Research Project – Mathematical Statistics

Anil Egin, Hamza Abdella Kedir

January 2023 - Bocconi University

1- INTRODUCTION

The movie industry has been growing a lot recently, with an increasing number of films produced and developing technologies that allow producers to bring screenwriters' imaginations to life more easily such as CGI, IMAX, 4DX or D-box can be shown examples. One reason for this growth is the rising budgets of movies, which have allowed for more extravagant and elaborate productions over the years. In addition, the gross income of movies has also been on the rise, with some movies earning billions of dollars at the box office. This growing trend has been drawing attention to invest more in movies since some productions like Avengers End Game, Avatar, Titanic had showed potential earnings in recent years. This is the motive behind why we wanted to analyze the most effective factors that contribute to gross revenues. At the end we will try to answer: does genre play a role in determining movie's expected gross revenue?

Our analysis is limited to IMDB data, and the movies that have a public budget and gross income available.

2- DATASET, DATA CLEANING AND DATA INFORMATION

2.1- Data Source

We found our dataset through a website called Kaggle which was scraped and re-designed using IMDB datasets. This dataset contains 7668 different movies between 1980-2020 and information on 15 features of each movie. The descriptions of the columns are below:

name – the name of each movie

rating – information about the content

genre – thematic categories

year /released – came out date

score – a relative score based on 10

votes – IMDB vote count

director – name of the director

writer – the name of the writer

star – the name of the main character

country – the country movie was filmed

budget – costs related to production

gross – income in dollars

company – production firm's name

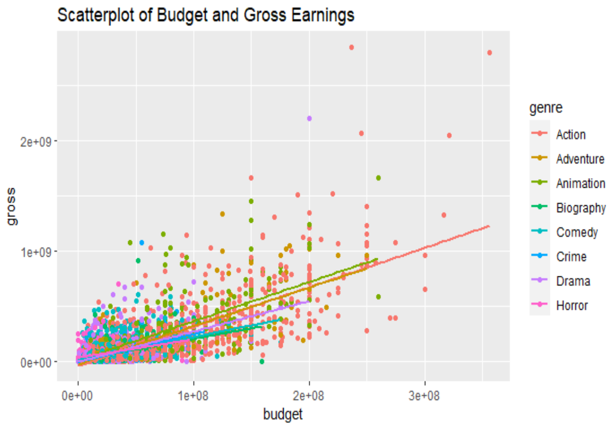
runtime – total duration in minutes

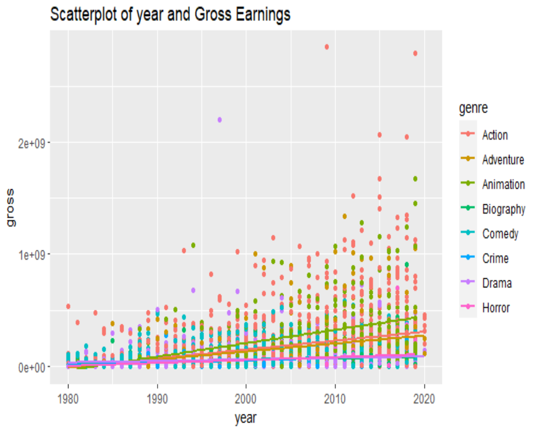
2.2- Data Cleaning

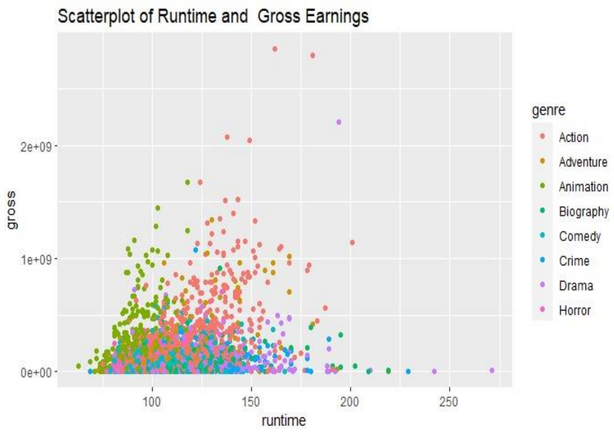
In order to make this data proper for regression and testing, we removed some of the genres from our dataset. As they have very little data which entailed wrong assumptions by pretending there is strong evidence for statistical significance in the multivariate linear regression test which we will apply. After we removed all the NULL rows, we ended up having 5435 different movies and then we excluded “Family”, “Fantasy”, “Mystery”, “Romance”, “Sci-Fi”, Thriller, “Western” genres. Finally, we have 5352 movies for analysis.

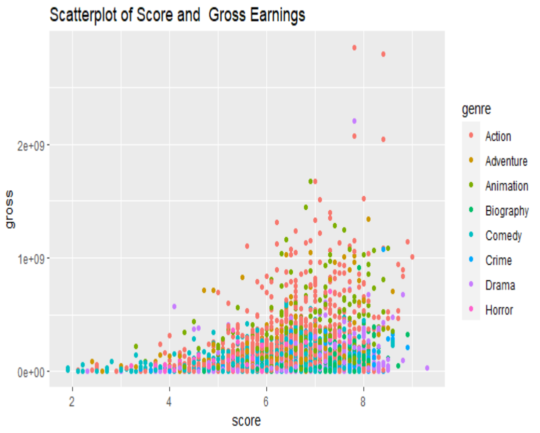
2.3- Data Visualization

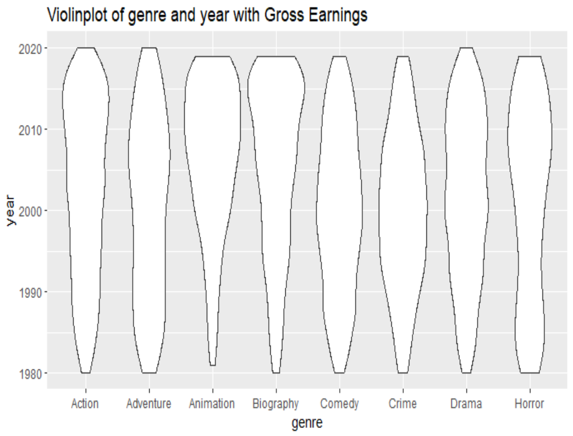
Before getting some background information about the data, we selected “year”, “score”, “votes”, “budget” and “runtime” factors to compare with the gross income of each movie and to have an idea of the relation between them.

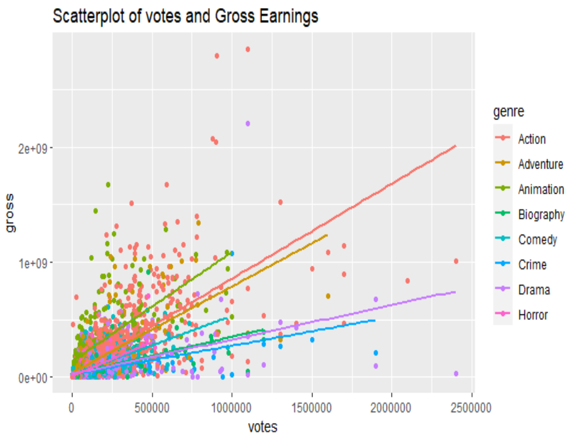










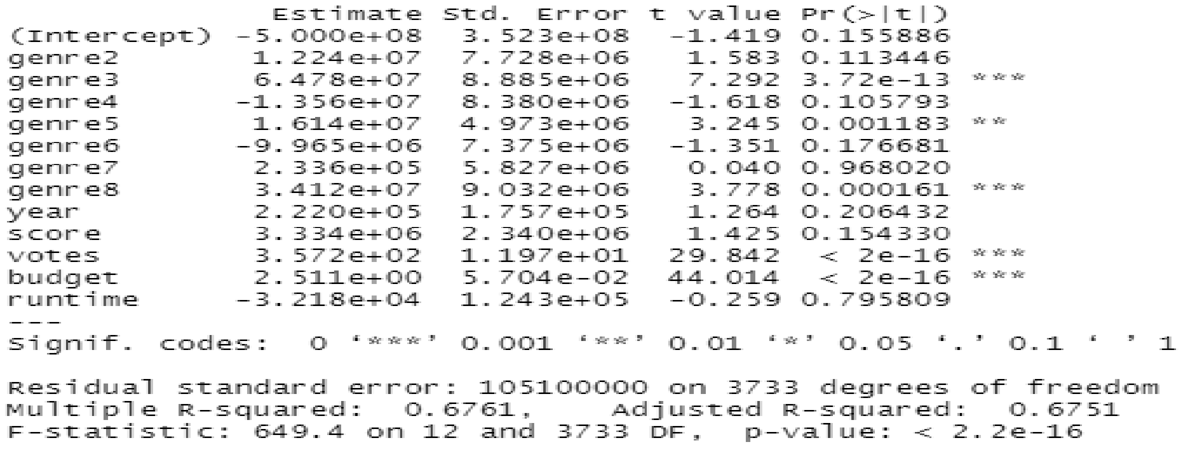


As we observe, it is easy to notice there is a linear relation between gross and budget as well as gross and votes. In terms of the genres, we can see some genres such as animation and action have shown some linear patterns in respective plots. Therefore, we can say our analysis show some promising progress as we might find statistically significant evidence that affects gross income. Additionally, there is an interesting fact that some genres have increasing trend of gross income over the years while other have exhibited different patterns. For instance, the crime genre reached its peak around 2000s and then started to lose its rise. We might touch on this specific relation if we able to find strong relation between them after the analysis.

3- Predictor Selecting and Multivariate Linear Regression

Now as we finished visual data observations, we start to find which are the best predictors for the response variable gross. Before starting forward selection, we convert the genre column into a categorical variable and in order to apply multivariate linear regression successfully we converted it into dummy variables with the first genre (action) as an intercept value. We thought there would be no need to change the intercept value as it is the most common genre and after applying to relevel to dummy variables we decided to stick with action and we separated the dataset into train and test partitions beforehand with setting %70 percent for train set to avoid overfitting during testing stage. To find the best predictors, we decided to perform forward selection, which is a method that involves adding a predictor at each step with the least p-values until it is not statistically significant evidence left, and to select the model with the highest accuracy. This method is useful for identifying the most important predictors.

Before starting we conducted K-fold cross-validation for forward regression which is a method used to evaluate the performance of a predictive model and improve its accuracy. It involves dividing the data into K folds or subsets, training the model on K-1 folds, and evaluating it on the remaining fold. This process is repeated K times, with a different fold used as the evaluation set each time. After performing cross-validation, and forward selection with max size 12 -there a reason why size being 12 is little chance of using every predictor being the optimal choice-it found the highest accuracy in the size 5 model with adding budget, vote, genre3(animation), genre5(Comedy) and genre8(horror), respectively since 5-sized predictor set has the max R-squared value. This suggests that these variables have a strong influence on movie gross income and are important to consider when predicting this outcome. To check this idea, we performed multivariate linear regression with all possible predictors and investigate their values.



```
               Estimate Std. Error t value Pr(>|t|)    
(Intercept) -5.000e+08  3.523e+08  -1.419 0.155886    
genre2       1.224e+07  7.728e+06   1.583 0.113446    
genre3       6.478e+07  8.885e+06   7.292 3.72e-13 ***
genre4      -1.356e+07  8.380e+06  -1.618 0.105793    
genre5       1.614e+07  4.973e+06   3.245 0.001183 ** 
genre6      -9.965e+06  7.375e+06  -1.351 0.176681    
genre7       2.336e+05  5.827e+06   0.040 0.968020    
genre8       3.412e+07  9.032e+06   3.778 0.000161 ***
year         2.220e+05  1.757e+05   1.264 0.206432    
score        3.334e+06  2.340e+06   1.425 0.154330    
votes        3.572e+02  1.197e+01  29.842  < 2e-16 ***
budget       2.511e+00  5.704e-02  44.014  < 2e-16 ***
runtime     -3.218e+04  1.243e+05  -0.259 0.795809    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 105100000 on 3733 degrees of freedom
Multiple R-squared:  0.6761,	Adjusted R-squared:  0.6751 
F-statistic: 649.4 on 12 and 3733 DF,  p-value: < 2.2e-16
```

After setting predictors to multivariate regression formula, we obtain **“gross = β0 + β1budget + β2votes + β3genre3+ β4genre5+ β5genre8 + ε “**. At first sight, we thought having a negative estimate was an error since we could not have negative gross income. Then we checked the correlation between predictors and found a high correlation coefficient and intercept mean for all the variables are 0 (which is an unrealistic scenario for the movie industry)**.** Continuing with further investigation, we deduce votes and budget of a movie have great p-values, namely very less than 0.05, indicating they state very significant evidence. Before statistical analysis, we thought the score would be a more appropriate choice than vote counts, however, after these significant results, we checked some data and found some movies which have very high scores with a few votes. This raised the opinion that some little-known movies were only scored by the people who loved them, and other people did not give an effort to criticize them. Also, it is well-known that IMDB scores do not reflect the popularity of the movie hence the gross income at all. The first movie that came to our mind was "The Shawshank Redemption". This movie is the highest-rated movie of all time but still, it had a terrible experience at the box office with 288M gross across the world, considerably low compared to less-rated movies. Then genre3, genre5, and genre8 follow budget and votes as greater predictors than others. We mentioned setting action as an intercept value causes every other genre to be compared to Action, one can comment that producing animation has a statistically significant relationship between the variables being analyzed and it produces higher gross income than action. With all these conclusions being said, adjusted R-squared ensures the countability of our analysis as showing %68 variance in the gross income.

As we found the best predictors, we finally performed multivariate linear regression with response variable gross; predictor variables genre, budget, votes. Now we want to check the normality and homoscedasticity of the residuals to see if the requirements of statistical significance are met. To check normality, we performed both Shapiro-Wilk and Kolmogorov-Smirnov tests and obtained great p-value scores. As p-values are less than 0.5 we reject the null hypothesis. Afterward, we made a plot of residuals against fitted values to see if residuals are normally distributed around 0 and most of the data is located close to 0 along with a straight line with some exceptions near the bottom right as we have a lot of data points it was expected. So far, the analysis we have carried out is satisfying for strong evidence.

To check again whether our model is a good fit for our data, we also conducted an ANOVA test, which is used when one has a categorical variables as predictor variables, in this case, it is very useful to assess our model. The results of the ANOVA test showed that our model has a high F-value and a satisfying p-value, indicating that the model is a good fit for the data. This suggests that the predictors included in the model jointly have a high impact on the response variable. Now our model has been confirmed by our findings with strong R-squared value of the ANOVA test which is quite same the one obtained by the model 0.66 indicating that, eventually, the budget of a movie has the strongest influence on its gross income, followed by its vote and genres. This suggests that genre should not be underestimated when predicting movie gross income, and votes also play a significant role which we assume is connected to popularity.

4- Limitations of the Study and Conclusion

Overall, the outcome of this project demonstrates that in the passing years of the movie industry, some parameters have always been more effective in the total revenue of the movie, the forward-step regression models developed in this project showed that certain variables, such as budget, votes, score, and some specific genres, had a significant impact on the gross revenue of movies. Also, that some variables did not have a remarkable effect on the gross revenue when considered over the dataset some variables like the year had an impact on specific genres, it is clear to see that on the scatterplot of year-to-gross revenue animation movies had a growth in revenue according to years.

The forward-step regression models that we fit the data were able to identify the variables that had the strongest influence on the response variable, and the ANOVA tests confirmed that there were significant differences between the levels of the predictor variables as discussed above. The data visualization techniques that we used, such as scatterplots, violin plots, and box plots, helped us to clearly see the results of the analysis and highlight patterns in the data that we found. For example, on the scatterplot of votes and gross earnings, we can clearly see the positive correlation between them, and our model already justifies it.

Based on the findings of our models and tests, there are several possible steps that could be taken. For example, we could conduct further investigations using additional models or hypotheses to better understand the underlying mechanisms driving the observed patterns and the relationships between the predictor variables and the response variable in more detail, or we could expand our data and work with more detailed variables to get better functioning parameters. Additionally, the results of this project could be used to inform movie production companies or make practical recommendations or even a business can be established just over analyzing timely data, since these models and testing are for predicting the success of a movie this could work like a new consulting branch. Overall, the results of this project provide valuable insights into the relationships between variables in the dataset and can inform future research and decision-making in the film industry.

5- Appendix

Bibliography

Dataset: https://www.kaggle.com/datasets/danielgrijalvas/movies?resource=download

https://datasets.imdbws.com

An Introduction to Mathematical Statistics, Bijma et al.

Wikipedia, the free encyclopedia

Genre Abbreviations

genre1 : Action

genre2 : Adventure

genre3 : Animation

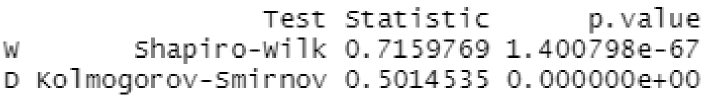
genre4 :Biography

genre5 :Comedy

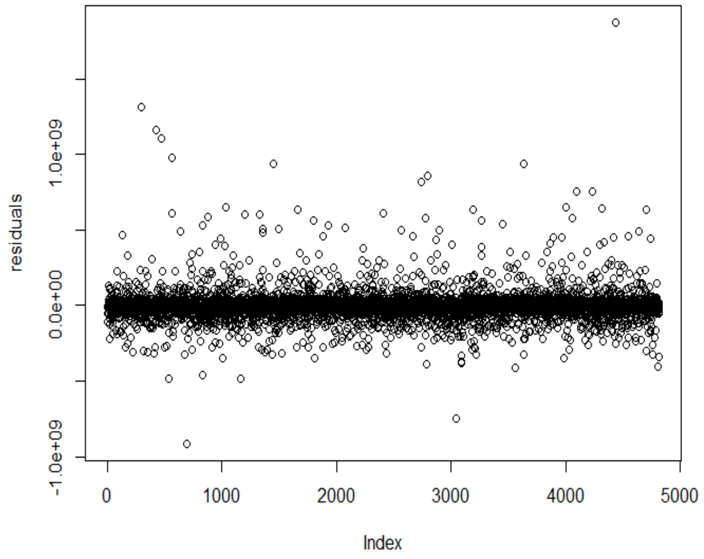
genre6 :Crime

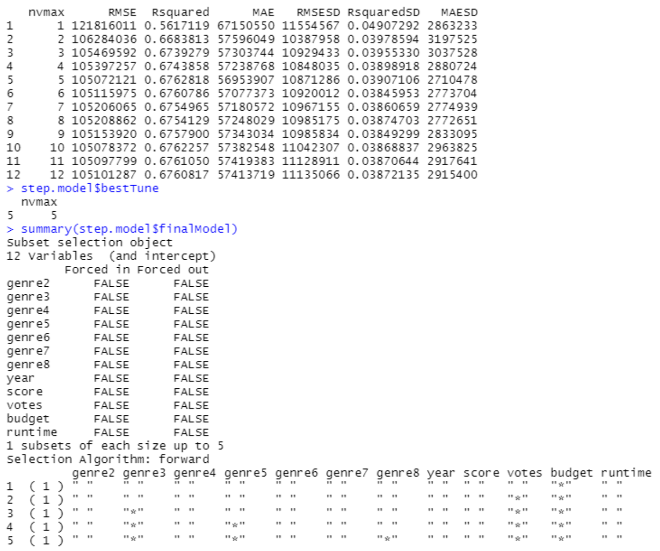
genre7 :Drama

genre8 : Horror

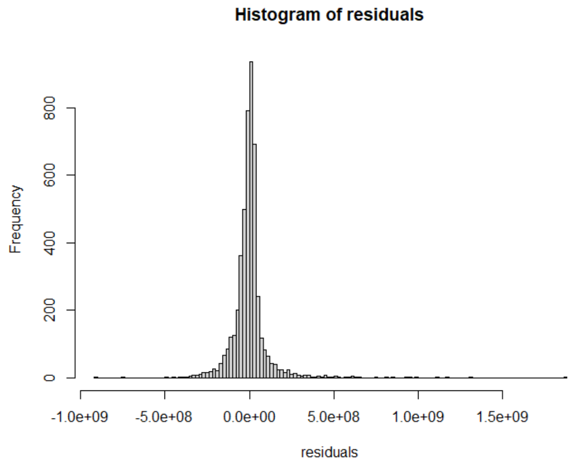


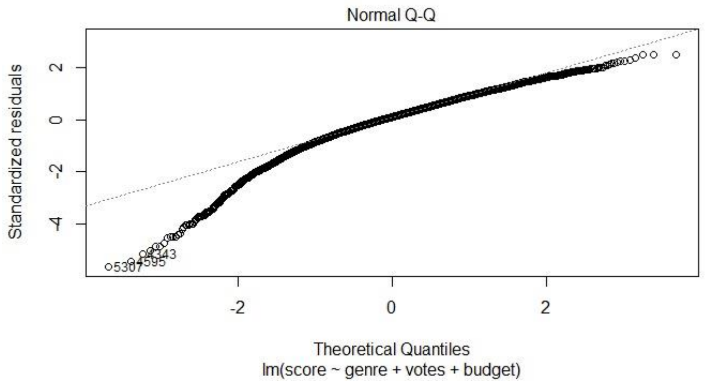
```
                  Test Statistic       p.value
W       Shapiro-Wilk 0.7159769 1.400798e-67
D Kolmogorov-Smirnov 0.5014535 0.000000e+00
```

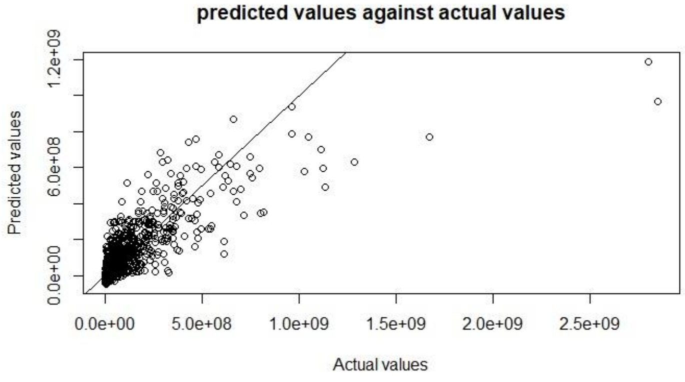


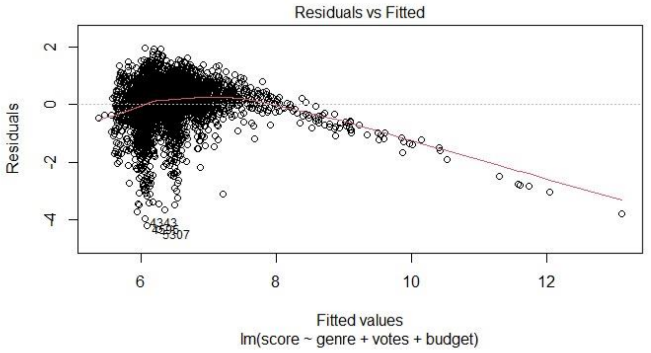


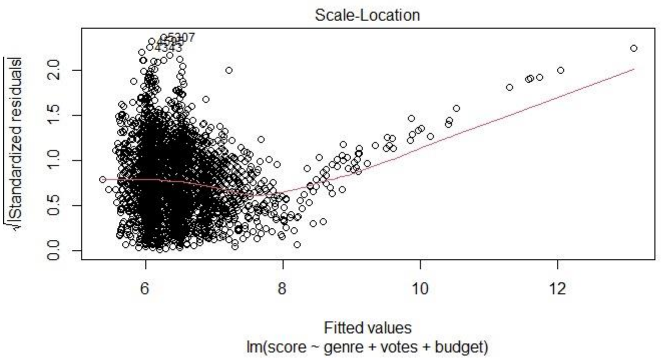
```
   nvmax      RMSE  Rsquared      MAE   RMSESD RsquaredSD   MAESD
1      1 121816011 0.5617119 67150550 11554567 0.04907292 2863233
2      2 106284036 0.6683813 57596049 10387958 0.03978594 3197525
3      3 105469592 0.6739279 57303744 10929433 0.03955330 3037528
4      4 105397257 0.6743858 57238768 10848035 0.03898918 2880724
5      5 105072121 0.6762818 56953907 10871286 0.03907106 2710478
6      6 105115975 0.6760786 57077373 10920012 0.03845953 2773704
7      7 105206065 0.6754965 57180572 10967155 0.03860659 2774939
8      8 105208862 0.6754129 57248029 10985175 0.03874703 2772651
9      9 105153920 0.6757900 57343034 10985834 0.03849299 2833095
10    10 105078372 0.6762257 57382548 11042307 0.03868837 2963825
11    11 105097799 0.6761050 57419383 11128911 0.03870644 2917641
12    12 105101287 0.6760817 57413719 11135066 0.03872135 2915400
> step.model$bestTune
  nvmax
5     5
> summary(step.model$finalModel)
Subset selection object
12 Variables  (and intercept)
        Forced in Forced out
genre2      FALSE      FALSE
genre3      FALSE      FALSE
genre4      FALSE      FALSE
genre5      FALSE      FALSE
genre6      FALSE      FALSE
genre7      FALSE      FALSE
genre8      FALSE      FALSE
year        FALSE      FALSE
score       FALSE      FALSE
votes       FALSE      FALSE
budget      FALSE      FALSE
runtime     FALSE      FALSE
1 subsets of each size up to 5
Selection Algorithm: forward
         genre2 genre3 genre4 genre5 genre6 genre7 genre8 year score votes budget runtime
1  ( 1 ) " "    " "    " "    " "    " "    " "    " "    " "  " "   " "   "*"    " "    
2  ( 1 ) " "    " "    " "    " "    " "    " "    " "    " "  " "   "*"   "*"    " "    
3  ( 1 ) " "    "*"    " "    " "    " "    " "    " "    " "  " "   "*"   "*"    " "    
4  ( 1 ) " "    "*"    " "    "*"    " "    " "    " "    " "  " "   "*"   "*"    " "    
5  ( 1 ) " "    "*"    " "    "*"    " "    " "    "*"    " "  " "   "*"   "*"    " "    
```


Histogram of residuals
Frequency
-1.0e+09
-5.0e+08
0.0e+00
5.0e+08
1.0e+09
1.5e+09
residuals

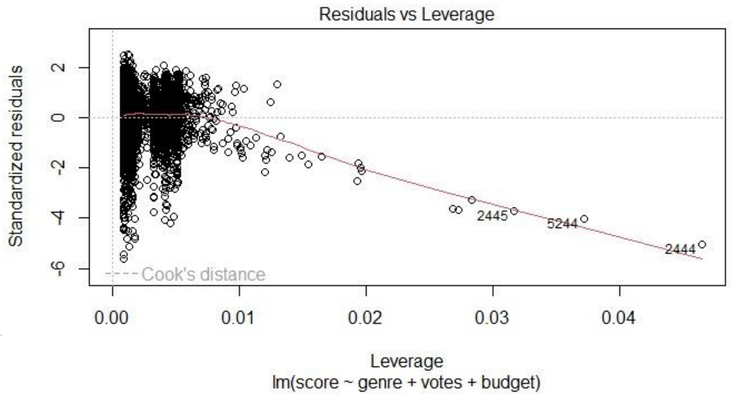

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(score ~ genre + votes + budget)


predicted values against actual values
Predicted values
0.0e+00
6.0e+08
1.2e+09
0.0e+00
5.0e+08
1.0e+09
1.5e+09
2.0e+09
2.5e+09
Actual values


Residuals vs Fitted
Residuals
Fitted values
lm(score ~ genre + votes + budget)



Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
5307
4343
6
8
10
12
Fitted values
lm(score ~ genre + votes + budget)


Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
lm(score ~ genre + votes + budget)