PSTAT 126 Final Project

Regression Analysis on Conventional and Social Media 2014 and 2015 dataset

Jasmine Kwok

6/11/2020

Abstract

In recent years, there is a growing interest for analysts and investors to assess the financial risk in film production. This study utilizes multiple linear regression analysis to predict the financial success of films and investigate the relationship between screens and year. The results demonstrate that a possible linear regression model consists of ratings, budget, screens, sequel, and aggregate followers as the predictors. Using this model, we achieved a mean gross income of $76,444,805 for predicting all films and only one film. Further, we found that year has no effect on screens in predicting mean gross income and there is a positive linear relationship between them.

Problem and Motivation

The primary goal of this project is to explore and analyze the gross income for movies in 2014 and 2015 as well as examine the relationship and significance of some explanatory variables. Ultimately, this project aims to develop an optimal regression model to predict the financial success of films in 2014 and 2015. The conventional and social media dataset used in this project is straightforward so individuals with only some basic knowledge on film would be able to comprehend. Readers who are interested or work in the film industry, however, would have better comprehension on this project.

The film industry is a significant contributor to a country's economy and a major employer in the United States. Due to the large costs involved in film production, it is crucial for analysts to research and understand major variables which contribute to a film's commercial and financial success. This project may provide insights into key features contributing to the financial success of films and spark future research to examine relationships between particularly unique explanatory variables in our dataset. This project may also be insightful to film producers to determine which features to focus on during the promotion phase to improve film success.

Data

The conventional and social media (CSM) 2014 and 2015 dataset was obtained from UC Irvine Machine Learning Repository. The original source of the data is from Youtube, Twitter, and IMDB (Ahmed, Jahangir, Afzal, Majeed,& Siddiqi, 2015). This dataset was provided by Mehreen Ahmed from the National University of Sciences and Technology, Islamabad, Pakistan (Ahmed et al., 2015). He is also the data collector who utilized the modules, Data Collector and Predictive Engine to obtain the data. Originally, the purpose of this dataset was for predictive analysis on the success of movies using machine learning algorithms (Ahmed et al., 2015). There are a total of 14 features in the dataset with some missing datas. The attributes include movie name, year, genre, budget, number of screens, sequel, ratings, gross income, sentiment score, number of comments, number of dislikes, number of likes, number of views, and aggregate actor followers. The categorical data include movie name and genre while the year belongs to ordinal data. There are a total of 11 features that are numerical data. There are only two unique variables to year which are 2014 and 2015. There are 231 unique variables for movie names and 15 unique variables representing different genres of the movie. This dataset represents only the movies that are within the diverse sources of IMDB, Wikipedia, Youtube, and Twitter (Ahmed et al., 2015). Hence, the data overrepresents American films and underrepresents movies produced and released in other countries.

Questions of Interest

1. What is a multiple linear regression model that can predict mean gross income of movies released in 2014 and 2015?
2. What is the estimated mean gross income for all movies and the predicted mean gross income for one movie with average values as their predictors?
3. What is the relationship between gross income and screens in 2014 and 2015?

Regression Methods

Outline

In the beginning of this project, exploratory analysis will be conducted by using scatterplots and added variable plots to observe some relationships that exist within the variables in my dataset. To answer the first question of interest a model selection using various methods such as backwards selection, regsubsets and summary table will be carried out. The influential index plot will be used to determine outliers, high leverage points, and ultimately, influential points. A diagnostics check using residuals and fitted plot and Q-Q plot was conducted to ensure there are no violations to the assumptions: linearity, equal variance, normality, and independence. To resolve the violations, the power transformation method and the box-Cox method will be used. To improve the regression model, the analysis of variance table may be used to determine the usefulness of interaction terms to be added to our model.To answer the second question, the model determined for the first question is used to calculate the confidence interval, prediction interval, and mean gross income. The anova table and scatter plot diagram is used to determine the relationship between mean gross income and screens in 2014 and 2015.

Exploratory Analysis

In the beginning of analysis, we used the scatterplots matrix and added variable plots to explore the relationships that exist between the variables. Since there are 13 attributes, only continuous variables were plotted to minimize predictors and ensure our scatterplot matrix is readable. From the scatterplots, there is a positive linear relationship between gross and ratings (Appendix B). There also seem to be a positive linear relationship between gross income and budget. We can observe an exponential relationship between gross income and screens as well as between budget and screens (Appendix B). These clear nonlinear relationships between variables suggests that future transformations on predictors may be needed for linear regression. This is also evidence that the linearity assumption may be violated. Other notable relationships include positive linear relationships between views and likes, views and dislikes, and views and comments. There seems to be a positive relationship between ratings and aggregate followers, and a negative relationship between budget and aggregate followers. From the added-variable plots, two notable predictors are budget and aggregate followers which has a clear positive linear relationship when other predictors are held constant (Appendix B). This suggests that our regression model should include these two predictors as it is useful in explaining gross income.

Model Selection

Given the large number of predictors, it is crucial to minimize our model by only including useful predictors. There are 13 predictors in our dataset which means that there are 8912 (2^{13}) possible regression models. To effectively select our model, a backward selection is carried out using the step function. This method iterates procedures to minimize Akaike's Information Criteria (AIC) and Bayesian Information Criteria (BIC). A backwards model selection in comparison to the forwards selection as it eliminates the possibility of a newly selected predictor having the same or more ability to explain parts of the response that is already explained by another predictor present in the model. We obtained the lowest AIC value as 6563.49, however, using the step function the minimized BIC value is not provided (Appendix C). The two methods selected different regression models. The backwards selection using AIC included 6 predictors, ratings, budget, screens, sequel, dislikes, and aggregate followers, for the response gross income. The model selected using BIC only includes 4 predictors ratings, budget, screens, aggregate followers. It is known that BIC places a higher penalty on the number of parameters due to the weight, so it tends to reward smaller models which is present in our observation.

An optimal variable selection takes into account coefficient of determination, adjusted R-squared, mean squared error, the two information criteria AIC and BIC, as well as Mallows' Cp statistic. The step function only captures one criteria which is the information criteria AIC and BIC. Hence, our analysis proceeds by carrying out a regression subsets method (regsubsets) on the first model obtained from backwards selection on AIC which is an exhaustive search on all other possible models based on existing predictors. The regression subsets method was not used prior to this as it is inefficient for our dataset when it has 13 predictors. Looking at the R-squared values, we select the model with the largest increase of 0.0397 which is model 2 that has the R-squared value 0.525 (Appendix D). This model only includes budget and screens as predictors for gross income. The criteria for smallest mean squared error is equivalent to the criteria of largest adjusted R-squared, hence, only one of the criterias has to be checked. The model with the highest adjusted R-squared is model 6 with 0.5878093 (Appendix D). Model 6 is the full model which has all the predictors chosen from the backward selection method. Based on Mallows' Cp statistic value, our choice would be model 5 which has the value of 8.715 closest to our q value 7(Appendix D). This model excludes dislikes and includes all other predictors. Our last criteria is to check for the model with the lowest Bayesian Information Criteria which was not provided by the step selection method. Similar to the step function, the model with the lowest BIC value of 134.128 is model 4 that includes the exact same predictors ratings, budget, screens, and aggregate followers. Based on all the criterias, the optimal model would be the model which includes the 6 predictors: ratings, budget, screens, sequel, dislikes, and aggregate followers (Appendix D). This is because it satisfies the most criterias, highest adjusted R-squared, lowest mean squared error, and lowest AIC.

Summary Table

```
##
## Call:
## lm(formula = Gross ~ Ratings + Budget + Screens + Sequel + Dislikes +
##     `Aggregate Followers`, data = csm_dataset)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -127511153  -32245753   -4850498   21292820  410362338
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -1.577e+08  3.268e+07  -4.825 3.02e-06 ***
## Ratings                2.088e+07  4.865e+06   4.292 2.91e-05 ***
## Budget                 7.121e-01  1.087e-01   6.549 6.13e-10 ***
## Screens                1.425e+04  3.993e+03   3.568 0.000463 ***
## Sequel                 1.175e+07  5.119e+06   2.295 0.022920 *
## Dislikes               6.967e+03  3.615e+03   1.927 0.055536 .
## `Aggregate Followers`  2.541e+00  9.263e-01   2.743 0.006708 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60270000 on 176 degrees of freedom
## Multiple R-squared:  0.6014, Adjusted R-squared:  0.5878
## F-statistic: 44.26 on 6 and 176 DF,  p-value: < 2.2e-16
```

From the summary table, a global F-test and a partial F-test for each single variable: Ratings (1), Budget (2), Screens (3), Sequel, (4) Dislikes(5), and Aggregate Followers(6).

1. $H_0 : B_1 = 0$ vs $H_1 : B_1 \neq 0$

For ratings, the test statistic was -4.825 and p-value was 2.91×10^{-5}. The p-value is significantly smaller than any conventional alpha values which indicates a strong evidence against null hypothesis so our decision is to reject the null hypothesis. We can conclude that ratings is a useful predictor for gross income when the predictors, budget, screens, sequel, dislikes, and aggregate followers are held constant.

2. $H_0 : B_2 = 0$ vs $H_1 : B_2 \neq 0$

The p-value of the partial f-test for the budget is 6.13×10^{-10} which is significantly smaller than all alpha values. This suggests a strong evidence against the null hypothesis. Our decision is to reject the null hypothesis where $B_2 = 0$. We can conclude that ratings is a useful predictor for gross income when the predictors, ratings, screens, sequel, dislikes, and aggregate followers are held constant.

3. $H_0 : B_3 = 0$ vs $H_1 : B_3 \neq 0$

The p-value of the partial f-test for the screens is 0.000463 which is significantly smaller than all alpha values. This suggests a strong evidence against the null hypothesis. Our decision is to reject the null hypothesis where $B_3 = 0$. We can conclude that screens is a useful predictor for gross income when all other predictors are held constant.

4. $H_0 : B_4 = 0$ vs $H_1 : B_4 \neq 0$

The p-value for sequel is 0.022920 which is smaller than the alpha value 0.05. This suggests a strong evidence against the null hypothesis. Our decision is to reject the null hypothesis stating $B_4 = 0$. We can conclude that sequel is a useful predictor for gross income when all other predictors are held constant.

5. $H_0 : B_5 = 0$ vs $H_1 : B_5 \neq 0$

The p-value of the partial f-test for dislikes is 0.055536 which is larger than alpha value 0.05, indicating a weak evidence against the null hypothesis. Our decision is to fail to reject the null hypothesis where $B_5 = 0$. We can conclude that dislikes is not a useful predictor for gross income when all other predictors are held constant. However, with alpha at 0.1, we do not have sufficient evidence to reject the null hypothesis since p-value is smaller than alpha. Our decision is to keep this predictor as we would like to check for influential points which may change the significance of our predictors.

6. $H_0 : B_6 = 0$ vs $H_1 : B_6 \neq 0$

The p-value for aggregate followers is 0.006708 which is smaller than the alpha value 0.05. This suggesting a strong evidence against the null hypothesis so our decision is to reject the null hypothesis stating $B_6 = 0$. We can conclude that aggregate followers is a useful predictor for our model when all other predictors are held constant.

Global F-test

$H_0 : B_1 = B_2 = B_3 = B_4 = B_5 = B_6 = 0$ vs H_1 :at least 1 predictor $B_i \neq 0$ where i = 1,2,3,4,5,6. The p-value for the global F-test is 2.2×10^{-16} which is much smaller than alpha 0.05, suggesting strong evidence against null hypothesis. This means our decision is to reject the null hypothesis. We can conclude that there is at least one useful predictor in our model.

Influential Points

Using the influence index plot, we are able to visualize the candidate points which are outlier and/or high leverage points to determine influential points. The goal of this step is to find and remove the influential points to improve our regression model. Looking at Cook's distance, the candidate points are 130 and 138 (Appendix F). For studentized residuals, the potential outliers are points 10 and 130 (Appendix F). From the hat values, the potential high leverage points are 132 and 138 (Appendix F). Only a few points are selected and removed from the candidates above and they are points 10 and 138 as removing many points at once changes the entire model. By removing the

points from the original dataset and constructing another summary table, we are able to determine if the points are influential points.

From our original summary table constructed with all the points in our dataset, we obtained residual standard error amount of 60270000, R-squared value of 0.6014 and adjusted R-squared value 0.5878 (Appendix E). This means that 58.78% of variability in gross income is explained by our current regression model. From our new summary table, the residual standard error is 57370000 which is significantly lower than our previous residual value (Appendix F). This means that the errors of prediction in our model is improved. The new R-squared value is 0.6234 and the adjusted R-squared value is 0.6105. This means that 61.05% of the variability in gross income is explained by our current regression model which is better than the original 58.78%. Through this method, we are able to determine that the points 10 and 138 are indeed influential points and should be removed from our model.

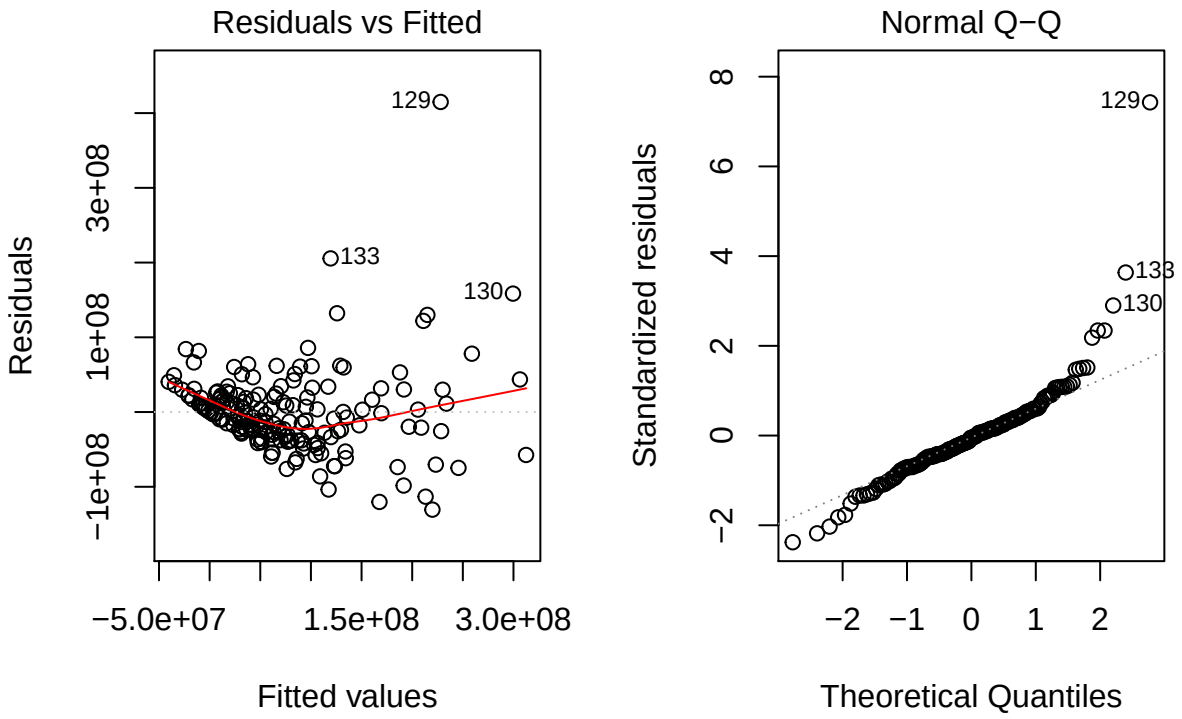
A partial f-test for dislikes was conducted again to determine its usefulness in predicting our model.

```
$H_{0}:B_{5}=0$ vs $H_{1}:B_{5}\not=0$
```

The p-value of the partial f-test is 0.580484 which is larger than alpha value 0.05 and 0.1 (Appendix F). This suggests weak evidence against the null hypothesis so our decision is to fail to reject the null hypothesis where $B_5 = 0$. We can conclude that dislikes is not a useful predictor for gross income when all other predictors are held constant and remove it from our current regression model. Hence, our current regression model would only include ratings, budget, screens, sequel, and aggregate followers to predict gross income.

Diagnostic Check

There are 4 key assumptions for multiple linear regression models: linearity, equal variance, normality, and independence. We assume that independence assumption is satisfied at the stage of data collection since the samples are collected independently using the modules Data Collector and Predictive Engine from diverse sources online. The other three assumptions are checked using the residuals vs fitted plot and Quantile-Quantile plot (Q-Q plot).

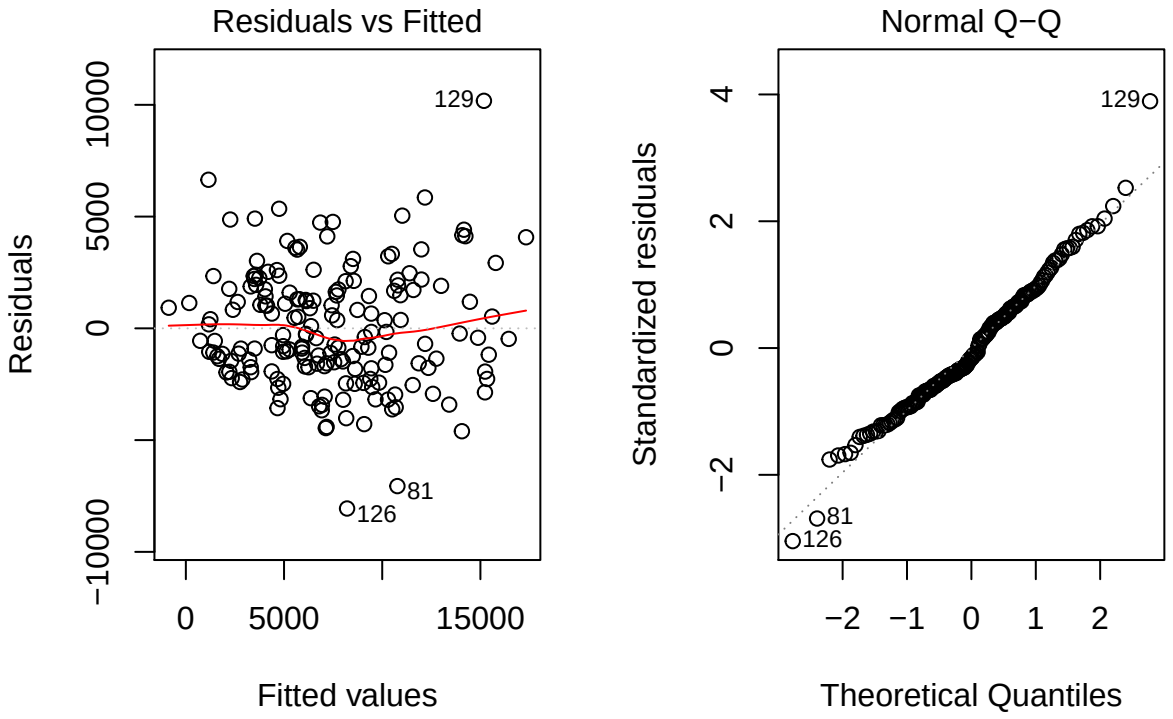


Looking at the residuals vs fitted plot, the points are mostly on the left side of the plot and not well scattered across the plot which violates the linearity assumption. The residuals also formed around the line $e_i = 0$ has a fanning pattern which violates the equal variance assumption. The Q-Q plot shows the right tail is above the y=x line and the left tail is below the y=x line. This indicates that there may be heavy-tailed distribution which violates the normality assumption (Appendix G).

Transformation

To resolve violations with non-normality, unequal variances, and nonlinear function, the power transform function is used to determine transformation needed on the predictors and the boxCox function is used to determine a transformation for response. According to the power transform function, we would transform the predictors budget, screens, sequel, and aggregate followers with lambda values 0.24, 0.79, -4.00 , and 0.16 (Appendix H). Lambda value of 1.00 for ratings indicates that no transformation has to be made for this predictor. To simplify future calculations, we round our lambda values to 0.5 , 1, -4.00, and 0 which means that no transformation will be carried out for screens as well. Two likelihood ratio tests are conducted. The first with the null hypothesis states that all the parameters are 0 which test that the transformation for all the predictors is logarithmic (lambda = 0). The p-value is 2.22×10^{-16} which is significantly smaller than any conventional alpha so our decision is to reject the null hypothesis. This means that not all transformations for the predictor are logarithmic. The second likelihood test states the null hypothesis as all lambda values is one indicating that no transformation is needed for any predictor. The p-value is 2.22×10^{-16} which is significantly smaller than any conventional alpha. Hence, our decision is to reject the null hypothesis. We conclude that it is better to transform the predictors using the

lambda values than have no transformation at all. From the boxCox method, we also obtained a rounded optimal lambda value of 0.5 from the graph (Appendix J). We choose this value to simplify future calculations although it is not within the 95% confidence interval as it is the closest rounded value to the peak of the graph.



A residuals and fitted plot and Q-Q plot is after the transformations on predictors and response are carried out to check if there are still violations of linearity, equal variance, and normality. From the residuals vs fitted plot, the points are randomly scattered across the plot with no pattern which satisfies linearity assumption. The residuals form a band from 5000 to -5000 around $e_i = 0$ with a few outliers which suggest that there is almost constant variance. Looking at the Q-Q plot, almost all the points fall on the y=x line with a few exceptions. However, real world data is imperfect and this is much better than the previous Q-Q plot. We can conclude that the normality assumption is not violated and all the assumptions for linear regression model are satisfied.

Interaction terms

```
## Analysis of Variance Table
##
## Model 1: (Gross^(0.5)) ~ Ratings + I(Budget^(0.5)) + Screens + I(Sequel^(-4)) +
##     log(`Aggregate Followers`)
## Model 2: (Gross^(0.5)) ~ Ratings + I(Budget^(0.5)) * Screens + I(Sequel^(-4)) +
##     log(`Aggregate Followers`)
##   Res.Df        RSS Df Sum of Sq     F Pr(>F)
```

```
## 1    175 1239951495
## 2    174 1239360395  1    591101 0.083 0.7736
```

```
## Analysis of Variance Table
##
## Model 1: (Gross^(0.5)) ~ Ratings + I(Budget^(0.5)) + Screens + I(Sequel^(-4)) +
##     log(`Aggregate Followers`)
## Model 2: (Gross^(0.5)) ~ Ratings * I(Budget^(0.5)) + Screens + I(Sequel^(-4)) +
##     log(`Aggregate Followers`)
##   Res.Df        RSS Df Sum of Sq      F   Pr(>F)
## 1    175 1239951495
## 2    174 1166375673  1  73575822 10.976 0.001123 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To improve our regression model, the Analysis of Variance Table was used to test for interaction terms between budget and screens, and budget and sequels (Appendix K).

$H_0 : B_{32} = 0$ vs $H_1 : B_{32} \neq 0$

The F value was 0.083 and p-value was 0.7736 which is much larger than alpha value 0.05. This indicates a weak evidence against the null hypothesis so our decision is to fail to reject the null hypothesis which is the reduced model without interaction term.Budget does not seem to interact with screens in its impact on gross income so the interaction term was not useful to explain our regression model with all other predictors held constant. Hence, we should not include it in our model.

$H_0 : B_{12} = 0$ vs $H_1 : B_{12} \neq 0$

The F value was 6.3073 and p-value was 0.0004392 which is much smaller than alpha value 0.05 which is a strong evidence against the null hypothesis. Our decision is to reject the null hypothesis. Budget appears to interact with screens so the interaction term was indeed useful to explain our regression model with all other predictors held constant. By the Hierarchy Principle, budget and ratings B_{12} should also be included in the model whether or not its coefficients are significant as its higher order term B_{12} is statistically significant for our predictor gross income.

Final regression model

```
##
## Call:
## lm(formula = (Gross^(0.5)) ~ Ratings * I(Budget^(0.5)) + Screens +
##     I(Sequel^(-4)) + log(`Aggregate Followers`), data = removed_dataset1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7834.2 -1781.1  -236.7  1744.4  9872.8
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)                 2.390e+02  2.520e+03   0.095  0.92454
## Ratings                     5.211e+01  3.672e+02   0.142  0.88732
## I(Budget^(0.5))            -6.349e-01  3.400e-01  -1.867  0.06352 .
## Screens                     1.157e+00  1.766e-01   6.551 6.23e-10 ***
## I(Sequel^(-4))             -1.685e+03  5.210e+02  -3.235  0.00146 **
## log(`Aggregate Followers`)  1.807e+02  9.048e+01   1.997  0.04739 *
## Ratings:I(Budget^(0.5))     1.664e-01  5.023e-02   3.313  0.00112 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2589 on 174 degrees of freedom
## Multiple R-squared:  0.7071, Adjusted R-squared:  0.697
## F-statistic: 70.01 on 6 and 174 DF,  p-value: < 2.2e-16
```

Looking at the summary table, the coefficient of the intercept (B_0) is 0.0239. This means that the predicted mean gross income when ratings, budget, screens, sequel, and aggregate followers are 0 is 0.0057 dollars. This is not meaningful as there is no movie produced without budget and other predictors. Hence we should not include it as we are extrapolating beyond our data. This is further supported by the partial f-test with null hypothesis (B_0)=0 and alternative hypothesis (B_0) not equal to 0. The p-value is 0.925 which is significantly larger than alpha value 0.05. This means the p-value serves as a very weak evidence against null hypothesis. Therefore, our decision is to fail to reject our null hypothesis. We can conclude that the mean gross income when all predictors are 0 is also 0. The adjusted R-squared for our model is 0.697. This indicates that 69.7% of variability in ($GrossIncome^{0.5}$) is explained by our regression model with all the predictors collectively. The value of (B_1) is 52.11 which means that for one unit increase in ratings, the value of predicted gross income is expected to increase by 2715.45 dollars (52.11^2) when all other predictors are held constant. The value of (B_2) is -0.635 which means that for one unit increase in ($budget^{0.5}$), the value of predicted gross income is expected to decrease by 0.40 dollars(0.635^(2)) when all other predictors are held constant. The value of (B_3) is 1.157 which means that for one unit increase in screens, the value of predicted gross income is expected to increase by 1.34 dollars (1.157^2) when all other predictors are held constant. The value of (B_4) is -0.0017 which means that for one unit increase in ($sequel^{-4}$), the value of predicted gross income is expected to decrease by 2.89×10^{-6} (0.0017^(2)) when all other predictors are held constant. The value of (B_5) is 18.077 which means that for one unit increase in log(Aggregate Followers), the value of predicted gross income is expected to increase by 325.78 (18.077^2) when all other predictors are held constant. The value of (B_6) is 0.166 which means that for one unit increase in ratings and budget, the value of predicted gross income is expected to decrease by 0.0276 (0.166^2) when all other predictors are held constant.Thus, our final multiple linear regression model is

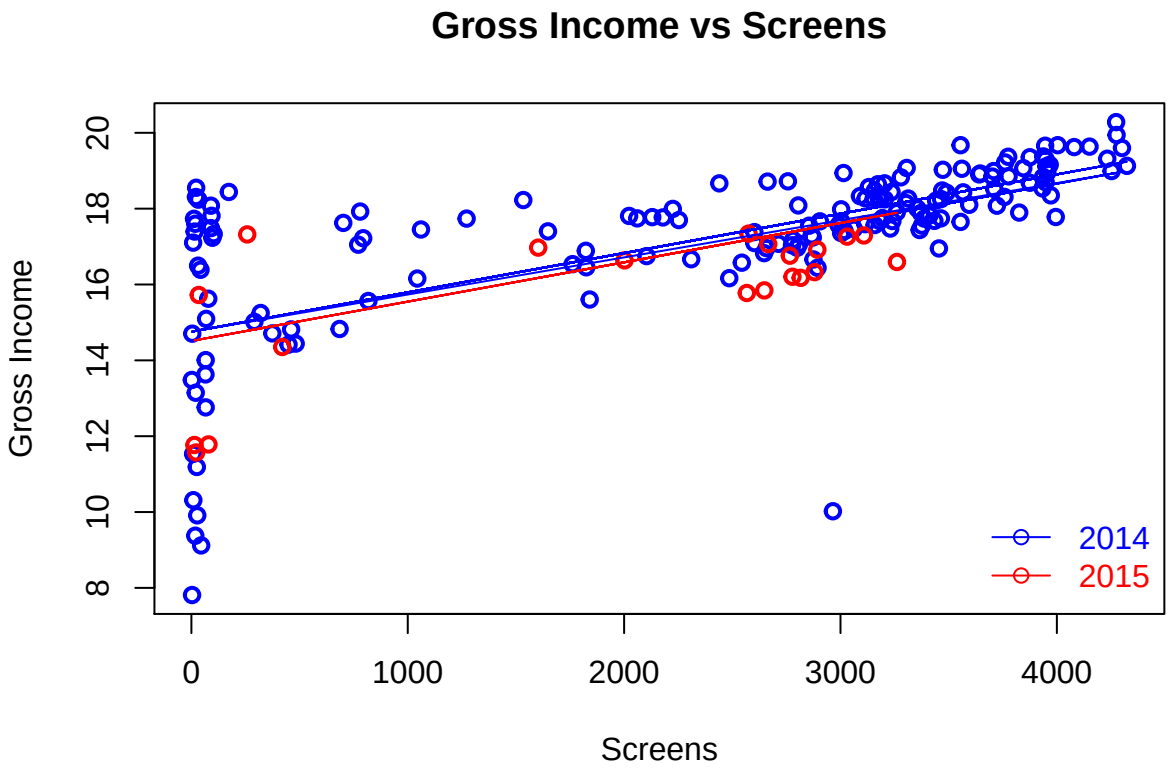
$$(Gross^{0.5}) = 52.11Ratings - 0.635Budget^{0.5} + 1.157Screens - 0.0017Sequel^{-4}$$
$$+18.077log(AggregateFollowers) + 0.166Ratings * Budget^{0.5}$$

Prediction interval and Confidence Interval

The mean gross income for all movies with the average values of ratings, budget, screens, sequel, and aggregate followers is 76,444,805 dollars (Appendix L). We are 95% confident that the mean gross income for all movies with these average values falls between \$64,639,124 and \$89,240,103. The predicted mean gross income for one movie with average values is 76,444,805 dollars. We

are 95% confident that the predicted gross income for one movie falls between $12,852,576 and $193,251,702 (Appendix L).

Relationship between gross income and screens in 2014 and 2015



Using the anova table, we are able to conclude that the non-parallel model is preferred to predict mean gross income with screens in 2014 and 2015 (Appendix M). Our null hypothesis represents the reduced model which is the main effects model while the alternative hypothesis represents the full model with interaction terms. The p-value is 0.3595 is larger than alpha 0.05 which indicates a weak evidence against the null hypothesis.This means that the interaction term is not useful in predicting mean gross income and there is no interaction between screens and year. The plot shows a positive linear relationship between screens and gross income. From the scatter plot, it is evident that the regression line for 2014 is above the regression line for 2015 which reflects a greater mean gross income for movies. Again, the absent interaction between the two terms are reflected through the two parallel slopes in the model. On average, movies in 2015 have a lower median gross income by 1.274 ($e^{0.24231}$)

Conclusion

Through regression analysis, a possible multilinear regression model to predict mean gross income of movies has ratings, budget, screens, sequel, and aggregate followers as the predictors. Our results

are accurate to a certain extent as the current regression model is only able to explain 69.7% of variability in mean gross income for movies and it does not violate any regression assumptions. The value of the predicted mean gross income for all movies and for one movie is the same which is $76,444,805. Further, there is a positive linear relationship between gross income, screens, and year, however, year does not have any effect on screens. This data is only generalizable to movies that are released in 2014 and 2015 with a main focus on movies that are produced and released in the United States. One possible extension to this analysis is to analyze the relationship between genres and screens, ratings, and budget. It is also possible to further improve our regression model by checking for other possible candidate outliers and high leverage points present in the dataset to detect and remove influential points.

Works Cited

Ahmed M, Jahangir M, Afzal H, Majeed A, Siddiqi I. Using Crowd-source based features from social media and Conventional features to predict the movies popularity. InSmart City/SocialCom/SustainCom (SmartCity), 2015 IEEE International Conference on 205 Dec 19 (pp. 273-278). IEEE.

Appendix 1: R code

Appendix A

```
library(readxl)
csm_dataset <- read_excel("2014 and 2015 CSM dataset.xlsx")
csm_dataset_ori<-csm_dataset
attach(csm_dataset)
dim(csm_dataset) # 231  14
```

```
## [1] 231  14
```

```
#check for categorical variables
# 1. Genre
is.factor(Genre) #False
```

```
## [1] TRUE
```

```
# 2.Movie
is.factor(Movie) #False
```

```
## [1] FALSE
```

```r
# 3. Year
is.factor(Year) #False
```

```
## [1] TRUE
```

```r
#modifications/cleaning
Genre <- as.factor(Genre)
Year <- as.factor(Year)
# renaming levels of Genre
levels(Genre) <- cbind("Action","Adventure","Drama","Mystery",
                       "Erotic","Thriller","Comedy","Romance","Historical fiction",
                       "Science fiction","Horror")
# find out where the missing values are
which(is.na(csm_dataset))
```

```
##  [1] 1276 1392 1411 1419 1425 1454 1471 1482 1502 1515 1616 3029 3047 3052 3058
## [16] 3062 3065 3068 3073 3080 3084 3098 3102 3113 3115 3119 3120 3133 3142 3148
## [31] 3160 3162 3221 3222 3223 3224 3225 3226 3227 3228 3229 3230 3231 3232 3233
## [46] 3234
```

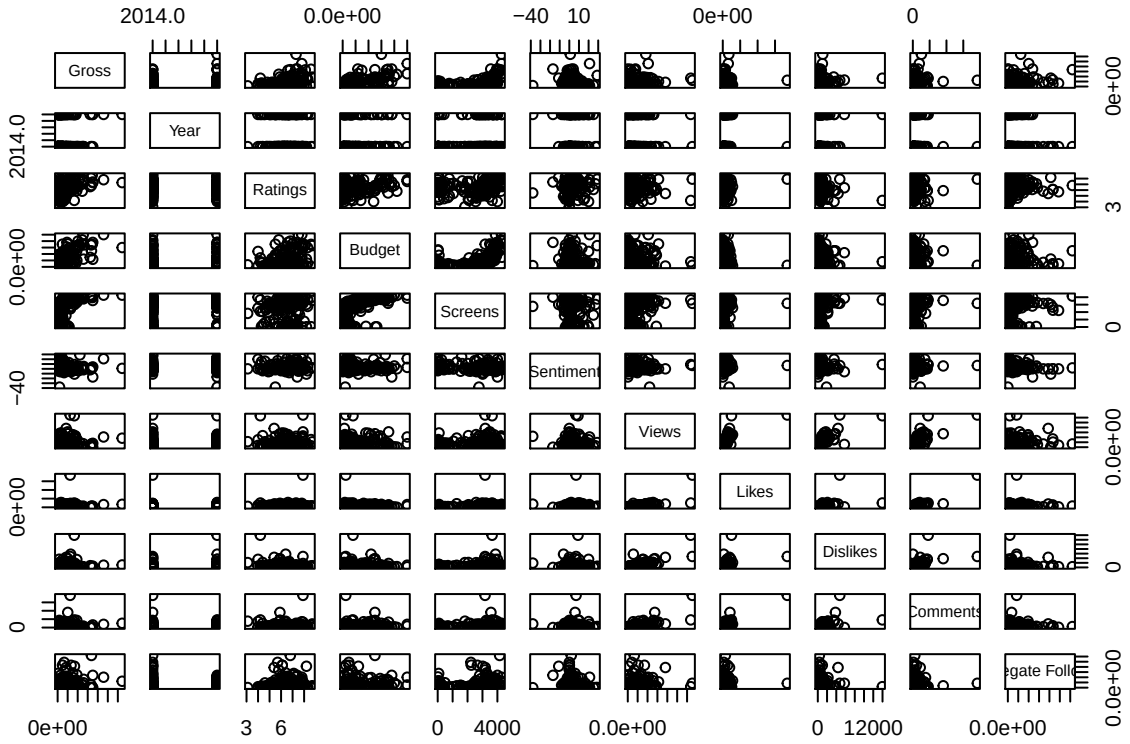
```r
#remove rows with NA
csm_dataset <- na.omit(csm_dataset) # no empty values

# looking through dataset, we see remove 0s from dislikes, comments, likes
library(dplyr)
csm_dataset <- filter(csm_dataset, Dislikes > 0, Likes > 0, Comments > 0,
                      `Aggregate Followers` >0, Screens >0)
attach(csm_dataset)
dim(csm_dataset)# 183  14
```

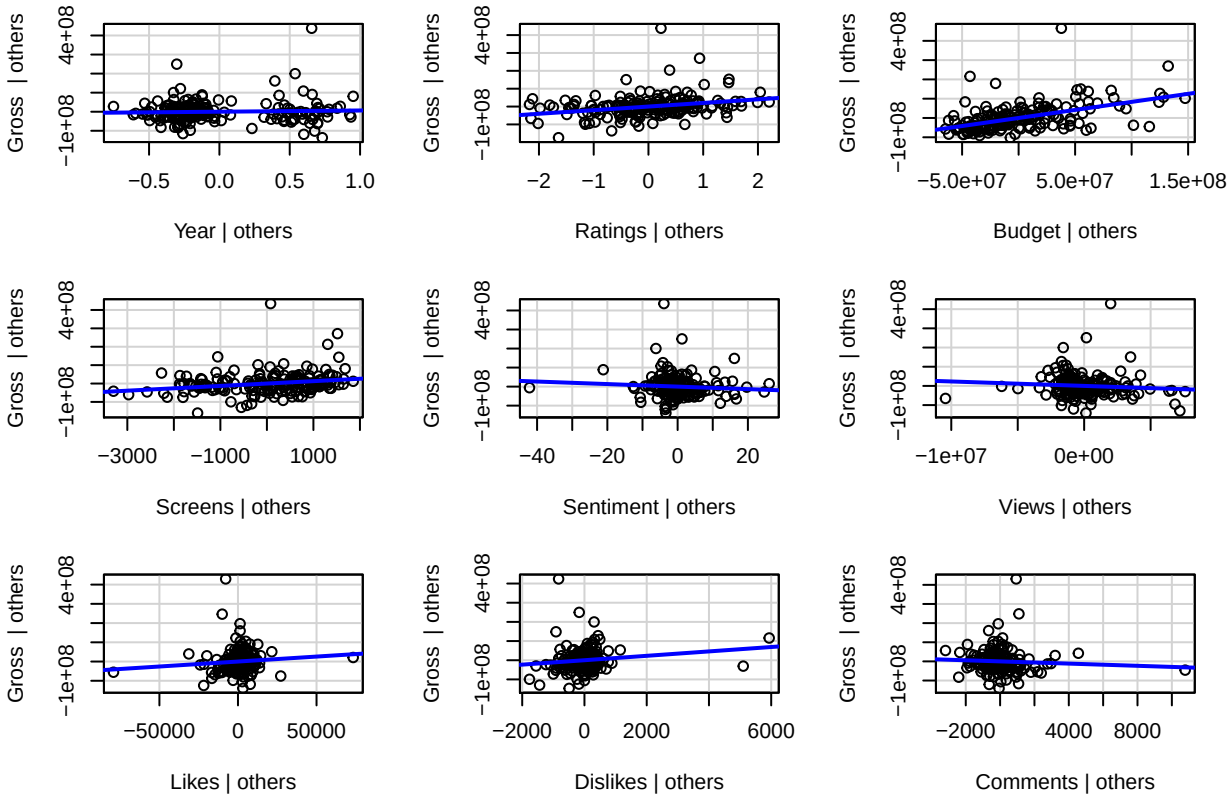
```
## [1] 183  14
```

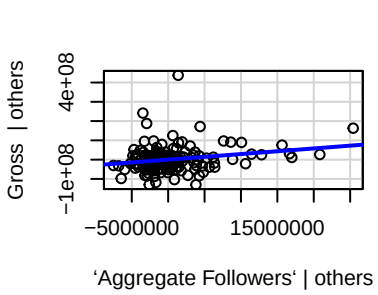
Appendix B

```r
# exploratory analysis
library(car)
full.lm <- lm(Gross ~ Year + Ratings + Genre + Budget + Screens + Sequel
              + Sentiment + Views + Likes + Dislikes + Comments +
                `Aggregate Followers`, data = csm_dataset) #except for Movie name
#summary(full.lm)
pairs(Gross ~  Year + Ratings + Budget + Screens + Sentiment + Views
      + Likes + Dislikes + Comments + `Aggregate Followers`, data = csm_dataset)
```




```
# AVPlots
num_data.lm<- lm(Gross ~  Year + Ratings + Budget + Screens + Sentiment
                 + Views + Likes + Dislikes + Comments + `Aggregate Followers`, data = csm_data
avPlots(num_data.lm, id=FALSE)
```


Gross | others
−1e+08
4e+08
−0.5
0.0
0.5
1.0
Year | others
−2
−1
0
1
2
Ratings | others
−5.0e+07
5.0e+07
1.5e+08
Budget | others
−3000
−1000
1000
Screens | others
−40
−20
0
20
Sentiment | others
−1e+07
0e+00
Views | others
−50000
0
50000
Likes | others
−2000
0
2000
6000
Dislikes | others
−2000
4000
8000
Comments | others



Appendix C

```
#USING BACKWARDS SELECTION TO FIND THE BEST MODEL/REDUCE PREDICTORS
mod.0 <- lm(Gross ~ 1, data = csm_dataset)
step(full.lm, scope = list(lower = mod.0, upper = full.lm),
     trace = 1, direction = "backward")
```

```
## Start:  AIC=6573.09
## Gross ~ Year + Ratings + Genre + Budget + Screens + Sequel +
##     Sentiment + Views + Likes + Dislikes + Comments + `Aggregate Followers`
##
##                         Df  Sum of Sq        RSS    AIC
## - Genre                  1 5.6559e+13 6.3097e+17 6571.1
## - Year                   1 3.0539e+14 6.3122e+17 6571.2
## - Comments               1 1.0237e+15 6.3194e+17 6571.4
## - Sentiment              1 1.7512e+15 6.3266e+17 6571.6
## - Likes                  1 3.6033e+15 6.3452e+17 6572.1
## - Views                  1 4.2689e+15 6.3518e+17 6572.3
## <none>                                6.3091e+17 6573.1
## - Sequel                 1 1.3401e+16 6.4431e+17 6574.9
## - Dislikes               1 1.4157e+16 6.4507e+17 6575.2
```

```
## - `Aggregate Followers`  1 2.7079e+16 6.5799e+17 6578.8
## - Screens                1 3.7986e+16 6.6890e+17 6581.8
## - Ratings                1 6.0352e+16 6.9126e+17 6587.8
## - Budget                 1 1.5264e+17 7.8356e+17 6610.7
##
## Step:  AIC=6571.11
## Gross ~ Year + Ratings + Budget + Screens + Sequel + Sentiment +
##     Views + Likes + Dislikes + Comments + `Aggregate Followers`
##
##                         Df  Sum of Sq        RSS    AIC
## - Year                   1 2.9628e+14 6.3127e+17 6569.2
## - Comments               1 1.1486e+15 6.3212e+17 6569.4
## - Sentiment              1 1.7274e+15 6.3270e+17 6569.6
## - Likes                  1 3.7274e+15 6.3470e+17 6570.2
## - Views                  1 4.2130e+15 6.3518e+17 6570.3
## <none>                                6.3097e+17 6571.1
## - Sequel                 1 1.3475e+16 6.4444e+17 6573.0
## - Dislikes               1 1.4106e+16 6.4508e+17 6573.2
## - `Aggregate Followers`  1 2.7211e+16 6.5818e+17 6576.8
## - Screens                1 3.8085e+16 6.6905e+17 6579.8
## - Ratings                1 6.0462e+16 6.9143e+17 6585.9
## - Budget                 1 1.5389e+17 7.8486e+17 6609.0
##
## Step:  AIC=6569.19
## Gross ~ Ratings + Budget + Screens + Sequel + Sentiment + Views +
##     Likes + Dislikes + Comments + `Aggregate Followers`
##
##                         Df  Sum of Sq        RSS    AIC
## - Sentiment              1 1.5096e+15 6.3277e+17 6567.6
## - Comments               1 1.5249e+15 6.3279e+17 6567.6
## - Views                  1 4.0726e+15 6.3534e+17 6568.4
## - Likes                  1 4.1855e+15 6.3545e+17 6568.4
## <none>                                6.3127e+17 6569.2
## - Sequel                 1 1.4489e+16 6.4575e+17 6571.3
## - Dislikes               1 1.4983e+16 6.4625e+17 6571.5
## - `Aggregate Followers`  1 2.6974e+16 6.5824e+17 6574.9
## - Screens                1 4.1172e+16 6.7244e+17 6578.8
## - Ratings                1 6.0408e+16 6.9167e+17 6583.9
## - Budget                 1 1.5365e+17 7.8492e+17 6607.1
##
## Step:  AIC=6567.63
## Gross ~ Ratings + Budget + Screens + Sequel + Views + Likes +
##     Dislikes + Comments + `Aggregate Followers`
##
##                         Df  Sum of Sq        RSS    AIC
## - Comments               1 1.4493e+15 6.3422e+17 6566.0
## - Likes                  1 4.0056e+15 6.3678e+17 6566.8
## - Views                  1 4.0367e+15 6.3681e+17 6566.8
```

```
## <none>                              6.3277e+17 6567.6
## - Dislikes               1 1.4650e+16 6.4742e+17 6569.8
## - Sequel                 1 1.6209e+16 6.4898e+17 6570.3
## - `Aggregate Followers`  1 2.8220e+16 6.6099e+17 6573.6
## - Screens                1 4.2516e+16 6.7529e+17 6577.5
## - Ratings                1 5.9023e+16 6.9180e+17 6582.0
## - Budget                 1 1.5228e+17 7.8506e+17 6605.1
##
## Step:  AIC=6566.05
## Gross ~ Ratings + Budget + Screens + Sequel + Views + Likes +
##     Dislikes + `Aggregate Followers`
##
##                          Df  Sum of Sq        RSS    AIC
## - Likes                   1 3.0962e+15 6.3732e+17 6564.9
## - Views                   1 4.8024e+15 6.3903e+17 6565.4
## <none>                              6.3422e+17 6566.0
## - Dislikes                1 1.3520e+16 6.4774e+17 6567.9
## - Sequel                  1 1.8438e+16 6.5266e+17 6569.3
## - `Aggregate Followers`   1 3.0300e+16 6.6452e+17 6572.6
## - Screens                 1 4.3857e+16 6.7808e+17 6576.3
## - Ratings                 1 6.3059e+16 6.9728e+17 6581.4
## - Budget                  1 1.5611e+17 7.9033e+17 6604.3
##
## Step:  AIC=6564.94
## Gross ~ Ratings + Budget + Screens + Sequel + Views + Dislikes +
##     `Aggregate Followers`
##
##                          Df  Sum of Sq        RSS    AIC
## - Views                   1 1.9108e+15 6.3923e+17 6563.5
## <none>                              6.3732e+17 6564.9
## - Dislikes                1 1.1370e+16 6.4869e+17 6566.2
## - Sequel                  1 1.8964e+16 6.5628e+17 6568.3
## - `Aggregate Followers`   1 2.9023e+16 6.6634e+17 6571.1
## - Screens                 1 4.6521e+16 6.8384e+17 6575.8
## - Ratings                 1 6.8290e+16 7.0561e+17 6581.6
## - Budget                  1 1.5302e+17 7.9034e+17 6602.3
##
## Step:  AIC=6563.49
## Gross ~ Ratings + Budget + Screens + Sequel + Dislikes + `Aggregate Followers`
##
##                          Df  Sum of Sq        RSS    AIC
## <none>                              6.3923e+17 6563.5
## - Dislikes                1 1.3493e+16 6.5272e+17 6565.3
## - Sequel                  1 1.9128e+16 6.5836e+17 6566.9
## - `Aggregate Followers`   1 2.7337e+16 6.6657e+17 6569.2
## - Screens                 1 4.6241e+16 6.8547e+17 6574.3
## - Ratings                 1 6.6915e+16 7.0615e+17 6579.7
## - Budget                  1 1.5579e+17 7.9502e+17 6601.4
```

```
##
## Call:
## lm(formula = Gross ~ Ratings + Budget + Screens + Sequel + Dislikes +
##     `Aggregate Followers`, data = csm_dataset)
##
## Coefficients:
##           (Intercept)                Ratings                 Budget
##            -1.577e+08              2.088e+07              7.121e-01
##               Screens                 Sequel               Dislikes
##             1.425e+04              1.175e+07              6.967e+03
## `Aggregate Followers`
##             2.541e+00
```

```
n<- length(Year)
n
```

```
## [1] 231
```

```
step(full.lm, scope = list(lower = mod.0, upper = full.lm),
     direction = 'backward', k = log(n), trace = 1)
```

```
## Start:  AIC=6617.84
## Gross ~ Year + Ratings + Genre + Budget + Screens + Sequel +
##     Sentiment + Views + Likes + Dislikes + Comments + `Aggregate Followers`
##
##                          Df  Sum of Sq        RSS    AIC
## - Genre                   1 5.6559e+13 6.3097e+17 6612.4
## - Year                    1 3.0539e+14 6.3122e+17 6612.5
## - Comments                1 1.0237e+15 6.3194e+17 6612.7
## - Sentiment               1 1.7512e+15 6.3266e+17 6612.9
## - Likes                   1 3.6033e+15 6.3452e+17 6613.4
## - Views                   1 4.2689e+15 6.3518e+17 6613.6
## - Sequel                  1 1.3401e+16 6.4431e+17 6616.2
## - Dislikes                1 1.4157e+16 6.4507e+17 6616.5
## <none>                                 6.3091e+17 6617.8
## - `Aggregate Followers`   1 2.7079e+16 6.5799e+17 6620.1
## - Screens                 1 3.7986e+16 6.6890e+17 6623.1
## - Ratings                 1 6.0352e+16 6.9126e+17 6629.1
## - Budget                  1 1.5264e+17 7.8356e+17 6652.1
##
## Step:  AIC=6612.42
## Gross ~ Year + Ratings + Budget + Screens + Sequel + Sentiment +
##     Views + Likes + Dislikes + Comments + `Aggregate Followers`
##
##                          Df  Sum of Sq        RSS    AIC
## - Year                    1 2.9628e+14 6.3127e+17 6607.1
```

```
## - Comments               1 1.1486e+15 6.3212e+17 6607.3
## - Sentiment              1 1.7274e+15 6.3270e+17 6607.5
## - Likes                  1 3.7274e+15 6.3470e+17 6608.1
## - Views                  1 4.2130e+15 6.3518e+17 6608.2
## - Sequel                 1 1.3475e+16 6.4444e+17 6610.8
## - Dislikes               1 1.4106e+16 6.4508e+17 6611.0
## <none>                                6.3097e+17 6612.4
## - `Aggregate Followers`  1 2.7211e+16 6.5818e+17 6614.7
## - Screens                1 3.8085e+16 6.6905e+17 6617.7
## - Ratings                1 6.0462e+16 6.9143e+17 6623.7
## - Budget                 1 1.5389e+17 7.8486e+17 6646.9
##
## Step:  AIC=6607.06
## Gross ~ Ratings + Budget + Screens + Sequel + Sentiment + Views +
##     Likes + Dislikes + Comments + `Aggregate Followers`
##
##                          Df  Sum of Sq        RSS    AIC
## - Sentiment               1 1.5096e+15 6.3277e+17 6602.1
## - Comments                1 1.5249e+15 6.3279e+17 6602.1
## - Views                   1 4.0726e+15 6.3534e+17 6602.8
## - Likes                   1 4.1855e+15 6.3545e+17 6602.8
## - Sequel                  1 1.4489e+16 6.4575e+17 6605.8
## - Dislikes                1 1.4983e+16 6.4625e+17 6605.9
## <none>                                 6.3127e+17 6607.1
## - `Aggregate Followers`   1 2.6974e+16 6.5824e+17 6609.3
## - Screens                 1 4.1172e+16 6.7244e+17 6613.2
## - Ratings                 1 6.0408e+16 6.9167e+17 6618.3
## - Budget                  1 1.5365e+17 7.8492e+17 6641.5
##
## Step:  AIC=6602.06
## Gross ~ Ratings + Budget + Screens + Sequel + Views + Likes +
##     Dislikes + Comments + `Aggregate Followers`
##
##                          Df  Sum of Sq        RSS    AIC
## - Comments                1 1.4493e+15 6.3422e+17 6597.0
## - Likes                   1 4.0056e+15 6.3678e+17 6597.8
## - Views                   1 4.0367e+15 6.3681e+17 6597.8
## - Dislikes                1 1.4650e+16 6.4742e+17 6600.8
## - Sequel                  1 1.6209e+16 6.4898e+17 6601.2
## <none>                                 6.3277e+17 6602.1
## - `Aggregate Followers`   1 2.8220e+16 6.6099e+17 6604.6
## - Screens                 1 4.2516e+16 6.7529e+17 6608.5
## - Ratings                 1 5.9023e+16 6.9180e+17 6612.9
## - Budget                  1 1.5228e+17 7.8506e+17 6636.1
##
## Step:  AIC=6597.03
## Gross ~ Ratings + Budget + Screens + Sequel + Views + Likes +
##     Dislikes + `Aggregate Followers`
```

```
##
##                          Df  Sum of Sq        RSS    AIC
## - Likes                   1 3.0962e+15 6.3732e+17 6592.5
## - Views                   1 4.8024e+15 6.3903e+17 6593.0
## - Dislikes                1 1.3520e+16 6.4774e+17 6595.4
## - Sequel                  1 1.8438e+16 6.5266e+17 6596.8
## <none>                                 6.3422e+17 6597.0
## - `Aggregate Followers`   1 3.0300e+16 6.6452e+17 6600.1
## - Screens                 1 4.3857e+16 6.7808e+17 6603.8
## - Ratings                 1 6.3059e+16 6.9728e+17 6608.9
## - Budget                  1 1.5611e+17 7.9033e+17 6631.9
##
## Step:  AIC=6592.48
## Gross ~ Ratings + Budget + Screens + Sequel + Views + Dislikes +
##     `Aggregate Followers`
##
##                          Df  Sum of Sq        RSS    AIC
## - Views                   1 1.9108e+15 6.3923e+17 6587.6
## - Dislikes                1 1.1370e+16 6.4869e+17 6590.3
## - Sequel                  1 1.8964e+16 6.5628e+17 6592.4
## <none>                                 6.3732e+17 6592.5
## - `Aggregate Followers`   1 2.9023e+16 6.6634e+17 6595.2
## - Screens                 1 4.6521e+16 6.8384e+17 6599.9
## - Ratings                 1 6.8290e+16 7.0561e+17 6605.7
## - Budget                  1 1.5302e+17 7.9034e+17 6626.4
##
## Step:  AIC=6587.59
## Gross ~ Ratings + Budget + Screens + Sequel + Dislikes + `Aggregate Followers`
##
##                          Df  Sum of Sq        RSS    AIC
## - Dislikes                1 1.3493e+16 6.5272e+17 6586.0
## - Sequel                  1 1.9128e+16 6.5836e+17 6587.5
## <none>                                 6.3923e+17 6587.6
## - `Aggregate Followers`   1 2.7337e+16 6.6657e+17 6589.8
## - Screens                 1 4.6241e+16 6.8547e+17 6594.9
## - Ratings                 1 6.6915e+16 7.0615e+17 6600.4
## - Budget                  1 1.5579e+17 7.9502e+17 6622.1
##
## Step:  AIC=6585.97
## Gross ~ Ratings + Budget + Screens + Sequel + `Aggregate Followers`
##
##                          Df  Sum of Sq        RSS    AIC
## - Sequel                  1 1.5606e+16 6.6833e+17 6584.8
## <none>                                 6.5272e+17 6586.0
## - `Aggregate Followers`   1 2.8573e+16 6.8130e+17 6588.4
## - Ratings                 1 5.6444e+16 7.0917e+17 6595.7
## - Screens                 1 6.1509e+16 7.1423e+17 6597.0
## - Budget                  1 1.5725e+17 8.0997e+17 6620.0
```

```
##
## Step:  AIC=6584.85
## Gross ~ Ratings + Budget + Screens + `Aggregate Followers`
##
##                          Df  Sum of Sq        RSS    AIC
## <none>                                6.6833e+17 6584.8
## - `Aggregate Followers`   1 3.6260e+16 7.0459e+17 6589.1
## - Ratings                 1 5.3198e+16 7.2153e+17 6593.4
## - Screens                 1 6.0295e+16 7.2863e+17 6595.2
## - Budget                  1 2.2221e+17 8.9054e+17 6631.9

##
## Call:
## lm(formula = Gross ~ Ratings + Budget + Screens + `Aggregate Followers`,
##     data = csm_dataset)
##
## Coefficients:
##           (Intercept)                Ratings                 Budget
##            -1.269e+08              1.803e+07              7.952e-01
##               Screens  `Aggregate Followers`
##             1.584e+04              2.891e+00
```

Appendix D

```
# Since it is only 6 predictors, we can use Regsubsets to conduct exhausitive search.
library(leaps)
csm_models<- regsubsets(Gross ~  Ratings + Budget + Screens + Sequel +
                          Dislikes + `Aggregate Followers`, data = csm_dataset)
summary.csm<-summary(csm_models)

#picking the best regression model using reg subsets
summary.csm$which
```

```
##   (Intercept) Ratings Budget Screens Sequel Dislikes `Aggregate Followers`
## 1        TRUE   FALSE   TRUE   FALSE  FALSE    FALSE                 FALSE
## 2        TRUE   FALSE   TRUE    TRUE  FALSE    FALSE                 FALSE
## 3        TRUE    TRUE   TRUE    TRUE  FALSE    FALSE                 FALSE
## 4        TRUE    TRUE   TRUE    TRUE  FALSE    FALSE                  TRUE
## 5        TRUE    TRUE   TRUE    TRUE   TRUE    FALSE                  TRUE
## 6        TRUE    TRUE   TRUE    TRUE   TRUE     TRUE                  TRUE
```

```
#Criteria
# model with largest increase in R^2 (equivalent to smallest MSE)
summary.csm$rsq
```

```
## [1] 0.4857658 0.5254611 0.5606424 0.5832531 0.5929844 0.6013980
```

```
0.5254611-0.4857658
```

```
## [1] 0.0396953
```

```
0.5606424-0.5254611
```

```
## [1] 0.0351813
```

```
# model with largest adjusted R^2
summary.csm$adjr2
```

```
## [1] 0.4829248 0.5201884 0.5532789 0.5738880 0.5814867 0.5878093
```

```
# model with smallest Mallow's Cp
summary.csm$cp
```

```
## [1] 48.056593 32.529430 18.995350 11.011760  8.714989  7.000000
```

```
# model with lowest bic
summary.csm$bic
```

```
## [1] -111.2900 -120.7819 -129.6689 -134.1281 -133.2425 -131.8555
```

Appendix E

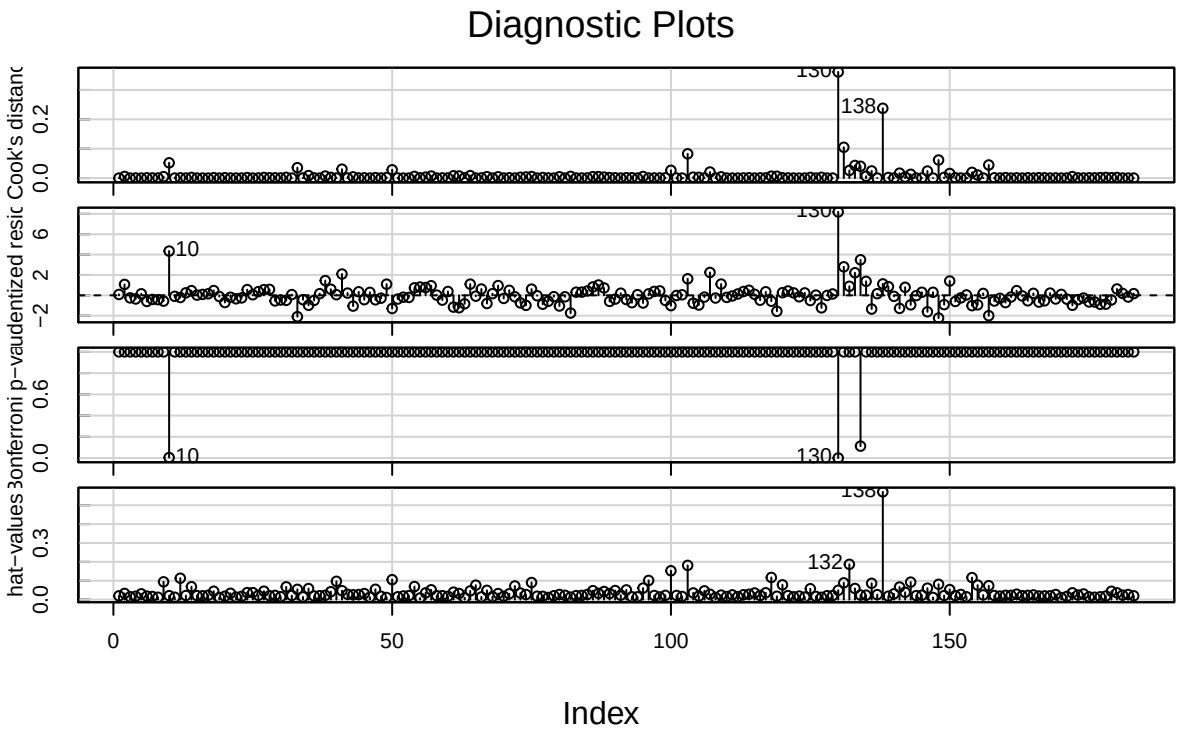
```
#SUMMARY TABLE
full2.lm <- lm(Gross ~  Ratings + Budget + Screens + Sequel + Dislikes +
                 `Aggregate Followers`, data = csm_dataset)
summary(full2.lm)
```

```
##
## Call:
## lm(formula = Gross ~ Ratings + Budget + Screens + Sequel + Dislikes +
##     `Aggregate Followers`, data = csm_dataset)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -127511153  -32245753   -4850498   21292820  410362338
##
## Coefficients:
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -1.577e+08  3.268e+07  -4.825 3.02e-06 ***
## Ratings                2.088e+07  4.865e+06   4.292 2.91e-05 ***
## Budget                 7.121e-01  1.087e-01   6.549 6.13e-10 ***
## Screens                1.425e+04  3.993e+03   3.568 0.000463 ***
## Sequel                 1.175e+07  5.119e+06   2.295 0.022920 *
## Dislikes               6.967e+03  3.615e+03   1.927 0.055536 .
## `Aggregate Followers`  2.541e+00  9.263e-01   2.743 0.006708 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60270000 on 176 degrees of freedom
## Multiple R-squared:  0.6014, Adjusted R-squared:  0.5878
## F-statistic: 44.26 on 6 and 176 DF,  p-value: < 2.2e-16
```

Appendix F

```
#CHECKING FOR OUTLIERS
influenceIndexPlot(full2.lm, id=TRUE)
```



```r
#remove the two data points
removed_dataset1<- csm_dataset[-c(10,138),]
#removed_dataset2<- csm_dataset[-c(130, 132, 138),]
full2rm_1.lm <- lm(Gross ~  Ratings + Budget + Screens + Sequel + Dislikes +
                     `Aggregate Followers`, data = removed_dataset1)

#compare the quadratic mean function
summary(full2.lm)
```

```
##
## Call:
## lm(formula = Gross ~ Ratings + Budget + Screens + Sequel + Dislikes +
##     `Aggregate Followers`, data = csm_dataset)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -127511153  -32245753   -4850498   21292820  410362338
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -1.577e+08  3.268e+07  -4.825 3.02e-06 ***
## Ratings                2.088e+07  4.865e+06   4.292 2.91e-05 ***
## Budget                 7.121e-01  1.087e-01   6.549 6.13e-10 ***
## Screens                1.425e+04  3.993e+03   3.568 0.000463 ***
## Sequel                 1.175e+07  5.119e+06   2.295 0.022920 *
## Dislikes               6.967e+03  3.615e+03   1.927 0.055536 .
## `Aggregate Followers`  2.541e+00  9.263e-01   2.743 0.006708 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60270000 on 176 degrees of freedom
## Multiple R-squared:  0.6014, Adjusted R-squared:  0.5878
## F-statistic: 44.26 on 6 and 176 DF,  p-value: < 2.2e-16
```

```r
summary(full2rm_1.lm) #no difference
```

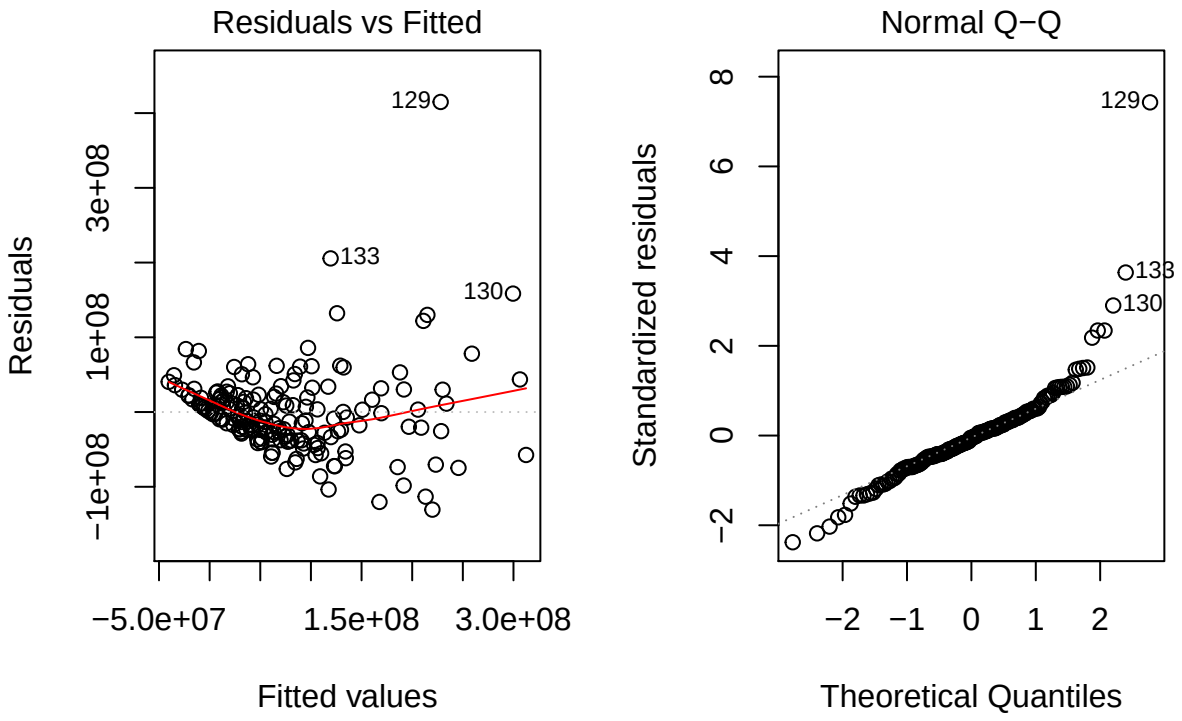
```
##
## Call:
## lm(formula = Gross ~ Ratings + Budget + Screens + Sequel + Dislikes +
##     `Aggregate Followers`, data = removed_dataset1)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -129417654  -29396772   -3158620   21205880  413563357
##
## Coefficients:
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -1.465e+08  3.127e+07  -4.685 5.62e-06 ***
## Ratings                1.929e+07  4.646e+06   4.152 5.16e-05 ***
## Budget                 7.343e-01  1.036e-01   7.086 3.31e-11 ***
## Screens                1.335e+04  3.849e+03   3.469 0.000659 ***
## Sequel                 1.182e+07  4.896e+06   2.414 0.016837 *
## Dislikes               2.847e+03  5.142e+03   0.554 0.580484
## `Aggregate Followers`  2.846e+00  8.865e-01   3.211 0.001578 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 57370000 on 174 degrees of freedom
## Multiple R-squared:  0.6234, Adjusted R-squared:  0.6105
## F-statistic: 48.01 on 6 and 174 DF,  p-value: < 2.2e-16
```

Appendix G

```
#CHECKING RESIDUALS VS FITTED AND QQ PLOT
full3.lm <- lm(Gross ~  Ratings + Budget + Screens + Sequel + `Aggregate Followers`
               , data = removed_dataset1)

#residuals vs fitted, Q-Q plot
par(mfrow = c(1,2))
plot(full3.lm, which=1)
plot(full3.lm,which=2)
```



Appendix H

```
#TRANSFORMING PREDICTORS
# Do not include categorical variable type
Trans.csm1 <- powerTransform(cbind(Ratings,Budget,Screens,Sequel,
                                   `Aggregate Followers`)~1,data =removed_dataset1)
summary(Trans.csm1)
```

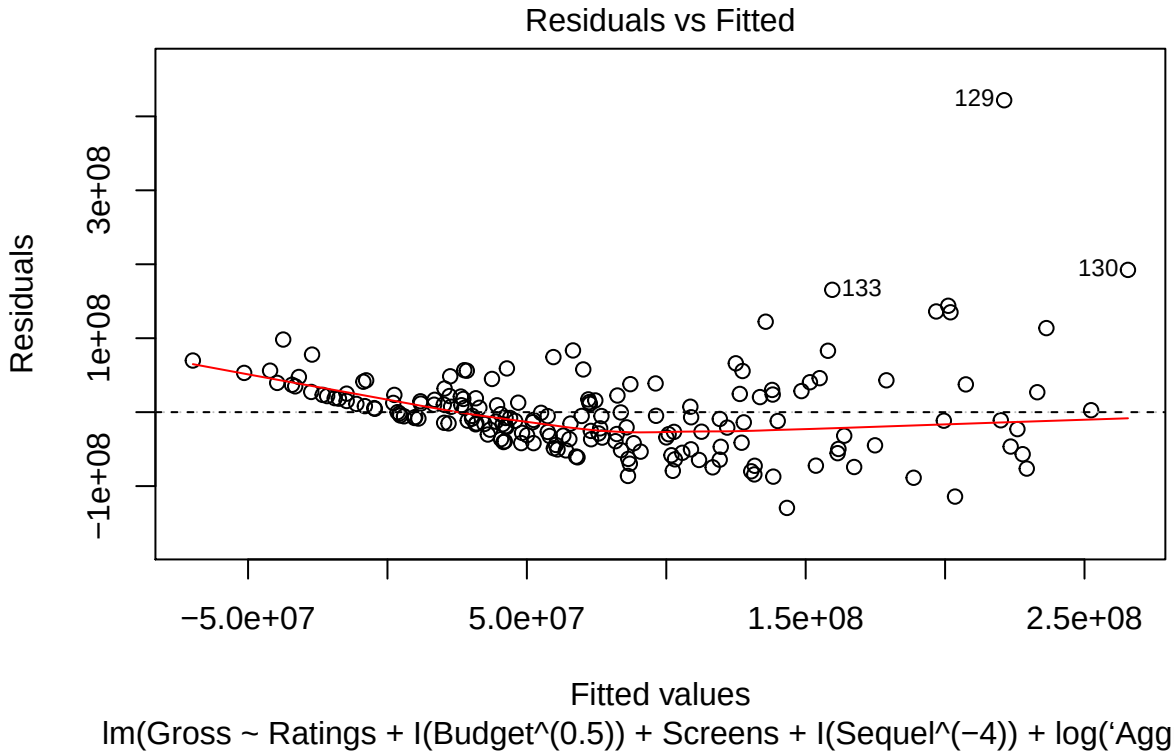
```
## bcPower Transformations to Multinormality
##                      Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Ratings                 1.6192        1.00       0.8996       2.3387
## Budget                  0.2436        0.24       0.1603       0.3268
## Screens                 0.7944        0.79       0.6545       0.9343
## Sequel                 -4.0035       -4.00      -4.7374      -3.2696
## Aggregate Followers     0.1606        0.16       0.0970       0.2243
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                     LRT df       pval
## LR test, lambda = (0 0 0 0 0) 482.2412  5 < 2.22e-16
##
```

```
## Likelihood ratio test that no transformations are needed
##                                       LRT df       pval
## LR test, lambda = (1 1 1 1 1) 1052.728  5 < 2.22e-16
```

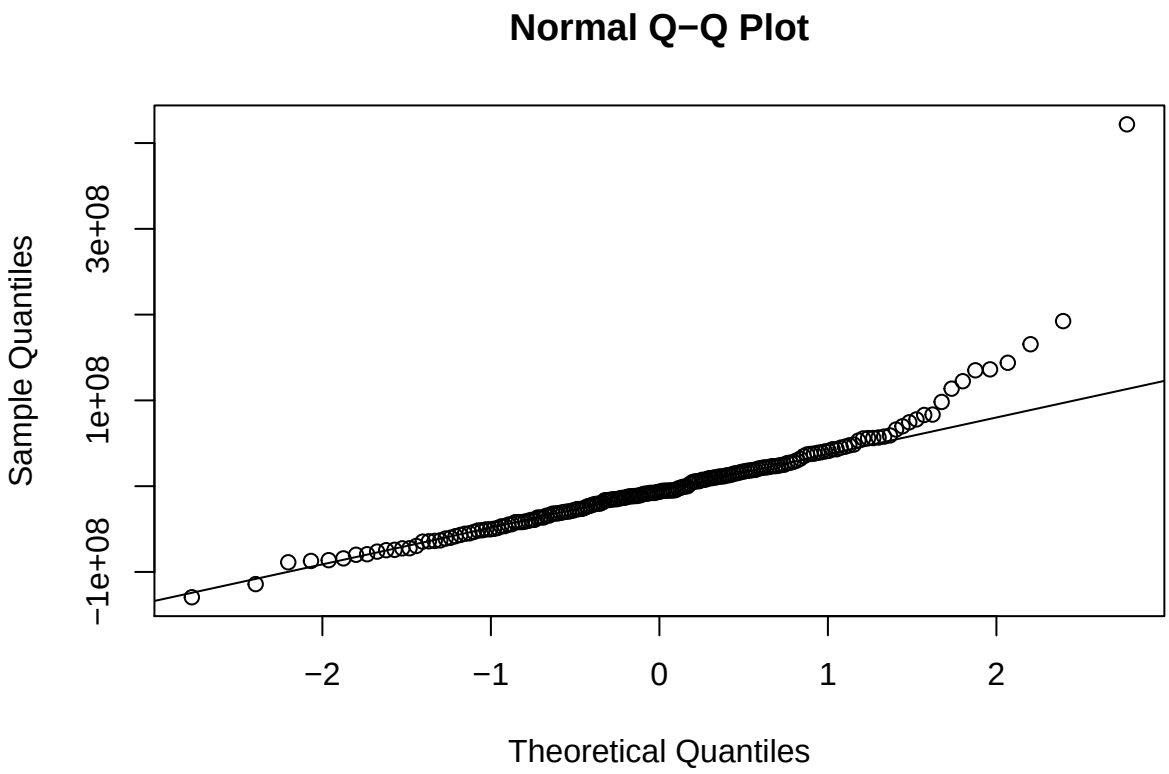
Appendix I

```
#Transform and check for if the assumptions are improved
csmtrans1.lm <- lm(Gross ~  Ratings + I(Budget**(0.5)) + Screens
                   + I(Sequel**(-4)) +  log(`Aggregate Followers`), data = removed_dataset1)

#checking the residuals and qqplot
plot(csmtrans1.lm, which=1)
abline(h = 0, lty = 2)
```

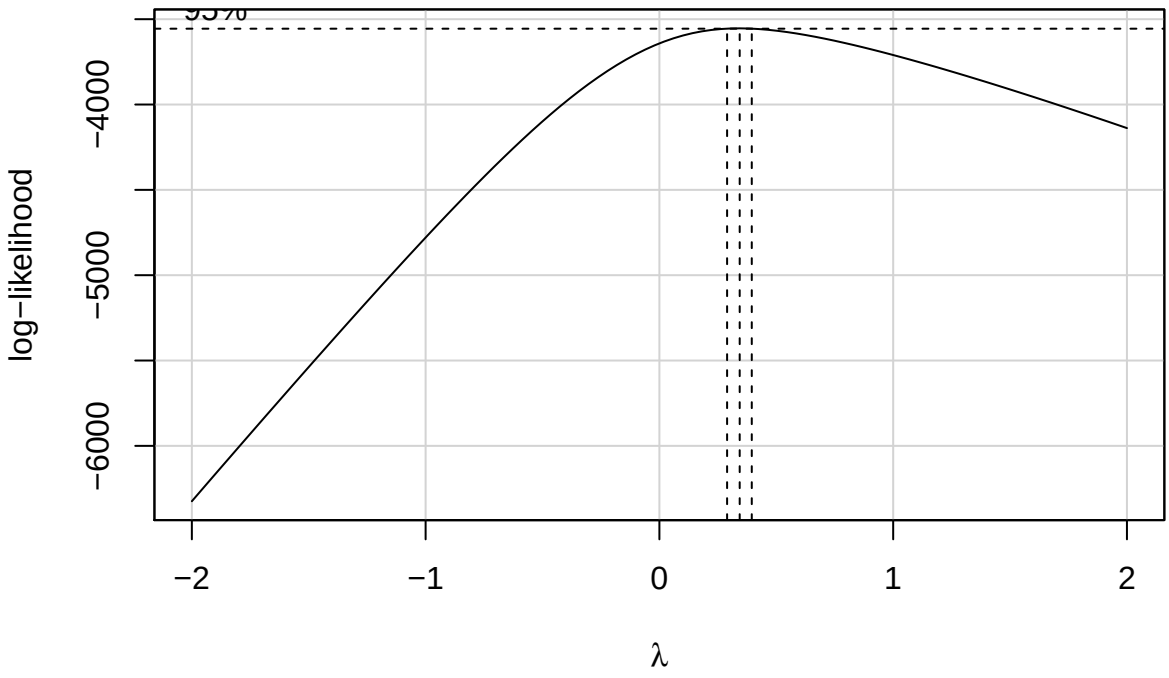



```
#QQ plot
qqnorm(resid(csmtrans1.lm))
qqline(resid(csmtrans1.lm))
```



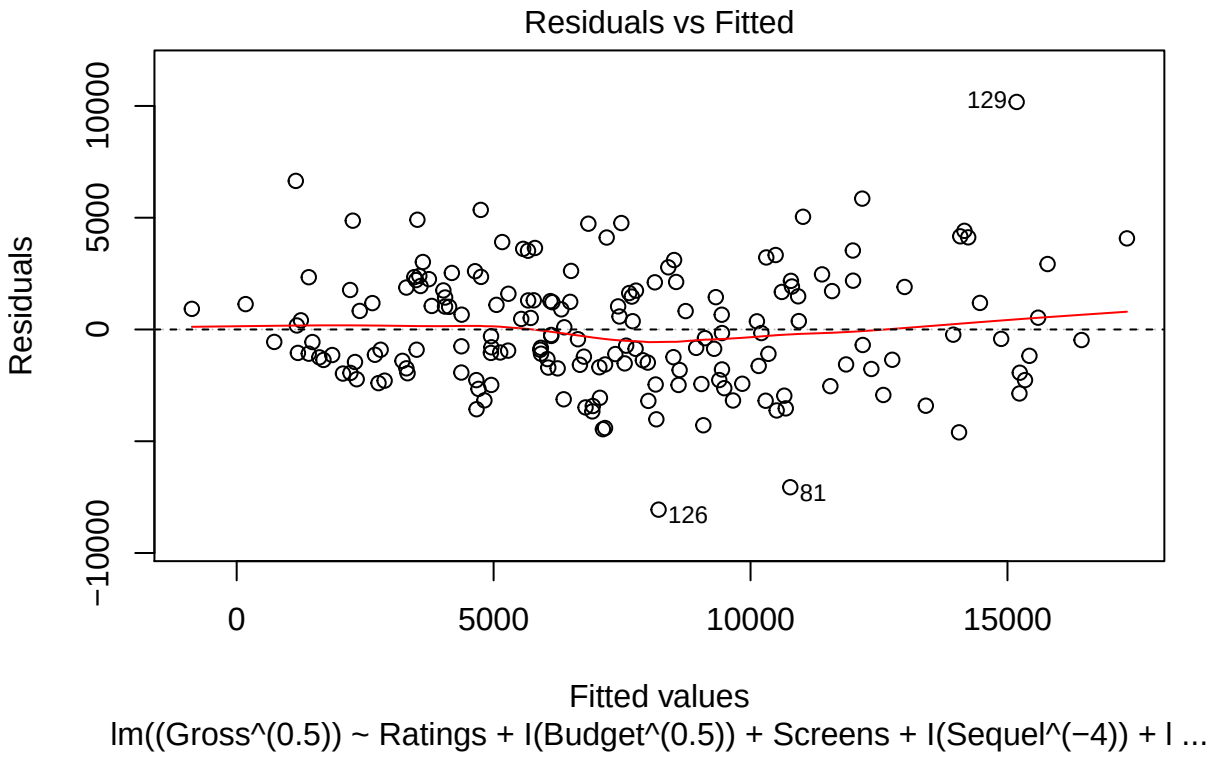
Appendix J

```
#TRANSFORMATION ON RESPONSE
Trans.csm2 <- boxCox(csmtrans1.lm)
```

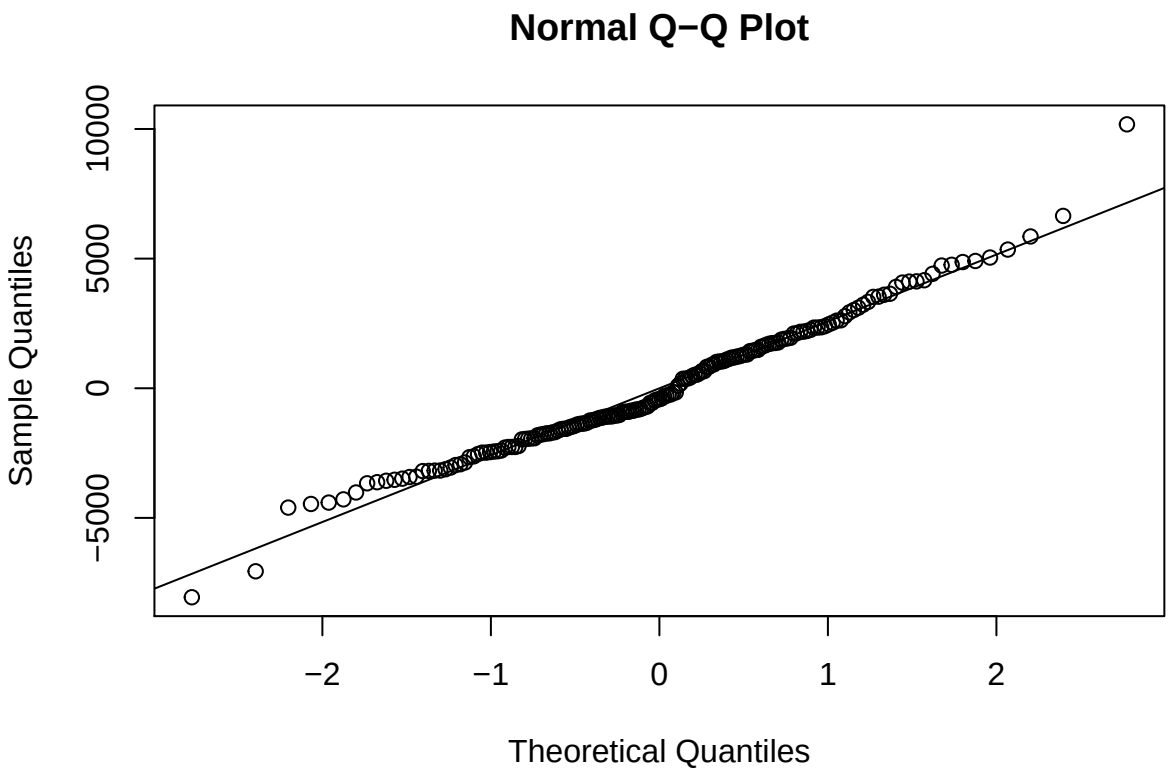



```
csmtrans2.lm <-  lm((Gross**(0.5)) ~  Ratings + I(Budget**(0.5)) + Screens
                    + I(Sequel**(-4)) + log(`Aggregate Followers`), data = removed_dataset1)

#checking the residuals and qqplot
plot(csmtrans2.lm, which=1)
abline(h = 0, lty = 2)
```




```
#QQ plot
qqnorm(resid(csmtrans2.lm))
qqline(resid(csmtrans2.lm))
```



Appendix K

```
#INTERACTION VARIABLES
# Full model
full4.lm <- lm((Gross**(0.5)) ~ Ratings + I(Budget**(0.5)) + Screens
               + I(Sequel**(-4)) + log(`Aggregate Followers`), data = removed_dataset1)

# We are interested in Genre and its interaction with other variables
# Budget and Screens
full4.1.lm <- lm((Gross**(0.5)) ~  Ratings + I(Budget**(0.5))*Screens
                 + I(Sequel**(-4)) + log(`Aggregate Followers`), data = removed_dataset1)
anova(full4.lm,full4.1.lm) # no interaction term
```

```
## Analysis of Variance Table
##
## Model 1: (Gross^(0.5)) ~ Ratings + I(Budget^(0.5)) + Screens + I(Sequel^(-4)) +
##     log(`Aggregate Followers`)
## Model 2: (Gross^(0.5)) ~ Ratings + I(Budget^(0.5)) * Screens + I(Sequel^(-4)) +
##     log(`Aggregate Followers`)
##   Res.Df        RSS Df Sum of Sq     F Pr(>F)
## 1    175 1239951495
```

```
## 2    174 1239360395  1    591101 0.083 0.7736
```

```
# Budget and Ratings
# Question 1
full4.2.lm <- lm((Gross**(0.5)) ~  Ratings*I(Budget**(0.5)) + Screens + I(Sequel**(-4)) + log(
anova(full4.lm,full4.2.lm)
```

```
## Analysis of Variance Table
##
## Model 1: (Gross^(0.5)) ~ Ratings + I(Budget^(0.5)) + Screens + I(Sequel^(-4)) +
##     log(`Aggregate Followers`)
## Model 2: (Gross^(0.5)) ~ Ratings * I(Budget^(0.5)) + Screens + I(Sequel^(-4)) +
##     log(`Aggregate Followers`)
##   Res.Df        RSS Df Sum of Sq      F   Pr(>F)
## 1    175 1239951495
## 2    174 1166375673  1  73575822 10.976 0.001123 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(full4.2.lm)
```

```
##
## Call:
## lm(formula = (Gross^(0.5)) ~ Ratings * I(Budget^(0.5)) + Screens +
##     I(Sequel^(-4)) + log(`Aggregate Followers`), data = removed_dataset1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7834.2 -1781.1  -236.7  1744.4  9872.8
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.390e+02  2.520e+03   0.095  0.92454
## Ratings                      5.211e+01  3.672e+02   0.142  0.88732
## I(Budget^(0.5))             -6.349e-01  3.400e-01  -1.867  0.06352 .
## Screens                      1.157e+00  1.766e-01   6.551 6.23e-10 ***
## I(Sequel^(-4))              -1.685e+03  5.210e+02  -3.235  0.00146 **
## log(`Aggregate Followers`)   1.807e+02  9.048e+01   1.997  0.04739 *
## Ratings:I(Budget^(0.5))      1.664e-01  5.023e-02   3.313  0.00112 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2589 on 174 degrees of freedom
## Multiple R-squared:  0.7071, Adjusted R-squared:  0.697
## F-statistic: 70.01 on 6 and 174 DF,  p-value: < 2.2e-16
```

Appendix L

```
# Answering Questions of interest
attach(removed_dataset1)
logaf <-log(`Aggregate Followers`)
full4.2_modified.lm<- lm((Gross**(0.5)) ~  Ratings*I(Budget**(0.5)) + Screens
                         + I(Sequel**(-4)) + logaf, data = removed_dataset1)
logafmean <- log(mean(`Aggregate Followers`))

#mean values for all predictors
avgvals<- data.frame(Ratings = mean(Ratings), Budget= mean(Budget), Screens= mean(Screens),
                     Sequel=mean(Sequel), logaf = logafmean)

# QUESTION 2.1
# confidence interval GrossIncome^(0.5) - for all movies
predict(full4.2_modified.lm, avgvals, interval = 'confidence', level = 0.95)
```

```
##        fit      lwr      upr
## 1 8743.272 8039.846 9446.698
```

```
# fit = 8743.272 lwr=8039.846 upr= 9446.698
# convert back to original Yi from transformed response
orifit <- 8743.272**2
# convert the confidence interval
ci_lwr <- 8039.846**2
ci_upr <- 9446.698**2
c(orifit, ci_lwr, ci_upr)
```

```
## [1] 76444805 64639124 89240103
```

```
# QUESTION 2.2
# prediction interval GrossIncome^(0.5) - for 1 movie
predict(full4.2_modified.lm, avgvals, interval = 'prediction', level = 0.95)
```

```
##        fit      lwr     upr
## 1 8743.272 3585.049 13901.5
```

```
# fit = 8743.272 lwr = 3585.049 upr = 13901.5
# convert the prediction interval
pi_lwr <- 3585.049**2
pi_upr <- 13901.5**2
c(orifit, pi_lwr, pi_upr)
```

```
## [1]  76444805  12852576 193251702
```

Appendix M

```
# QUESTION 3
# We are interested in categorical variables year and genre
# Year
np.csm_year.lm<- lm(log(Gross) ~ Year *Screens, data = removed_dataset1)
par.csm_year.lm<- lm(log(Gross) ~ Year + Screens, data = removed_dataset1)
# which model is better?
anova(par.csm_year.lm,np.csm_year.lm) # parallel model preferred
```

```
## Analysis of Variance Table
##
## Model 1: log(Gross) ~ Year + Screens
## Model 2: log(Gross) ~ Year * Screens
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    178 487.90
## 2    177 485.59  1    2.3159 0.8442 0.3595
```

```
yhat <- fitted(par.csm_year.lm)
yhatA <- yhat[Year == 2014]
yhatB <- yhat[Year == 2015]

plot(Screens, log(Gross), type = 'n', xlab = 'Screens', ylab = 'Gross Income',
     main = 'Gross Income vs Screens')
points(Screens[Year == 2014], log(Gross)[Year == 2014], col = 'blue', lwd=2)
points(Screens[Year == 2015], log(Gross)[Year == 2015], col = 'red', lwd=2)
lines(Screens[Year == 2014], yhatA, col = 'blue')
lines(Screens[Year == 2015], yhatB, col = 'red')
legend('bottomright', bty = 'n', col = c('blue','red'), c('2014', '2015'),
       lty = c(1,1), text.col =c('blue','red'), pch = c(1, 1))
```


```
par.csm_year.lm # Year: 0 - 2014; 1- 2015
```

```
##
## Call:
## lm(formula = log(Gross) ~ Year + Screens, data = removed_dataset1)
##
## Coefficients:
## (Intercept)         Year      Screens
##   502.76366     -0.24231      0.00103
```