# IMDB Movie Project Report

*Xiaofan Xia*

*12/7/2019*

## 1. Overview

### 1.1 Introduction

This project focuses on digging into the dataset with the information of more than 5000 movies including IMDB scores, the number of movies' Facebook likes and basic information about movies' casts. My main objective of this project is to find out how much movies' reputation is influenced by other related variables such as the popularity of the casts, and movies' budget. The reputation of the movie would be quantified as IMDB rating scores, and the popularity of the movie would be quantified as the number of Facebook likes and the number of voted users. In this project, I firsly cleaned the dataset and did an Exploratory Data Analysis before digging into the modeling analysis. The major model I would use is the multi-level linear regression model.

### 1.2 Motivation

The initial idea of choosing the IMDB movie topic comes from my personal interest. I am a very big fan of many kinds of movies. During my spare time, I always choose to watch movies to relax. Also, many friends of mine are working in the film industry right now. Therefore, I generate some questions about movies and their rating system, and that is why I choose this dataset and this project. I hope that I could solve my own questions about movies during this project as well as help my friends to learn their own industries from a different perspective with data analysis.

### 1.3 Dataset: IMDB 5000 Movie Dataset from Kaggle.

The IMDB 5000 Movie Dataset contains 28 variables such as movie names, director names, IMDB scores, and movie Facebook likes, etc. This is an open database from Kaggle.

Source: https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset

### 1.4 Research Questions

a. Is there any relationship between the popularity of casts (Facebook likes) and the income of the movie (gross)?

b. Is there any relationship between the popularity of casts (Facebook likes) and the reputation of the movie(IMDB rating scores)?

c. What are the patterns of gross and reputation of movies regarding different movie genres?

## 2. Exploratory Data Analysis

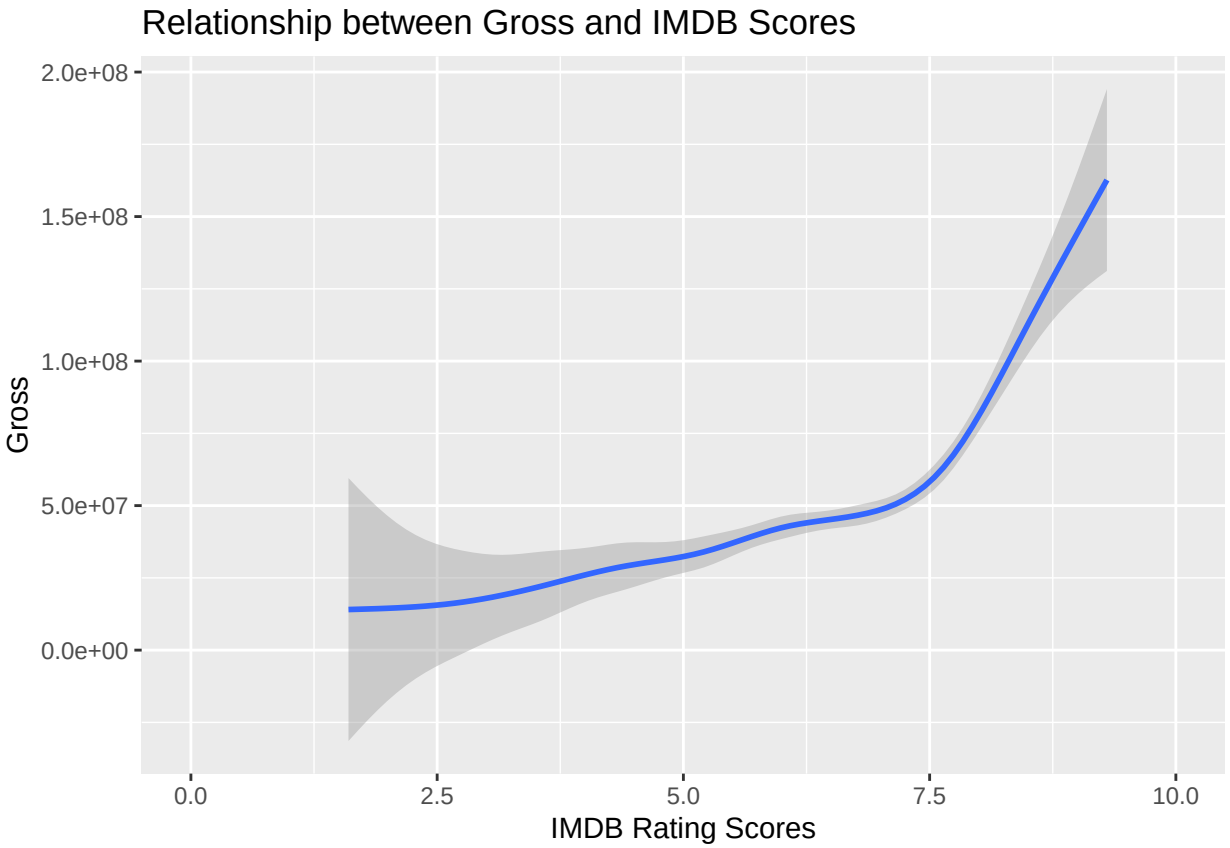
### 2.1 Basic Statistics of Major Variables

After cleaning the dataset, I first explore the basic statistics of major variables I am interested in. In this part, I am assuming the gross and budget are in the same currency with the same unit. There are 66 different countries and 48 different languages in this dataset, which shows a diverse background of movies. Here's a table of statistics (values of mean, maximum, minimum and median) for variables including the number of Facebook likes, IMDB rating scores and gross.

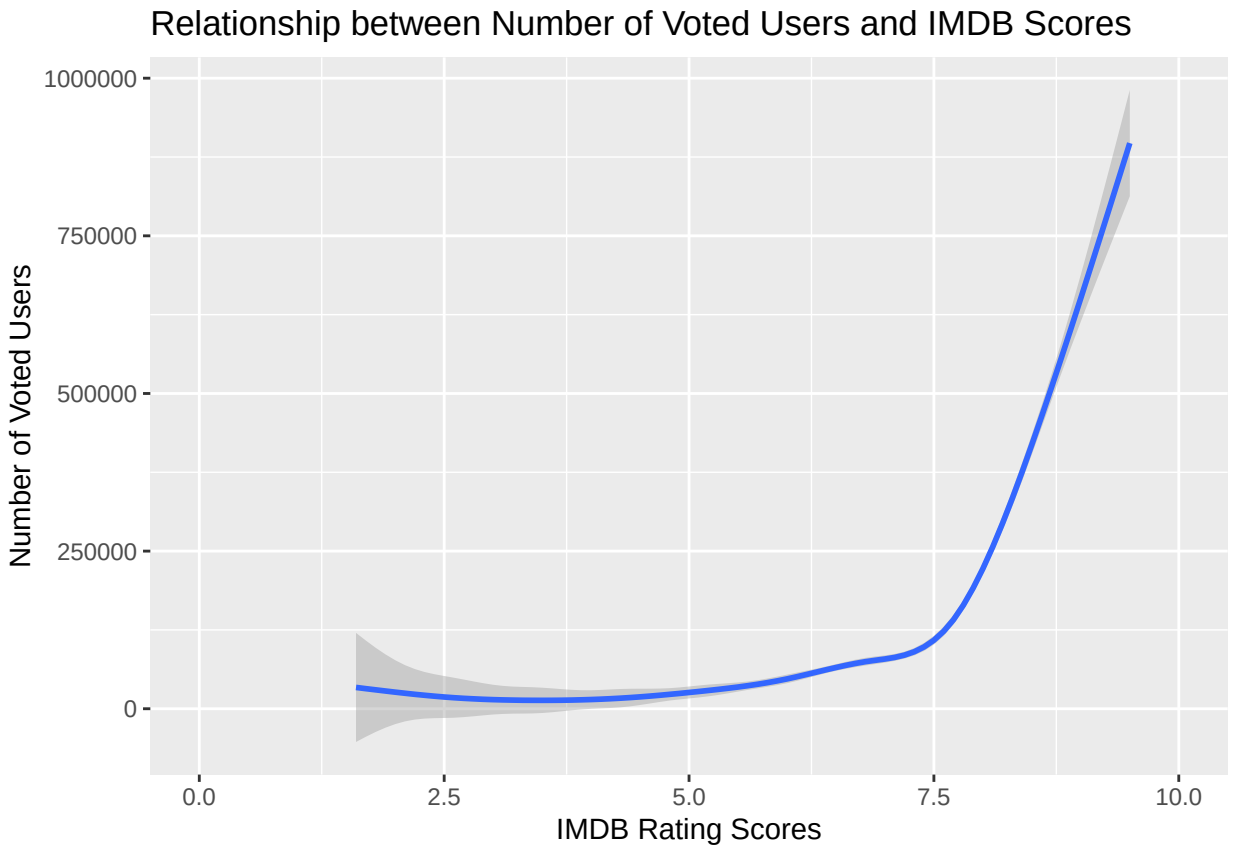
| | Mean | Maxium | Minium | Median |
|---|---|---|---|---|
| Number of Facebook likes | 7.525965e+03 | 3.490000e+05 | 0.0 | 166.0 |
| Number of Cast Total Facebook likes | 9.699064e+03 | 6.567300e+05 | 0.0 | 3090.0 |
| Number of voted users | 8.366816e+04 | 1.689764e+06 | 5.0 | 34359.0 |
| IMDB rating scores | 6.442138e+00 | 9.500000e+00 | 1.6 | 6.6 |
| Gross | 4.846841e+07 | 7.605058e+08 | 162.0 | 25517500.0 |
| Budget | 3.975262e+07 | 1.221550e+10 | 218.0 | 20000000.0 |
| Duration Time in minutes | 1.072011e+02 | 5.110000e+02 | 7.0 | 103.0 |
| Number of Critic Reviews | 1.401943e+02 | 8.130000e+02 | 1.0 | 110.0 |

## 2.2 Graph Presentation

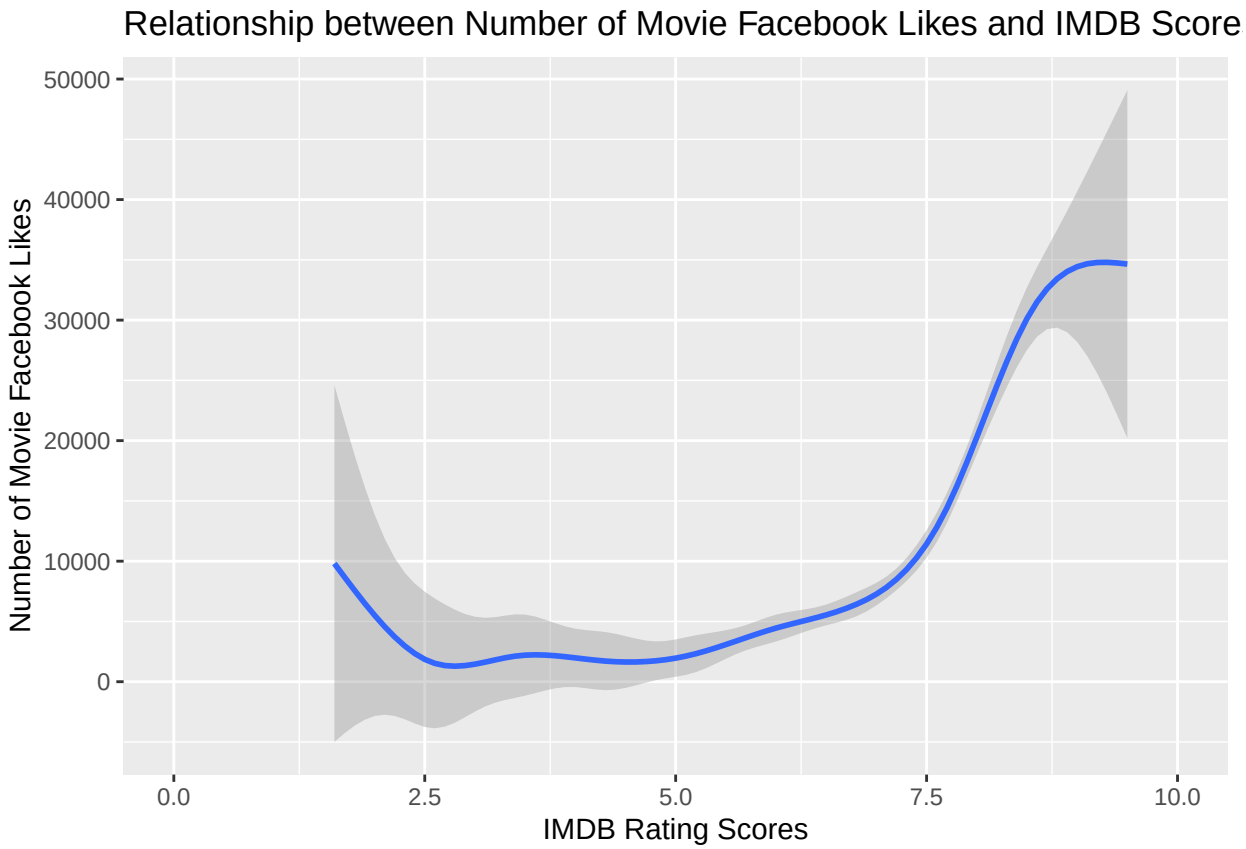
Here are some graphs showing the relationship between major variables.



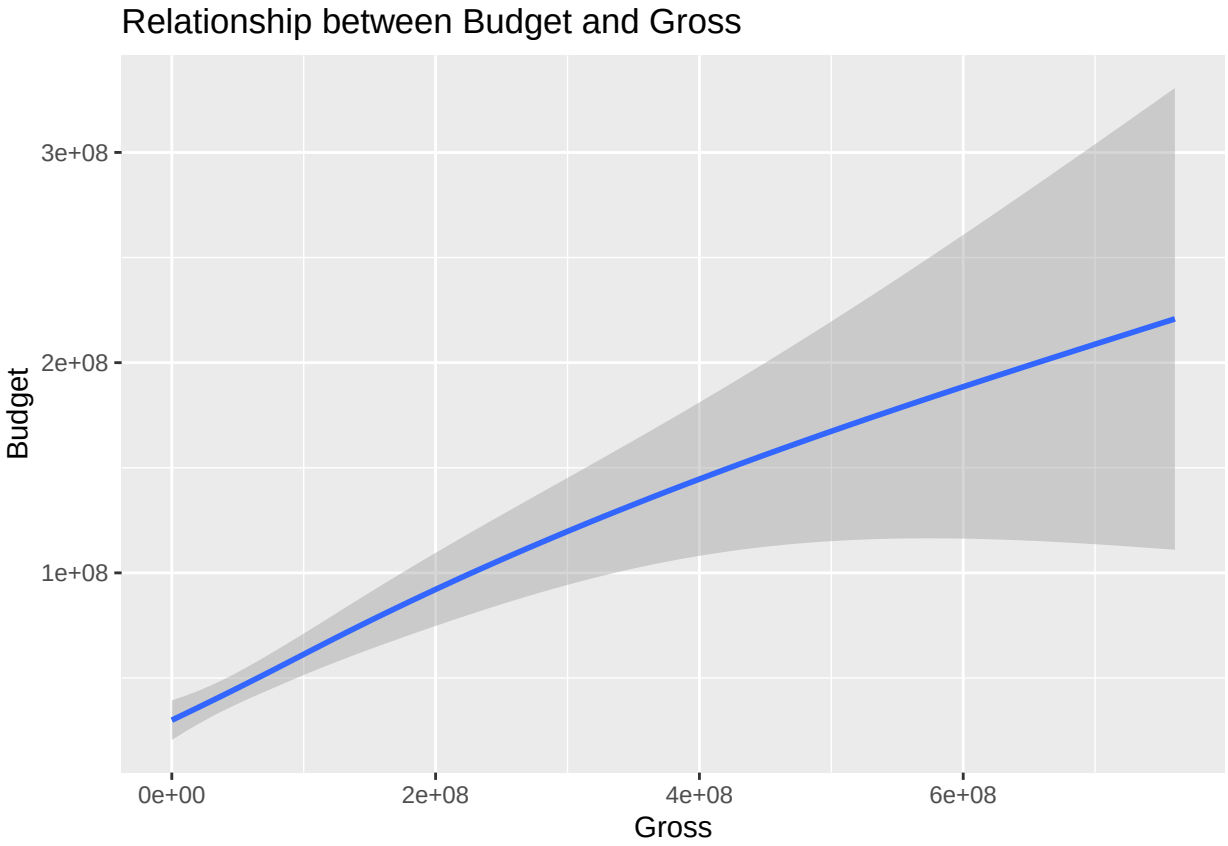

The first graph shows the relationship between movie gross and movie IMDB rating scores. Generally speaking, with a higher IMDB rating score, a movie tends to have a higher gross. Around IMDB score 7.5, the range of gross is relatively small, which indicates that movies with IMDB score 7.5 tend to have similar profit. We can see that people would spend money on movies with a good reputation and movies with a good reputation could gain more and more profit. Keeping a great reputation is kind of significant for a movie to survive in the movie market today. Indeed, more and more filming production teams put great effort into advertising their work to the public than before.



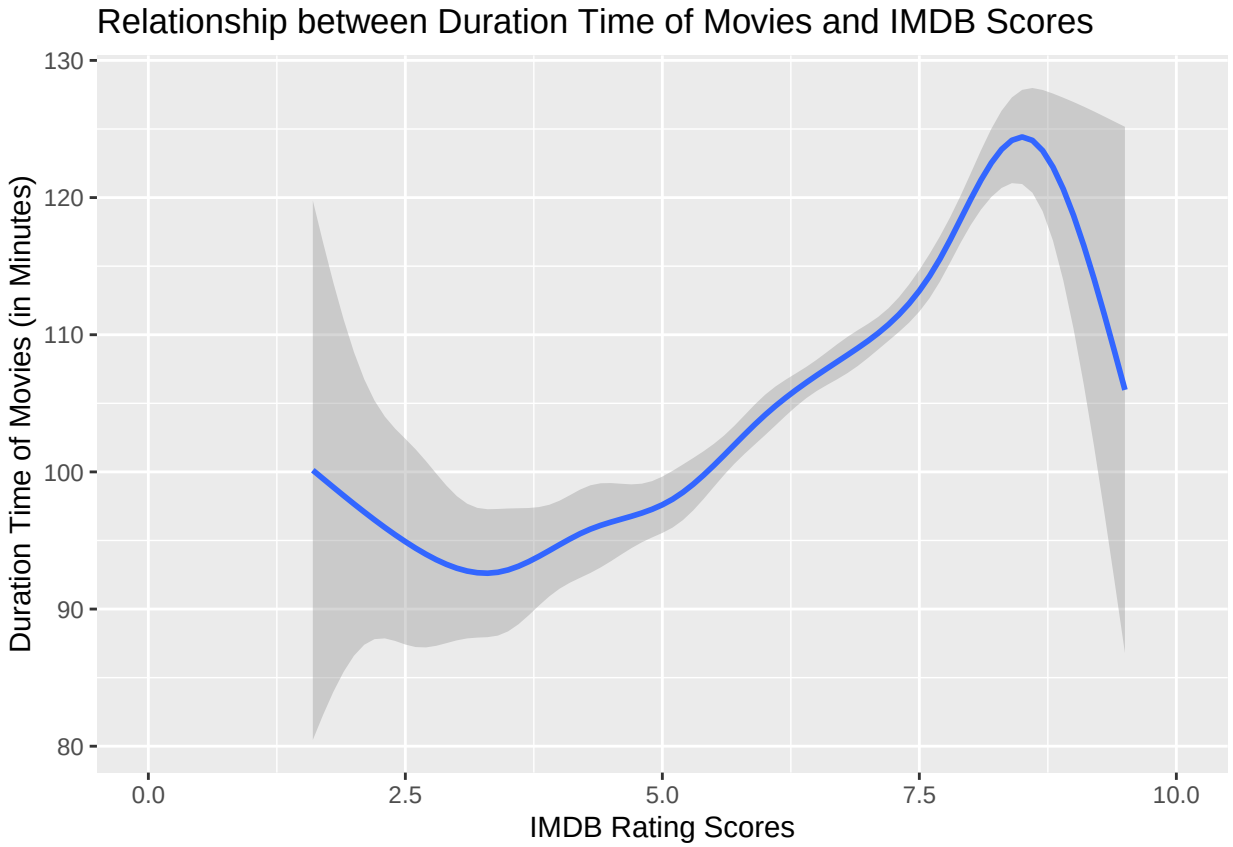

The second graph shows the relationship between IMDB rating scores and the number of voted users. It is pretty obvious that movies with higher IMDB scores tend to have more voted users, which indicates that most users have similar tastes in movies and higher-rated movies have more audiences willing to vote for them.




The third graph gives a visual interpretation of the relationship between the number of movie Facebook likes and IMDB rating scores, which also could indicate the relationship between movies' popularity and their reputation. Although some movies with lower IMDB scores receive relatively more Facebook likes, we can see that movies with higher IMDB scores generally have more Facebook likes.




From the fourth graph, we can see that generally, movies with higher gross tend to have higher budgets in the beginning phase of filming production. Furthermore, movies with higher gross tend to have a wider range of budgets than movies with lower gross. Generally speaking, better scripts of movies could earn more budgets at the beginning phase of filming production and better scripts have larger chances to gain more gross. Indeed, these days, movies with low budgets could also be famous and win many big awards with novel ideas and cutting-edge shooting techniques, which should explain the wide range of budgets in higher gross movies.




The fifth graph here interprets the relationship between the movie duration time and the IMDB rating scores visually. We can see that movies with most high scores and most low scores tend to have a wider range of duration time, but movies with scores from around 5.0 to 8.0 have relatively stable patterns in duration time. Behind the graph, duration time is not the main factor affecting IMDB rating scores since the duration time of a movie is generally between 90 minutes to 120 minutes.

## 3. Medthod

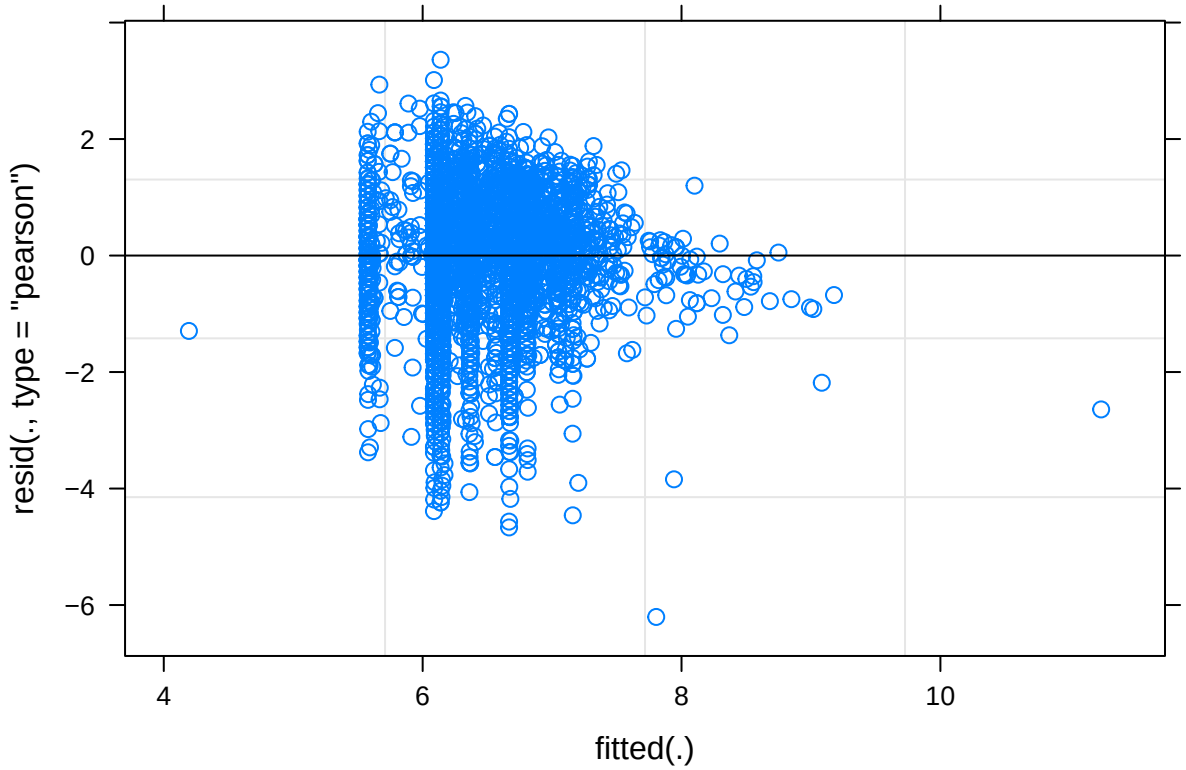
In this project, I focus on the relationship between the popularity of movies and the reputation of movies. Therefore, I use the multi-level linear regression model to analyze that if there is any relationship between the numbers of movie Facebook likes and the IMDB rating scores regarding different movie genres. I assume the first movie genre of each movie listing in the dataset to be the primary genre of the movie.

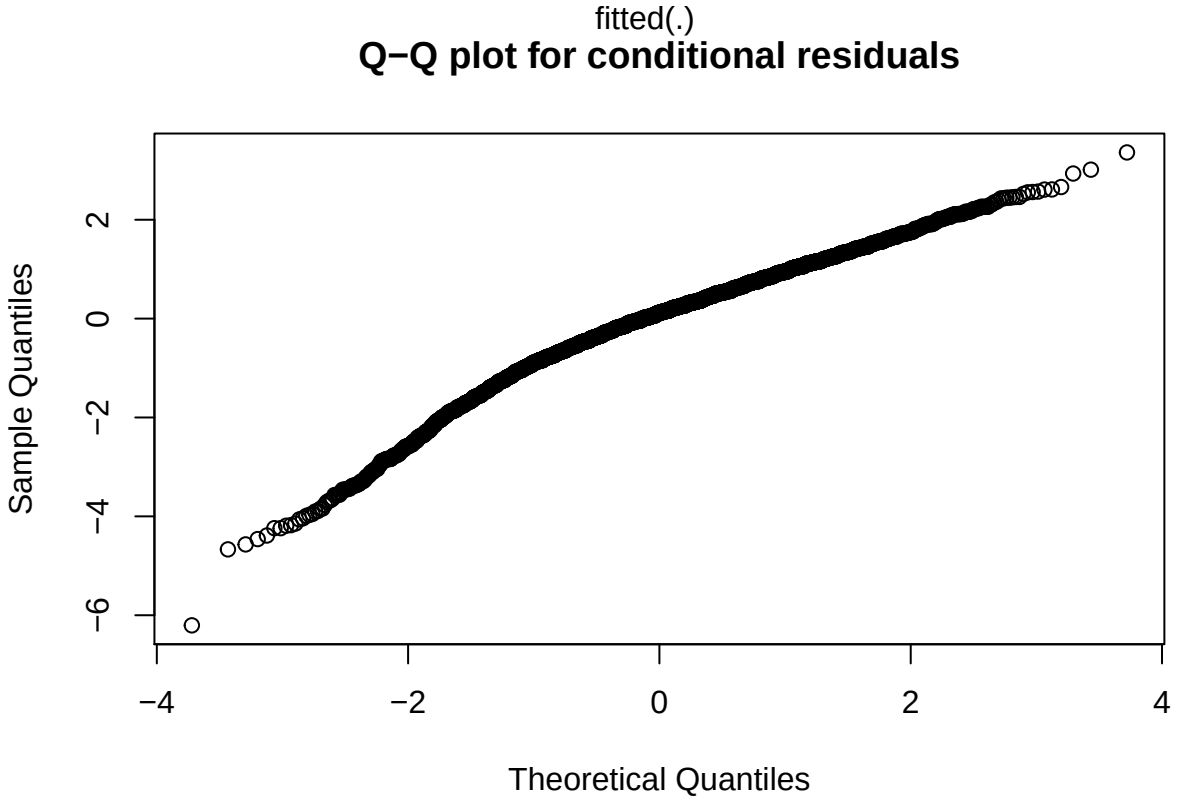
The multi-level linear regression model is also known as the hierarchical linear regression model, which is a statistical model of parameters that vary at more than one level. Multilevel models provide an alternative type of analysis for univariate or multivariate analysis of repeated measures.

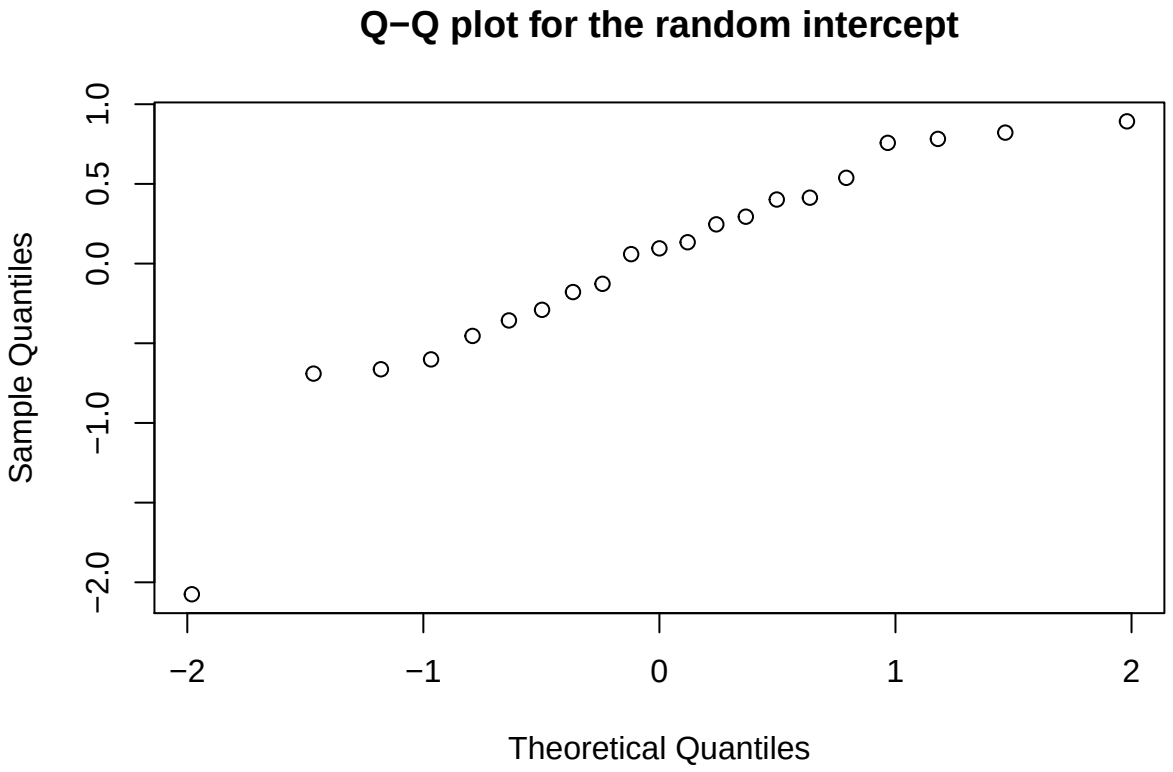
## 4. Modeling and Analysis

Here is the model I use to analyze the relationship between IMDB score and the number of movies' Facebook likes regarding different movie genres. I use correlated random intercept and random slope in this model. I believe that there might be some differences in movies' reputation and popularity regarding different movie genres. After I run this model, I check the fitted residual plot, QQ plot for conditional residuals and QQ plot for random effects. The results look quite good from the plot, so I think the model is fitted. Every movie genre has its own intercept and slope in this model and each of them is quite different, which verifies my guess. For instance, when no one likes an action movie on Facebook, this movie's IMDB score might

be around 6.09. Also, if the movie has one unit of likes on Facebook (one like), the IMDB score of this movie tends to increase in 1.52e-05 units. Besides, the change rate and intercept of each movie genre are also different from the change rate and the intercept of all of the movies.



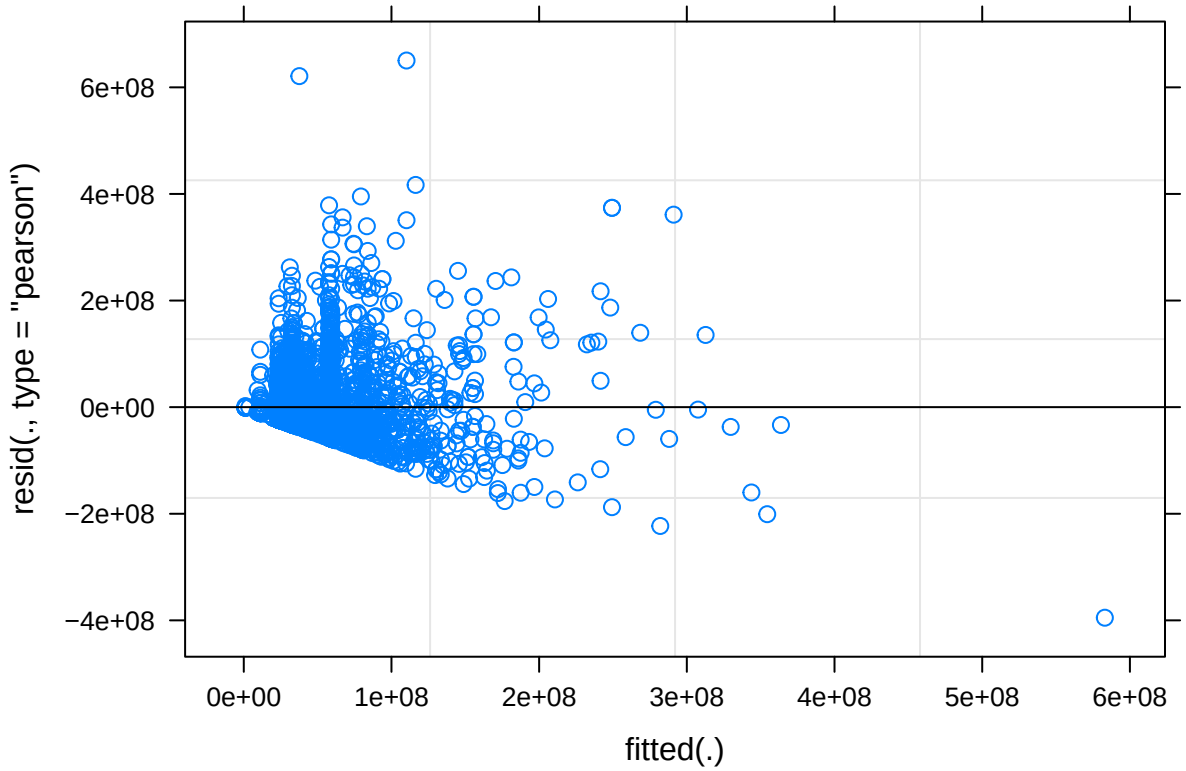



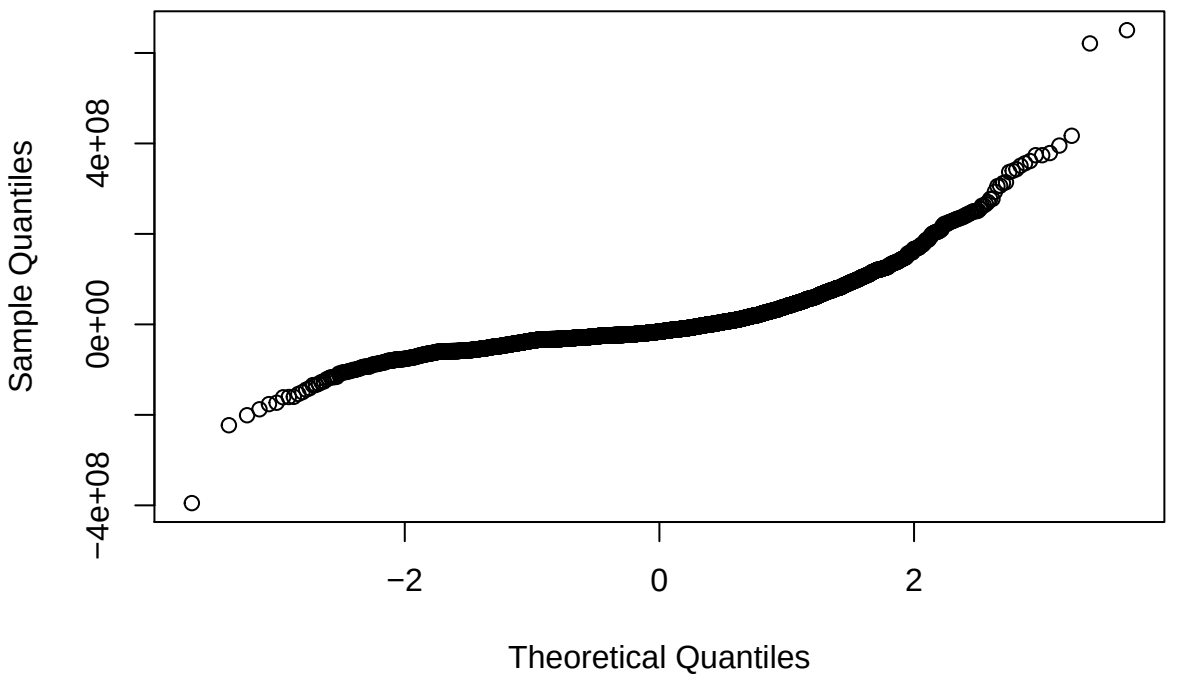


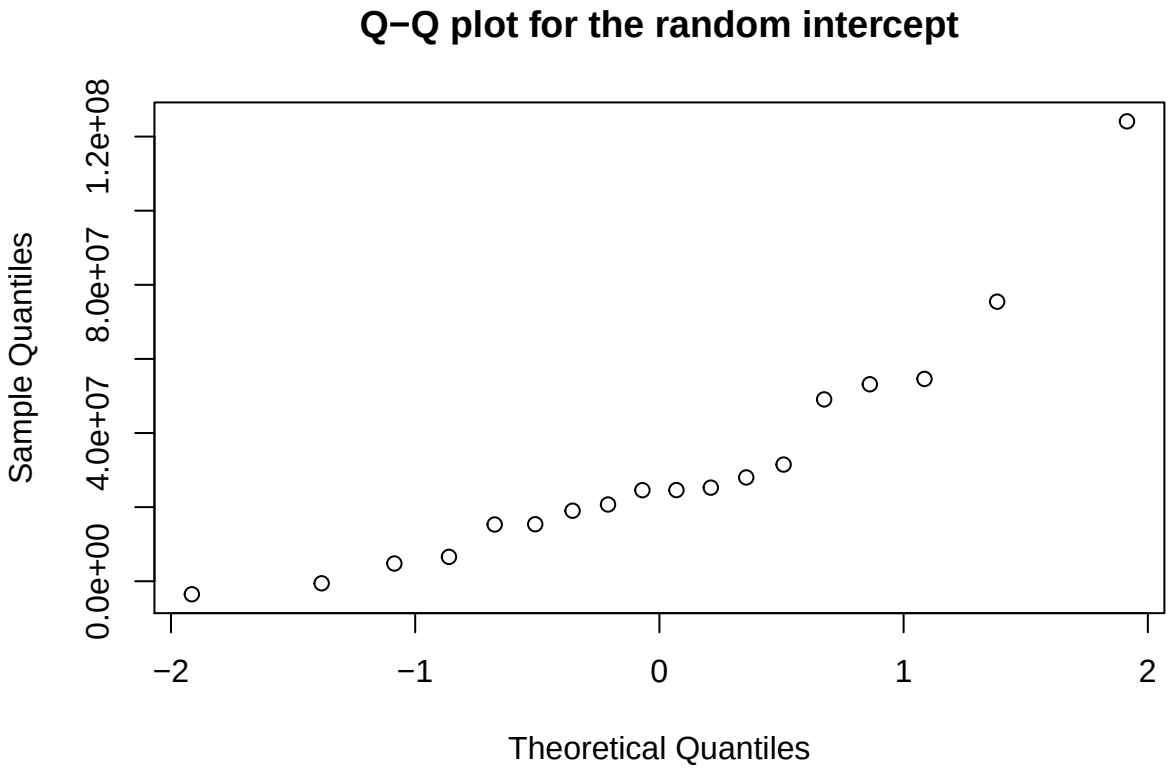

| | (Intercept) | movie_facebook_likes |
|---|---|---|
| Action | 6.087209 | 1.52e-05 |
| Adventure | 6.361377 | 1.40e-05 |
| Animation | 6.560027 | 1.31e-05 |
| Biography | 7.048167 | 1.09e-05 |
| Comedy | 6.138920 | 1.50e-05 |
| Crime | 6.803946 | 1.20e-05 |
| Documentary | 7.158380 | 1.04e-05 |
| Drama | 6.667752 | 1.26e-05 |
| Family | 5.664661 | 1.71e-05 |
| Fantasy | 6.325101 | 1.41e-05 |
| Film-Noir | 7.087716 | 1.07e-05 |
| Game-Show | 4.190923 | 2.37e-05 |
| History | 7.023388 | 1.10e-05 |
| Horror | 5.575405 | 1.75e-05 |
| Music | 6.679546 | 1.26e-05 |
| Musical | 5.975747 | 1.57e-05 |
| Mystery | 6.399792 | 1.38e-05 |
| Romance | 5.811943 | 1.64e-05 |
| Sci-Fi | 5.909130 | 1.60e-05 |
| Thriller | 5.602739 | 1.74e-05 |
| Western | 6.512019 | 1.33e-05 |

The model below is what I use to analyze the relationship between movies' gross and the number of movies' Facebook likes regarding different movie genres. I use correlated random intercept and random slope in this model. I believe that there also should be some differences in movies' profit and popularity regarding different movie genres since the popular movie genres like action and comedy tend to have more audiences. After I run this model, I check the fitted residual plot, QQ plot for conditional residuals and QQ plot for random effects. The results look quite good from the plot, so I think the model is fitted. Every movie genre has its own

intercept and slope in this model and each of them is quite different, which verifies my guess. For instance, assuming all the movies' gross are in the same currency and the same unit, when no one likes an action movie on Facebook, this movie's profit would be around 59193285. Also, if the movie has one unit of likes on Facebook (one like), the gross of this movie tends to increase in 1545.763 units. Besides, the change rate and intercept of each movie genre are also different from the change rate and the intercept of all of the movies.






| | movie_facebook_likes | (Intercept) |
|---|---|---|
| Action | 1545.76266 | 59193285 |
| Adventure | 1505.08167 | 57756314 |
| Animation | 2136.04138 | 80043648 |
| Biography | 695.88040 | 29172970 |
| Comedy | 792.95471 | 32601917 |
| Crime | 586.39900 | 25305768 |
| Documentary | 186.34537 | 11174710 |
| Drama | 538.83669 | 23625729 |
| Family | 3513.51518 | 128700032 |
| Fantasy | 696.69840 | 29201864 |
| Horror | 715.18899 | 29855006 |
| Music | -15.58185 | 4042053 |
| Musical | 1389.32312 | 53667386 |
| Mystery | 891.53907 | 36084204 |
| Romance | 435.58123 | 19978446 |
| Sci-Fi | 433.60595 | 19908673 |
| Thriller | -99.26823 | 1086007 |
| Western | 135.46400 | 9377432 |

# 5. Conclusion and Discussion

## 5.1 Results

After comparing several multi-level linear regression models, I find out which model would best fit in this project and answer my research questions. There is a positive relationship between IMDB rating scores and movies' Facebook likes but it quite varies within different movie genres. Also, the profit of movies is also related to the popularity of movies. People tend to have different rating patterns and consumption patterns regarding varied movie genres. Apparently, action movies and comedy movies attract more audiences to vote

for them and spend money on them. Usually, more audiences could lead to higher reputations of movies. Indeed, the filming industry people should put great effort into attracting audiences and advertising their works.

## 5.2 Limitations and Concerns

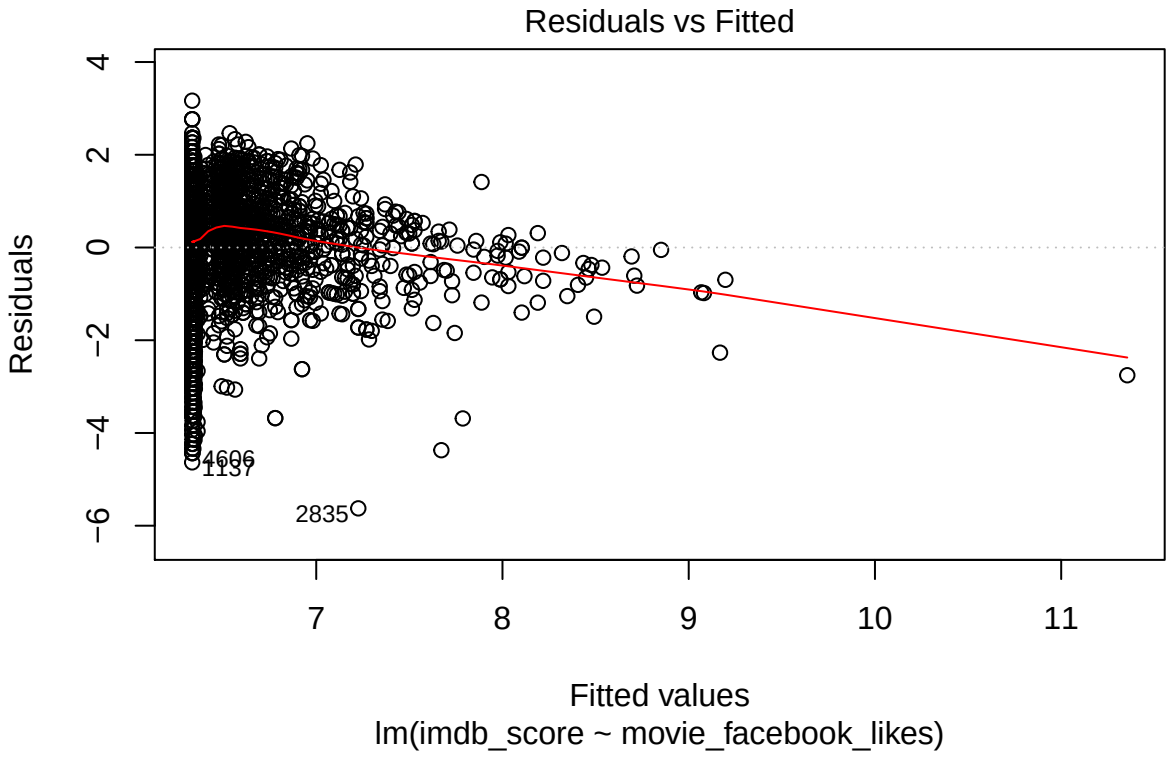
For now, I could not tell that if the movies' Facebook likes influence the IMDB rating scores or the opposite. Since I have no information about the time range of Facebook likes recording, the number of likes might be from people who watch the movie after the release time of the movie. Therefore, some of the likes are not correlated with profit.

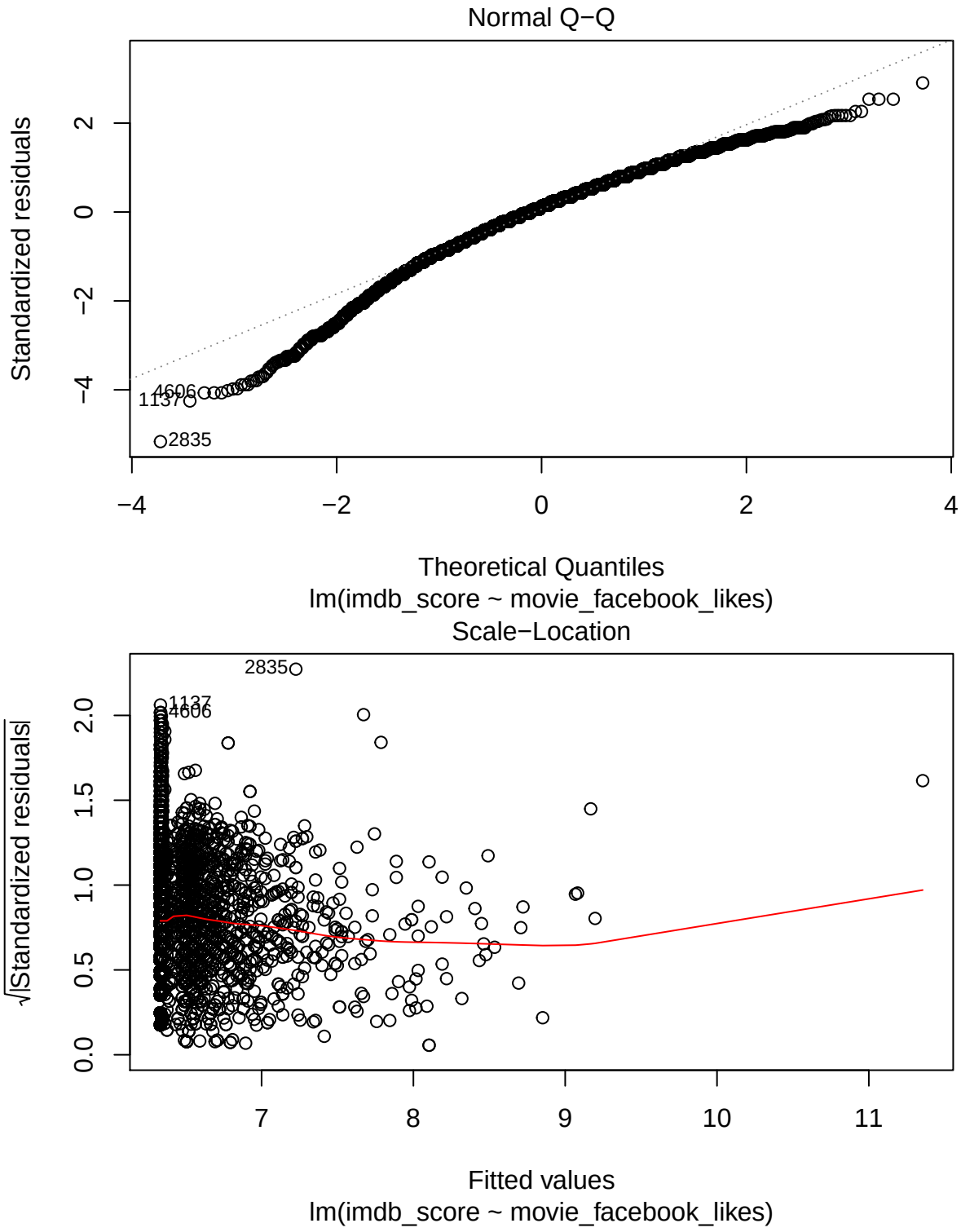
# 6. Appendix and Reference

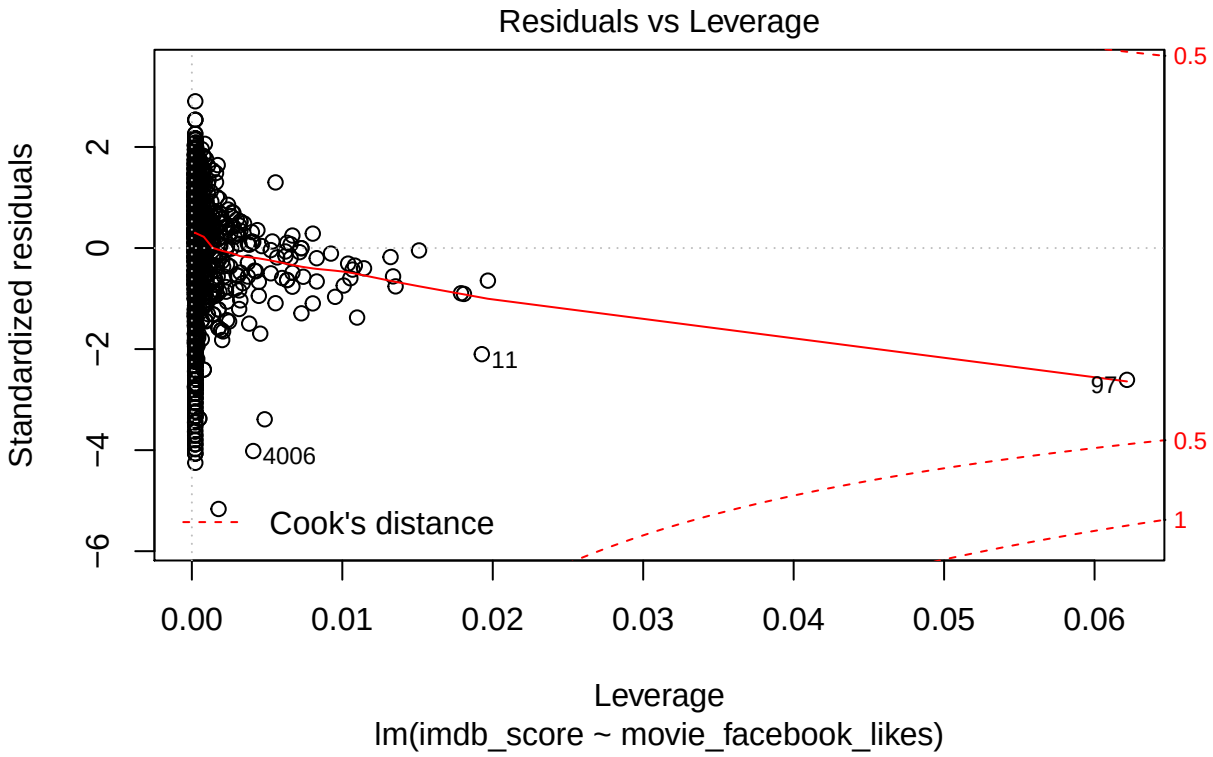
## 6.1 Appendix

Here are some models not being used in this project.

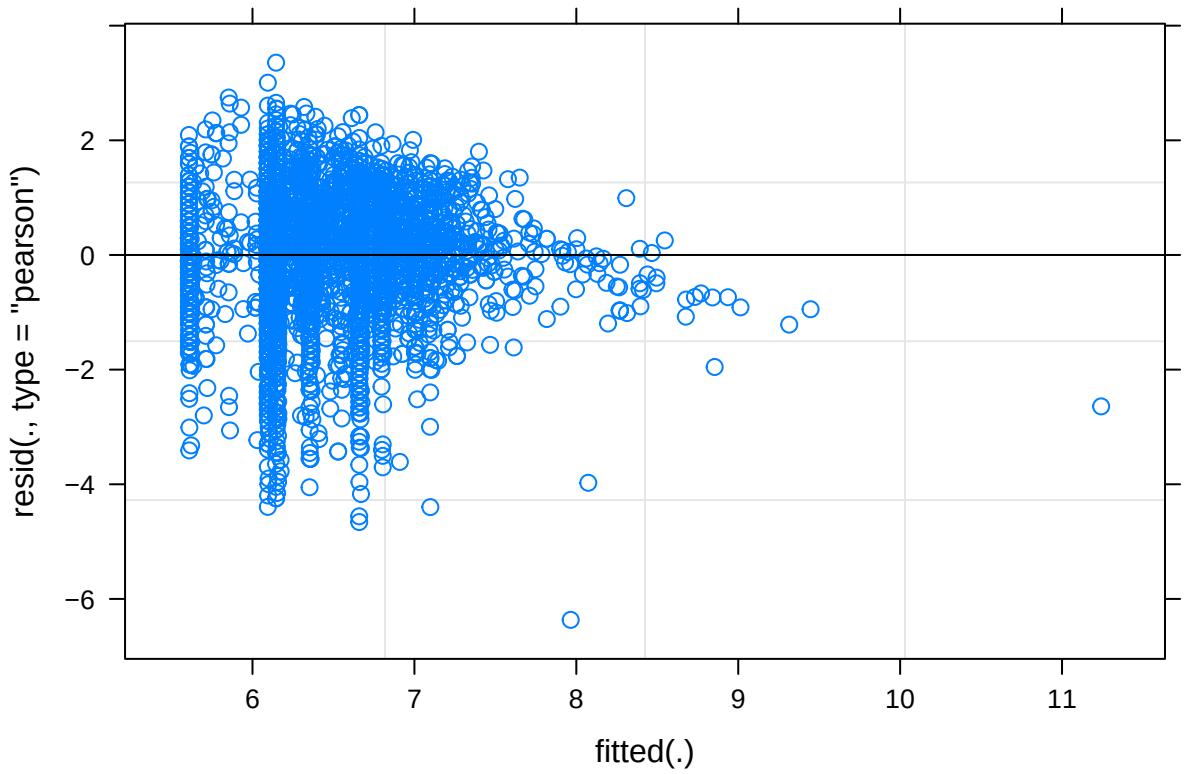
```
##
## Call:
## lm(formula = imdb_score ~ movie_facebook_likes, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6258 -0.6339  0.1359  0.7661  3.1661
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          6.334e+00  1.648e-02   384.4   <2e-16 ***
## movie_facebook_likes 1.439e-05  7.948e-07    18.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.09 on 5041 degrees of freedom
## Multiple R-squared:  0.06103,    Adjusted R-squared:  0.06085
## F-statistic: 327.7 on 1 and 5041 DF,  p-value: < 2.2e-16
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(imdb_score ~ movie_facebook_likes)
Scale−Location
√|Standardized residuals|
Fitted values
lm(imdb_score ~ movie_facebook_likes)



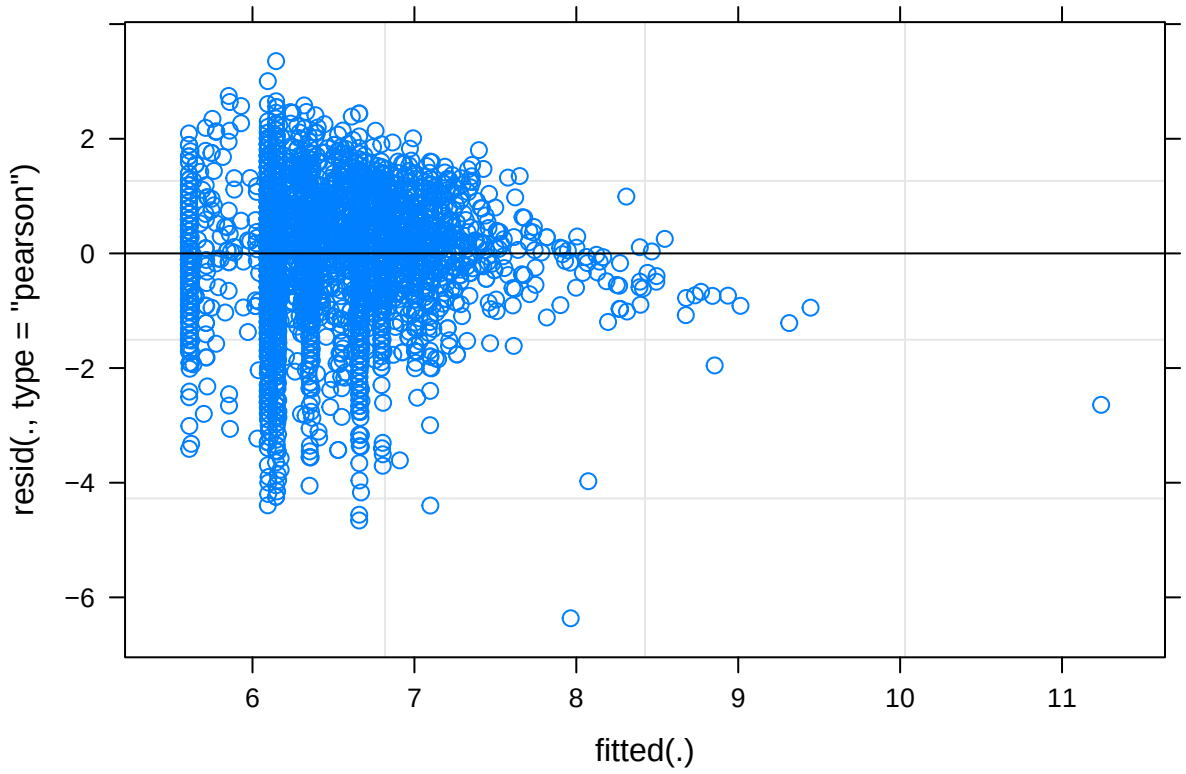

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: imdb_score ~ movie_facebook_likes + (1 | primarygenre)
##    Data: df
##
## REML criterion at convergence: 14723.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -6.1578 -0.5363  0.1019  0.6645  3.2450
##
## Random effects:
##  Groups       Name        Variance Std.Dev.
##  primarygenre (Intercept) 0.2323   0.482
##  Residual                 1.0685   1.034
## Number of obs: 5043, groups:  primarygenre, 21
##
## Fixed effects:
##                       Estimate Std. Error t value
## (Intercept)          6.306e+00  1.231e-01   51.21
## movie_facebook_likes 1.401e-05  7.628e-07   18.37
##
## Correlation of Fixed Effects:
##             (Intr)
## mv_fcbk_lks -0.039
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
```




```
##              (Intercept) movie_facebook_likes
## Action          6.094673           1.4009e-05
## Adventure       6.352393           1.4009e-05
## Animation       6.546038           1.4009e-05
## Biography       7.003474           1.4009e-05
## Comedy          6.145667           1.4009e-05
## Crime           6.795879           1.4009e-05
## Documentary     7.096607           1.4009e-05
## Drama           6.659036           1.4009e-05
## Family          5.852339           1.4009e-05
## Fantasy         6.315162           1.4009e-05
## Film-Noir       6.536951           1.4009e-05
## Game-Show       5.697259           1.4009e-05
## History         6.518250           1.4009e-05
## Horror          5.607247           1.4009e-05
## Music           6.417984           1.4009e-05
## Musical         6.140864           1.4009e-05
## Mystery         6.407384           1.4009e-05
## Romance         6.018550           1.4009e-05
## Sci-Fi          6.026566           1.4009e-05
## Thriller        5.710566           1.4009e-05
## Western         6.479654           1.4009e-05
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: imdb_score ~ movie_facebook_likes + ((1 | primarygenre) + (0 +
##     movie_facebook_likes | primarygenre))
##    Data: df
##
## REML criterion at convergence: 14723.8
##
```

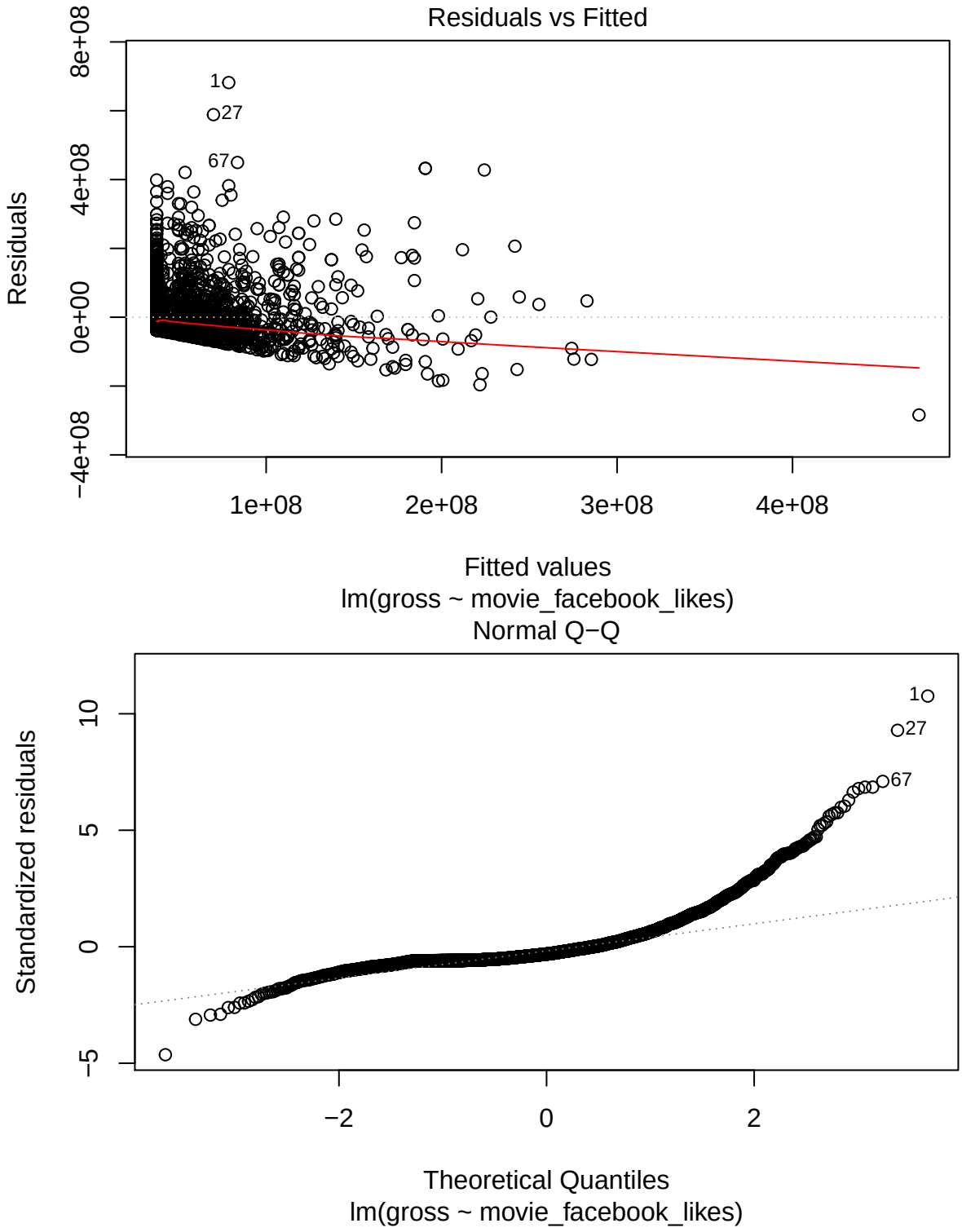
```
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -6.1578 -0.5363  0.1019  0.6645  3.2450
##
## Random effects:
##  Groups         Name                 Variance Std.Dev.
##  primarygenre   (Intercept)          0.2323   0.482
##  primarygenre.1 movie_facebook_likes 0.0000   0.000
##  Residual                            1.0685   1.034
## Number of obs: 5043, groups:  primarygenre, 21
##
## Fixed effects:
##                       Estimate Std. Error t value
## (Intercept)          6.306e+00  1.231e-01   51.21
## movie_facebook_likes 1.401e-05  7.628e-07   18.37
##
## Correlation of Fixed Effects:
##             (Intr)
## mv_fcbk_lks -0.039
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

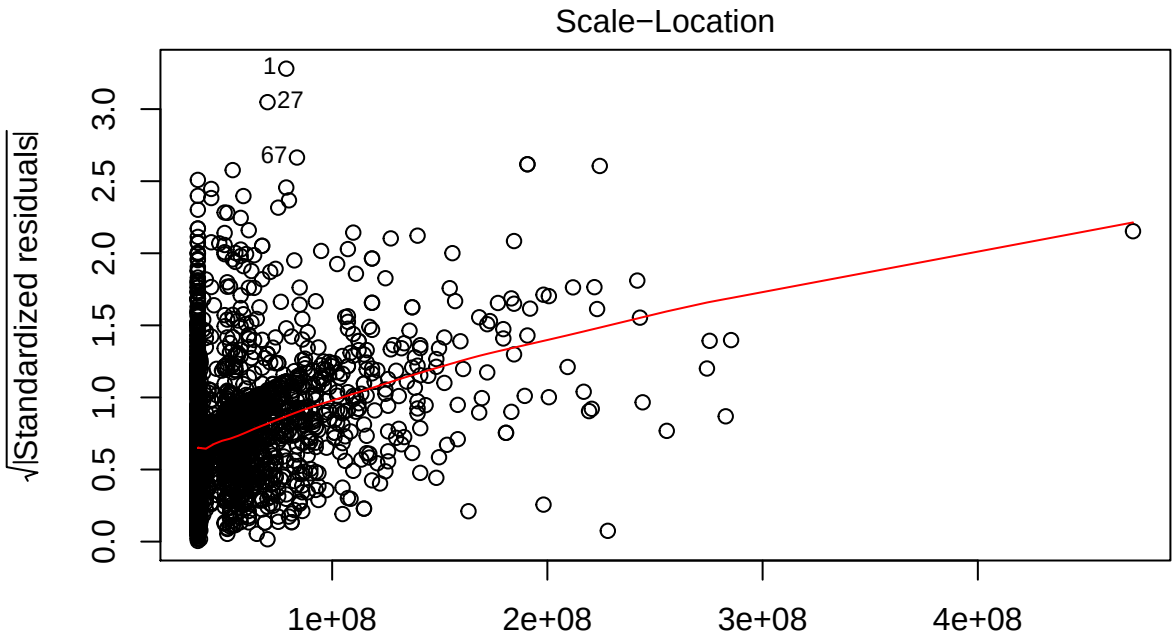


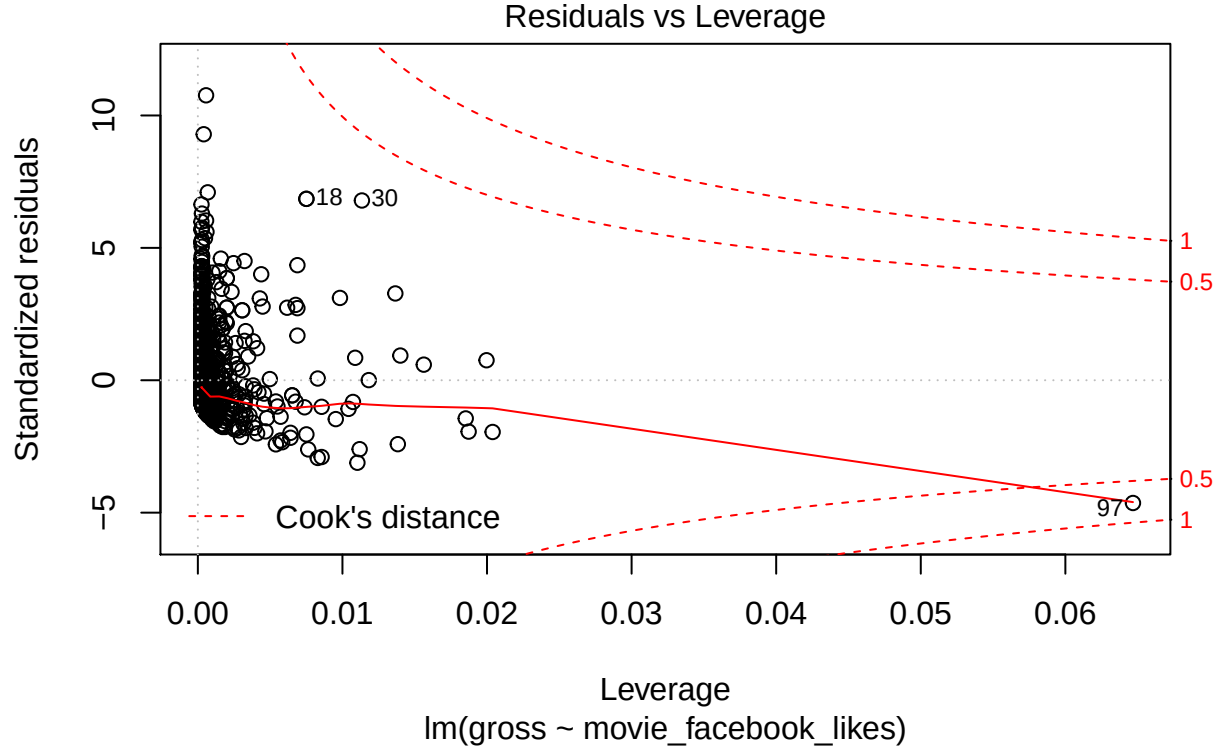

```
##             (Intercept) movie_facebook_likes
## Action         6.094673           1.4009e-05
## Adventure      6.352393           1.4009e-05
## Animation      6.546038           1.4009e-05
## Biography      7.003474           1.4009e-05
## Comedy         6.145667           1.4009e-05
```

```
## Crime          6.795879           1.4009e-05
## Documentary    7.096607           1.4009e-05
## Drama          6.659036           1.4009e-05
## Family         5.852339           1.4009e-05
## Fantasy        6.315162           1.4009e-05
## Film-Noir      6.536951           1.4009e-05
## Game-Show      5.697259           1.4009e-05
## History        6.518250           1.4009e-05
## Horror         5.607247           1.4009e-05
## Music          6.417984           1.4009e-05
## Musical        6.140864           1.4009e-05
## Mystery        6.407384           1.4009e-05
## Romance        6.018550           1.4009e-05
## Sci-Fi         6.026566           1.4009e-05
## Thriller       5.710566           1.4009e-05
## Western        6.479654           1.4009e-05
```

```
##
## Call:
## lm(formula = gross ~ movie_facebook_likes, data = df)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -284103775  -36171382  -19864613   13606771  681822670
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.760e+07  1.066e+06   35.27   <2e-16 ***
## movie_facebook_likes 1.245e+03  4.728e+01   26.33   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63380000 on 4157 degrees of freedom
##   (884 observations deleted due to missingness)
## Multiple R-squared:  0.1429, Adjusted R-squared:  0.1427
## F-statistic: 693.3 on 1 and 4157 DF,  p-value: < 2.2e-16
```

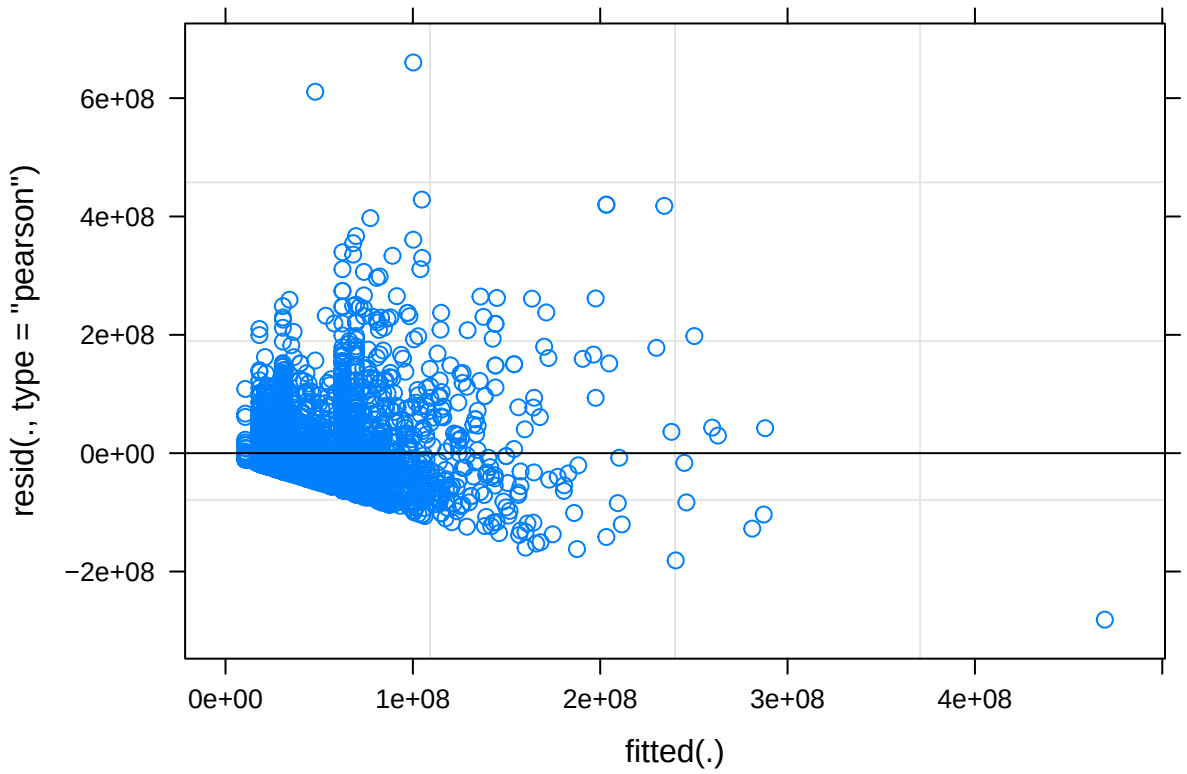

Residuals vs Fitted
Residuals
Fitted values
lm(gross ~ movie_facebook_likes)
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(gross ~ movie_facebook_likes)




```
## Linear mixed model fit by REML ['lmerMod']
## Formula: gross ~ movie_facebook_likes + (1 | primarygenre)
##    Data: df
##
## REML criterion at convergence: 160797.5
##
```

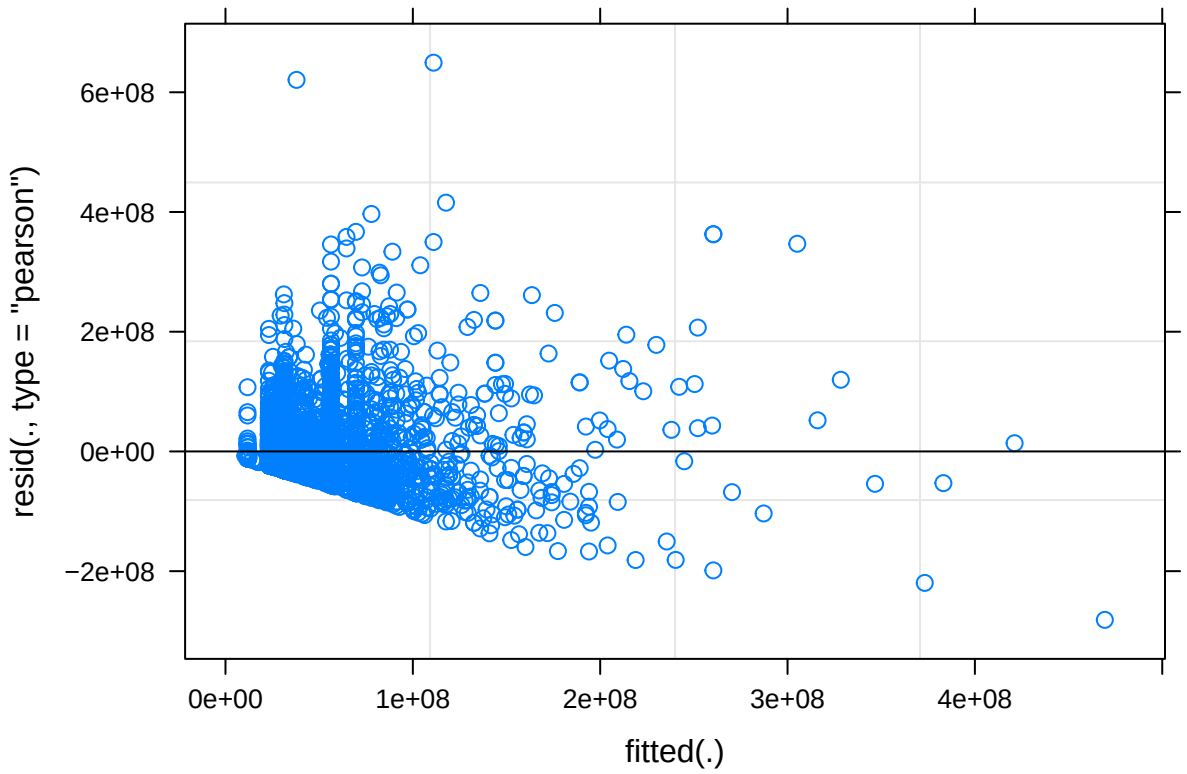
```
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.6751 -0.5032 -0.2446  0.2430 10.9721
##
## Random effects:
##  Groups       Name        Variance  Std.Dev.
##  primarygenre (Intercept) 5.403e+14 23243311
##  Residual                 3.622e+15 60178915
## Number of obs: 4159, groups:  primarygenre, 18
##
## Fixed effects:
##                       Estimate Std. Error t value
## (Intercept)          3.487e+07  6.524e+06   5.345
## movie_facebook_likes 1.145e+03  4.550e+01  25.166
##
## Correlation of Fixed Effects:
##             (Intr)
## mv_fcbk_lks -0.059
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
```




```
##             (Intercept) movie_facebook_likes
## Action         62423478             1145.174
## Adventure      69667718             1145.174
## Animation      73554369             1145.174
## Biography      22559298             1145.174
## Comedy         30618322             1145.174
## Crime          21071846             1145.174
## Documentary    10556093             1145.174
## Drama          18117752             1145.174
```

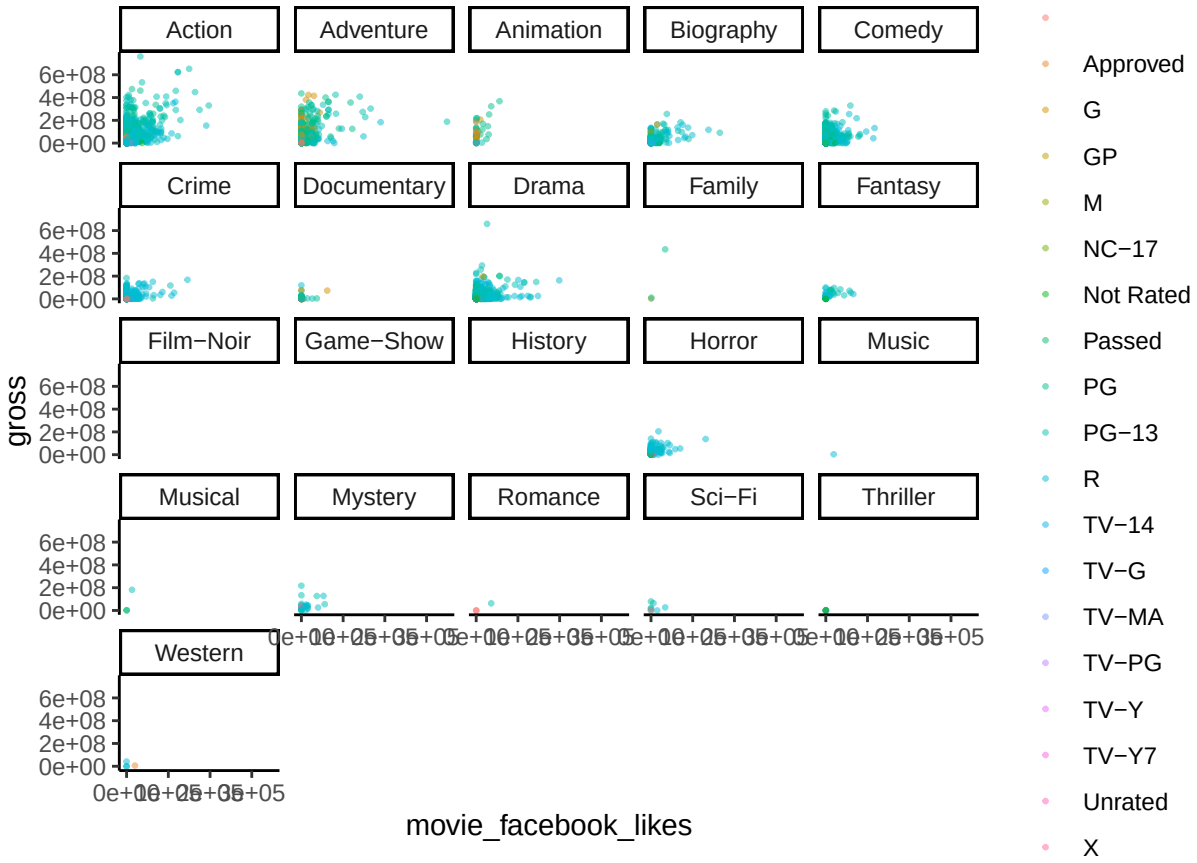
```
## Family         66047073             1145.174
## Fantasy        26375667             1145.174
## Horror         27305243             1145.174
## Music          27963445             1145.174
## Musical        41537661             1145.174
## Mystery        35082694             1145.174
## Romance        26259176             1145.174
## Sci-Fi         25794394             1145.174
## Thriller       18558409             1145.174
## Western        24179821             1145.174
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## gross ~ movie_facebook_likes + ((1 | primarygenre) + (0 + movie_facebook_likes |
##     primarygenre))
##    Data: df
##
## REML criterion at convergence: 160729
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.7636 -0.5061 -0.2645  0.2463 10.9950
##
## Random effects:
##  Groups         Name                Variance  Std.Dev.
##  primarygenre   (Intercept)         5.282e+14 22982206
##  primarygenre.1 movie_facebook_likes 7.779e+07     8820
##  Residual                           3.489e+15 59066639
## Number of obs: 4159, groups:  primarygenre, 18
##
## Fixed effects:
##                      Estimate Std. Error t value
## (Intercept)          32195769    6609366   4.871
## movie_facebook_likes     1946       2180   0.893
##
## Correlation of Fixed Effects:
##             (Intr)
## mv_fcbk_lks -0.041
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
## convergence code: 0
## unable to evaluate scaled gradient
## Model failed to converge: degenerate  Hessian with 1 negative eigenvalues
```



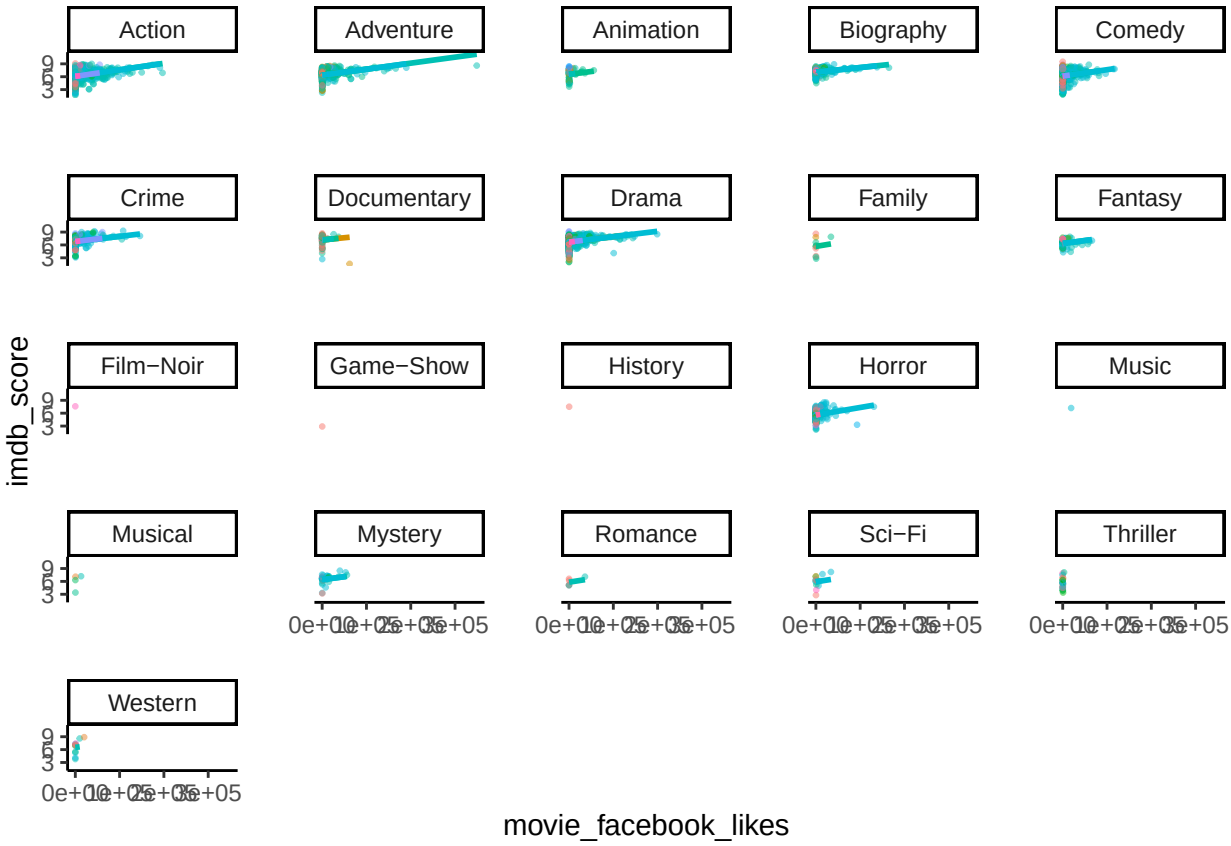

```
##               (Intercept) movie_facebook_likes
## Action           56308047            1659.3078
## Adventure        69629082            1145.3614
## Animation        55527645            4652.1488
## Biography        25499514             868.0314
## Comedy           31454029             946.2798
## Crime            25340503             595.7528
## Documentary      11773885             507.5150
## Drama            23187510             567.1202
## Family           26095651           11620.6477
## Fantasy          30827714             640.9175
## Horror           29999601             719.3135
## Music            31113706           -1081.3815
## Musical          27817075            9682.5152
## Mystery          38041254             767.7981
## Romance          24595996            1075.3529
## Sci-Fi           29429167            -155.8830
## Thriller         16997699            1520.1942
## Western          25885769            -694.0463
```

The graph below shows the relationship between movies' gross and movies' Facebook likes in different movie genres.




```
## List of 2
##  $ legend.position: chr "none"
##  $ panel.spacing  : 'unit' num 2lines
##   ..- attr(*, "valid.unit")= int 3
##   ..- attr(*, "unit")= chr "lines"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

The graph below shows the relationship between IMDB rating scores and movies' Facebook likes in different movie genres. Also, there are more action movies and comedy movies releasing from 1916 to 2016 than history movies or game-shows.




## 6.2 Reference

https://en.wikipedia.org/wiki/Multilevel_model

https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset