Analysis of IMDb Movies Data Set: Estimating U.S. Gross Revenue

TJ Pavaritpong, Vedaant Agarwal, Jacob Razdolsky, Jay Lim
University of Illinois Urbana-Champaign

19 October 2023

1 Introduction

This report aims to summarize our results of the case study, where we propose a regression model for estimating the total U.S. gross revenue of a movie using various predictors.

Our movies data set is from IMDB and consists of movies between 1986 and 2016. Each row in the data set represents a movie and the columns (variables) are: budget, company, country, director, genre, gross, name, rating, released, runtime, score, star, votes, writer, and year. Our model's response variable will be 'gross', which is the total U.S. gross revenue (in U.S. dollars).

Beyond serving as an exercise in analyzing data sets and constructing linear regression models, estimating the gross revenue of a movie is a relevant problem in the real world because of its implications for investors and professionals within the film industry. As we can see in the data set, movies cost a substantial amount of time and money to produce, so it is in an investor's or company's best interest to determine what factors contribute to a movie's success before committing large amounts of resources to producing one. Existing firms may also seek to analyze how they can improve their current movies in production. These are just a few among many examples of why this case study is a relevant problem to consider.

The report's sections include data cleaning and exploratory data analysis, model selection, diagnostics, model results (in terms of prediction and estimation), conclusion, and appendix.

2 Data Cleaning and Exploratory Data Analysis

We inspected the summary statistics of the variables, and found that the 1st quartile of budget was 0. This means that at least 25% of the observations of budget are 0, which must signify missing data because a movie with 0 budget does not make sense. Since it is expected that budget will be used as a predictor, these rows should be dropped. 2182 of the initial total 6820 rows were dropped by removing the 0 budget rows. The new summary statistics is shown in Table[1].

We found out the frequency of every level of each categorical variable. We noticed that most categorical variables had thousands of levels. The maximum frequency for each level is much lower than the total number of rows in the data set. This might lead to our model being inaccurate because there might not be enough observations in each level. Based on the excessive number of levels in categorical variables and the low number of observations in each level, we decided to group some of the levels together. We also dropped movie names and release date as predictors as they are mostly unique without any logical grouping schema; and release date is somewhat redundant with release year.

In terms of numerical variables, there appeared to be a non-linear relationship between the predictors 'score' and 'votes'. Besides that, there did not appear to be strong linear relationships that would suggest co-linearity between the predictors, based on the correlation matrix and scatterplots, so we left the numerical variables as is. The scatterplots for all the numerical columns is shown in Figure[1]. We also looked at the distribution of the numerical columns, and noticed that most of them were highly skewed.

3 Multiple Linear Regression Model

3.1 Significance Level

Because there is no major consequence of a Type I error, but we still want to be conservative with our estimates, so we will use default significance level $\alpha = 0.05$.

3.2 Full Model

The predictors we used in our full model to estimate the Total U.S. gross revenue (in U.S. dollars) are: budget, company, country, genre , rating, run time, score, votes, and writer as seen in the summary output for the full model.

The R^2 is 0.6282, meaning that the predictors in the final model explain 62.82 percent of the variance in the response variable, gross revenue of the movie in U.S. dollars.

3.3 Final Model

3.3.1 Backwards Elimination Algorithm

To determine the final model, a series of F-tests had to be conducted. We manually performed a Backwards Selection Algorithm, which entails testing the significance of a single predictor at a time by removing said predictor from the full model and looking at the F-test from the ANOVA table output of the full and reduced model. The F-test for all our tests for single predictors looks like:

$$\begin{cases} H_0 : \beta_i = 0 \text{ (reduced model is better)} \\ H_\alpha : \beta_i \neq 0 \text{ (full model is better)} \end{cases}$$

In the first iteration, it was found that the slopes for director and star were not significant (which is to say $\beta_{\text{director}} = \beta_{\text{star}} = 0$). However, the p-value for star was higher than that for director, so at the end of the first iteration, we dropped star as a predictor. An example R output ANOVA table for the following hypothesis test during the first iteration is shown in Table[2]:

$$\begin{cases} H_0 : \beta_{\text{budget}} = 0 \\ H_\alpha : \beta_{\text{budget}} \neq 0 \end{cases}$$

The second iteration had its 'full' model as the original full model, but without star as a predictor. It was found that the slope for director was not significant, so at the end of the second iteration, we dropped director as a predictor.

The second iteration had its 'full' model as the original full model, but without star and director as predictors. The significance of all remaining slopes was tested one by one, but it was found all the other slopes were significant. Therefore, the best model according to the Backwards Elimination Algorithm was the full model but without star or director as predictors.

3.3.2 Identical Predictors Tests

It should be noted, however, that the Backwards Elimination Algorithm does not necessarily give the best model that is possible. Thus, we perform a few more tests to check if any pair of slopes (and so, the predictors) are identical. The hypothesis test can be formulated as:

$$\begin{cases} H_0 : \beta_i = \beta_j \\ H_\alpha : \beta_i \neq \beta_j \end{cases}$$

for $i \neq j$. Again, since we are dealing with primarily categorical variables, we use the F-test results from the ANOVA table for the two models. An example R output ANOVA

table for the following hypothesis test is shown in Table[3]:

$$\begin{cases} H_0 : \beta_{\text{budget}} = \beta_{\text{votes}} \\ H_\alpha : \beta_{\text{budget}} \neq \beta_{\text{votes}} \end{cases}$$

For all pairs of numerical variables excluding year (which was not tested), it was found that the slopes were not equal. This means that the predictors are unique, and we cannot eliminate any further predictors through this test.

At this point, we have performed some extra tests (identical predictor tests) to check if we can get anything better, but the model selected by Backwards Elimination Algorithm is still the best. So, we take that to be our final model. Lastly, the summary table of the final model is shown in Table[4].

4 Model Diagnostics

4.1 Unusual Observations

4.1.1 High Leverages

From extracting the leverage point of the final model and putting the threshold of $2p/n$, we could see that there was 160 high leverage points that exceeded the threshold. These high leverage points represent about 3.45% of the total observations. The half-normal plot is shown in Figure[2]. We also noted that 85 of the leverages can be considered as "bad-leverage points". We also noted that since none of Cook's Distances ≥ 1, we can say that there are no highly influential points in the data. The plot for the Cook's Distance is shown in Figure[4].

4.1.2 Outliers

By doing the outlier test under the null hypothesis H_0 that no outliers are present, we computed the critical T_{n-p-1} value with Bonferroni correction and got -4.405751 with significance level $\alpha = 0.05$. The decision rule states that if an observation's studentized residual is higher than the absolute value of critical T_{n-p-1} value with Bonferroni correction, then observation would be considered an outlier. Using this decision rule, we conclude that there are 31 outliers in out final model.

4.2 Model Assumptions

4.2.1 Normality

For the Normality test, we used the Kolomogorov-Smirnov Normality test, since the final model has $n = 4638$ which is $n > 50$. The hypothesis to test is:

$$\begin{cases} H_0 : \text{The distribution is Normal} \\ H_\alpha : \text{The distribution is not Normal} \end{cases}$$

Through the *ks.test* function in R, putting the first parameter as final model and the second parameter as cdf of the Normal. The test statistics, D_n, is 0.55886, and the p-value is stated as less than $2.2 \cdot 10^{-16}$. Since the p-value is less than the significance level of $\alpha = 0.05$, we reject the null hypothesis of normality and conclude that the normality assumption is not satisfied.

4.2.2 Variance

From the plot of residuals vs. fitted values it is seen that constant variance is violated because the residuals vs. fitted plot is not randomly scattered. It seems as fitted values get larger, residuals get larger as well. This means that the constant variance assumption is violated. Additionally, from the studentized Breusch-Pagan test, the null hypothesis is that the variance is constant, and the alternative hypothesis is that that variance is not constant. From the output, our p-value is less than $2.2 \cdot 10^{-16}$. This means that we reject the null at significance level of $\alpha = 0.05$ and conclude that the variance is not constant.

5 Results

5.1 Estimation

5.1.1 Movie: Little Shop of Horrors

The first movie in the dataset we selected (randomly) was "Little Shop of Horrors". The actual gross sales of the movie was \$38,747,385, and the estimated gross sale value based on the final model was \$42,285,315. The 95% confidence interval for the gross sales of this movie was (\$37,771,264 , \$46,799,367). Note that the actual value of gross sales does fall within the 95% confidence interval, however the interval itself is very wide, which might indicate that the model might not be precise enough.

5.1.2 Movie: Armed and Dangerous

The second movie in the dataset we selected (randomly) was "Armed and Dangerous". The actual gross sales of the movie was $15,945,534, and the estimated gross sale value based on the final model was $17,181,965. The 95% confidence interval for the gross sales of this movie was ($12,750,846 , $21,613,085). Note that the actual value of gross sales does fall within the 95% confidence interval, however the interval itself is very wide, which might indicate that the model might not be precise enough.

5.2 Prediction

5.2.1 Movie: Wonder Woman 1984

The first movie not included in the dataset that we chose was "Wonder Woman 1984". We found out the relevant values for our predictors by searching up the movie on IMDb and Wikipedia. The actual gross revenue for the movie was $169,600,000 and the predicted gross revenue by the model was $189,334,469. The 95% prediction interval for the gross sales of this movie was ($109,817,513 , $268,851,424). While the actual value for the gross sales does fall within the 95% confidence interval, the interval itself is extremely wide, which sort of invalidates any predictions because such a high range is not useful in practice.

5.2.2 Movie: Fantastic Beasts: The Secrets of Dumbledore

The second movie not included in the dataset that we chose was "Fantastic Beasts: The Secrets of Dumbledore". We found out the relevant values for our predictors by searching up the movie on IMDb and Wikipedia. The actual gross revenue for the movie was $407,200,000 and the predicted gross revenue by the model was $198,066,631. The 95% prediction interval for the gross sales of this movie was ($118,446,615 , $277,686,648). The actual value for the gross sales does not fall within the 95% confidence interval; in fact, the actual gross sales is about twice the predicted gross revenue. The prediction interval is also very wide, but it still was not able to include the actual gross revenue, indicating very poor performance.

6 Conclusion

The goal is this analysis was to create a statistical model for estimating the total U.S. gross revenue of a movie using various predictors.

During our exploratory data analysis, we checked our predictor variables and made sure they did not affect each other. We also removed some entries that had incomplete data

such as budget=0, and we grouped some of the categorical variables' values together into smaller groups, such as keeping one large genre group such as "PG-13" and then grouping all other smaller genre groups into an "other" group. After that, we created a full model and repeatedly performed tests to see if we can remove some variables that do not have significant impacts on the model's ability to estimate/predict total gross revenue of a movie such as the star and director variables.

From the model we created, we found that some of the factors that had the most impact on estimating the total U.S. gross revenue of movies are *rating*, *country*, and *genre*. Movies rated 'R', on average, tend to have a gross revenue of $1.98 million lower than movies with other ratings, excluding 'PG-13'. On average, movies produced in the USA tend to have a gross revenue of $957,300 than movies produced outside the USA. Movies from certain genres such as comedy also tend to have a higher gross revenue than movies from other genres.

Unfortunately, the data we used to create the model violates the assumptions needed for our model to make accurate predictions. Due to the limited scope of this case study, we did not perform any remedial measures to make the data fit the model's assumptions. Because of this, the model's performance of predicting the total gross revenue of new movies is very poor, and the model gives very broad estimates, which is not useful to real-world use. However, the model works decently for estimating the gross revenue of movies that was used to create the model. Given more time, we suggest transforming the data or using different modeling techniques apart from MLR so that our model's assumptions are met. This would likely improve the model's ability to predict new movies' total gross revenue to a more useful level.

7 Appendix

7.1 Exploratory Data Analysis

7.1.1 Summary Statistics

```
##      budget             company             country             director
##  Min.   :     6000   Length:4638        Length:4638        Length:4638
##  1st Qu.: 10000000   Class :character   Class :character   Class :character
##  Median : 23000000   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 36145602
##  3rd Qu.: 46000000
##  Max.   :300000000
##     genre               gross               name               rating
##  Length:4638        Min.   :      309   Length:4638        Length:4638
##  Class :character   1st Qu.:  6290905   Class :character   Class :character
##  Mode  :character   Median : 23455506   Mode  :character   Mode  :character
##                     Mean   : 46074694
##                     3rd Qu.: 57782434
##                     Max.   :936662225
##    released            runtime          score           star
##  Length:4638        Min.   : 69.0   Min.   :1.500   Length:4638
##  Class :character   1st Qu.: 96.0   1st Qu.:5.800   Class :character
##  Mode  :character   Median :104.0   Median :6.400   Mode  :character
##                     Mean   :107.6   Mean   :6.356
##                     3rd Qu.:117.0   3rd Qu.:7.100
##                     Max.   :280.0   Max.   :9.300
##      votes             writer               year
##  Min.   :    183   Length:4638        Min.   :1986
##  1st Qu.:  16110   Class :character   1st Qu.:1996
##  Median :  43940   Mode  :character   Median :2003
##  Mean   :  95702                      Mean   :2002
##  3rd Qu.: 109393                      3rd Qu.:2010
##  Max.   :1861666                      Max.   :2016
```

Table 1: The new summary statistics of the columns of the dataset after removing the rows with 0 budget.

7.1.2 Scatterplots

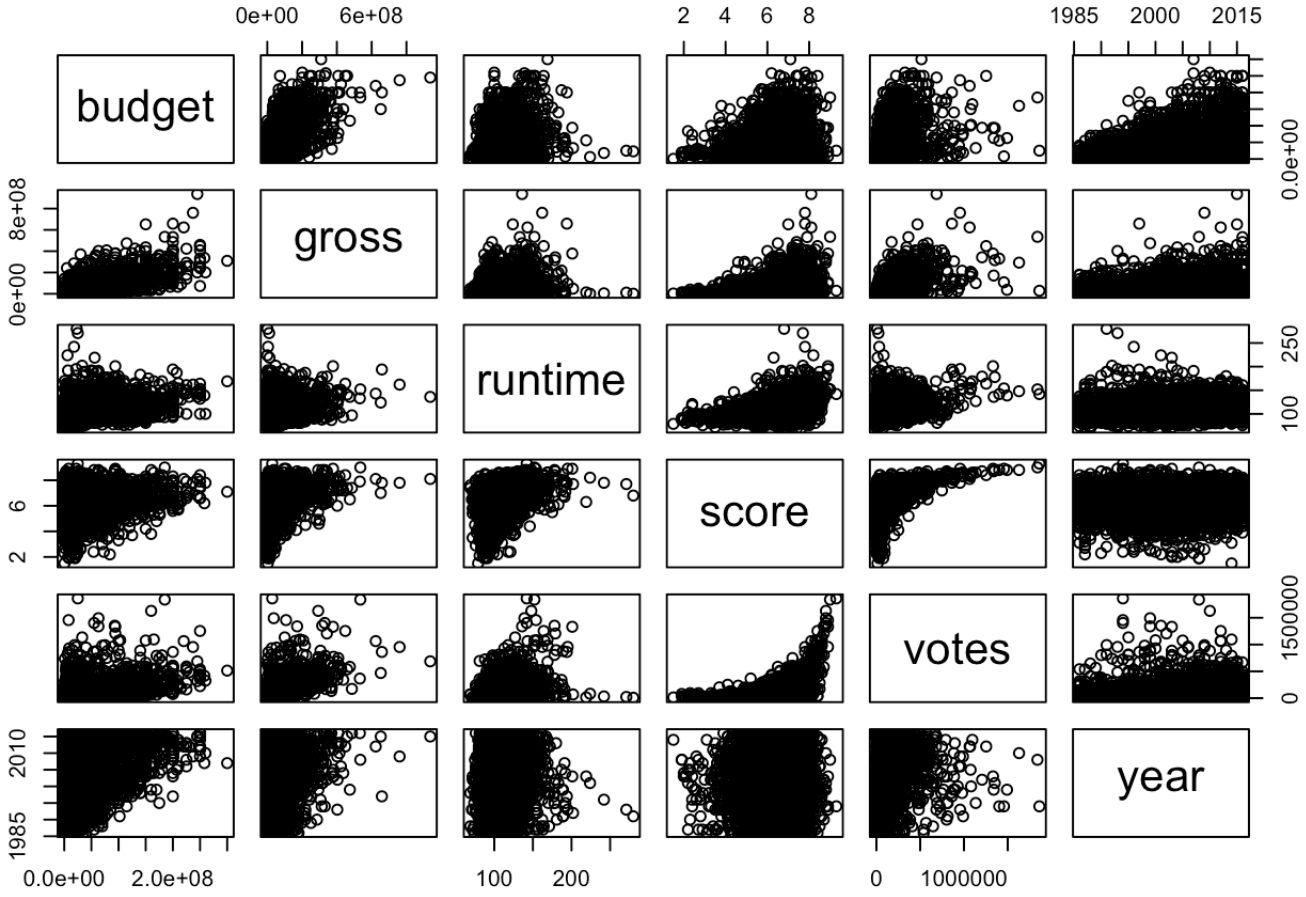

Figure 1: The scatterplots for each numerical column pair in the dataset.

7.2 Model Selection

7.2.1 Example ANOVA Output Table for Backwards Elimination Algorithm

```
## Analysis of Variance Table
##
## Model 1: gross ~ (budget + company + country + director + genre + rating +
##     runtime + score + star + votes + writer + year) - budget
## Model 2: gross ~ budget + company + country + director + genre + rating +
##     runtime + score + star + votes + writer + year
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   4622 9.8263e+18
## 2   4621 7.5537e+18  1 2.2726e+18 1390.3 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 2: The ANOVA output table for testing for the significance of slope of the budget predictor in the first iteration of Backwards Elimination Algorithm. As seen from the p-value, we reject the null and conclude that the budget predictor slope is statistically significant.

7.2.2 Example ANOVA Output Table for Identical Predictors Test

```
## Analysis of Variance Table
##
## Model 1: gross ~ company + country + genre + rating + runtime + score +
##     I(budget + votes) + writer + year
## Model 2: gross ~ (budget + company + country + director + genre + rating +
##     runtime + score + star + votes + writer + year) - star -
##     director
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   4625 9.6970e+18
## 2   4624 7.5556e+18  1 2.1414e+18 1310.5 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 3: The ANOVA output table for testing whether the slopes for the predictors budget and votes is the same. As seen from the p-value, we reject the null and conclude that the budget and votes slopes are not the same.

7.2.3 Final Model

```
##
## Call:
## lm(formula = gross ~ ., data = movie_final)
##
## Residuals:
##       Min         1Q     Median         3Q        Max
## -349206139  -17316712   -2631787   10780858  616992453
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    5.065e+08  1.496e+08   3.385 0.000717 ***
## budget                         7.736e-01  2.035e-02  38.025  < 2e-16 ***
## companySmall Company          -2.818e+06  1.288e+06  -2.188 0.028737 *
## countryUSA                     9.573e+06  1.567e+06   6.111 1.07e-09 ***
## genreComedy                    7.384e+06  1.787e+06   4.131 3.67e-05 ***
## genreDrama                    -2.051e+05  2.044e+06  -0.100 0.920070
## genreOther                     4.464e+06  1.739e+06   2.567 0.010302 *
## ratingPG-13                   -9.045e+06  1.837e+06  -4.923 8.82e-07 ***
## ratingR                       -1.980e+07  1.774e+06 -11.164  < 2e-16 ***
## runtime                       -1.233e+05  3.991e+04  -3.090 0.002013 **
## score                          2.121e+06  7.458e+05   2.844 0.004476 **
## votes                          1.911e+02  5.251e+00  36.394  < 2e-16 ***
## writer2 or More Movies Written  3.401e+06  1.235e+06   2.753 0.005933 **
## year                          -2.527e+05  7.465e+04  -3.385 0.000717 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40420000 on 4624 degrees of freedom
## Multiple R-squared:  0.6292, Adjusted R-squared:  0.6282
## F-statistic: 603.7 on 13 and 4624 DF,  p-value: < 2.2e-16
```

Table 4: The summary output of the final model

7.3 Model Diagnostics

7.3.1 Half-Normal Plot

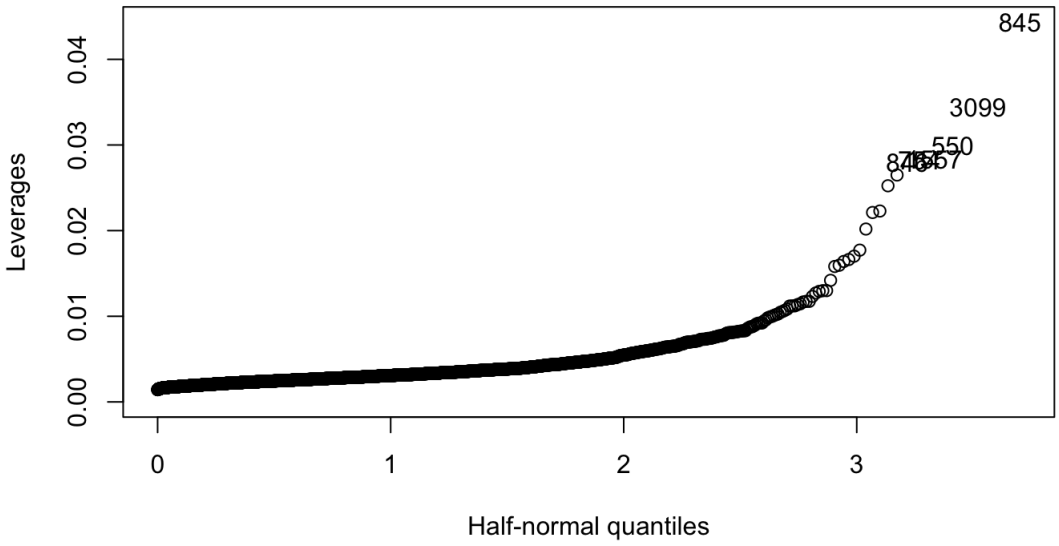

Figure 2: There do appear to be leverages that are unusually large at the right end of the plot.

7.3.2 Q-Q Plot

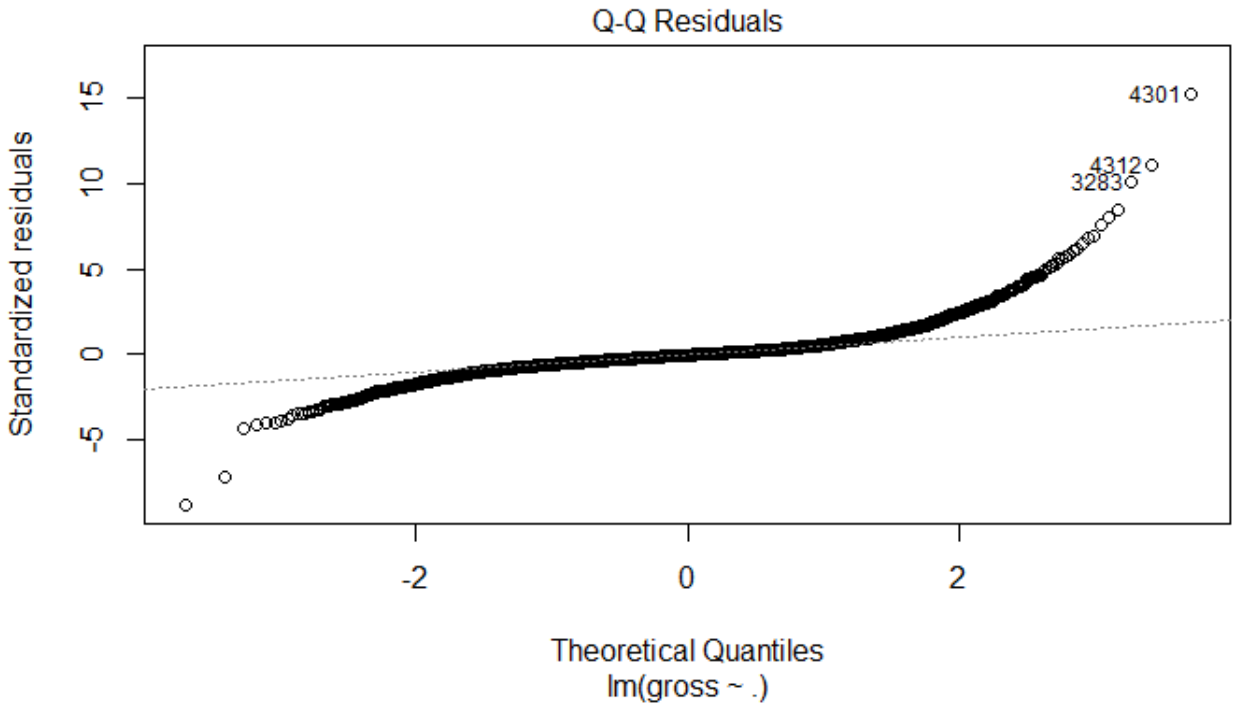

Figure 3: The shape of Q-Q plot shown illustrates that there is more variance than you would expect in a normal distribution.

7.3.3 Cook's Distance Plot

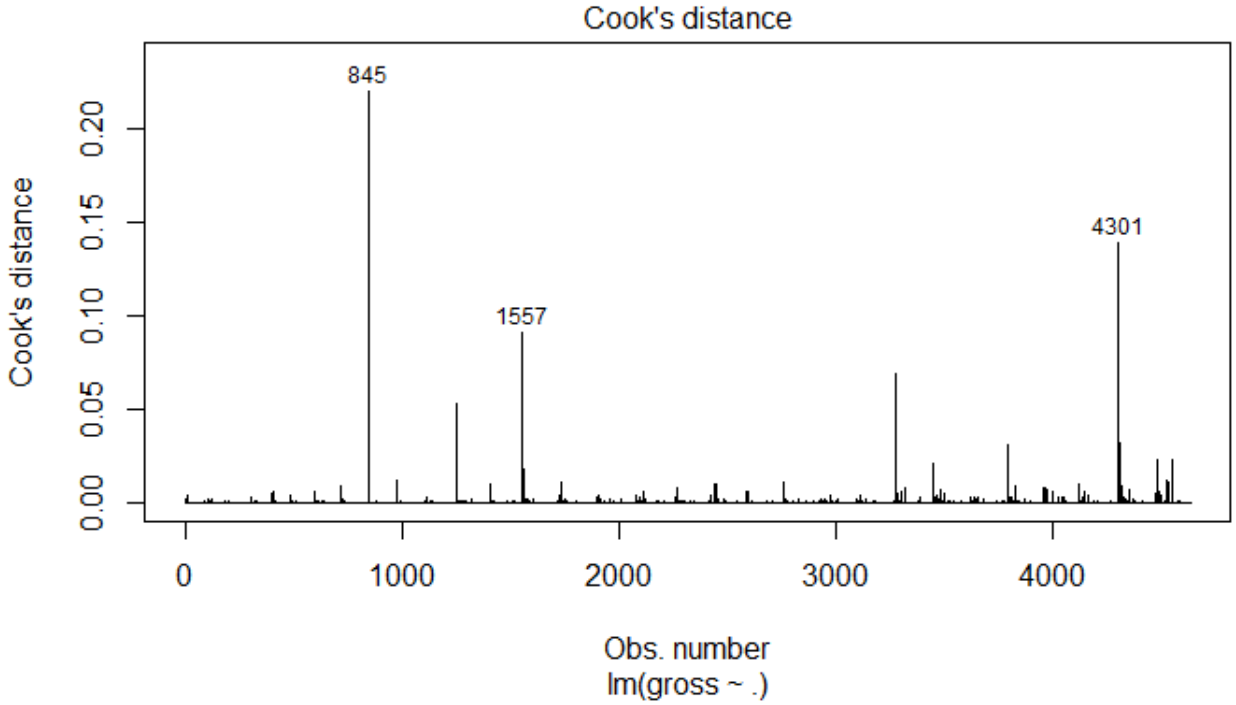

Figure 4: The graphs show that there are no Cook's Distances above 1, indicating there are no highly influential points in the data.

7.3.4 Residuals vs. Fitted

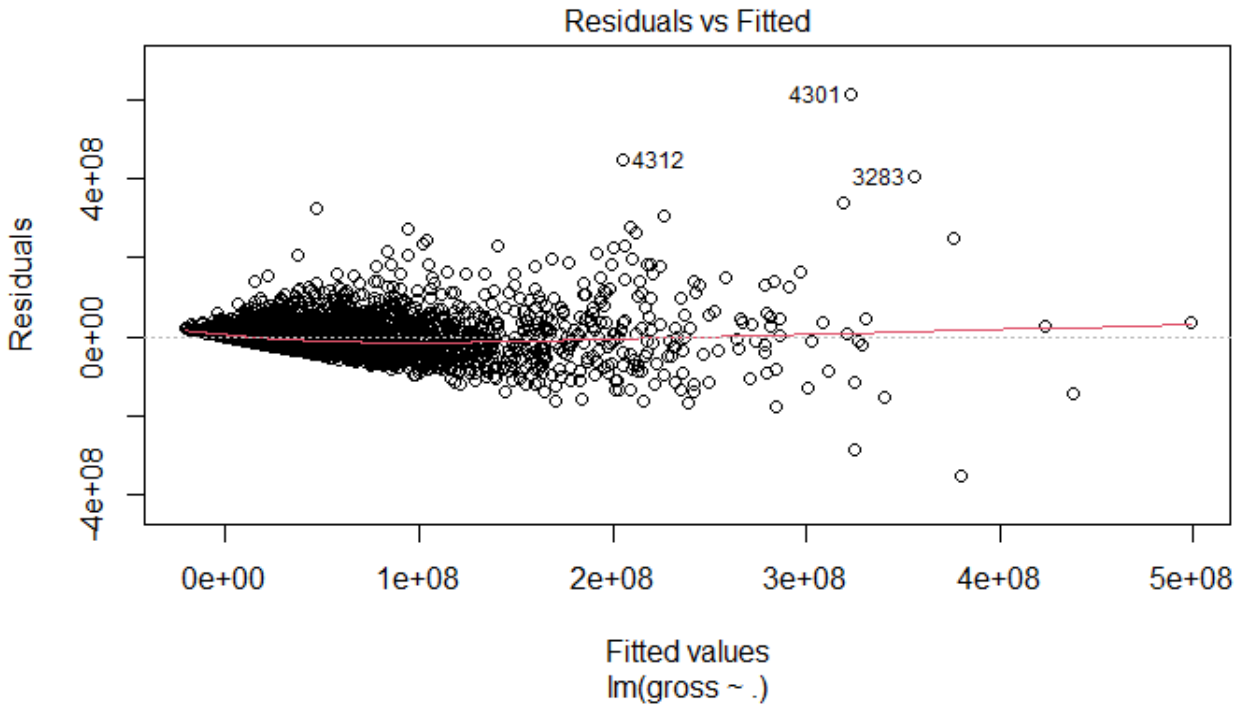

Figure 5: The variance increases as we go from left to right, and the points do not fall nicely within two parallel lines but rather they fall within two sloped lines. This indicates the violation of the constant variance assumption.

7.3.5 Histogram of Residuals

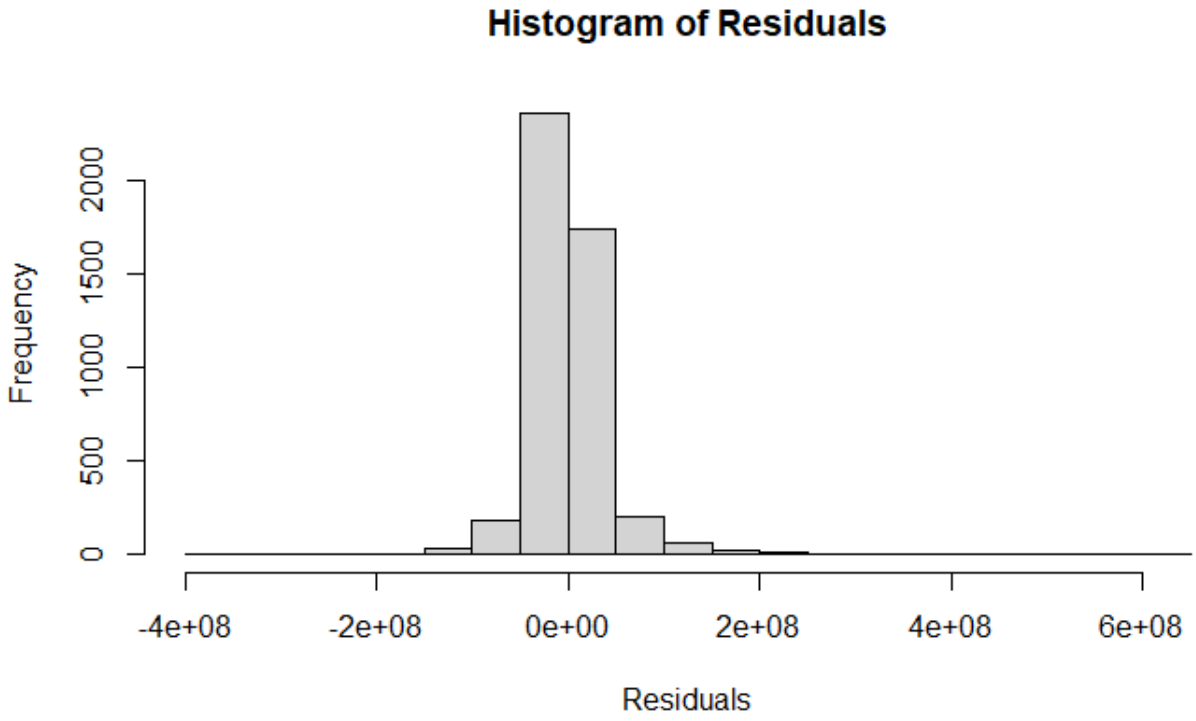

Figure 6: The overall frequency of the residuals are shown to be skewed to the right.