movie rating. All else held equal, compared to documentaries, feature movies and TV movies tend to score lower. What's more, all else held equal, compared to a movied rated as "spilled", an "upright" movie has a lower score.