Modeling and prediction for movies

Amin Ghaderi

February 15, 2018

Setup

Load packages

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.3
```

```
library(statsr)
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 3.4.3
```

Load data

We use load command to import the movies data.

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

Part 1: Data

Acquisition: This data is randomly selected from IMDB and Rotten Tompato APIs from movies produced before 2016.

Population: To be included in this data set, the movie needs to be (1) in the Rotten Tomatoes and IMDB databases, (2) produced before 2016.

Causality/Generalization: Since the data is randomly sampled from the discussed population and no ***random assignment*** is performed, the results of this study does not demonstrate any causality. Any results could be merely used to demonstrate correlation. The results is also only generalizable to the poplation discussed above, which are movies in IMDB and RT databases, produced before 2016. * * *

Part 2: Research question

In this study, we are choosing our target (dependant variable) as the IMDB score (`imdb_rating`). We would like to see how this score is affected by different factors. We are mostly interested in the effect of `critics_score`, `audience_rating`, and whether movie is on Top 200 Office Box list (`top200_box`). We will also study the effect of rest of the parameters and finally will build a model that can predict the IMDB popularity of the movies based on the available parameters.

Part 3: Exploratory data analysis

3.1 EDA on `critics_score` vs `imdb_rating`:

This is the first pair that we are studying. Let's first see the data type of these features to decide the required statistics.

3.1.1 Scatter Plot and Fitting Linear Model

```
lapply(movies, class)$critics_score
```

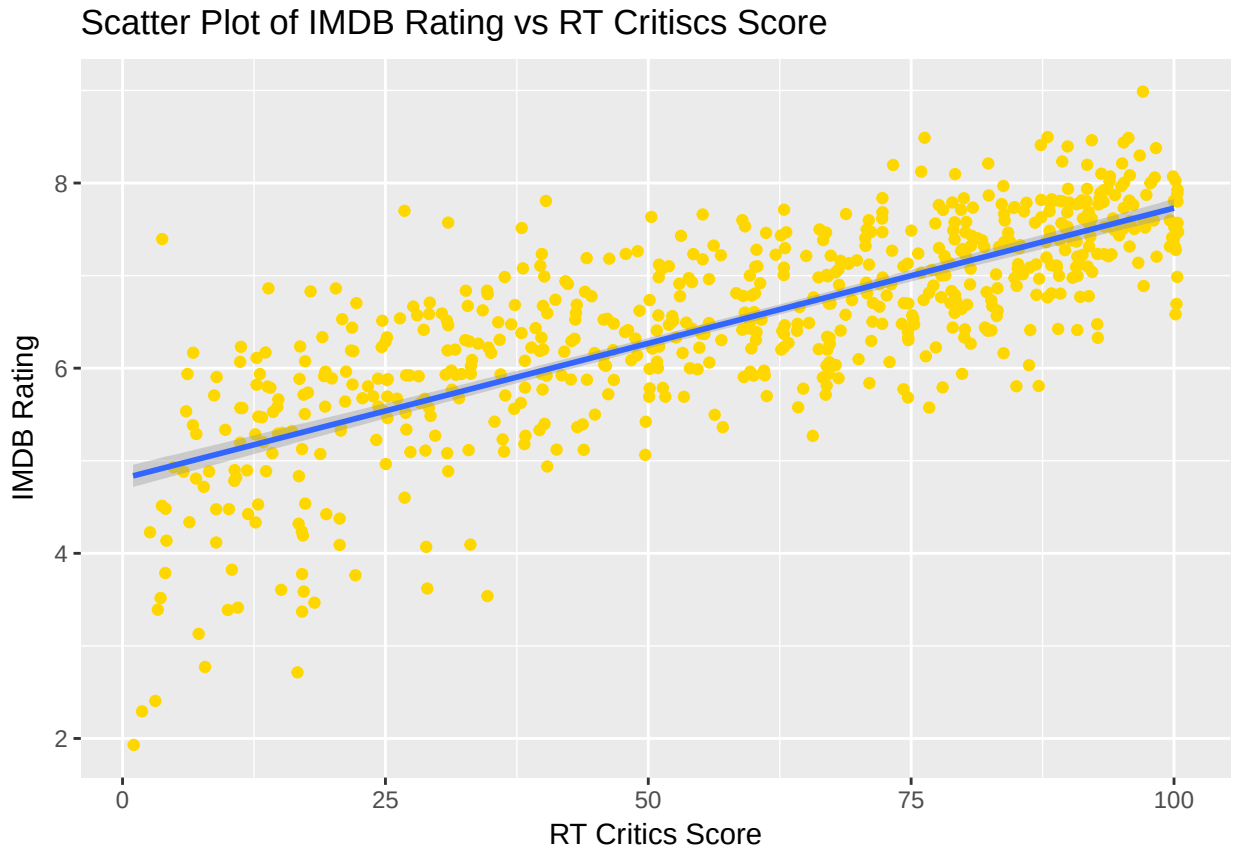
```
## [1] "numeric"
```

```
lapply(movies, class)$imdb_rating
```

```
## [1] "numeric"
```

Both variables are numeric. Therefore, we go with a scatter plot to first get an understanding of the data. Explanatory variable (`critics_score`) will be on the x-axis and the target (`imdb_rating`) on the y-axis.

```
ggplot(movies, aes(x=critics_score, y=imdb_rating))+geom_jitter(color='gold')+
  labs(x="RT Critics Score", y="IMDB Rating",
       title = 'Scatter Plot of IMDB Rating vs RT Critiscs Score')+
  stat_smooth(method = "lm", se = TRUE)
```



There is a very obvious linear relationship between these two values. The relationship is positive, linear, and strong. Let's quantify this relationship. we are going to fit a linear model. Before that however, we need to see the correlation between these two variables.

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

```
## Warning: package 'bindrcpp' was built under R version 3.4.3
```

```
## # A tibble: 1 x 1
##   `cor(critics_score, imdb_rating)`
##                               <dbl>
## 1                             0.765
```

The correlaton also is quite large. Therefore, we can conclude that the relationship between these two values is quite strong.

Let's fit the linear model and see the R-squared score.

```
imdb_critics <- lm (imdb_rating~critics_score, data=movies)
summary(imdb_critics)
```

```
##
## Call:
## lm(formula = imdb_rating ~ critics_score, data = movies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.93679 -0.39499  0.04512  0.43875  2.47556
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.8075715  0.0620690   77.45   <2e-16 ***
## critics_score 0.0292177  0.0009654   30.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6991 on 649 degrees of freedom
## Multiple R-squared:  0.5853, Adjusted R-squared:  0.5846
## F-statistic: 915.9 on 1 and 649 DF,  p-value: < 2.2e-16
```

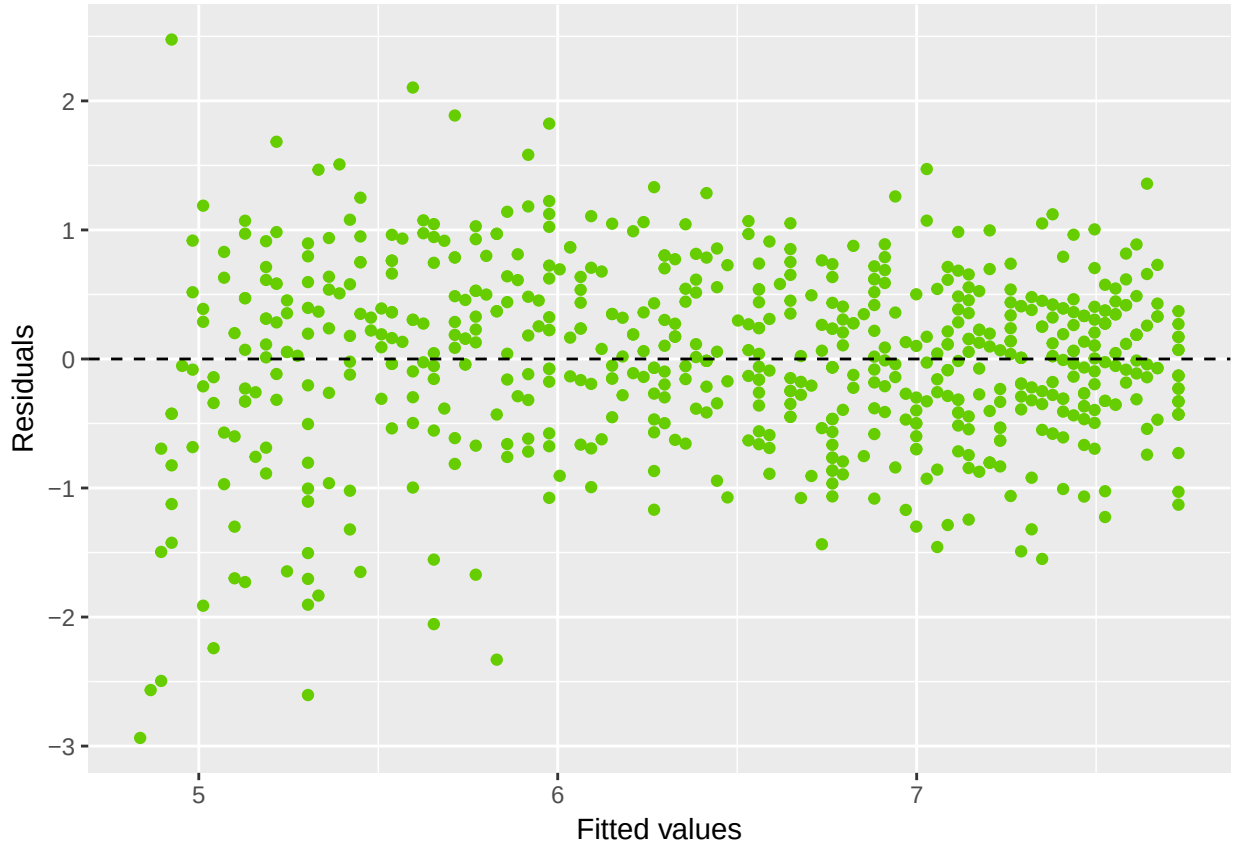
p-value of the relationship is very small and the adjusted R-squred demonstrate that 58.5% of the variability of the data can be explained by this feature, which is very impressive. However, we need to also assure that this linear regression model is credible.

3.1.2 Model Diagnostics

We want to assure that this linear model is relaiable. We need to chec for (1) linearity, (2) nearly normal distribution, and (3) constant variability.

** Linearity **: We already checked this using a scatter plot. We can also verify this using a plot of the residuals vs. fitted (predicted) values.

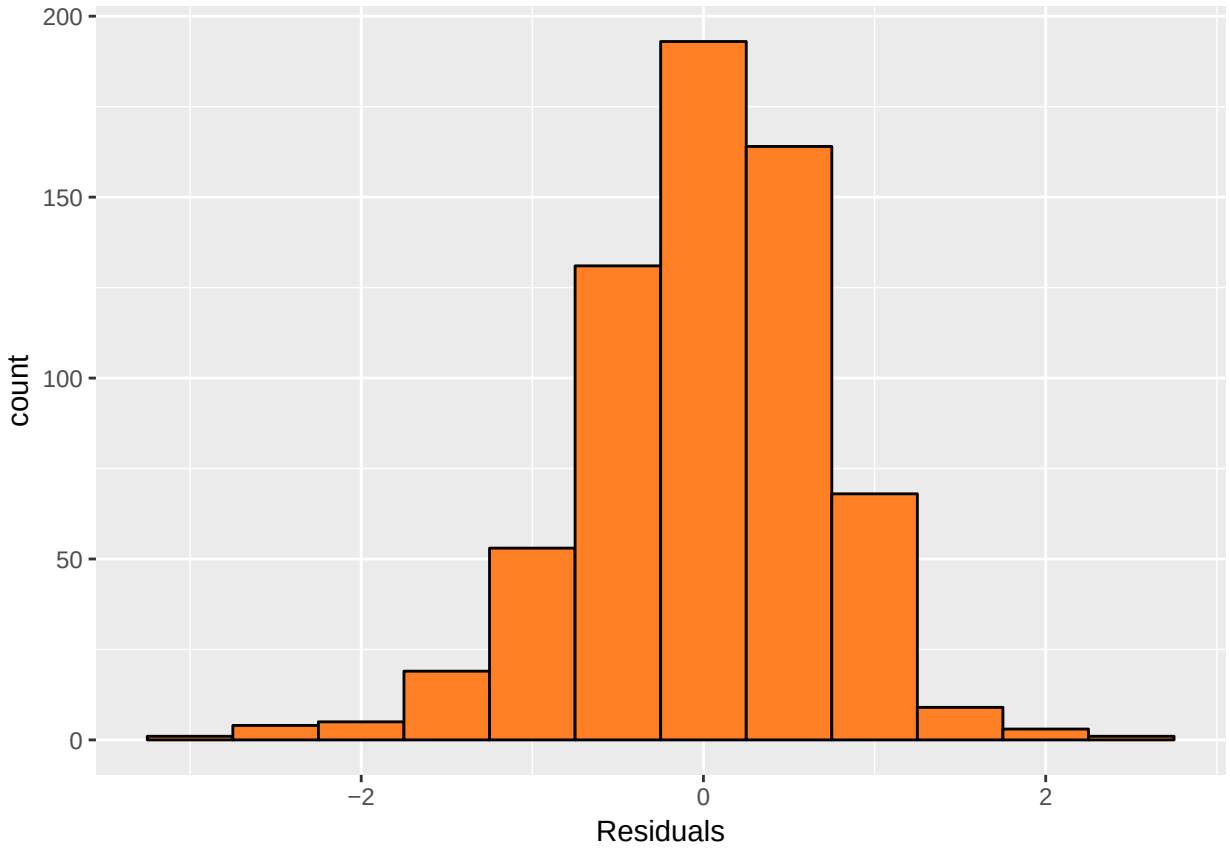
```
ggplot(data = imdb_critics, aes(x = .fitted, y = .resid)) +
  geom_point(color = 'chartreuse3') +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```



The plot seems to be randomly distributed. However, there seems to be a bit more data in higher rating s and more scatter around lower ratings.

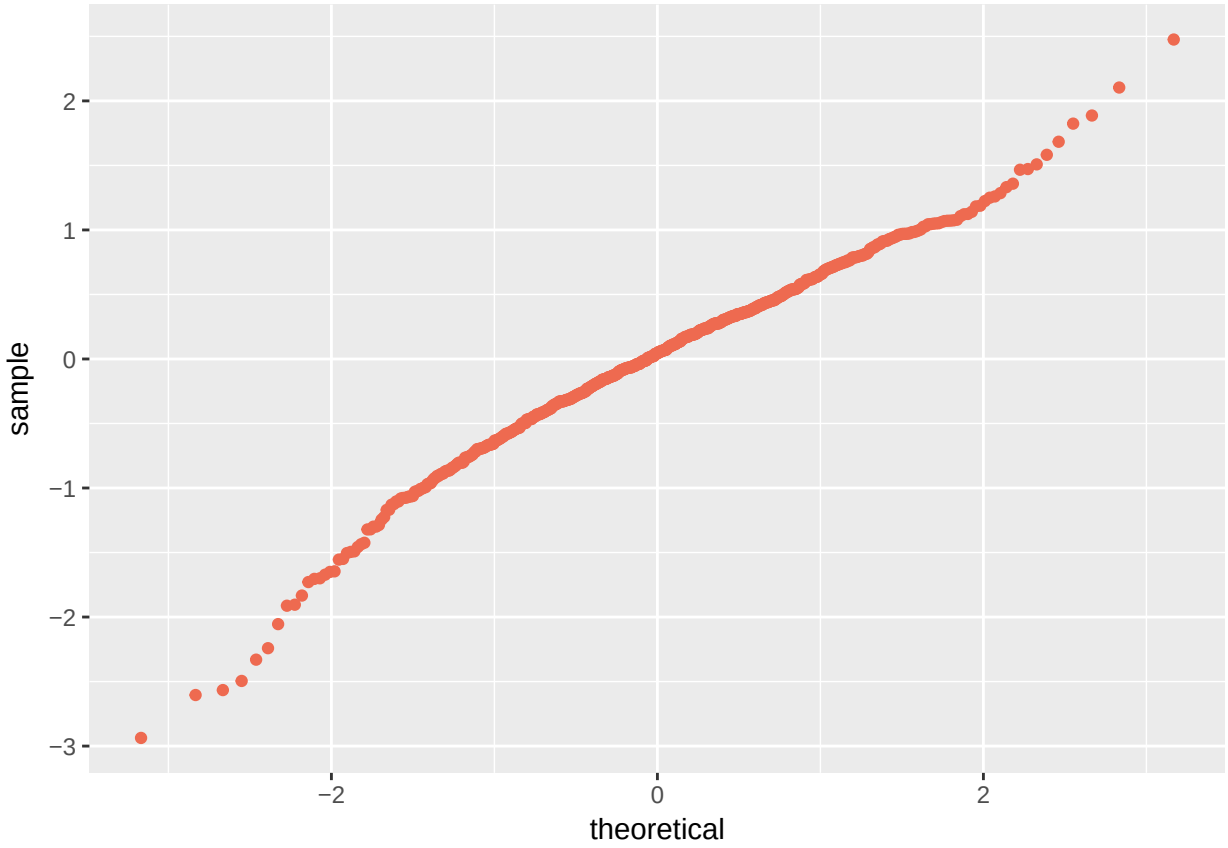
** Nearly Normal Residuals **: To check this condition, we will check the histograms.

```
ggplot(data = imdb_critics, aes(x = .resid)) +
  geom_histogram(binwidth = .5, fill = 'chocolate1', color = 'black' ) +
  xlab("Residuals")
```



There seems to be a symettry around 0 and data could be roughly considered normal. We can also use the normal probablity plot of residuals.

```
ggplot(data = imdb_critics, aes(sample = .resid)) +
  stat_qq(color = 'coral2')
```



This relationship also seems to be linear which assures us that the residuals are distributed normally. Therefore, considering that the model is credible, we can say that the critics score can be a very good predictor of the movie popularity (IMDB rating).

Part 4: Modeling

NOTE: Insert code chunks as needed by clicking on the "Insert a new code chunk" button above. Make sure that your code is visible in the project you submit. Delete this note when before you submit your work.

Part 5: Prediction

NOTE: Insert code chunks as needed by clicking on the "Insert a new code chunk" button above. Make sure that your code is visible in the project you submit. Delete this note when before you submit your work.

Part 6: Conclusion