Understanding Impact of Movie Characteristics on Number of Votes Received

Lucas Mazza mazzal@purdue.edu
Nikhil Mirpuri nmirpuri@purdue.edu
Rucha Soman rssoman@purdue.edu
Sanjana Kaushik kaushik7@purdue.edu
Varun Gannavarapu vgannava@purdue.edu

***Abstract*—Viewership of movies has increasingly become a common pastime since the late 19^{th} with the introduction of the projector. As technology has continuously advanced, viewership has grown to an extent where it has claimed a substantial share of the entertainment industry. While the growth of this industry remains undeterred overall, the impact of an individual movie can be best surmised by the number of individuals it left a lasting impact on, to the extent they are willing to publicly discuss their opinion of the movie. Our study presents a systematic approach to unveiling how different characteristics of movies may lead to an increased amount of discussion, thereby leading to the existence of votes, whether positive or negative. We use a variety of statistical tests to measure for among others, statistical significance, variability, normality, influential points and multicollinearity issues. We conclude by discussing the overarching impact of our results, and outline some potential future research questions in this area.**

***Index Terms*—k-fold validation, multicollinearity, normality, weighted-least squares**

I. Background and Related Work

As movies continue to have an increased influence on the entertainment industry, there are multiple ways to gauge how important they are with respect to humans. An understanding of the importance of movies can be best gauged by user interaction. Studies have suggested that likeability helps improve interpersonal relationships [1] among others. As such, this paper looks to build off this understanding. In particular, we observe what is the relationship between movies and voting on movies. That is, how many people will watch a movie, and then vote on a movie, based on a variety of characteristics in the movie.

Although there isn't a lot of research done pertaining to movies and votes about the movie specifically, there are still a few studies of note that relate to movies in general. In this section we discuss said papers, and what impact they may have had on this field of study.

Public Opinion There exist a few studies that look at how viewership of movies correlate with public opinion. For instance, Riley [2] looks at how public response differs with professional film reviews. A similar study was done by Rai [3], in which they look at how Hollywood reviews impact American viewing audiences. Our study is related in that we also look at professional film reviews as one of many characteristics we consider. However, we don't take public response in terms of viewership as our dependent variable. Instead, we look at it as how many individuals interact with the movie via voting, as our dependent variable. This provides a different lens of context, in that we consider a movie as having a profound impact on someone if they take the time to cast a vote about the movie, not just watch it.

Movie Ratings There are various ways that exist to rate a movie. For example, Penthey [4] observes the impact of how movie reviews impacts consumers and their interest in watching a given movie. Another study of note was conducted by Moon [5], in which they look to study the dynamic effect of multiple factors on viewer satisfaction, one of which was rating. Similar to what was noted above, we again consider movie ratings, rather two types of ratings, as just a couple of many factors that contribute to voting on movies. In addition, we aren't directly looking at a positive or negative reaction as the response to movies, rather we just want to see if individuals are interacting with the movie in general, and what drives these interactions.

II. Experimental Design

Our overarching research question that we operated under was: **How do different characteristics of movies impact the number of votes for a particular movie?** To observe this, we first identify our Y variable being the number of votes, as to best support the research question we defined.

A. Alteration of Data

From the initial dataset we worked with, we first began with removing some variables. These included title, description, director, actors, and rank. Title and description did not provide any information we deemed useful, and rank we felt was an arbitrary scale that would not be useful for our analysis. Impact of director and actors was left for discussion as a means of future exploration. From here, we encoded some of the data values provided to help streamline the methodology behind making our model.

Genre Genre depicts the list of genres the movie falls under. Because many if not all of the movies in the dataset are classified as having multiple genres, this may make it difficult to come up with a consistent measurement scale. To account for this, we first identify the first genre the movie falls under, and classify this as its overarching genre. From there, we filtered movies such that only the 3 most prominent genres were left, those being action, drama and comedy. These three were then encoded to have numeric values representing the text values, allowing us to make use of them within a numeric model.

Year Year depicts the movie the year was made in. The original range for the dataset was defined as $2006 - 2016$. We instead mapped this to be a categorical variable, where movies made between $2006 - 2011$ are mapped to a value of 0, and movies made between $2011 - 2016$ are mapped to a value of 1.

Metascore Metascore is defined traditionally as a weighted average rating for a movie, derived from reputed critics. The scale normally is defined as $0-100$, but to compare it against a standard rating scale of $0 - 10$, we normalized the range such that the metascore rating followed the same upper/lower bounds of the traditional rating scale.

B. Main Model Design

Our individual characteristics are given as follows:

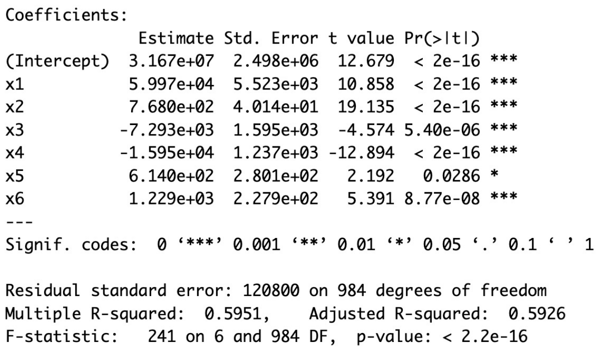
- $X1$: Rating of the movie. This is a continuous variable, rated on a scale of $1 - 10$
- $X2$: Revenue of the movie. This is given as some number of millions, and is a continuous variable.
- $X3$: Genre of the movie. This is given as a categorical variable, with a number that maps to an individual genre.
- $X4$: Year of the movie. Given as a categorical variable, with a value that maps to a time frame in which the movie was made.
- $X5$: Metascore of the movie. This is a weighted average rating of the movie, given as a continuous variable with a scale of $0 - 10$
- $X6$: Runtime of the movie. As the name suggests, it gives the length of the movie in minutes, and is a continuous variable.

For our overarching model, we used an ANOVA test on a simple linear model to map out what we would expect our results to look like. The model can best be summarized by Equation 1

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 \quad (1)$$

Each X_i corresponds to a given variable as discussed above. Each β_i corresponds to some weight, with β_0 being the intercept of our linear model.

Figure 1 provides us with a summary of the output derived from the linear model in R.



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.167e+07  2.498e+06  12.679  < 2e-16 ***
x1           5.997e+04  5.523e+03  10.858  < 2e-16 ***
x2           7.680e+02  4.014e+01  19.135  < 2e-16 ***
x3          -7.293e+03  1.595e+03  -4.574 5.40e-06 ***
x4          -1.595e+04  1.237e+03 -12.894  < 2e-16 ***
x5           6.140e+02  2.801e+02   2.192   0.0286 *
x6           1.229e+03  2.279e+02   5.391 8.77e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 120800 on 984 degrees of freedom
Multiple R-squared:  0.5951,	Adjusted R-squared:  0.5926
F-statistic:   241 on 6 and 984 DF,  p-value: < 2.2e-16
```

Fig. 1. Full Model Summary Table

From an initial glance, we notice that the p-value for each characteristic falls begin a significance threshold of $\alpha = 0.05$, and we have an $R-$squared value of $\sim 60\%$. This suggests that at least one of these variables has a significant contribution to the number of votes.

III. Research Questions

To gather a further understanding of our overarching model, we first look for some relationships that we deemed interesting between each of the characteristics mentioned. Of the 6, each is explored in at least one question, with some being explored as having an impact on one another with respect to Y.

A. Question 1

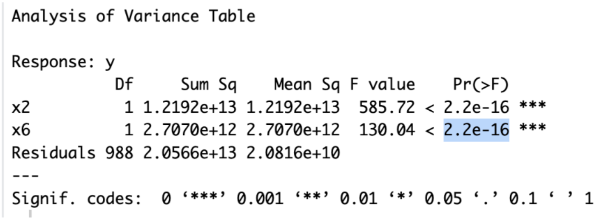
Question 1 explores the idea: **Do the revenue and runtime have a significant linear impact on the number of votes that a movie receives?** For this, we take a smaller subset of our model, making use of Y, $X2$, and $X6$. The null hypothesis and by extension the reduced model are given as follows:

$$H_0 : \beta_2 = \beta_6 = 0$$
$$Y = \beta_0 + \epsilon$$

We also consider the alternative hypothesis and full model as follows:

$$H_a : \beta_2 \neq \beta_6$$
$$Y = \beta_0 + \beta_2 X_2 + \beta_6 X_6 + \epsilon$$

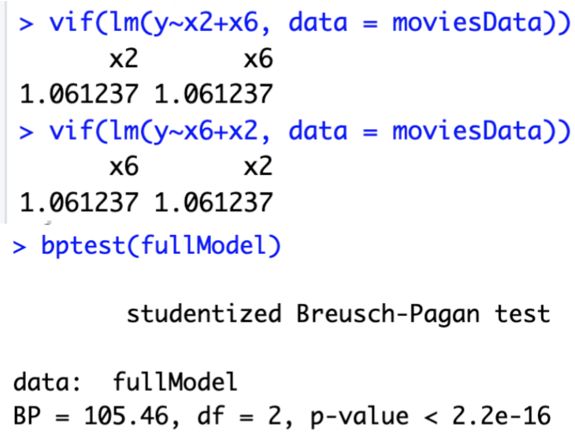
We first run a linear summary of the model, given by Figure 2. An interesting point of note was that both of these variables have a significant impact on the number of votes received, as evidence by both having a value of $\alpha > 0.5$.



```
Analysis of Variance Table

Response: y
           Df     Sum Sq    Mean Sq F value    Pr(>F)
x2          1 1.2192e+13 1.2192e+13  585.72 < 2.2e-16 ***
x6          1 2.7070e+12 2.7070e+12  130.04 < 2.2e-16 ***
Residuals 988 2.0566e+13 2.0816e+10
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 2. Variance Table Summary for Question 1

Beyond the initial test, we next look to analyze trends in the data, specifically any relationships that exist with respect to variability, normality, etc. When we run the Shapiro-Wilk normality test, we find that $W = 0.873$, with the p-value being less than 2.2×10^{-16}. This suggests that we have a violation in normality with respect to the residuals. To test for multicollinearity and heteroscedasticity, we use the variance inflation factor (VIF) and Breusch-Pagan tests, with the results reflected in Figure 3 We also look to graph the residuals to get an idea of if there are any interesting. Specifically, we conduct Cook's Distance and DFBETAS tests to look at where we are regarding the presence of multiple influential points. Figure 4 depicts the graphical output of running the Cook's Distance, and Figure 5 depicts the graphical output



```
> vif(lm(y~x2+x6, data = moviesData))
       x2       x6 
1.061237 1.061237 
> vif(lm(y~x6+x2, data = moviesData))
       x6       x2 
1.061237 1.061237 
> bptest(fullModel)

	studentized Breusch-Pagan test

data:  fullModel
BP = 105.46, df = 2, p-value < 2.2e-16
```

Fig. 3. Breusch-Pagan and VIF Output

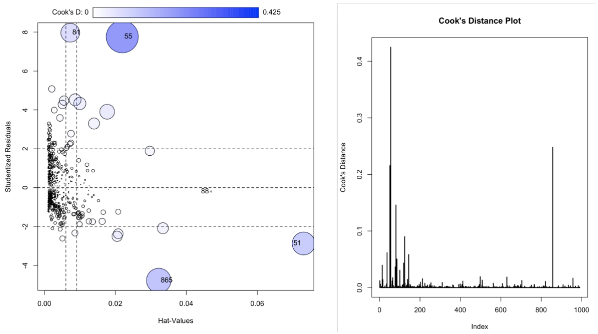


Fig. 4. Cook's Distance Test

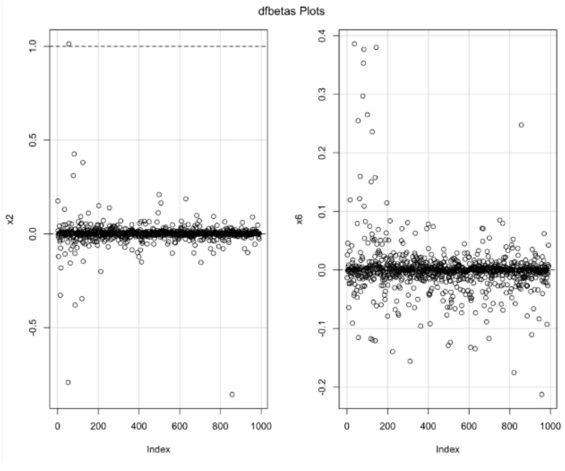
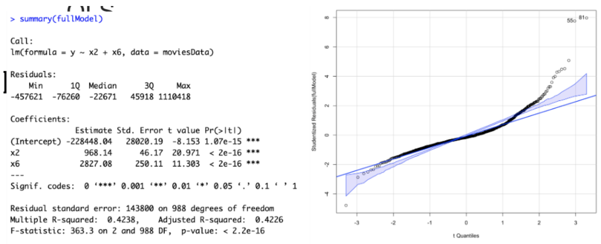


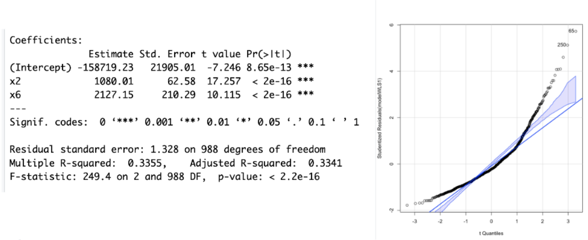
Fig. 5. DFBETAS Test



```
> summary(fullModel)

Call:
lm(formula = y ~ x2 + x6, data = moviesData)

Residuals:
    Min      1Q  Median      3Q     Max 
-457621  -76260  -22671   45918 1110418 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -228448.04   28020.19  -8.153 1.07e-15 ***
x2              968.14      46.17  20.971  < 2e-16 ***
x6             2827.08     250.11  11.303  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 143800 on 988 degrees of freedom
Multiple R-squared:  0.4238,	Adjusted R-squared:  0.4226 
F-statistic: 363.3 on 2 and 988 DF,  p-value: < 2.2e-16
```

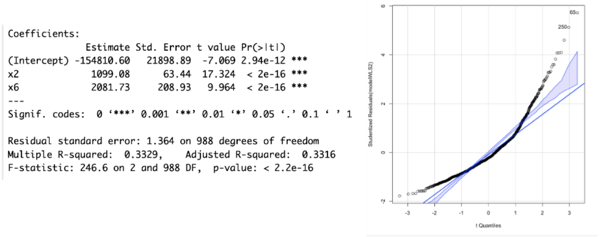
Fig. 6. OLS Analysis

of running the DFBETAS test. Based on these two tests, it appears that we may have multiple influential points of interest. To remediate, we look to apply both Ordinary Least Squares (OLS) and Weighted Least Squares (WLS). Figure 6 gives the results of applying this analysis in R. Figure 7 gives the results of applying the WLS analysis in R. We also have the given confidence intervals for each of the weights in our model. To better supplement the WLS method, we look to first apply bootstrapping to counter both influential points and outliers. Figure 8 depicts what the output of this method. Upon applying the bootstrapping method, we find that this process was inconclusive. In fact, we find that applying the bootstrapping method may have proved detrimental in this case. The biggest factor supporting this would be the fact that R^2 has decreased in value, even though the spread of residuals has improved slightly. This in turn suggests that overfitting the model may have occurred. From these tests, we conclude that we reject the null hypothesis. This gives us that we accept the alternative hypothesis, $Y = \beta_0 + \beta_2 X_2 + \beta_6 X_6 + \epsilon$. Therefore, we say that both revenue and runtime have a significant linear impact on the number of votes that any given movies receive.



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -158719.23   21905.01  -7.246 8.65e-13 ***
x2             1080.01      62.58  17.257  < 2e-16 ***
x6             2127.15     210.29  10.115  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.328 on 988 degrees of freedom
Multiple R-squared:  0.3355,	Adjusted R-squared:  0.3341 
F-statistic: 249.4 on 2 and 988 DF,  p-value: < 2.2e-16
```

Fig. 7. WLS Analysis



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -154810.60   21898.89  -7.069 2.94e-12 ***
x2             1099.08      63.44  17.324  < 2e-16 ***
x6             2081.73     208.93   9.964  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.364 on 988 degrees of freedom
Multiple R-squared:  0.3329,	Adjusted R-squared:  0.3316 
F-statistic: 246.6 on 2 and 988 DF,  p-value: < 2.2e-16
```

Fig. 8. WLS Bootstrapping

B. *Question 2*

Question 2 explores the idea: **Is the impact of rating on votes the same for the top three genres of movies?**. For this, we take a smaller subset of our model, making use of Y,

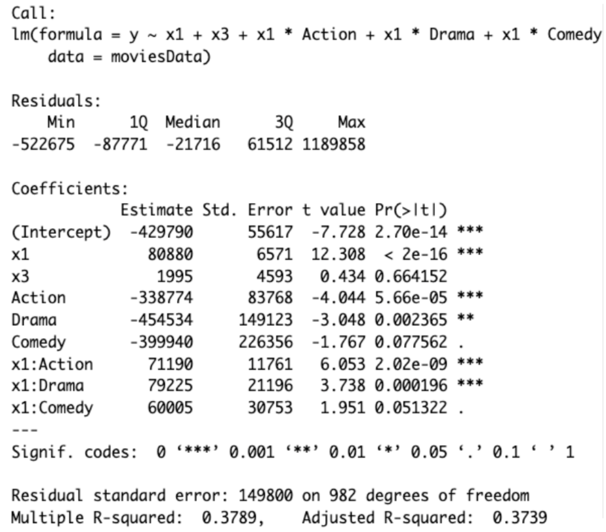
$X1$, and $X3$. From $X3$, we isolate three specific categories, those being *Action, Drama,* and *Comedy*. The null hypothesis and by extension the reduced model are given as follows:

$$\begin{aligned} H_0 &: \beta_{13} = 0 \\ Y &= \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \epsilon \end{aligned}$$

We also consider the alternative hypothesis and full model as follows:

$$\begin{aligned} H_a &: \beta_{13} \neq 0 \\ Y &= \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_{13} X_1 X_3 + \epsilon \end{aligned}$$

We first run a linear summary of the model, given by Figure 9. From this we notice that we have a very low R-squared value as the main takeaway. This suggests that the choice of independent variable doesn't have much of an impact on the dependent variable, i.e. the impact of rating on votes isn't necessarily the same for all three genres.



```
Call:
lm(formula = y ~ x1 + x3 + x1 * Action + x1 * Drama + x1 * Comedy
    data = moviesData)

Residuals:
    Min      1Q  Median      3Q     Max 
-522675  -87771  -21716   61512 1189858 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -429790      55617  -7.728 2.70e-14 ***
x1              80880       6571  12.308  < 2e-16 ***
x3               1995       4593   0.434 0.664152    
Action        -338774      83768  -4.044 5.66e-05 ***
Drama         -454534     149123  -3.048 0.002365 ** 
Comedy        -399940     226356  -1.767 0.077562 .  
x1:Action       71190      11761   6.053 2.02e-09 ***
x1:Drama        79225      21196   3.738 0.000196 ***
x1:Comedy       60005      30753   1.951 0.051322 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 149800 on 982 degrees of freedom
Multiple R-squared:  0.3789,	Adjusted R-squared:  0.3739
```

Fig. 9. lm() Summary for Question 2

Beyond our initial test, we look to make use of tests to understand factors such as variability and normality. When we apply the Shapiro-Wilk normality test, we get that $W = 0.89113$, and that the p value is less than 2.2×10^{-16}. We also plotted residuals for this set of data, as shown in Figure 10 After observing the residual plots, our big takeaway is that there are a series of violations that must be considered. In particular we have that there is non-constant variance, there is a definite normality violation, and that outliers are present. The outliers point in particular is very apparent after observing these residual graphs, notably the second where we have points towards the top right that are significantly different from what we expect along the general trend-line.

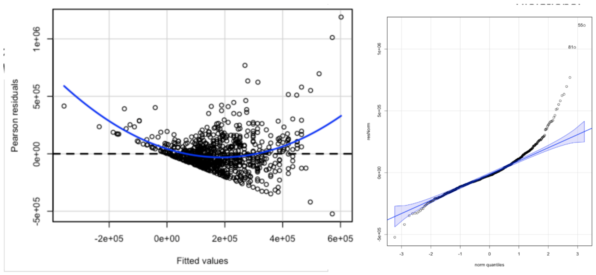


Fig. 10. Residual Plots

To address this issue, we look to rectify this issue by attempting to apply Weighted Least-Squares Regression (WLS). The following code snippet is an example of what this script looked like in R:

```
wts1 <- 1/fitted(lm(abs(residuals(lm(
    y~x1+x3+x1*Actions+x1*Drama
    +x1*Comedy,
    data=data)))~x1+x3+x1*Action
    +x1*Drama+x1+Comedy, data=data))^2
)))
modelWLS<-lm(
    y~x1+x3+x1*Actions+x1*Drama+x1*Comedy,
    data=data, weights=wts1)
```

Because we are looking to address the concern of non-constant variance, we should in theory see this reflected in residual graphs after applying this methodology. Figure 11 reflects what the result of these changes would look like. We can see the impact of applying the WLS model as it helps address the issue of non-constant variance. In particular, we notice how there are fewer outliers, and the plotted points much better resemble the expected trendlines on each graph. From these results, and running another ANOVA test in R we conclude that we should still reject our reduced model in favor of our full model, that is $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_{13} X_1 X_3 + \epsilon$. Therefore, the impact of rating on the number of votes is different for each of the top three genres.

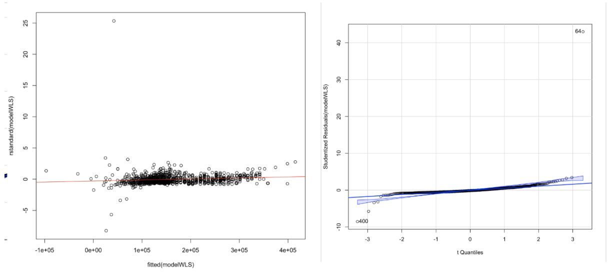


Fig. 11. Results from applying WLS

C. *Question 3*

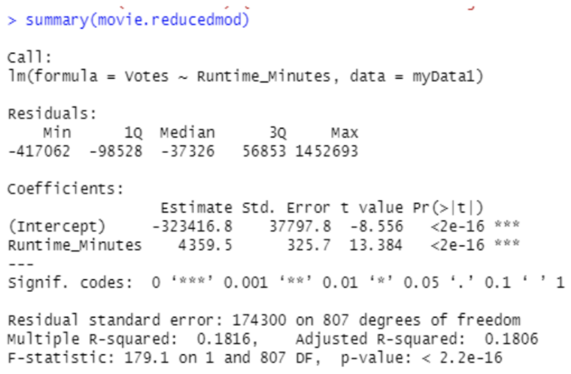
Question 1 explores the idea: **Given the runtime, does year have a significant impact on Y?** For this, we take a smaller subset of our model, making use of Y, $X4$, and $X6$. The null hypothesis and by extension the reduced model are given as follows:

$$\begin{aligned} H_0 &: \beta_4 = 0 \\ Y &= \beta_0 + \beta_6 X_6 + \epsilon \end{aligned}$$

We also consider the alternative hypothesis and full model as follows:

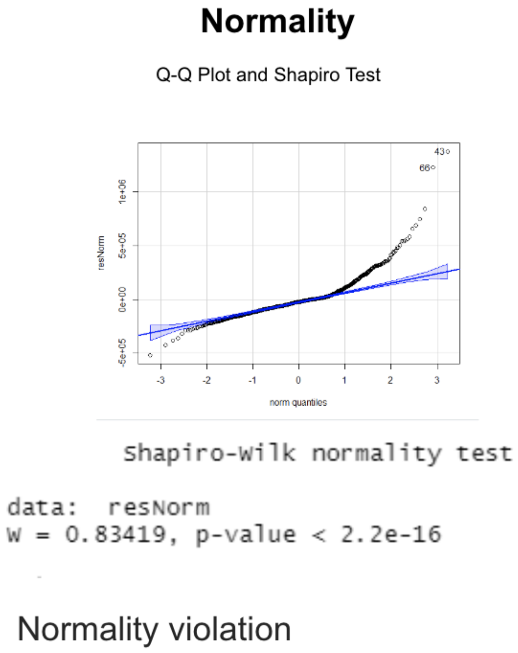
$$\begin{aligned} H_a &: \beta_4 \neq 0 \\ Y &= \beta_0 + \beta_4 X_4 + \beta_6 X_6 + \epsilon \end{aligned}$$

We first run a linear summary of the model, given by Figure 2. From the initial analysis, we can see that the adjusted R square is very, very small, at around 0.18. This is a point of concern, as it tells us that the independent and dependent variables are probably not that closely related. Beyond the initial test, we next look to analyze trends in the data, specifically any relationships that exist with respect to variability, normality, etc. When we run the Shapiro-Wilk normality test, we find that $W = 0.83419$, with the p-value being less than 2.2×10^{-16}, as reflected in Figure 13. This suggests that we have a violation in normality with respect to the residuals. To test for multicollinearity and heteroscedasticity, we use the variance inflation factor (VIF) and Breusch-Pagan tests, with the results being that VIF < 5, meaning there is no issue with multicollinearity. However, the BP test gives us that there is a constant variance issue, as shown in Figure 14. We also look to graph the residuals to get an idea of if there are any interesting. Specifically, we conduct Cook's Distance and DFBETAS tests to look at where we are regarding the presence of multiple influential points. Figure 15 depicts the graphical output of running the Cook's Distance, and Figure 16 depicts the graphical output of running the DFBETAS test. Based on these two tests, it appears that we may have multiple influential points of interest. To handle this, we apply bootstrapping, reflected in Figure 17. Our conclusion remains that the year has a significant impact on the number of votes for a movie.



```
> summary(movie.reducedmod)

Call:
lm(formula = Votes ~ Runtime_Minutes, data = myData1)

Residuals:
    Min      1Q  Median      3Q     Max 
-417062  -98528  -37326   56853 1452693 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -323416.8    37797.8  -8.556   <2e-16 ***
Runtime_Minutes    4359.5      325.7  13.384   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 174300 on 807 degrees of freedom
Multiple R-squared:  0.1816,	Adjusted R-squared:  0.1806 
F-statistic: 179.1 on 1 and 807 DF,  p-value: < 2.2e-16
```

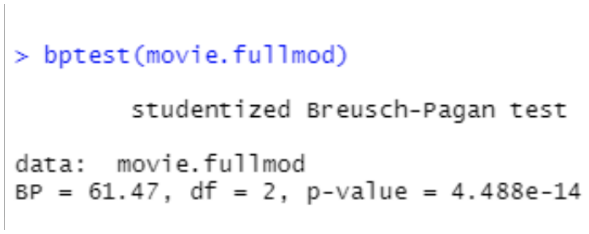
Fig. 12. Variance Table Summary for Question 3



```
        Shapiro-Wilk normality test

data:  resNorm
W = 0.83419, p-value < 2.2e-16
```

Normality violation

Fig. 13. Shapiro-Wilk Test



```
> bptest(movie.fullmod)

        studentized Breusch-Pagan test

data:  movie.fullmod
BP = 61.47, df = 2, p-value = 4.488e-14
```

Fig. 14. Breusch-Pagan Test

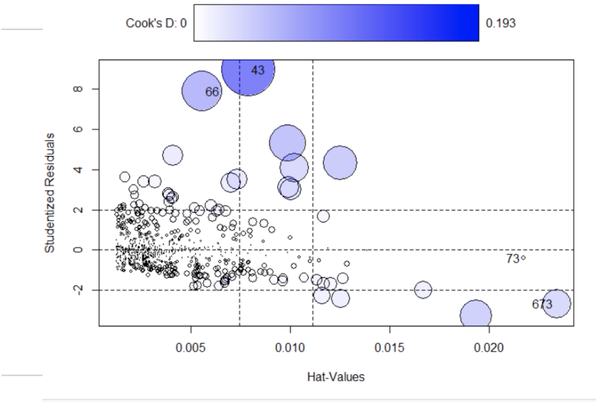


Fig. 15. Cook's Distance Test

D. Question 4

Question 4 explores the idea: **Do rating and metascore have the same impact on number of votes?** For this, we take a smaller subset of our model, making use of Y, $X1$, and $X5$. The null hypothesis and by extension the reduced model are given as follows:

$$\begin{aligned} H_0 &: \beta_1 = \beta_5 = \beta \\ Y &= \beta_0 + \beta(X1 + X5) + \epsilon \end{aligned}$$

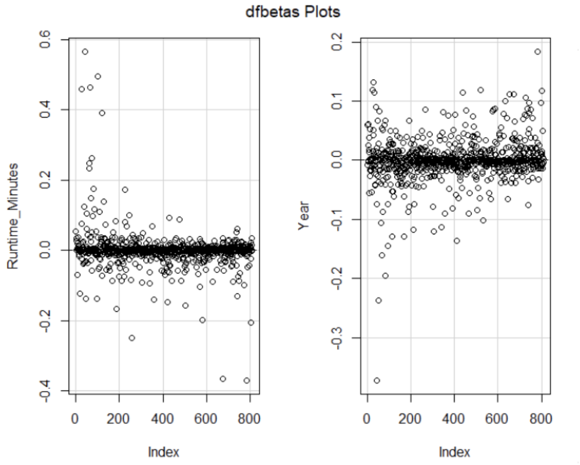
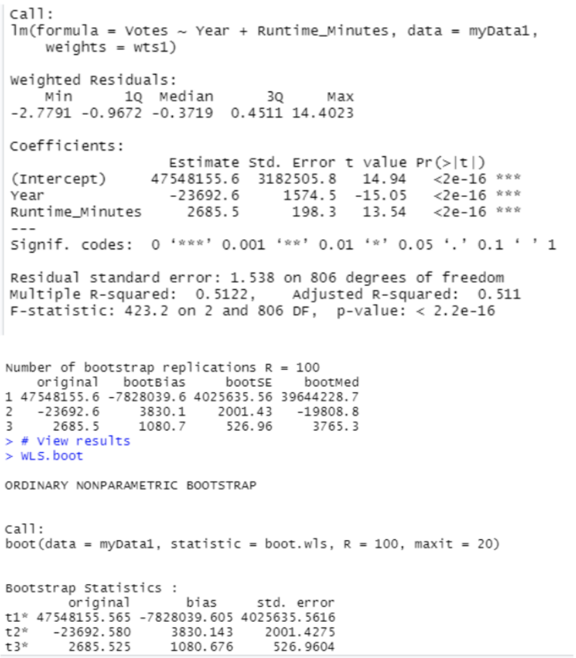


Fig. 16. DFBETAS Test



```
Call:
lm(formula = Votes ~ Year + Runtime_Minutes, data = myData1,
    weights = wts1)

Weighted Residuals:
    Min      1Q  Median      3Q     Max 
-2.7791 -0.9672 -0.3719  0.4511 14.4023 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     47548155.6  3182505.8   14.94   <2e-16 ***
Year              -23692.6     1574.5  -15.05   <2e-16 ***
Runtime_Minutes     2685.5      198.3   13.54   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.538 on 806 degrees of freedom
Multiple R-squared:  0.5122,	Adjusted R-squared:  0.511 
F-statistic: 423.2 on 2 and 806 DF,  p-value: < 2.2e-16


Number of bootstrap replications R = 100 
    original   bootBias     bootSE    bootMed
1 47548155.6 -7828039.6 4025635.56 39644228.7
2   -23692.6     3830.1    2001.43   -19808.8
3     2685.5     1080.7     526.96     3765.3
> # View results
> WLS.boot

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = myData1, statistic = boot.wls, R = 100, maxit = 20)


Bootstrap Statistics :
        original        bias     std. error
t1* 47548155.565 -7828039.605 4025635.5616
t2*   -23692.580     3830.143    2001.4275
t3*     2685.525     1080.676     526.9604
```

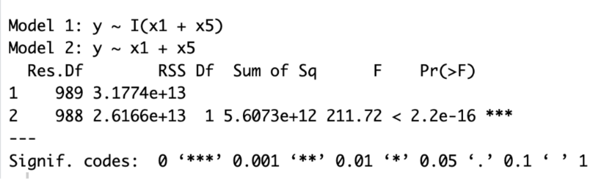
Fig. 17. WLS Bootstrapping

We also consider the alternative hypothesis and full model as follows:

$$
\begin{aligned}
H_a &: \beta_1 \neq \beta_5 \\
Y &= \beta_0 + \beta_1 X_1 + \beta_5 X_5 + \epsilon
\end{aligned}
$$

We first run a linear summary of the model, given by Figure 18. Our initial conclusion would be that our characteristics don't have the same impact on the number of votes. That said, its important to consider that metascore is a weighted average of reviews, so that may cause a distinction in the data, suggesting further analysis is required.

Beyond the initial test, we next look to analyze trends in the data, specifically any relationships that exist with respect to variability, normality, etc. When we run the Shapiro-Wilk normality test, we find that $W = 0.88755$, with the p-value being less than 2.2×10^{-16}. This suggests that we have a violation in normality with respect to the residuals. To test for multicollinearity we use the variance inflation factor (VIF), and find that we have a VIF value of 1.537908 and 1.537908 for $X1$ and $X5$ respectively. Because both are less than 10, there is no multicollinearity issue. We also look to graph the residuals to get an idea of if there are any interesting. Specifically, we conduct Cook's Distance and DFBETAS tests to look at where we are regarding the presence of multiple influential points. Figure 19 depicts the graphical output of running the Cook's Distance, and Figure 20 depicts the graphical output of running the DFBETAS test. To remediate, we look to apply Weighted Least Squares (WLS). Figure 21 gives the results of applying the WLS analysis in R. To better supplement the WLS method, we look to first apply robust bootstrapping to counter both influential points and outliers. Figure 22 depicts what the output of this method. Upon applying the bootstrapping method, we find that this process was inconclusive. In fact, we find that applying the bootstrapping method may have proved detrimental in this case. Overall, we saw no major change as a result, meaning that rating and metascore don't have the same overarching impact on the number of votes that a movie receives.



```
Model 1: y ~ I(x1 + x5)
Model 2: y ~ x1 + x5
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)    
1    989 3.1774e+13                                  
2    988 2.6166e+13  1 5.6073e+12 211.72 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 18. Variance Table Summary for Question 4

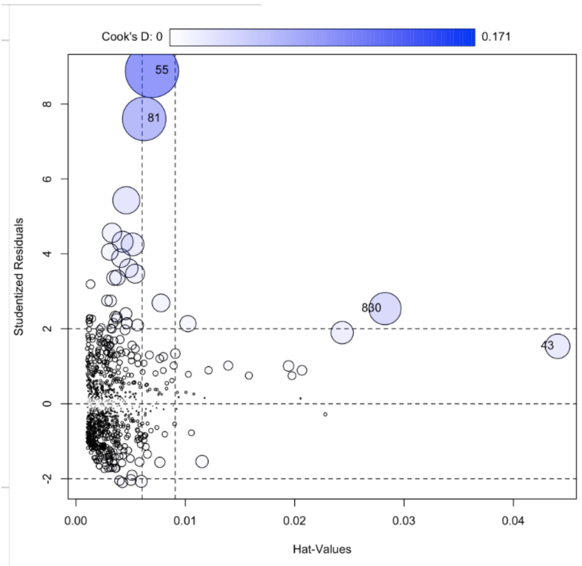


Fig. 19. Cook's Distance Test

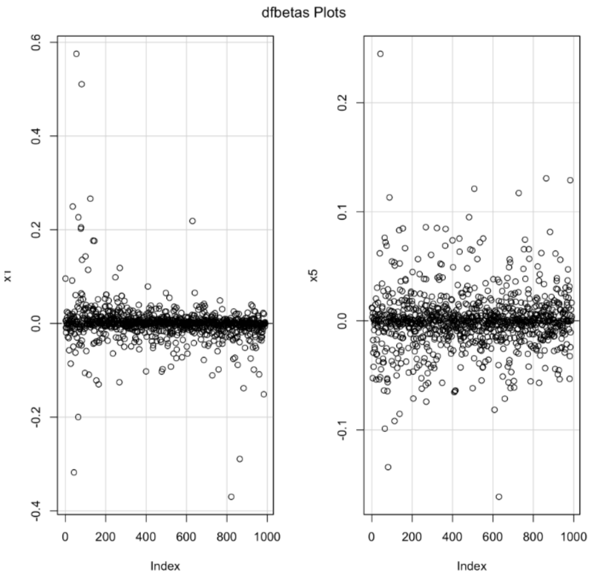
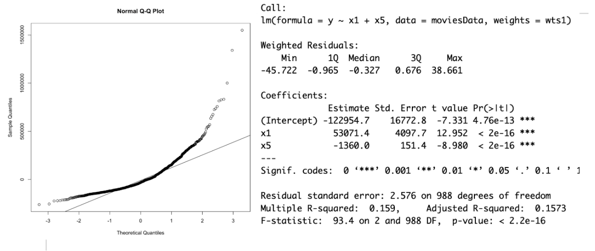
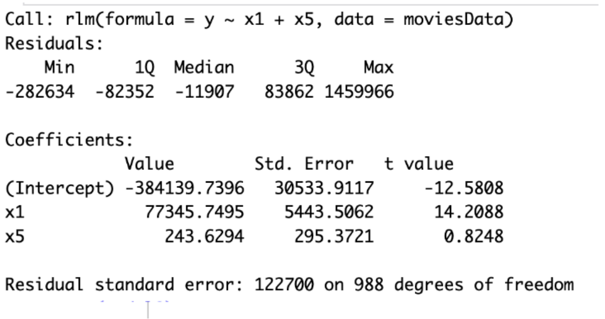


Fig. 20. DFBETAS Test



```
Call:
lm(formula = y ~ x1 + x5, data = moviesData, weights = wts1)

Weighted Residuals:
    Min      1Q  Median      3Q     Max 
-45.722  -0.965  -0.327   0.676  38.661 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -122954.7    16772.8  -7.331 4.76e-13 ***
x1            53071.4     4097.7  12.952  < 2e-16 ***
x5            -1360.0      151.4  -8.980  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.576 on 988 degrees of freedom
Multiple R-squared:  0.159,	Adjusted R-squared:  0.1573 
F-statistic:  93.4 on 2 and 988 DF,  p-value: < 2.2e-16
```

Fig. 21. WLS Analysis



```
Call: rlm(formula = y ~ x1 + x5, data = moviesData)
Residuals:
    Min      1Q  Median      3Q     Max 
-282634  -82352  -11907   83862 1459966 

Coefficients:
            Value        Std. Error   t value 
(Intercept) -384139.7396   30533.9117    -12.5808
x1            77345.7495    5443.5062     14.2088
x5              243.6294     295.3721      0.8248

Residual standard error: 122700 on 988 degrees of freedom
```

Fig. 22. WLS Bootstrapping

E. Question 5

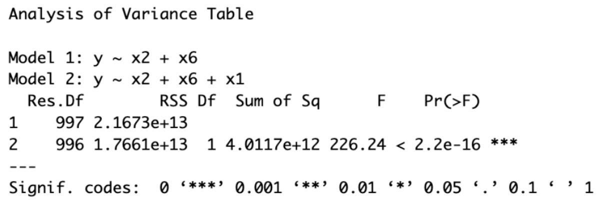
Question 5 explores the idea: **Do the revenue and runtime have a significant linear impact on the number of votes that a movie receives?** For this, we take a smaller subset of our model, making use of Y, $X1$, $X2$, and $X6$. The null hypothesis and by extension the reduced model are given as follows:

$$H_0 : \beta_1 = 0$$
$$Y = \beta_0 + \beta_2 X_2 + \beta_6 X_6 + \epsilon$$

We also consider the alternative hypothesis and full model as follows:

$$H_0 : \beta_1 \neq 0$$
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_6 X_6 + \epsilon$$

We first run a linear summary of the model, given by Figure 23. The biggest takeaway is that the p-value is less than 0.05, which suggests that in a model with runtime and revenue, the rating of the movie still has a significant impact on the number of votes for a movie. Another important takeaway however, is that $R^2 = 0.185$. Because this value is very low, the independent variables may not explain a lot with respect to the dependent variables.



```
Analysis of Variance Table

Model 1: y ~ x2 + x6
Model 2: y ~ x2 + x6 + x1
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)    
1    997 2.1673e+13                                   
2    996 1.7661e+13  1 4.0117e+12 226.24 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 23. Variance Table Summary for Question 5

We next conduct a series of tests for normality, variance, etc. From running the Shapiro test, we get a result of $W = 0.90869$, and conclude that residuals in the dataset are normally distributed. We also run a VIF test, and find that because the results for each variable are all under 10, there are no multicollinearity issues. However, when we run the Brown-Forsythe test, we find that $p < 0.05$, which showed that we had a violation on constant variance. We also look to graph the residuals to get an idea of if there are any interesting. Specifically, we conduct Cook's Distance and DFBETAS tests to look at where we are regarding the presence of multiple influential points. Figure 24 depicts the graphical output of running the Cook's Distance, and Figure 25 depicts the graphical output of running the DFBETAS test. Based on these two tests, it appears that we may have multiple influential points of interest. To fix this, we attempt to apply both the WLS and OLS tests. Figure 26 gives the results of applying the OLS analysis in R. Figure 27 gives the results of applying the WLS analysis in R. From there, we decide to apply bootstrapping as a means for remediation. Figure 28 reflects this We can see that the residual standard

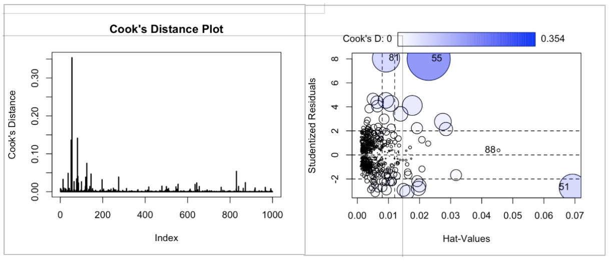


Fig. 24. Cook's Distance Test

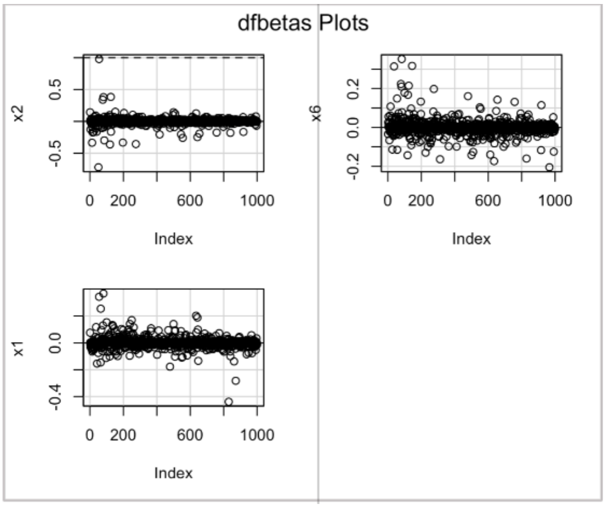
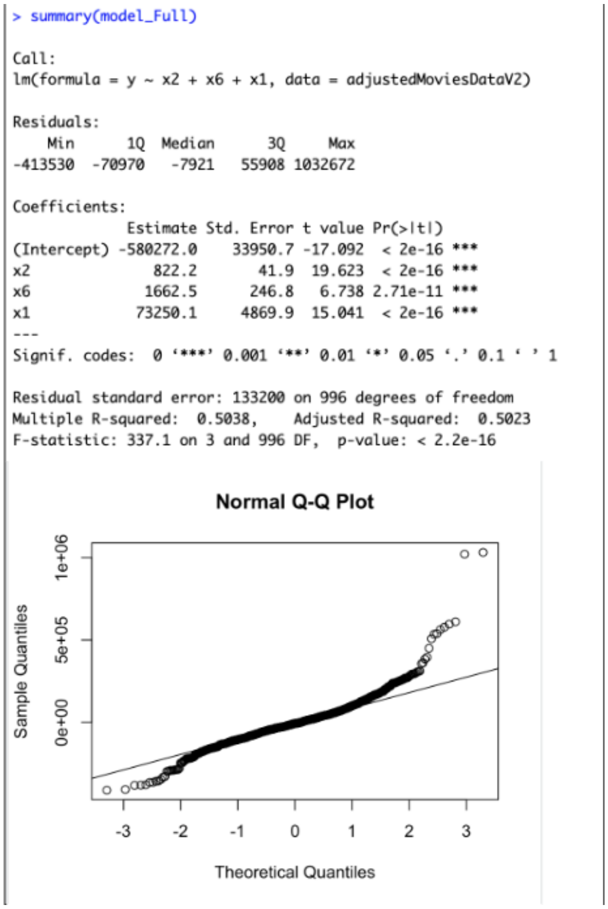
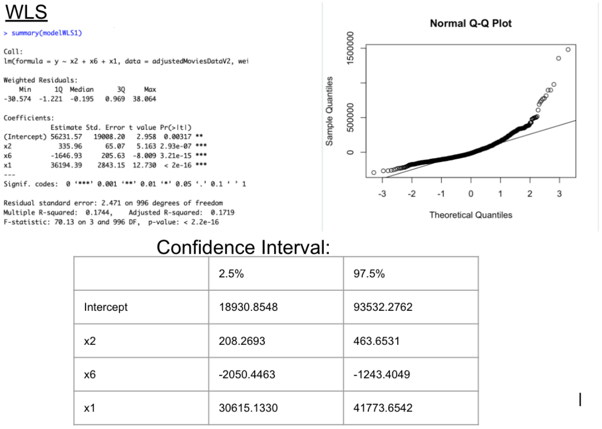


Fig. 25. DFBETAS Test



```
> summary(model_Full)

Call:
lm(formula = y ~ x2 + x6 + x1, data = adjustedMoviesDataV2)

Residuals:
    Min      1Q  Median      3Q     Max
-413530  -70970   -7921   55908 1032672

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -580272.0    33950.7 -17.092  < 2e-16 ***
x2              822.2       41.9  19.623  < 2e-16 ***
x6             1662.5      246.8   6.738 2.71e-11 ***
x1            73250.1     4869.9  15.041  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 133200 on 996 degrees of freedom
Multiple R-squared:  0.5038,	Adjusted R-squared:  0.5023
F-statistic: 337.1 on 3 and 996 DF,  p-value: < 2.2e-16
```

Fig. 26. OLS Analysis



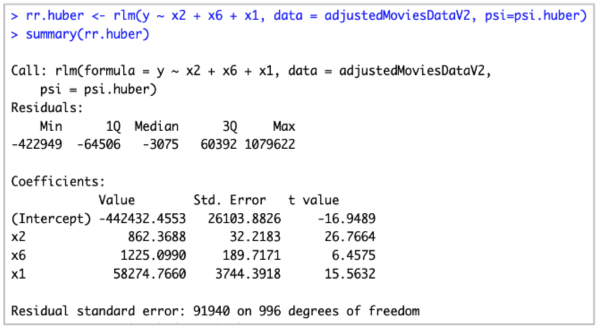
WLS

```
> summary(modelWLS1)

Call:
lm(formula = y ~ x2 + x6 + x1, data = adjustedMoviesDataV2, wei

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-30.574  -1.221  -0.195   0.969  38.064

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 56231.57   19008.20   2.958  0.00317 **
x2            335.96      65.07   5.163 2.93e-07 ***
x6          -1646.93     205.63  -8.009 3.21e-15 ***
x1          36194.39    2843.15  12.730  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.471 on 996 degrees of freedom
Multiple R-squared:  0.1744,	Adjusted R-squared:  0.1719
F-statistic: 70.13 on 3 and 996 DF,  p-value: < 2.2e-16
```

Confidence Interval:

	2.5%	97.5%
Intercept	18930.8548	93532.2762
x2	208.2693	463.6531
x6	-2050.4463	-1243.4049
x1	30615.1330	41773.6542

Fig. 27. WLS Analysis



```
> rr.huber <- rlm(y ~ x2 + x6 + x1, data = adjustedMoviesDataV2, psi=psi.huber)
> summary(rr.huber)

Call: rlm(formula = y ~ x2 + x6 + x1, data = adjustedMoviesDataV2,
    psi = psi.huber)
Residuals:
    Min      1Q  Median      3Q     Max
-422949  -64506   -3075   60392 1079622

Coefficients:
            Value        Std. Error   t value
(Intercept) -442432.4553  26103.8826   -16.9489
x2              862.3688     32.2183    26.7664
x6             1225.0990    189.7171     6.4575
x1            58274.7660   3744.3918    15.5632

Residual standard error: 91940 on 996 degrees of freedom
```

Fig. 28. WLS Bootstrapping

error has decreased as a result, meaning that bootstrapping has actually improved our outcome. Therefore, making use of bootstrapping was a good decision, and in fact made our results more viable. Overall, we conclude that in a model with rating and revenue, the rating will still have a significant impact on the number of votes for a movie.

IV. Discussion

A. K-Fold Validation

K-fold validation was used to identify what the predictive power is of our model. We apply this across each of our sub-questions, followed by the final model to see what our results would be. In doing this, we found a RMSE value of $1.4 - 1.6$ for each of the questions, suggesting that there is a noticeable amount of prediction power with the model we have constructed. That is, we think we would be able to predict reasonably accurately the number of votes that would be cast on a new movie given certain characteristics. That said, due to a variety of normality concerns and our inability to fix them with WLS and bootstrapping, we find that there isn't a definite conclusion. As such, we can say that there might be some relationship between a lot of these characteristics and the votes cast. Even if we feel that certain characteristics were more impactful, there is no definite way to confirm this without further investigation and better data. In doing this, we may be able to support the K-fold validation claim that our model has a reasonable amount of predicting power.

B. Future Works

The first area which our study could be expanded, is an understanding of what we have concluded based on the chosen dataset. As discussed in our alterations, we made a lot of changes to categorical variables, in some cases completely removing categories altogether. Specific points of interest would be as follows:

Genre We removed all but the primary genre for each movie, after which we only analyzed the top three genres as a whole. As such, we acknowledge there may have been some points of classification in the beginning of our study, which may have impact our overarching results. In figuring out how we might be able to account for all genres, or finding a more accurate method of classifying movies if we were to stick with just a single genre, this may have given us more accurate results.

Director We didn't account for director in our study. If we peruse the dataset, we notice some very popular names may appear when working with the dataset, specifically how they might impact the overall presence of votes. For example, a movie in the dataset with a large quantity of votes would be *The Dark Knight*, directed by Christopher Nolan. Nolan's name appears across multiple movies, so perhaps looking at how his involvement with films, and how many of them tend to have large quantities of votes from a surface perspective, might be an interesting point of study in predicting which future movies may garner large quantities of votes.

Actor Similar to directors, we didn't account for actors. Actors however, falls into a similar problem as with observing genres, in that many if not all of the movies within the dataset tend to have multiple actors listed. Further analysis on this would require some form of isolation for which actors played the most important role. This would be crucial, as some movies listed what are dubbed as "A-list actors" but have them in minor roles, i.e. cameos. For example, a prominent movie featured in the dataset is *The Wolf of Wall Street*. Leonardo DiCaprio is listed as the primary actor for the movie, but a prominent name also listed is Matthew McConaughey. The provided dataset doesn't inform the user which actor is the "main actor" or the length of time they appear in the movie, so analyzing this without some form of context wouldn't be appropriate. That said, adding more context to the dataset to allow for this time of analysis would better explore the idea of relationships between viewers and actors they tend to favor.

Likeability The remaining characteristics in the dataset still seem as though they may not be of particular interest. Other exploration of this work could be some further investigation of likeability index. Because we have a large quantity of votes, we could instead flip the study to understand what is the correlation between large quantities of votes, and how much people are actually invested in the movie. Our study does leave some hints towards this trend existing, but given what we chose as our characteristics and dependent variable, we can not make this claim with statistical backing. As such, exploring this might be of particular interest in that we may find more reasons as to why there may be larger/smaller quantities of votes for movies, and how they directly correlate with opinions on a given movie.

V. Conclusion

Given the increasing amount of influence that movies have on society, its important to recognize how these movies interact with different viewer groups. Our study recognizes the impact of different characteristics on movies, and works to identify important relationship that exist between these movies and other works. Through statistical analysis, we surmise which factors may be most important with respect to voting on movies, as well as which factors have little to no impact. Through K-fold analysis, we validate our model and outline to what extent we have prediction power, as well as discuss limitations of our study. Finally, we lay groundwork for future exploration of these relationships, as well as provide multiple avenues that may allow for greater insights.

References

[1] N. J. Pulles and P. Hartman, "Likeability and its effect on outcomes of interpersonal interaction," *Industrial Marketing Management*, vol. 66, pp. 56–63, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0019850117304765

[2] R. C. Riley, "Public taste: A comparison of movie popularity and critical opinion," *William and Mary Scholar Works*, 1982. [Online]. Available: https://scholarworks.wm.edu/cgi/viewcontent.cgi?article=4603

[3] S. Rai, "Impact of hollywood online film reviews on american viewers' perception," *n.d.*, 2020. [Online]. Available: https://tinyurl.com/25aej9sn

[4] J. R. Penthey, "The influence of movie reviews on consumers," *University of New Hampshire Scholars' Repository*, 2015. [Online]. Available: https://scholars.unh.edu/cgi/viewcontent.cgi?article=1267

[5] S. Moon, P. K. Bergey, and D. Iacobucci, "Dynamic effects among movie ratings, movie revenues, and viewer satisfaction," *Journal of Marketing*, vol. 74, no. 1, pp. 108–121, 2010. [Online]. Available: https://doi.org/10.1509/jmkg.74.1.108