Regression Analysis of IMDB 5000 Movies Dataset

Meierhaba Rexiti

California State University of East Bay

Abstract

This paper explores which variables among the given 28, are critical in telling the IMDb rating of a movie. By doing regression analysis on the IMDb dataset applying machine leaning algorithms, this article also examine how predictors, such as duration, number of critics for review, and face numbers in posters, correlate to the IMDb score. It turns out that the most important factor that effects movie rating is the duration. In general, the longer the duration, the higher the score is. But if length of duration is too long, it ends up decreasing the score. The more the number of critics reviewing a movie, the higher the score will be. Lastly, a poster with many face numbers tends to have lower score. The study predicts the IMDb score using the best fitted model. By calculating the mean absolute error, verifying that 51.14% of the IMDb scores in test dataset had been correctly predicted by our model.

Keywords: regression, machine learning, means absolute error

Regression Analysis of IMDb 5000 Movies Dataset

How can we tell the goodness of a movie and make suggestion to friends without relying on our instinct? Is there any universal way to claim the quality of movies? This is something that triggers my partner's and my interest. The purpose of this study is to explore among the given 28 variables, which of them are critical in telling the IMDb rating of a movie. Do people think longer movies are generally considered as good movies to which the answer is yes after our analysis, which is surprising. But also, the scores starts decreasing as duration keeps increasing to certain point. How other predictors, such as number of critics for review and face number in posters, correlate to the movie IMDb score, respectively, will also give us some interesting insights. Last but not the least, the study will predict IMDb score with a model we created and evaluate model performance by calculating prediction accuracy.

Materials and Methods

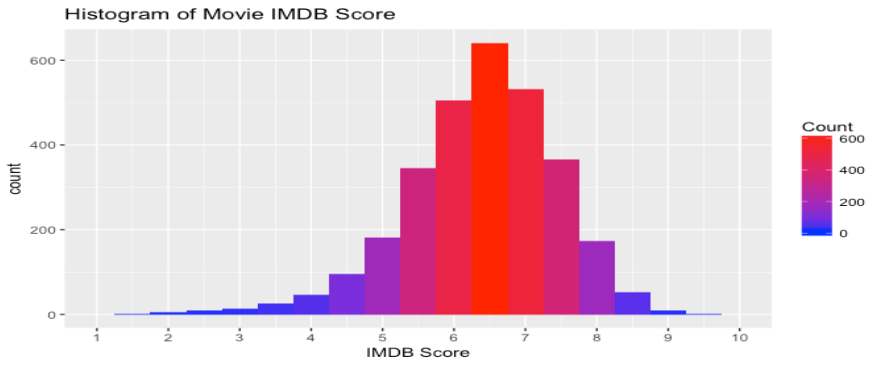
This data set was obtained from Kaggle. The author scraped 5000+ movies from IMDB website and obtained all needed 28 variables for 5043 movies and 4906 posters (998MB), spanning across 100 years in 66 countries. There are 2399 unique director names, and thousands of actors/actresses. Below are the 28 variables: movie title, color, number of critic for reviews, movie facebook likes, duration, director name, director facebook likes, actor 3 name, actor 3 facebook likes, actor 2 name, actor 2 facebook likes, actor 1 name, actor 1 facebook likes, gross, genres, number of voted users, cast total facebook likes, face number in posters, plot keywords, movie imdb link, number of user for reviews, languages, country, content rating, budget, title year (year of release), IMDB score and aspect ratio.

This dataset is a proof of concept. It can be used for experimental and learning purpose. For comprehensive movie analysis and accurate movie ratings predictions, a decent data set and

a lot more attributes are needed.

The Analysis starts with data cleaning. First of all, since each movie can belong to many genres, we only assign the first genre appeared in the "genres" column. Next, not all budgets are in US dollars. It is beyond our scope of interest to convert all to same currency since inflation and other economic factors would have to be taken into consideration. Therefore, we only keep movies from USA and thus, predictor "language" with only 10 observations different from English is no longer useful in prediction and been excluded. IMDB links are also removed with little prediction value. After that, a missing value plot separate by predictors is created and all missing values are then removed.

After data cleaning, a histogram of dependent variable, IMDb score, is plotted as below:



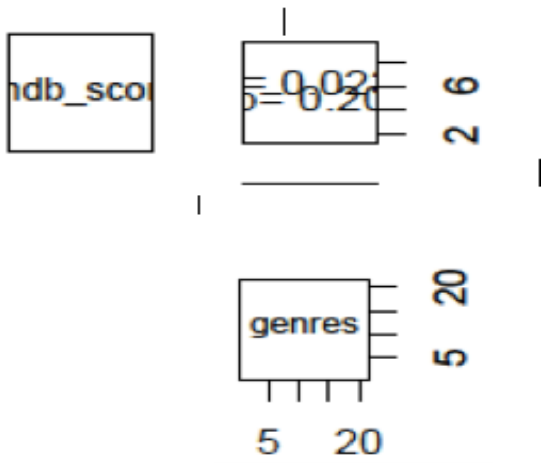
From the plot, IMDb score follows a normal distribution but a little skewed to the left. Number of observations with lower scores between 1 and 5 is greater than generally considered good movies with scores between 8 and 10.

Next, title year is examined. We don't think that IMDb scores are time dependent and this has been verified with a time series plot with no trend nor seasonal component in it. Therefore, this will just be a regression analysis without time series effect. In addition, in another analysis we created with everything else held the same, but treating "title year" as categorical

variable, due to the large amount of levels in title year, it is hard to make predictions while not all levels of years appeared in test data set have appeared and been estimated in the training data, because observations in train and test data sets are randomly selected. Therefore, for model precision purpose, we decide to treat title year as numerical variable.

The last step in data exploration is looking at the correlation between variables and IMDb scores. A correlation plot and a correlation test with p-values associated for all numerical variables are created. The p-values for correlation test is shown below:

	imdb_score
num_critic_for_reviews	0.00
duration	0.00
director_facebook_likes	0.00
actor_3_facebook_likes	0.00
actor_1_facebook_likes	0.00
gross	0.00
num_voted_users	0.00
cast_total_facebook_likes	0.00
facenumber_in_poster	0.00
num_user_for_reviews	0.00
budget	0.00
title_year	0.00
actor_2_facebook_likes	0.00
imdb_score	0.00
aspect_ratio	0.04
movie_facebook_likes	0.00



From the picture on the left, all numerical variables are significant, but aspect of ratio is excluded from our model since p-value is very close to 0.05. From the graph on the right, genre has a p-value of correlation of 0.20, but still will be include in the model just for practice purpose. At the same time, from the pairs.panels plot, all the actors and director names are excluded since none of them are significant. To wrap up, now there are only 12 columns are kept, those are:

```
'imdb_score','num_voted_users','num_critic_for_reviews','num_user_for_reviews
','duration','facenumber_in_poster','gross','movie_facebook_likes','director_
facebook_likes','cast_total_facebook_likes','budget','title_year','genres')
```

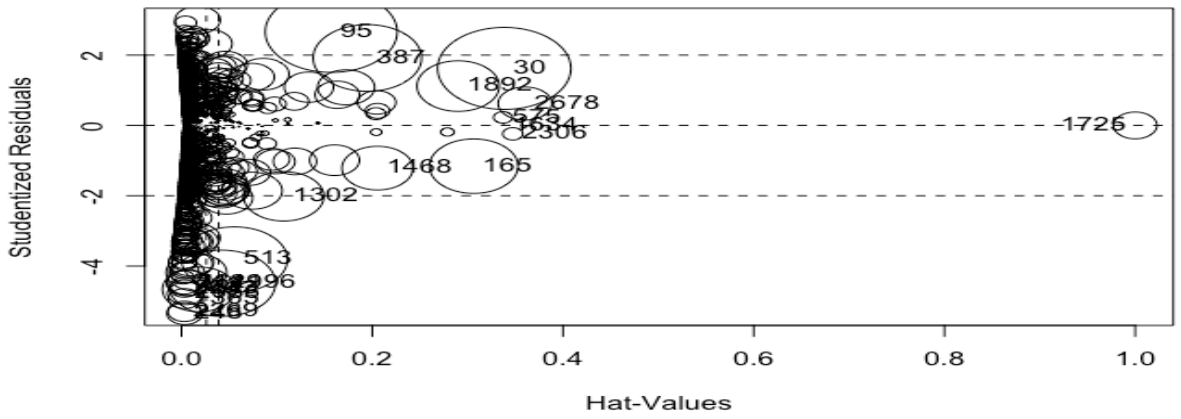
Next step is model selection. The study starts with step function with AIC criteria. Considering balancing of AIC number and model simplicity, among the three full models, the

later analysis is based on the linear additive model with the interaction terms of number voted users and number of user for reviews, gross and budget, duration and number voted users. The three interaction terms are picked, because they appear to have very significant correlation from the correlation plot with p-values smaller than 0.05. The null model, full model and the step function chosen are as below:

- null=lm(imdb_score~1,data=movie.sig)
- full3=lm(imdb_score~num_voted_users+num_critic_for_reviews+num_user_for_reviews+duration+facenumber_in_poster+gross+movie_facebook_likes+director_facebook_likes+cast_total_facebook_likes+budget+title_year+factor(genres)+duration*num_voted_users+ num_voted_users* num_user_for_reviews+ gross* budget, data=movie.sig)
- Step:AIC=-1535.83
  imdb_score~num_voted_users+factor(genres)+title_year+num_critic_for_reviews+budget+duration+num_voted_users:duration

However, the model suggested from the step function has an Adjusted R^2 of 0.4629, which is lower than that of the full model (0.4778) and with all terms significant. Therefore, we take the approach to study from the full model. After splitting 90% of the data into train and the rest into test data set, we fit a model excluding insignificant terms: director facebook likes, movie facebook likes, and cast total facebook likes. Lack of fit test on those two models have a p-value of 0.3276, suggesting dropping the terms have contribution in improving model performance. Yet, the residuals of this model have a general trend of curvature, so we fit another model adding quadratic terms to all except “face number in posters” and “title year”. This increases the adjusted R^2 to 0.5116. The third model lm.fit3, excluded the insignificant terms,

quadratic term for 'gross' and interaction of budget*gross, based on the result of the previous model. The partial F-test has a p-value of 0.1063, confirming lm.fit3 is a good fit. After checking the residuals, we thought maybe a quadratic term should be added to title year as well. But, lack of fit test with no p-value tells us adding the term makes no change to the model. Last, we produce an influence plot to check for residual outliers for model lm.fit3:

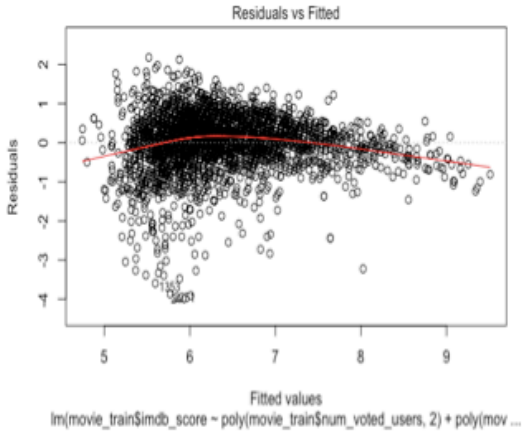


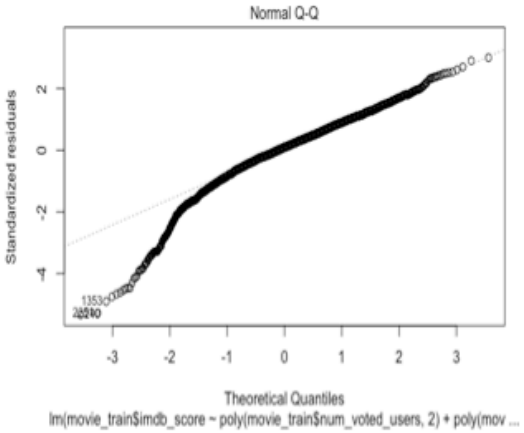
From the plot, we remove 10 points: 1725,165,30,95,387,1892,2678,1468,634 and 57, this is fitted as lm.fit5. Although the coefficients have not been changed much based on the coefficients comparison, the residuals plots show no improvement and stays the same. Since lm.fit5 has an adjusted R^2 of 0.5116, 51% of the variation can be explained by this model, which is very close to lm.fit3. For the convenience of later analysis, the final model for this study is decided to be:

lm.fit5=lm(imdb_score~poly(num_voted_users,2)+poly(num_critic_for_reviews,2)+poly(num_user_for_reviews,2)+poly(duration,2)+facenumber_in_poster+gross+poly(budget,2)+title_year+factor(genres)+duration*num_voted_users+num_voted_users*num_user_for_reviews,data=movie_train).

Results

The diagnostics for our model:





The first plot shows an obvious pattern, points are more clustering towards left, together with the QQ-plot, implying normality assumption been violated. However, from the first plot, majority of points are around zero, indicating equal variance should pass. This prediction has been verified by formal assumption test: residuals failed Shapiro test, and reject null hypothesis of none-constant variance test, meaning residuals have equal variance.

The R output for lm.fit5 shows that all terms are significant except for the quadratic term for number of user for reviews. But the result from another lack of fit indicates it's not necessary to exclude the term, so it's been kept. To interpret the result, I will take just the number voted users as an example:

```
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    5.112e+01  3.781e+00  13.520  < 2e-16 ***
poly(num_voted_users, 2)1      3.996e+01  4.995e+00   8.000 1.84e-15 ***
poly(num_voted_users, 2)2     -1.712e+01  2.142e+00  -7.994 1.93e-15 ***
```

The model can be written as:

IMDb score = 51.12 + 39.96*number voted users -17.12*(number voted users) ^{2}. When there's 0 number of voted users, the slope is such that the IMDb score will increase by 39.96 for an additional number of voted users if the slope remains unchanged. However, the quadratic term indicates that actually for additional number of voted users, the IMDb score will decrease by 17.12, making the fitted plot a concave shape.

From our result, we found some interesting facts. The more the number of voted users, the higher the score will be. The more people are voting meaning that the movie is capturing attention. By common sense we know that this is true. What is more surprising is that if more critics are reviewing the movies, the higher the score. The critics can influence whether readers decide to see a film. The result indicates that regardless of the content of critics' review, as long as more critics left reviews, people generally consider the movie as good movie. Another observation is that when duration is 0, IMDb score will increase by 13.76 for one unit increase in movie duration. But the slope will keep changing and thus for additional unit increase in duration, IMDB score will decrease by 2.5. This means that if movie duration pass certain length, people might be less likely to like it. This is true from personal experience. If a movie is too long, it's harder to keep concentrating or might end up not finishing the movie, both might yield a lower score. In addition, putting too many faces into one poster does not seem to be a good idea for advertising. The more the faces, the lower the score will be. The interaction between duration and number voted users is negatively correlated with movie score, which make sense that as we illustrated previously, for longer movies, people could be end up not finishing it and thus less number of voted users, leading a lower score.

In the end, we predict the IMDb score for observations in the test data set, and calculated the absolute mean square error to be 0.5174. This implies that, on average, the difference

between our model's predictions and the true quality score was about 0.5174 on a quality scale from zero to 10. Our model is doing fairly well.

Discussions

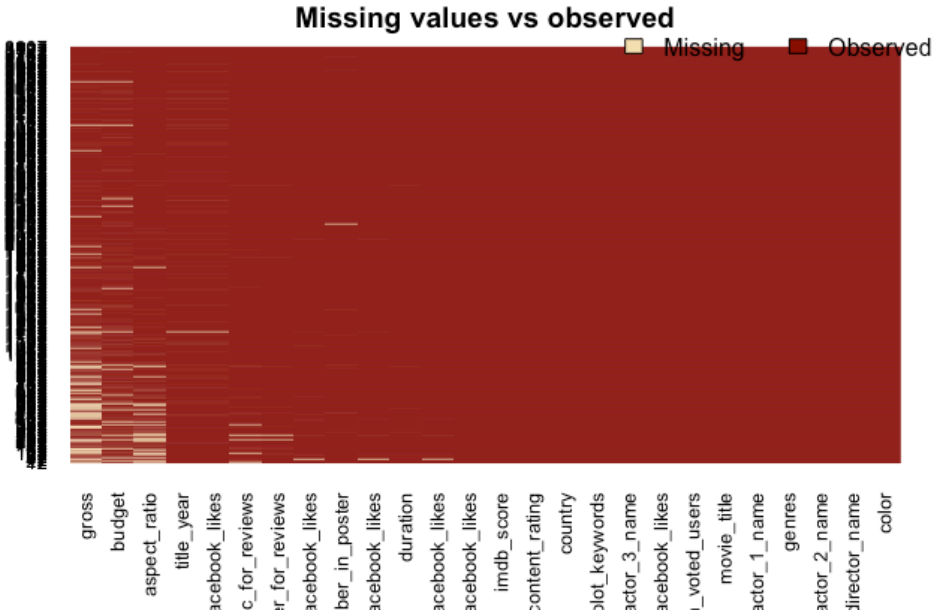
The model fails the normality assumption. In the residual plots, there are many points on the left, and sparse on the right. Maybe a Z-score standardization to the data set before modeling would help. Or we can do a Boxcox transformation on the dependent variable after fitting a model.

There are also some possibilities for improvement in model selection procedure. Actor 1 and actor 3 facebook likes also having significant correlation with IMDb score from the pairwise correlation test results. They should also be included in the model. On the other hand, genre does not have significant correlation, not necessarily to be kept in the model. In addition, instead of creating correlation plot and pairwise correlation test that can only used for numerical variables, creating correlation plot with p-value that works for both numerical and categorical variables is more effective way. More significant interaction terms could be added to be model to be studied. This might help to get a higher adjusted R^2 and get more insights. What's more, there is big mistake in the report. I forgot to use set.seed function before randomly selecting training and test data sets. Which results in slightly different each time of processing. Therefore, for report reproduction, set.seed function should be added before splitting.

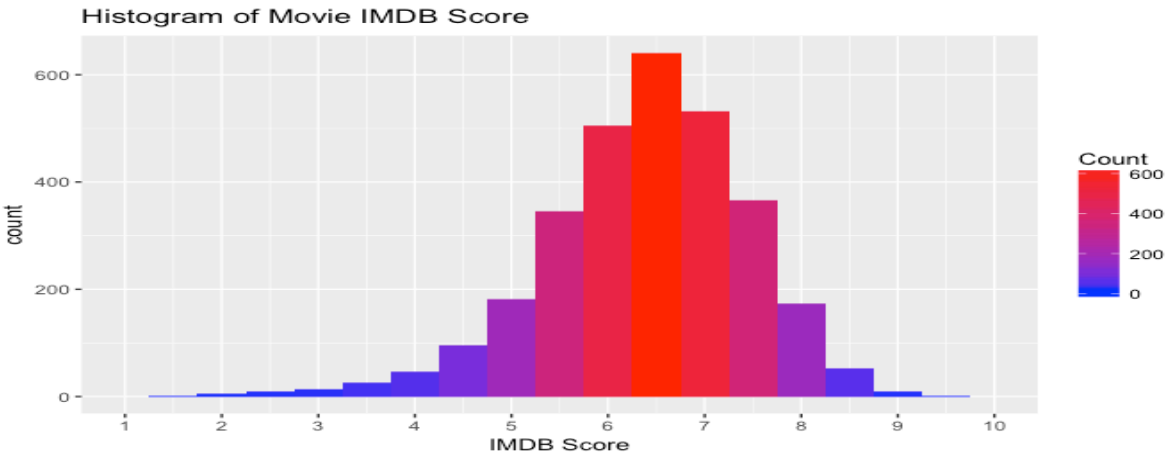
Additionally, we could also use Random Forest, a very powerful machine learning algorithms for making predictions, in the later study to gain more insights. Comparing the results from regression analysis and the output from random forest, we should be able to know which is more accurate in making predictions.

APPENDIX

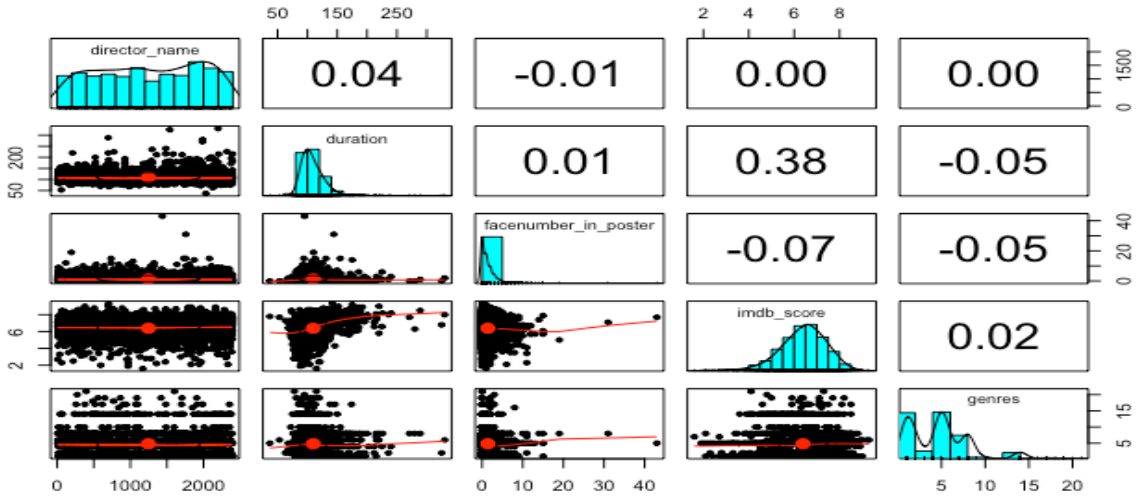
```
## Data exploration
m<- read.csv('movie_metadata copy.csv')
which(colnames(m)=='genres')
which(colnames(m)=='X.8')
m<-m[,-c(11:19)]
movie.usa<-m[which(m['country']=='USA'),]
movie.df= data.frame(movie.usa)
mm<-movie.df[, -which(names(movie.df)=='movie_imdb_link')]
# Check for missing value:
library(Amelia)
missmap(mm, main = "Missing values vs observed")
sapply(mm,function(x) sum(is.na(x))) # number of missing values for each variable
```



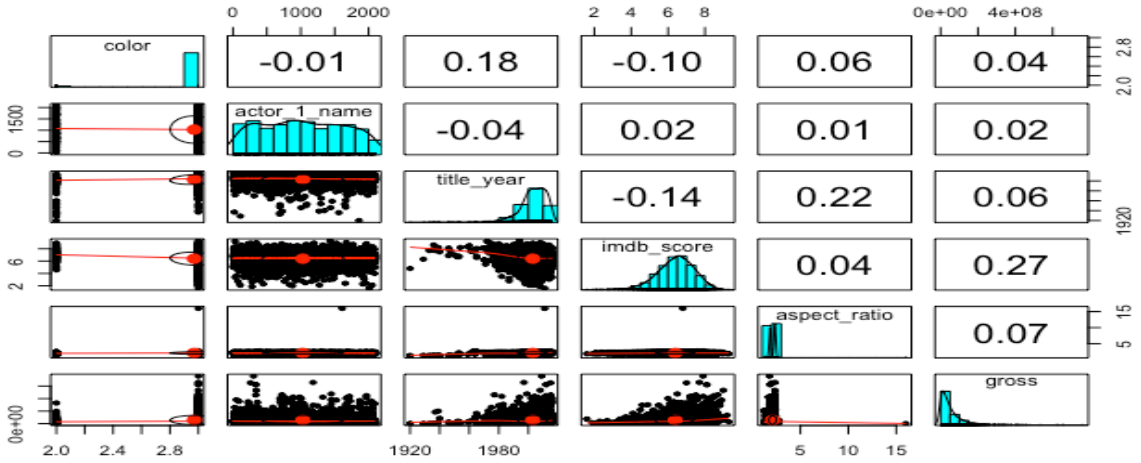

```
# omit missing values:
movie<-na.omit(mm)
sapply(movie,function(x) sum(is.na(x))) # double check for missing values
## libraries needed
library(psych)
library(car)
library(RColorBrewer)
library(corrplot)
library(ggplot2)
## histogram for IMDb Score:
max(movie$imdb_score) # 9.4
ggplot(movie, aes(x = imdb_score)) +
        geom_histogram(aes(fill = ..count..), binwidth =0.5) +
        scale_x_continuous(name = "IMDB Score",
                           breaks = seq(0,10),
                           limits=c(1, 10)) +
        ggtitle("Histogram of Movie IMDB Score") +
        scale_fill_gradient("Count", low = "blue", high = "red")
```



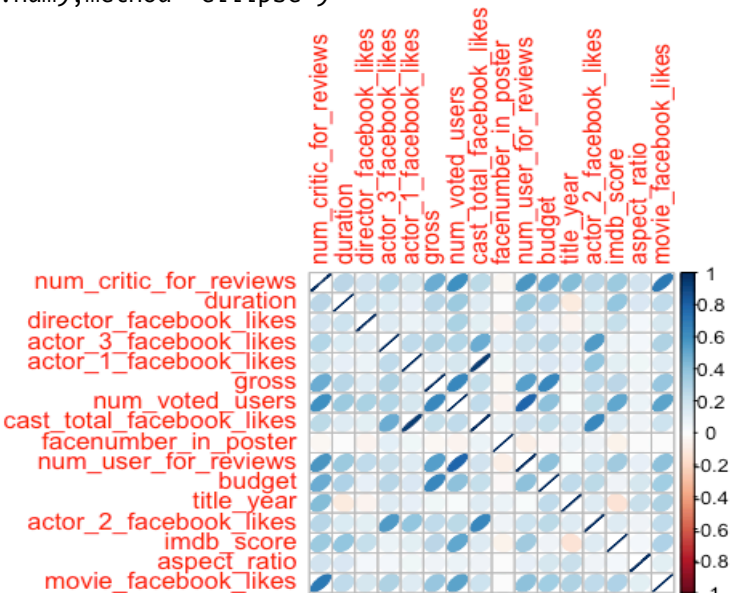

```
## Explore Correlation:
pairs.panels(movie[c('director_name','duration','facenumber_in_poster','imdb_score','genres')])
```



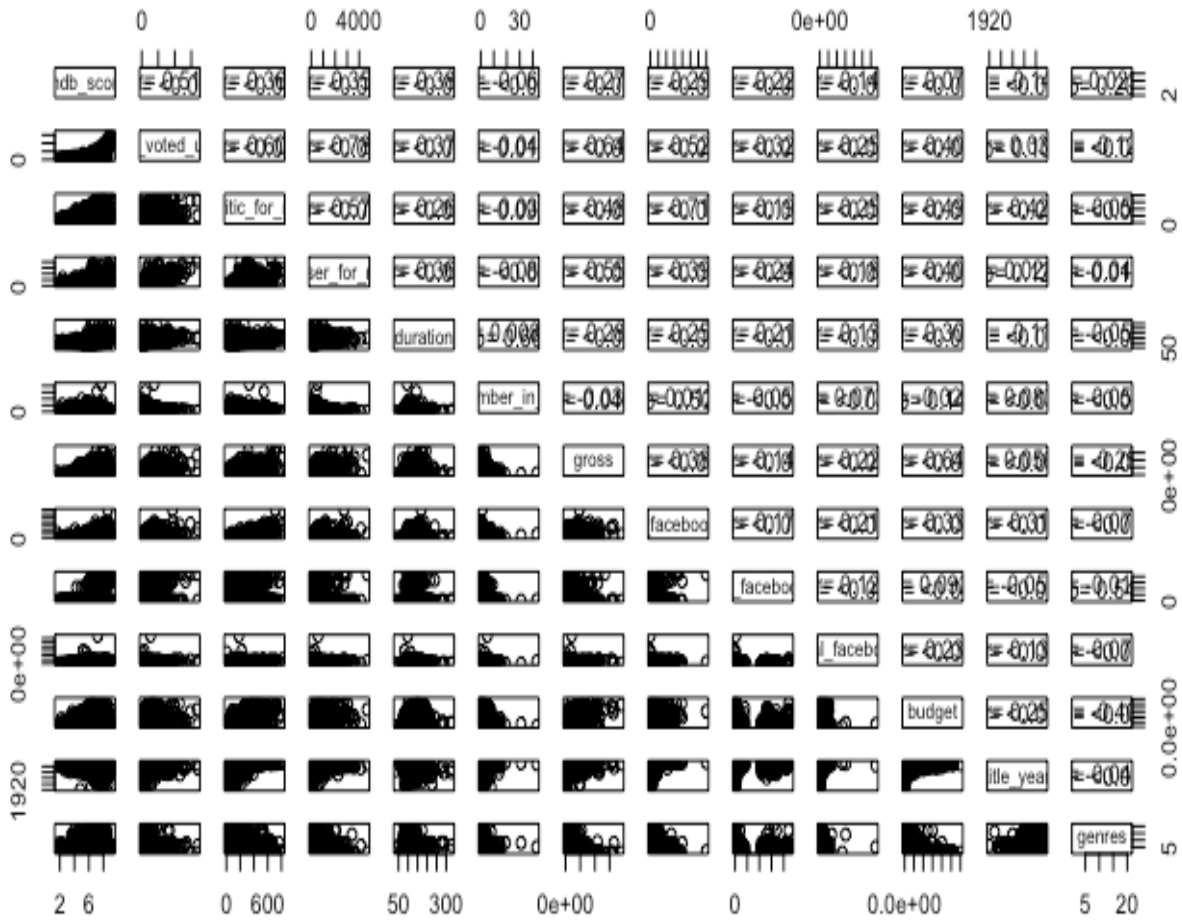

```
pairs.panels(movie[c('color','actor_1_name','title_year','imdb_score','aspect_ratio','gross')])
```



```
# Corplot for all numerical variables:
nums<- sapply(movie,is.numeric) # select numeric columns
movie.num<- movie[,nums]
corrplot(cor(movie.num),method='ellipse')
```




```
# corplot with p-value:
panel.cor <- function(x, y, digits = 2, cex.cor, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  # correlation coefficient
  r <- cor(x, y)
  txt <- format(c(r, 0.123456789), digits = digits)[1]
  txt <- paste("r= ", txt, sep = "")
  text(0.5, 0.6, txt)

 # p-value calculation
 p <- cor.test(x, y)$p.value
  txt2 <- format(c(p, 0.123456789), digits = digits)[1]
  txt2 <- paste("p= ", txt2, sep = "")
  if(p<0.01) txt2 <- paste("p= ", "<0.01", sep = "")
  text(0.5, 0.4, txt2)
}
pairs(movie.sig, upper.panel = panel.cor)
```



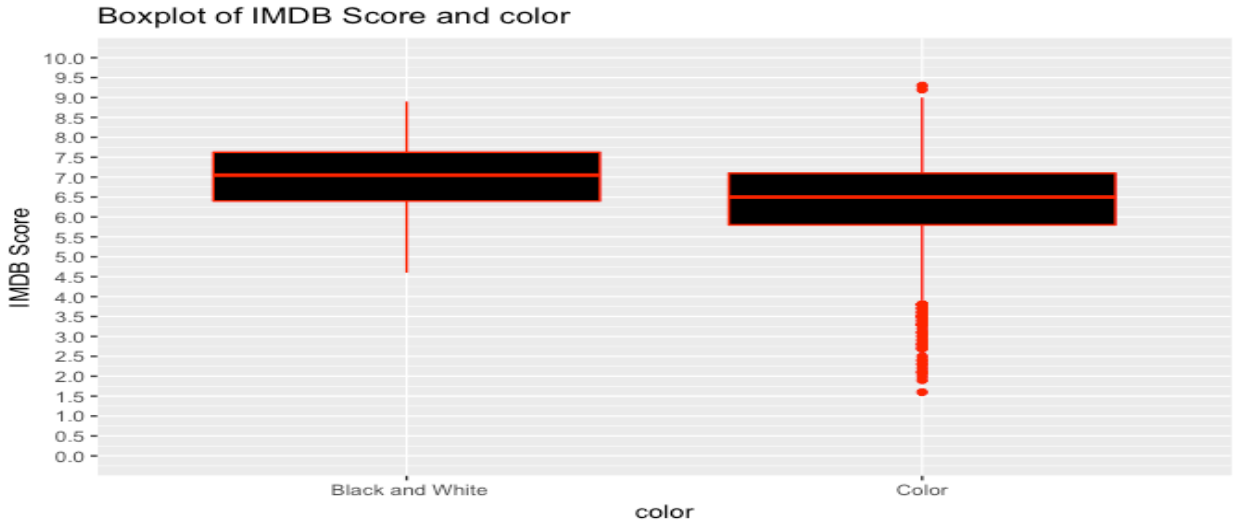

```
# cor.test:
corr.test(movie.num,y=NULL,use='pairwise',method='pearson',adjust='holm',alpha=0.05)  #
x must be numeric:
Call:corr.test(x = movie.num, y = NULL, use = "pairwise", method = "pearson",
                                                        adjust = "holm", alpha = 0.05)
Correlation matrix
                            num_critic_for_reviews duration director_facebook_likes
num_critic_for_reviews                        1.00     0.26                     0.19
duration                                      0.26     1.00                     0.21
director_facebook_likes                       0.19     0.21                     1.00
actor_3_facebook_likes                        0.28     0.14                     0.12
actor_1_facebook_likes                        0.17     0.09                     0.09
gross                                         0.48     0.28                     0.14
num_voted_users                               0.60     0.37                     0.32
cast_total_facebook_likes                     0.25     0.13                     0.12
facenumber_in_poster                         -0.03     0.01                    -0.05
num_user_for_reviews                          0.57     0.36                     0.24
budget                                        0.49     0.30                     0.09
title_year                                    0.42    -0.11                    -0.06
```

```
actor_2_facebook_likes                      0.28     0.15                     0.12
imdb_score                                  0.36     0.38                     0.22
aspect_ratio                                0.18     0.16                     0.05
movie_facebook_likes                        0.71     0.25                     0.17
                           actor_3_facebook_likes actor_1_facebook_likes gross
num_critic_for_reviews                       0.28                     0.17  0.48
duration                                     0.14                     0.09  0.28
director_facebook_likes                      0.12                     0.09  0.14
actor_3_facebook_likes                       1.00                     0.25  0.30
actor_1_facebook_likes                       0.25                     1.00  0.13
gross                                        0.30                     0.13  1.00
num_voted_users                              0.28                     0.17  0.64
cast_total_facebook_likes                    0.48                     0.95  0.22
facenumber_in_poster                         0.10                     0.05 -0.04
num_user_for_reviews                         0.22                     0.12  0.55
budget                                       0.27                     0.15  0.64
title_year                                   0.13                     0.09  0.06
actor_2_facebook_likes                       0.55                     0.38  0.25
imdb_score                                   0.09                     0.12  0.27
aspect_ratio                                 0.05                     0.05  0.07
movie_facebook_likes                         0.31                     0.12  0.38
                          num_voted_users cast_total_facebook_likes facenumber_in_post
er
num_critic_for_reviews               0.60                     0.25              -0.03
duration                             0.37                     0.13               0.01
director_facebook_likes              0.32                     0.12              -0.05
actor_3_facebook_likes               0.28                     0.48               0.10
actor_1_facebook_likes               0.17                     0.95               0.05
gross                                0.64                     0.22              -0.04
num_voted_users                      1.00                     0.25              -0.04
cast_total_facebook_likes            0.25                     1.00               0.07
facenumber_in_poster                -0.04                     0.07               1.00
num_user_for_reviews                 0.78                     0.18              -0.09
budget                               0.40                     0.23              -0.03
title_year                           0.03                     0.13               0.08
actor_2_facebook_likes               0.25                     0.63               0.07
imdb_score                           0.51                     0.14              -0.07
aspect_ratio                         0.09                     0.07               0.01
movie_facebook_likes                 0.52                     0.21               0.01
                         num_user_for_reviews budget title_year actor_2_facebook_likes
num_critic_for_reviews                   0.57   0.49       0.42                   0.28
duration                                 0.36   0.30      -0.11                   0.15
director_facebook_likes                  0.24   0.09      -0.06                   0.12
actor_3_facebook_likes                   0.22   0.27       0.13                   0.55
actor_1_facebook_likes                   0.12   0.15       0.09                   0.38
gross                                    0.55   0.64       0.06                   0.25
num_voted_users                          0.78   0.40       0.03                   0.25
cast_total_facebook_likes                0.18   0.23       0.13                   0.63
facenumber_in_poster                    -0.09  -0.03       0.08                   0.07
num_user_for_reviews                     1.00   0.40       0.03                   0.20
budget                                   0.40   1.00       0.25                   0.25
title_year                               0.03   0.25       1.00                   0.13
actor_2_facebook_likes                   0.20   0.25       0.13                   1.00
```

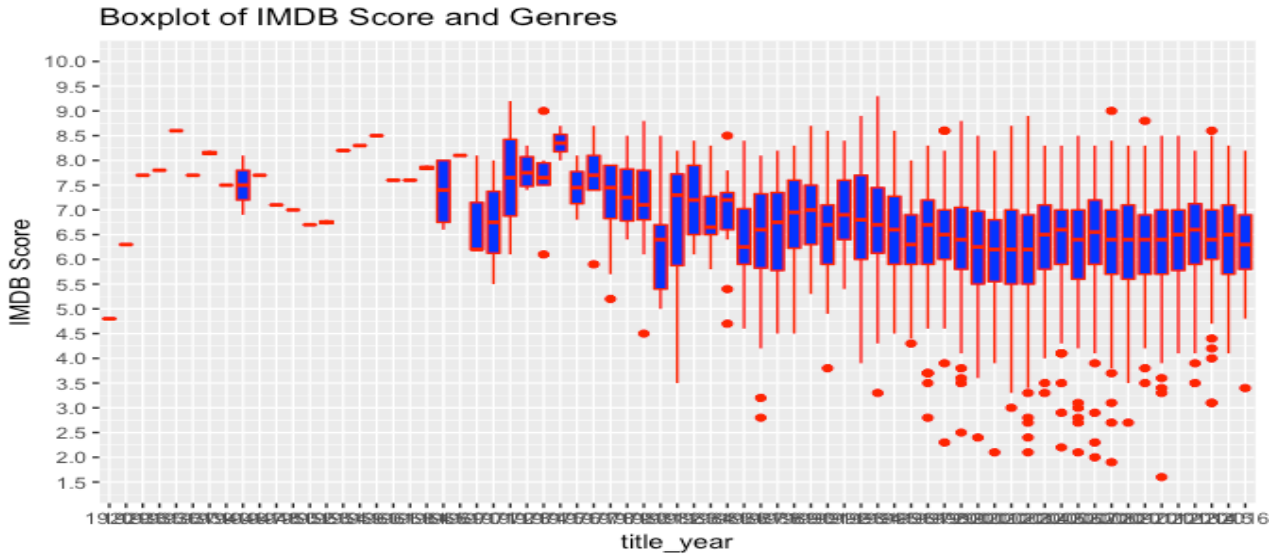
```
imdb_score                          0.35   0.07     -0.14                  0.13
aspect_ratio                        0.10   0.18       0.22                 0.07
movie_facebook_likes                0.39   0.33       0.31                 0.25
                                  imdb_score aspect_ratio movie_facebook_likes
num_critic_for_reviews              0.36            0.18                  0.71
duration                            0.38            0.16                  0.25
director_facebook_likes             0.22            0.05                  0.17
actor_3_facebook_likes              0.09            0.05                  0.31
actor_1_facebook_likes              0.12            0.05                  0.12
gross                               0.27            0.07                  0.38
num_voted_users                     0.51            0.09                  0.52
cast_total_facebook_likes           0.14            0.07                  0.21
facenumber_in_poster               -0.07            0.01                  0.01
num_user_for_reviews                0.35            0.10                  0.39
budget                              0.07            0.18                  0.33
title_year                         -0.14            0.22                  0.31
actor_2_facebook_likes              0.13            0.07                  0.25
imdb_score                          1.00            0.04                  0.29
aspect_ratio                        0.04            1.00                  0.11
movie_facebook_likes                0.29            0.11                  1.00
Sample Size
[1] 3005
Probability values (Entries above the diagonal are adjusted for multiple tests.)
                                  num_critic_for_reviews duration director_facebook_likes
num_critic_for_reviews                              0.00     0.00                     0.00
duration                                            0.00     0.00                     0.00
director_facebook_likes                             0.00     0.00                     0.00
actor_3_facebook_likes                              0.00     0.00                     0.00
actor_1_facebook_likes                              0.00     0.00                     0.00
gross                                               0.00     0.00                     0.00
num_voted_users                                     0.00     0.00                     0.00
cast_total_facebook_likes                           0.00     0.00                     0.00
facenumber_in_poster                                0.09     0.66                     0.00
num_user_for_reviews                                0.00     0.00                     0.00
budget                                              0.00     0.00                     0.00
title_year                                          0.00     0.00                     0.00
actor_2_facebook_likes                              0.00     0.00                     0.00
imdb_score                                          0.00     0.00                     0.00
aspect_ratio                                        0.00     0.00                     0.01
movie_facebook_likes                                0.00     0.00                     0.00
                                  actor_3_facebook_likes actor_1_facebook_likes gross
num_critic_for_reviews                              0.00                     0.00  0.00
duration                                            0.00                     0.00  0.00
director_facebook_likes                             0.00                     0.00  0.00
actor_3_facebook_likes                              0.00                     0.00  0.00
actor_1_facebook_likes                              0.00                     0.00  0.00
gross                                               0.00                     0.00  0.00
num_voted_users                                     0.00                     0.00  0.00
cast_total_facebook_likes                           0.00                     0.00  0.00
facenumber_in_poster                                0.00                     0.01  0.05
num_user_for_reviews                                0.00                     0.00  0.00
budget                                              0.00                     0.00  0.00
title_year                                          0.00                     0.00  0.00
```

```
actor_2_facebook_likes                        0.00                      0.00  0.00
imdb_score                                     0.00                     0.00  0.00
aspect_ratio                                   0.01                     0.00  0.00
movie_facebook_likes                          0.00                      0.00  0.00
                         num_voted_users cast_total_facebook_likes facenumber_in_post
er
num_critic_for_reviews              0.00                        0               0.65
duration                            0.00                        0               1.00
director_facebook_likes             0.00                        0               0.06
actor_3_facebook_likes              0.00                        0               0.00
actor_1_facebook_likes              0.00                        0               0.07
gross                               0.00                        0               0.37
num_voted_users                     0.00                        0               0.17
cast_total_facebook_likes           0.00                        0               0.00
facenumber_in_poster                0.02                        0               0.00
num_user_for_reviews                0.00                        0               0.00
budget                              0.00                        0               0.14
title_year                          0.10                        0               0.00
actor_2_facebook_likes              0.00                        0               0.00
imdb_score                          0.00                        0               0.00
aspect_ratio                        0.00                        0               0.55
movie_facebook_likes                0.00                        0               0.50
                        num_user_for_reviews budget title_year actor_2_facebook_likes
num_critic_for_reviews                  0.00   0.00       0.00                  0.00
duration                                0.00   0.00       0.00                  0.00
director_facebook_likes                 0.00   0.00       0.04                  0.00
actor_3_facebook_likes                  0.00   0.00       0.00                  0.00
actor_1_facebook_likes                  0.00   0.00       0.00                  0.00
gross                                   0.00   0.00       0.04                  0.00
num_voted_users                         0.00   0.00       0.65                  0.00
cast_total_facebook_likes               0.00   0.00       0.00                  0.00
facenumber_in_poster                    0.00   0.65       0.00                  0.01
num_user_for_reviews                    0.00   0.00       0.65                  0.00
budget                                  0.00   0.00       0.00                  0.00
title_year                              0.12   0.00       0.00                  0.00
actor_2_facebook_likes                  0.00   0.00       0.00                  0.00
imdb_score                              0.00   0.00       0.00                  0.00
aspect_ratio                            0.00   0.00       0.00                  0.00
movie_facebook_likes                    0.00   0.00       0.00                  0.00
                                    imdb_score aspect_ratio movie_facebook_likes
num_critic_for_reviews            0.00            0.00                         0
duration                             0.00            0.00                      0
director_facebook_likes           0.00            0.10                         0
actor_3_facebook_likes            0.00            0.07                         0
actor_1_facebook_likes            0.00            0.05                         0
gross                                0.00            0.00                      0
num_voted_users                     0.00            0.00                       0
cast_total_facebook_likes         0.00            0.00                         0
facenumber_in_poster                0.00            1.00                      1
num_user_for_reviews                0.00            0.00                      0
budget                               0.00            0.00                      0
title_year                          0.00            0.00                       0
actor_2_facebook_likes            0.00            0.00                         0
```

```
imdb_score                    0.00          0.34                    0
aspect_ratio                  0.04          0.00                    0
movie_facebook_likes          0.00          0.00                    0
 To see confidence intervals of the correlations, print with the short=FALSE option
# Boxplots for significant categorical predictors
fill <- "Black"
line <- "Red"
ggplot(movie, aes(x = color, y =imdb_score)) +geom_boxplot(fill = fill, colour = line)
+scale_y_continuous(name = "IMDB Score",breaks = seq(0, 10, 0.5),limits=c(0, 10))
+scale_x_discrete(name = "color") +ggtitle("Boxplot of IMDB Score and color")
```




```
# boxplot for title-year
library(ggplot2)
fill <- "Blue"
line <- "Red"
ggplot(movie, aes(x = as.factor(title_year), y =imdb_score)) +geom_boxplot(fill =
fill, colour = line) +scale_y_continuous(name = "IMDB Score",breaks = seq(1.5, 10,
0.5),limits=c(1.5, 10)) +scale_x_discrete(name = "title_year") +ggtitle("Boxplot of
IMDB Score and Genres")
```



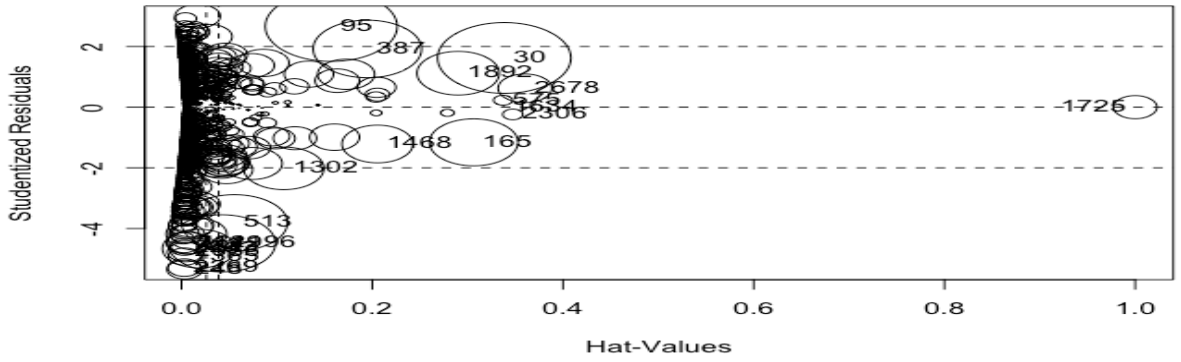
```
## Fitting regression model:
movie.sig<-
movie[,c('imdb_score','num_voted_users','num_critic_for_reviews','num_user_for_reviews
','duration','facenumber_in_poster','gross','movie_facebook_likes','director_facebook_
likes','cast_total_facebook_likes','budget','title_year','genres')]
# Step function:
null=lm(movie.sig$imdb_score~1) # set null model
summary(null)
Call:
lm(formula = movie.sig$imdb_score ~ 1)

Residuals:
            Min                  1Q      Median                  3Q             Max
-4.7873 -0.5873                0.1127                  0.7127                2.9127
Coefficients:                                Estimate Std. Error t value Pr(>|t|)
(Intercept)          6.3873                   0.0192          332.6          <2e-16 ***
---
Signif. codes:                             0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 1.053 on 3004 degrees of freedom

full3=lm(movie.sig$imdb_score
~movie.sig$num_voted_users+movie.sig$num_critic_for_reviews+movie.sig$num_user_for_rev
iews+movie.sig$duration+movie.sig$facenumber_in_poster+movie.sig$gross+movie.sig$movie
_facebook_likes+movie.sig$director_facebook_likes+movie.sig$cast_total_facebook_likes+
movie.sig$budget+movie.sig$title_year+factor(movie.sig$genres)+movie.sig$duration*movi
e.sig$num_voted_users+movie.sig$num_voted_users*movie.sig$num_user_for_reviews+movie.s
ig$gross*movie.sig$budget,data=movie.sig)
summary(full3)

Call:
lm(formula = movie.sig$imdb_score ~ movie.sig$num_voted_users +
                    movie.sig$num_critic_for_reviews + movie.sig$num_user_for_reviews +
                movie.sig$duration + movie.sig$facenumber_in_poster + movie.sig$gross +
                   movie.sig$movie_facebook_likes + movie.sig$director_facebook_likes +
                                movie.sig$cast_total_facebook_likes + movie.sig$budget +
                movie.sig$title_year + factor(movie.sig$genres) + movie.sig$duration *
     movie.sig$num_voted_users + movie.sig$num_voted_users * movie.sig$num_user_for_rev
iews +
                                  movie.sig$gross * movie.sig$budget, data = movie.sig)

Residuals:
            Min                  1Q      Median                  3Q             Max
-5.0519 -0.3700                0.0863                  0.4828                2.0996

Coefficients:
                                                      Estimate Std. Error t value
(Intercept)                                          4.748e+01  3.592e+00  13.218
movie.sig$num_voted_users                            7.890e-06  4.790e-07  16.472
movie.sig$num_critic_for_reviews                     2.427e-03  2.275e-04  10.669
movie.sig$num_user_for_reviews                      -3.039e-04  6.998e-05  -4.343
movie.sig$duration                                   1.277e-02  9.200e-04  13.882
movie.sig$facenumber_in_poster                      -1.858e-02  6.806e-03  -2.730
```

```
movie.sig$gross                                          -1.469e-09  4.191e-10  -3.505
movie.sig$movie_facebook_likes                           -2.370e-06  9.659e-07  -2.454
movie.sig$director_facebook_likes                         3.969e-06  4.482e-06   0.885
movie.sig$cast_total_facebook_likes                       7.641e-07  7.181e-07   1.064
movie.sig$budget                                         -5.900e-09  5.917e-10  -9.971
movie.sig$title_year                                     -2.154e-02  1.790e-03 -12.032
factor(movie.sig$genres)Adventure                         3.308e-01  5.338e-02   6.196
factor(movie.sig$genres)Animation                         7.426e-01  1.319e-01   5.629
factor(movie.sig$genres)Biography                         6.551e-01  7.512e-02   8.720
factor(movie.sig$genres)Comedy                            1.515e-01  4.284e-02   3.537
factor(movie.sig$genres)Crime                             4.496e-01  6.353e-02   7.077
factor(movie.sig$genres)Documentary                       8.960e-01  1.579e-01   5.676
factor(movie.sig$genres)Drama                             4.965e-01  4.835e-02  10.269
factor(movie.sig$genres)Family                            3.329e-01  4.432e-01   0.751
factor(movie.sig$genres)Fantasy                          -1.544e-01  1.419e-01  -1.089
factor(movie.sig$genres)Horror                           -3.577e-01  7.638e-02  -4.683
factor(movie.sig$genres)Musical                          -2.616e-01  5.459e-01  -0.479
factor(movie.sig$genres)Mystery                           1.263e-01  1.939e-01   0.652
factor(movie.sig$genres)Romance                           5.476e-01  5.392e-01   1.016
factor(movie.sig$genres)Sci-Fi                            1.673e-01  2.900e-01   0.577
factor(movie.sig$genres)Thriller                         -4.858e-01  7.627e-01  -0.637
factor(movie.sig$genres)Western                          -1.277e-01  5.408e-01  -0.236
movie.sig$num_voted_users:movie.sig$duration             -3.052e-08  3.447e-09  -8.852
movie.sig$num_voted_users:movie.sig$num_user_for_reviews -3.752e-10  9.851e-11  -3.809
movie.sig$gross:movie.sig$budget                          1.411e-17  2.887e-18   4.886
                                                                        Pr(>|t|)
(Intercept)                                                              < 2e-16 ***
movie.sig$num_voted_users                                                < 2e-16 ***
movie.sig$num_critic_for_reviews                                         < 2e-16 ***
movie.sig$num_user_for_reviews                                          1.46e-05 ***
movie.sig$duration                                                       < 2e-16 ***
movie.sig$facenumber_in_poster                                          0.006371 **
movie.sig$gross                                                         0.000463 ***
movie.sig$movie_facebook_likes                                           0.014175 *
movie.sig$director_facebook_likes                                       0.376035
movie.sig$cast_total_facebook_likes                                     0.287447
movie.sig$budget                                                         < 2e-16 ***
movie.sig$title_year                                                     < 2e-16 ***
factor(movie.sig$genres)Adventure                                       6.60e-10 ***
factor(movie.sig$genres)Animation                                       1.98e-08 ***
factor(movie.sig$genres)Biography                                        < 2e-16 ***
factor(movie.sig$genres)Comedy                                          0.000411 ***
factor(movie.sig$genres)Crime                                           1.83e-12 ***
factor(movie.sig$genres)Documentary                                     1.51e-08 ***
factor(movie.sig$genres)Drama                                            < 2e-16 ***
factor(movie.sig$genres)Family                                          0.452648
factor(movie.sig$genres)Fantasy                                         0.276414
factor(movie.sig$genres)Horror                                           2.95e-06 ***
factor(movie.sig$genres)Musical                                         0.631791
factor(movie.sig$genres)Mystery                                         0.514773
factor(movie.sig$genres)Romance                                         0.309947
factor(movie.sig$genres)Sci-Fi                                          0.563982
factor(movie.sig$genres)Thriller                                        0.524230
```

```
factor(movie.sig$genres)Western                                          0.813336
movie.sig$num_voted_users:movie.sig$duration                              < 2e-16 ***
movie.sig$num_voted_users:movie.sig$num_user_for_reviews 0.000143 ***
movie.sig$gross:movie.sig$budget                                         1.08e-06 ***
---
Signif. codes:                          0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 0.7607 on 2974 degrees of freedom
Multiple R-squared: 0.483, Adjusted R-squared: 0.4778
F-statistic: 92.63 on 30 and 2974 DF, p-value: < 2.2e-16

```
step(null,scope=list(lower=null,upper=full3),direction='forward')
# Choosen AIC output
Step:  AIC=-1535.83  movie.sig$imdb_score ~ movie.sig$num_voted_users +
factor(movie.sig$genres) + movie.sig$title_year +
movie.sig$num_critic_for_reviews + movie.sig$budget + movie.sig$duration +
movie.sig$num_voted_users:movie.sig$duration                                Df
Sum of Sq    RSS     AIC + movie.sig$num_user_for_reviews  1   26.4426 1748.7 -
1578.9 + movie.sig$facenumber_in_poster  1    2.9576 1772.2 -1538.8 +
movie.sig$cast_total_facebook_likes  1    1.1823  1774.0  -1535.8  <none>
1775.1 -1535.8 + movie.sig$movie_facebook_likes  1    0.9446 1774.2 -1535.4 +
movie.sig$director_facebook_likes  1    0.3854 1774.8 -1534.5 + movie.sig$gross
1    0.0191 1775.1 -1533.9

# Split data into Train and Test
indx = sample(1:nrow(movie.sig), as.integer(0.9*nrow(movie.sig)))
indx # ramdomize rows, save 90% of data into index

movie_train = movie.sig[indx,]
movie_test = movie.sig[-indx,]

lm.fit3<-
lm(movie_train$imdb_score~poly(movie_train$num_voted_users,2)+poly(movie_train$num_cri
tic_for_reviews,2)+poly(movie_train$num_user_for_reviews,2)+poly(movie_train$duration,
2)+movie_train$facenumber_in_poster+movie_train$gross+poly(movie_train$budget,2)+movie
_train$title_year+factor(movie_train$genres)+movie_train$duration*movie_train$num_vote
d_users+movie_train$num_voted_users*movie_train$num_user_for_reviews)

# Check outliers for lm.fit3
library(car)
influencePlot(lm.fit3, id.n=10)
```

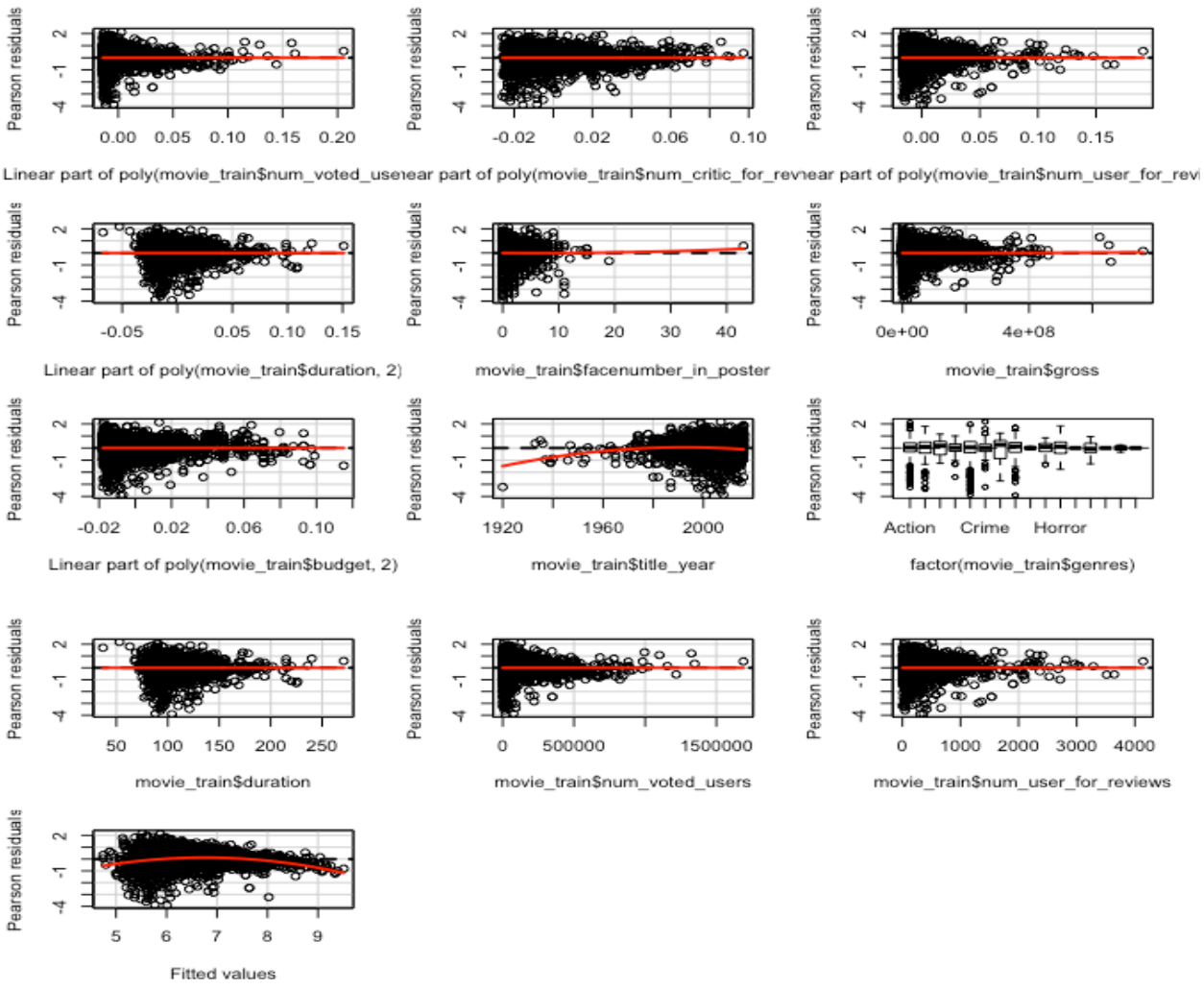


```
# Final model: lm.fit 5
# lm.fit5: model based on lm.fit3 removing 10 outliers.
movie_train<-movie_train[-c(1725,165,30,95,387,1892,2678,1468,634,57),]

lm.fit5<-
lm(movie_train$imdb_score~poly(movie_train$num_voted_users,2)+poly(movie_train$num_cri
tic_for_reviews,2)+poly(movie_train$num_user_for_reviews,2)+poly(movie_train$duration,
2)+movie_train$facenumber_in_poster+movie_train$gross+poly(movie_train$budget,2)+movie
_train$title_year+factor(movie_train$genres)+movie_train$duration*movie_train$num_vote
d_users+movie_train$num_voted_users*movie_train$num_user_for_reviews)
summary(lm.fit5)


Call:
lm(formula = movie_train$imdb_score ~ poly(movie_train$num_voted_users,
    2) + poly(movie_train$num_critic_for_reviews, 2) + poly(movie_train$num_user_for_r
eviews,
               2) + poly(movie_train$duration, 2) + movie_train$facenumber_in_poster +
           movie_train$gross + poly(movie_train$budget, 2) + movie_train$title_year +
    factor(movie_train$genres) + movie_train$duration * movie_train$num_voted_users +
                        movie_train$num_voted_users * movie_train$num_user_for_reviews)

Residuals:
             Min                1Q      Median                3Q             Max
-3.9076 -0.3552            0.0688                0.4625              2.1776


Coefficients: (3 not defined because of singularities)
                                                                    Estimate Std. Er
ror
(Intercept)                                                        5.111e+01  3.788e+00
poly(movie_train$num_voted_users, 2)1                              3.936e+01  5.016e+00
poly(movie_train$num_voted_users, 2)2                             -1.704e+01  2.181e+00
poly(movie_train$num_critic_for_reviews, 2)1                       1.277e+01  1.325e+00
poly(movie_train$num_critic_for_reviews, 2)2                      -7.141e+00  8.217e-01
poly(movie_train$num_user_for_reviews, 2)1                        -1.851e+01  2.393e+00
poly(movie_train$num_user_for_reviews, 2)2                         2.427e+00  1.578e+00
poly(movie_train$duration, 2)1                                     1.372e+01  1.122e+00
poly(movie_train$duration, 2)2                                    -2.483e+00  8.049e-01
movie_train$facenumber_in_poster                                  -2.261e-02  7.124e-03
movie_train$gross                                                 -7.074e-10  3.240e-10
poly(movie_train$budget, 2)1                                      -1.022e+01  1.167e+00
poly(movie_train$budget, 2)2                                       7.221e+00  8.067e-01
movie_train$title_year                                            -2.234e-02  1.895e-03
factor(movie_train$genres)Adventure                                3.559e-01  5.554e-02
factor(movie_train$genres)Animation                                7.284e-01  1.386e-01
factor(movie_train$genres)Biography                                6.334e-01  7.655e-02
factor(movie_train$genres)Comedy                                   1.408e-01  4.423e-02
factor(movie_train$genres)Crime                                    4.674e-01  6.446e-02
factor(movie_train$genres)Documentary                              1.312e+00  1.634e-01
factor(movie_train$genres)Drama                                    4.938e-01  4.976e-02
factor(movie_train$genres)Family                                   2.018e-01  4.273e-01
factor(movie_train$genres)Fantasy                                 -2.120e-01  1.421e-01
```

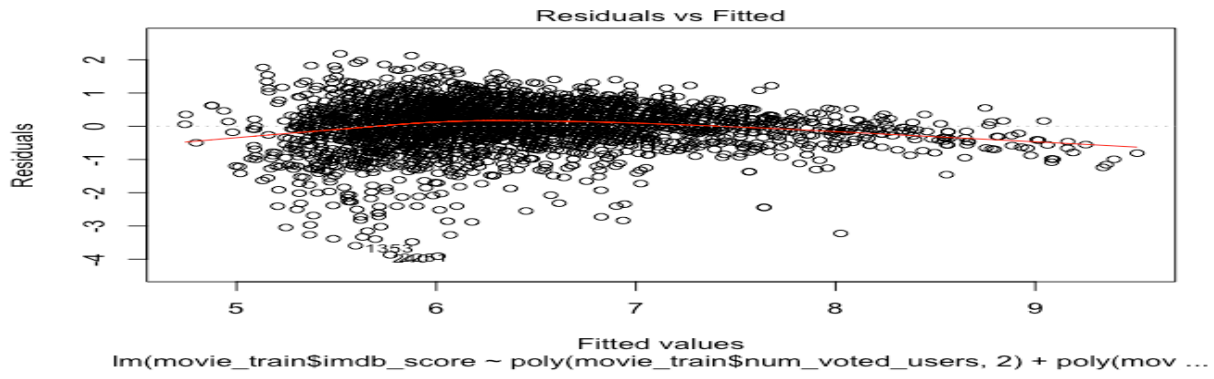
```
factor(movie_train$genres)Horror                                  -3.848e-01  8.040e-02
factor(movie_train$genres)Musical                                 -1.028e-01  7.377e-01
factor(movie_train$genres)Mystery                                  1.728e-01  2.000e-01
factor(movie_train$genres)Romance                                  8.931e-01  7.365e-01
factor(movie_train$genres)Sci-Fi                                   1.134e-01  3.701e-01
factor(movie_train$genres)Western                                  9.152e-01  7.361e-01
movie_train$duration                                                      NA         NA
movie_train$num_voted_users                                               NA         NA
movie_train$num_user_for_reviews                                          NA         NA
movie_train$duration:movie_train$num_voted_users                  -1.770e-08  4.096e-09
movie_train$num_voted_users:movie_train$num_user_for_reviews      1.246e-09   3.616e-10
                                                                  t value Pr(>|t|)
(Intercept)                                                        13.492  < 2e-16 ***
poly(movie_train$num_voted_users, 2)1                               7.846 6.21e-15 ***
poly(movie_train$num_voted_users, 2)2                              -7.813 7.98e-15 ***
poly(movie_train$num_critic_for_reviews, 2)1                        9.639  < 2e-16 ***
poly(movie_train$num_critic_for_reviews, 2)2                       -8.690  < 2e-16 ***
poly(movie_train$num_user_for_reviews, 2)1                         -7.737 1.44e-14 ***
poly(movie_train$num_user_for_reviews, 2)2                          1.537 0.124298
poly(movie_train$duration, 2)1                                     12.229  < 2e-16 ***
poly(movie_train$duration, 2)2                                     -3.085 0.002058 **
movie_train$facenumber_in_poster                                   -3.174 0.001523 **
movie_train$gross                                                  -2.184 0.029077 *
poly(movie_train$budget, 2)1                                       -8.755  < 2e-16 ***
poly(movie_train$budget, 2)2                                        8.950  < 2e-16 ***
movie_train$title_year                                            -11.790  < 2e-16 ***
factor(movie_train$genres)Adventure                                 6.408 1.74e-10 ***
factor(movie_train$genres)Animation                                 5.257 1.58e-07 ***
factor(movie_train$genres)Biography                                 8.273  < 2e-16 ***
factor(movie_train$genres)Comedy                                    3.184 0.001471 **
factor(movie_train$genres)Crime                                     7.251 5.40e-13 ***
factor(movie_train$genres)Documentary                               8.028 1.48e-15 ***
factor(movie_train$genres)Drama                                     9.925  < 2e-16 ***
factor(movie_train$genres)Family                                    0.472 0.636816
factor(movie_train$genres)Fantasy                                  -1.492 0.135833
factor(movie_train$genres)Horror                                   -4.787 1.79e-06 ***
factor(movie_train$genres)Musical                                  -0.139 0.889184
factor(movie_train$genres)Mystery                                   0.864 0.387742
factor(movie_train$genres)Romance                                   1.213 0.225378
factor(movie_train$genres)Sci-Fi                                    0.306 0.759288
factor(movie_train$genres)Western                                   1.243 0.213871
movie_train$duration                                                   NA        NA
movie_train$num_voted_users                                            NA        NA
movie_train$num_user_for_reviews                                       NA        NA
movie_train$duration:movie_train$num_voted_users                   -4.322 1.60e-05 ***
movie_train$num_voted_users:movie_train$num_user_for_reviews        3.447 0.000576 ***
---
Signif. codes:                          0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7344 on 2634 degrees of freedom
Multiple R-squared:                 0.5167, Adjusted R-squared:                 0.5112
F-statistic: 93.85 on 30 and 2634 DF,  p-value: < 2.2e-16
```

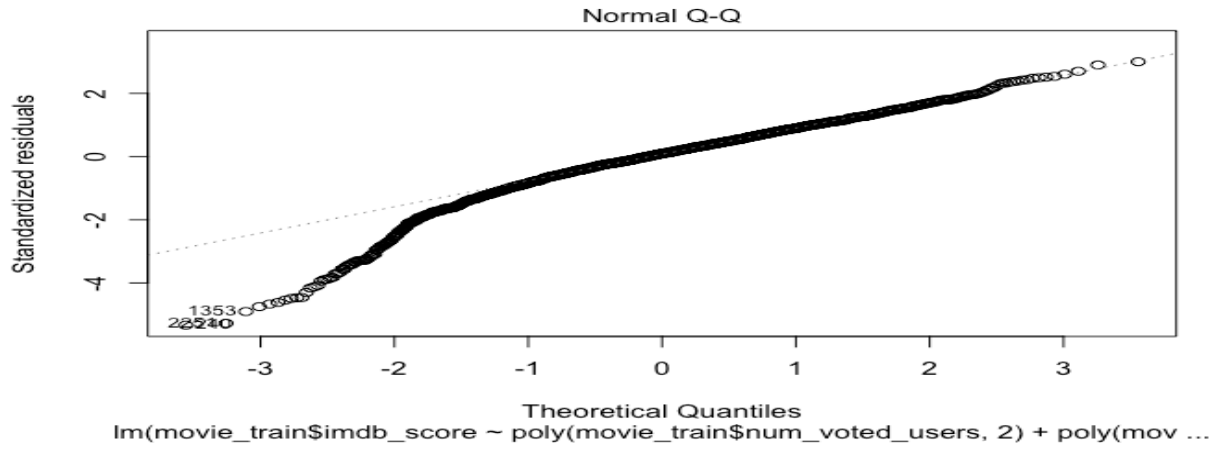
```
# compare coefficients before and after removing outliers:
compareCoefs(lm.fit3, lm.fit5)
all:
1: lm(formula = movie_train$imdb_score ~ poly(movie_train$num_voted_users, 2) +
   poly(movie_train$num_critic_for_reviews, 2) + poly(movie_train$num_user_for_reviews,

               2) + poly(movie_train$duration, 2) + movie_train$facenumber_in_poster +
           movie_train$gross + poly(movie_train$budget, 2) + movie_train$title_year +
     factor(movie_train$genres) + movie_train$duration * movie_train$num_voted_users +
                         movie_train$num_voted_users * movie_train$num_user_for_reviews)
2: lm(formula = movie_train$imdb_score ~ poly(movie_train$num_voted_users, 2) +
   poly(movie_train$num_critic_for_reviews, 2) + poly(movie_train$num_user_for_reviews,

               2) + poly(movie_train$duration, 2) + movie_train$facenumber_in_poster +
           movie_train$gross + poly(movie_train$budget, 2) + movie_train$title_year +
     factor(movie_train$genres) + movie_train$duration * movie_train$num_voted_users +
                         movie_train$num_voted_users * movie_train$num_user_for_reviews)
                                                                    Est. 1      SE 1
(Intercept)                                                       5.12e+01  3.79e+00
poly(movie_train$num_voted_users, 2)1                             3.98e+01  5.00e+00
poly(movie_train$num_voted_users, 2)2                            -1.71e+01  2.14e+00
poly(movie_train$num_critic_for_reviews, 2)1                      1.28e+01  1.33e+00
poly(movie_train$num_critic_for_reviews, 2)2                     -7.11e+00  8.22e-01
poly(movie_train$num_user_for_reviews, 2)1                       -1.85e+01  2.36e+00
poly(movie_train$num_user_for_reviews, 2)2                        2.57e+00  1.56e+00
poly(movie_train$duration, 2)1                                    1.38e+01  1.12e+00
poly(movie_train$duration, 2)2                                   -2.50e+00  8.04e-01
movie_train$facenumber_in_poster                                 -2.24e-02  7.12e-03
movie_train$gross                                                -6.93e-10  3.23e-10
poly(movie_train$budget, 2)1                                     -1.03e+01  1.17e+00
poly(movie_train$budget, 2)2                                      7.13e+00  8.07e-01
movie_train$title_year                                           -2.24e-02  1.89e-03
factor(movie_train$genres)Adventure                               3.60e-01  5.53e-02
factor(movie_train$genres)Animation                               7.27e-01  1.39e-01
factor(movie_train$genres)Biography                               6.33e-01  7.65e-02
factor(movie_train$genres)Comedy                                  1.39e-01  4.41e-02
factor(movie_train$genres)Crime                                   4.66e-01  6.44e-02
factor(movie_train$genres)Documentary                             1.31e+00  1.63e-01
factor(movie_train$genres)Drama                                   4.96e-01  4.97e-02
factor(movie_train$genres)Family                                  1.99e-01  4.27e-01
factor(movie_train$genres)Fantasy                                -2.12e-01  1.42e-01
factor(movie_train$genres)Horror                                 -3.85e-01  8.04e-02
factor(movie_train$genres)Musical                                -1.06e-01  7.38e-01
factor(movie_train$genres)Mystery                                 1.73e-01  2.00e-01
factor(movie_train$genres)Romance                                 8.92e-01  7.37e-01
factor(movie_train$genres)Sci-Fi                                  1.13e-01  3.70e-01
factor(movie_train$genres)Western                                 9.13e-01  7.36e-01
movie_train$duration
movie_train$num_voted_users
movie_train$num_user_for_reviews
movie_train$duration:movie_train$num_voted_users                 -1.79e-08  4.09e-09
movie_train$num_voted_users:movie_train$num_user_for_reviews      1.23e-09  3.53e-10
                                                                    Est. 2      SE 2
```

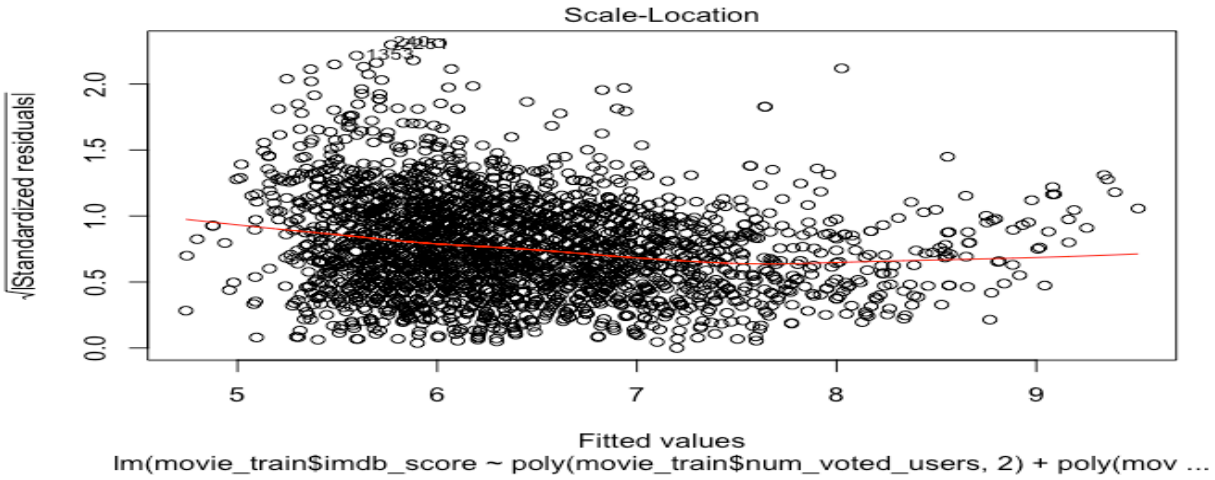
```
(Intercept)                                                      5.11e+01  3.79e+00
poly(movie_train$num_voted_users, 2)1                            3.94e+01  5.02e+00
poly(movie_train$num_voted_users, 2)2                           -1.70e+01  2.18e+00
poly(movie_train$num_critic_for_reviews, 2)1                     1.28e+01  1.33e+00
poly(movie_train$num_critic_for_reviews, 2)2                    -7.14e+00  8.22e-01
poly(movie_train$num_user_for_reviews, 2)1                      -1.85e+01  2.39e+00
poly(movie_train$num_user_for_reviews, 2)2                       2.43e+00  1.58e+00
poly(movie_train$duration, 2)1                                   1.37e+01  1.12e+00
poly(movie_train$duration, 2)2                                  -2.48e+00  8.05e-01
movie_train$facenumber_in_poster                                -2.26e-02  7.12e-03
movie_train$gross                                               -7.07e-10  3.24e-10
poly(movie_train$budget, 2)1                                    -1.02e+01  1.17e+00
poly(movie_train$budget, 2)2                                     7.22e+00  8.07e-01
movie_train$title_year                                          -2.23e-02  1.90e-03
factor(movie_train$genres)Adventure                              3.56e-01  5.55e-02
factor(movie_train$genres)Animation                              7.28e-01  1.39e-01
factor(movie_train$genres)Biography                              6.33e-01  7.66e-02
factor(movie_train$genres)Comedy                                 1.41e-01  4.42e-02
factor(movie_train$genres)Crime                                  4.67e-01  6.45e-02
factor(movie_train$genres)Documentary                            1.31e+00  1.63e-01
factor(movie_train$genres)Drama                                  4.94e-01  4.98e-02
factor(movie_train$genres)Family                                 2.02e-01  4.27e-01
factor(movie_train$genres)Fantasy                               -2.12e-01  1.42e-01
factor(movie_train$genres)Horror                                -3.85e-01  8.04e-02
factor(movie_train$genres)Musical                               -1.03e-01  7.38e-01
factor(movie_train$genres)Mystery                                1.73e-01  2.00e-01
factor(movie_train$genres)Romance                                8.93e-01  7.37e-01
factor(movie_train$genres)Sci-Fi                                 1.13e-01  3.70e-01
factor(movie_train$genres)Western                                9.15e-01  7.36e-01
movie_train$duration
movie_train$num_voted_users
movie_train$num_user_for_reviews
movie_train$duration:movie_train$num_voted_users                -1.77e-08  4.10e-09
movie_train$num_voted_users:movie_train$num_user_for_reviews  1.25e-09  3.62e-10

# Diagnostics for lm.fit5
library(car)
residualPlots(lm.fit5)
```

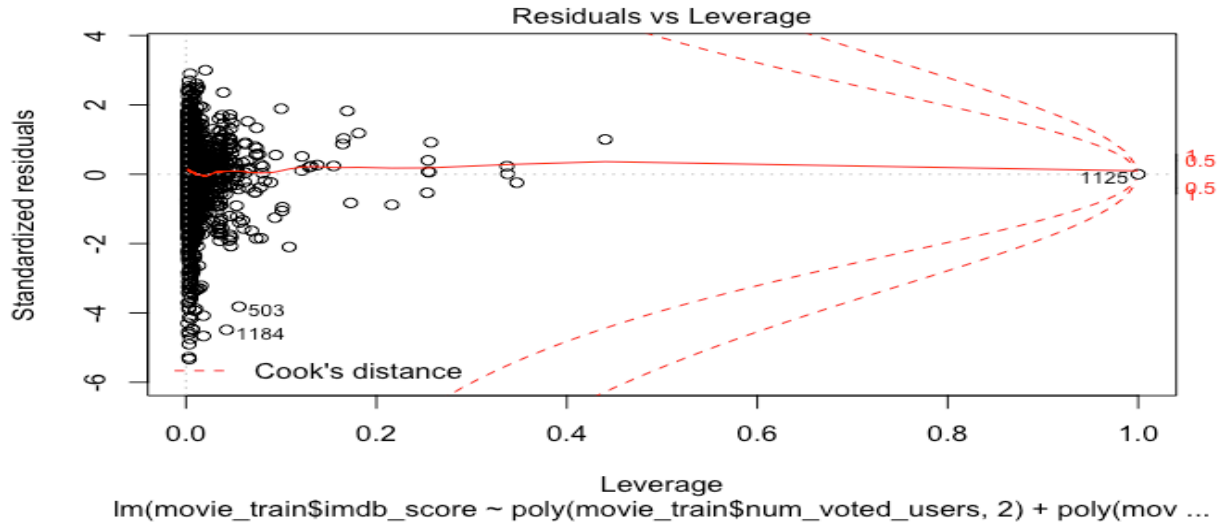



```
# another polt
plot(lm.fit5)
```










```
# formal assumption test
shapiro.test(lm.fit3$residuals)
```

```
shapiro.test(lm.fit5$residuals)

Shapiro-Wilk normality test
data: lm.fit3$residuals
W = 0.94761, p-value < 2.2e-16
 Shapiro-Wilk normality test
data: lm.fit5$residuals
W = 0.94742, p-value < 2.2e-16

ncvTest(lm.fit3)
ncvTest(lm.fit5)

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 172.5443                    Df = 1                        p = 2.058198e-39
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 172.5443    Df = 1     p = 2.058198e-39

## Make prediction
# lm.fit6 =lm.fit 5 using difference writing
lm.fit6<-
lm(imdb_score~poly(num_voted_users,2)+poly(num_critic_for_reviews,2)+poly(num_user_for
_reviews,2)+poly(duration,2)+facenumber_in_poster+gross+poly(budget,2)+title_year+genr
es+duration*num_voted_users+num_voted_users*num_user_for_reviews,data=data.frame(movie
_train))
summary(lm.fit6)

pr<-predict.lm(lm.fit6,newdata = data.frame(movie_test),interval = 'confidence')
pr

# Check prediction accuracy
MAE <- function(actual, predicted) {
mean(abs(actual - predicted))
}
MAE(pr, movie_test$imdb_score)
[1] 0.5174271
```

REFERENCES

IMDB 5000 movie dataset [Data file]. Available from

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