Examining the effectiveness of possible predictors for IMDb ratings

CROSS-EVALUATION OF METRICS TO PREDICT THE IMPORTANCE OF CREATIVE WORKS

Regression and Multivariate Data Analysis

Anna Skarpalezou

Assignment 3

Movies undoubtedly constitute a critical part of modern pop-culture. With international productions gaining popularity and the affordable, ubiquitous streaming services available today, we are now consuming more content than ever. The industry is overflowing with creative works. Therefore, an effective system in filtering out the important works from the meaningless one, would be of great assistance to the general public in deciding amongst them.

IMDB is one of the most widely used platforms for cinephiles, offering information ranging from the financial statistics of a movie's release, to random trivia and, of course, the movie's IMDb score. Registered users get to vote on every released title available in the database, on a scale of 1 to 10. Individual votes are then aggregated and summarized as a single IMDb rating, visible on the title's main page. Another popular website with movie data is Metacritic. It similarly assigns scores to films however the voting isn't open to the public. The website curates a group of respected critics, assigns scores to their reviewing abilities, and then applies a weighted average to their reviews. For a score to become available on the Metacritic website, at least 4 accredited critics must have reviewed it.

The goal of this study was to create a model that relates some of each movie's features to its popularity, as expressed by its IMDb score. The hypothesis of this study was that movies which are more favorably reviewed by the accredited critics, as seen through their meta-score, whose gross income was higher, which received a higher number of votes on IMDb and which had a longer runtime, would also be highly valued by the wider audience as reflected in their IMDb score. Following this hypothesis, a simple linear regression model. The primary goal of this study is to test the relationship between the critic reviews, runtime, number of votes, movie runtime and audience preference, modeled by the following simple regression:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i.$$

$$\boxed{\text{IMDb score} = \beta_0 + \beta_1 * \text{Metascore} + \beta_2 * \text{Gross Income} + \beta_3 * \text{number of votes} + \beta_4 * \text{runtime} + \varepsilon}$$

β_1 represents the estimated associated change in a film's IMDb score as its Meta-score increases by one percent, holding all else in the model fixed. Similarly, β_2 represents the estimated associated change in a film's IMDb score as its Gross Income increases by one unit, holding all else in the model fixed. Most importantly, these β_s are not to be interpreted marginally and do not show the relationship of the target variable to a single predictor alone.

Data Collection

This analysis is based on the first 150 most voted on feature films for 2019 on IMDB[1].
I decided to use web-scrapping as my data collection technique, to ensure the reliability and consistency of the data. To do so, I used python, using a third-party library named 'requests'. I sent a HTTP request to the URL of the IMDB webpage, the server responding with the HTML content of the webpage. I proceeded to parse the data, however since most of it is nested, I

[1] The IMDB page I used can be found at: https://www.imdb.com/search/title/?title_type=feature&year=2019-01-01,2019-12-31&sort=num_votes,desc

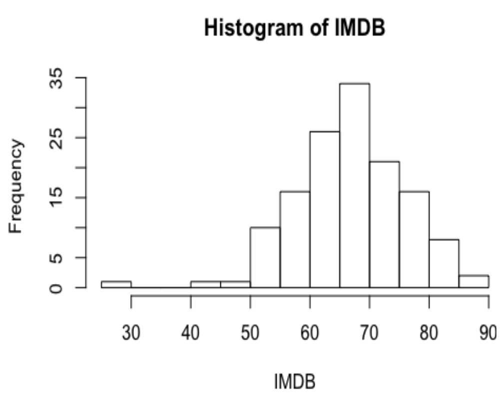
resorted to a few string processing techniques. I used another third-party python library, Beautiful Soup, to create a data tree and parse through it, able to pull out the pieces of information that were of importance for the analysis. From there, I used the 'find_all' function, to separate the html content that referred to each specific movie and continued pulling out data from each section. I was able to do so due to the consistency of the html code. I exported that to an excel sheet and then to a csv, to more easily process the information in R. A detailed view of the code I wrote for web-scrapping can be found in appendix A. The csv file created from the web-scraping can be found in appendix B.

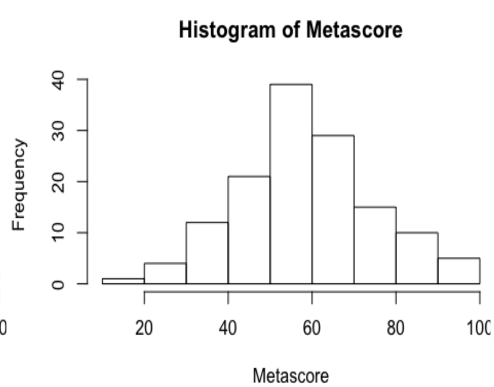
Data Analysis

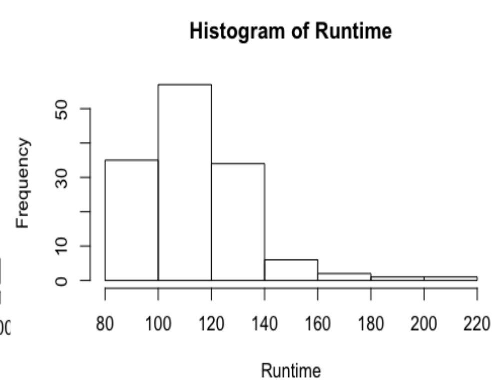
Lets first take a look at the data:
Descriptive Statistics: IMDb, Metascore, Runtime, Votes, Gross Income

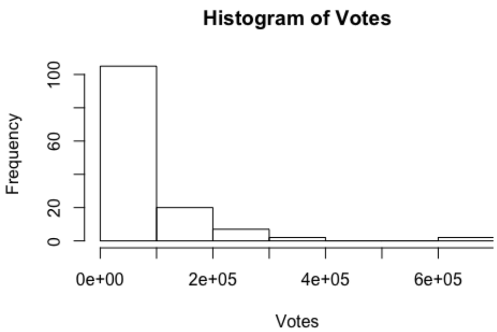
IMDB	Metascore	Runtime	Votes	Gross
Min. :28.00	Min. :19.00	Min. : 81.0	Min. : 14407	Min. : 0.35
1st Qu.:61.75	1st Qu.:48.00	1st Qu.:100.0	1st Qu.: 28046	1st Qu.: 31.86
Median :67.00	Median :58.50	Median :113.0	Median : 47234	Median : 173.01
Mean :67.12	Mean :59.01	Mean :114.3	Mean : 83385	Mean : 28302.48
3rd Qu.:74.00	3rd Qu.:68.25	3rd Qu.:123.2	3rd Qu.: 88808	3rd Qu.: 31414.00
Max. :86.00	Max. :96.00	Max. :209.0	Max. :695789	Max. :283621.00

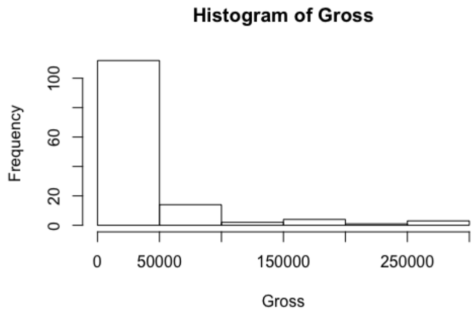
A "typical" movie has an IMDb score of around 65-70, with the most beloved movies in the dataset being 'Joker' and 'Parasite' with an IMDb score of 86. The least liked one was 'Cats', with an IMDb score of 28. On the other hand, a "typical" movie has an Metascore of around 57-63, with the most well performing movie amongst critics being 'Parasite' with an Metascore of 96. The least critically acclaimed movie in the dataset was 'Polar', with a score of 19. The "typical" movie had a runtime of around 115 minutes. Unsurprisingly, the longest movie in the dataset was Scorsese's acclaimed 'Irishman' running for a total of 209 minutes, arguably boosting the movie's popularity due to the traction it gained with people on the internet joking about having trouble not falling asleep. The shortest film was 'I lost my body' an animated French film running for only 81 minutes. The most voted on movie was blockbuster 'Joker' with 695,789 votes, while a typical movie only received around 83 thousand votes. The highest grossing film was 'Parasite', with a Gross Income of $283621 million USD, with an average film earning around $28-29 million USD.
Below are frequency histograms for each of the Variables:



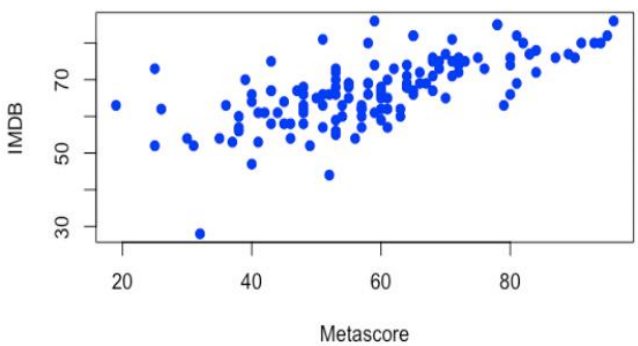


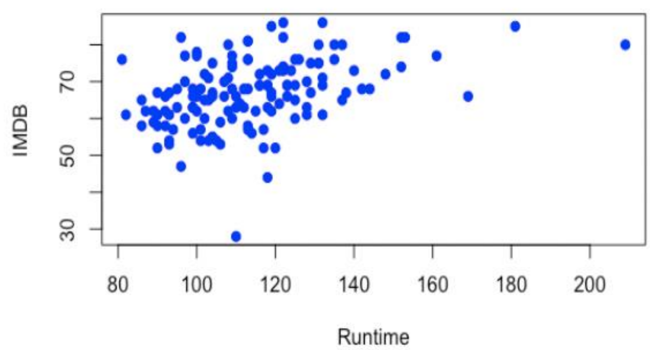


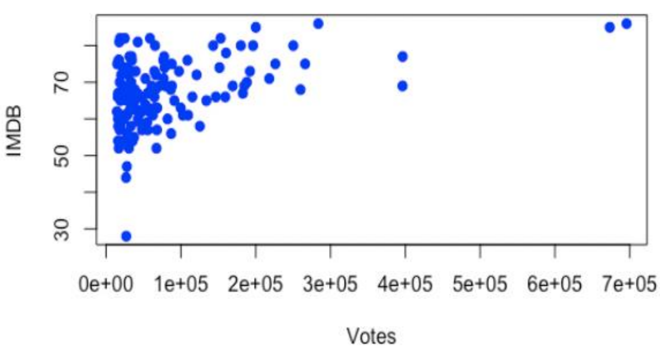


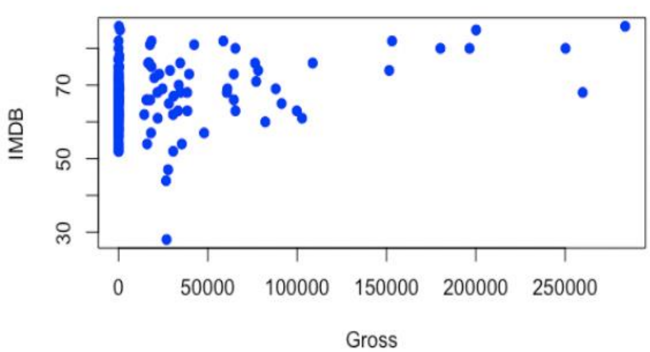


To explore the relationship of the predictors to the target variable on an preliminary level, I used scatter plot of the response variable versus each predictor. These describe the independent relationships however and not how the predictors will work together in our model.
As expected, the movies that receive higher Metascores in general also tend to receive higher metascores. Surprisingly, it seems that runtime might also be a factor, with longer running movies tending to fractionally receive better IMDb scores. The weakest relationship seems to be that with Votes and Gross Income.









We can see both from the histograms and the scatter plots above that, while Metascore seems to be fairly symmetric, the other predictors seem to be skewed to the right (long right-tailed). This may suggest that logarithms might be applicable in modeling.

Below are the results of a regression of IMDb score on the 4 predictors:

```
Call:
lm(formula = IMDB ~ Votes + Gross + Runtime + Metascore)

Residuals:
     Min       1Q   Median       3Q      Max
-29.0518  -3.8062   0.3679   3.4572  17.6529

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.022e+01  4.038e+00   9.960  < 2e-16 ***
Votes       1.641e-05  6.389e-06   2.569   0.0113 *
Gross       9.660e-06  1.164e-05   0.830   0.4080
Runtime     5.067e-02  3.589e-02   1.412   0.1603
Metascore   3.298e-01  3.815e-02   8.646 1.64e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.482 on 131 degrees of freedom
Multiple R-squared:  0.5241,    Adjusted R-squared:  0.5095
F-statistic: 36.06 on 4 and 131 DF,  p-value: < 2.2e-16
```

Regression Equation

$$\boxed{\begin{aligned}\text{IMDb score} &= 40.22 + 1.641(10^{-5})Votes + 9.66(10^{-6})Gross \\ &+0.05067\ Runtime + .3298 Metascore\end{aligned}}$$

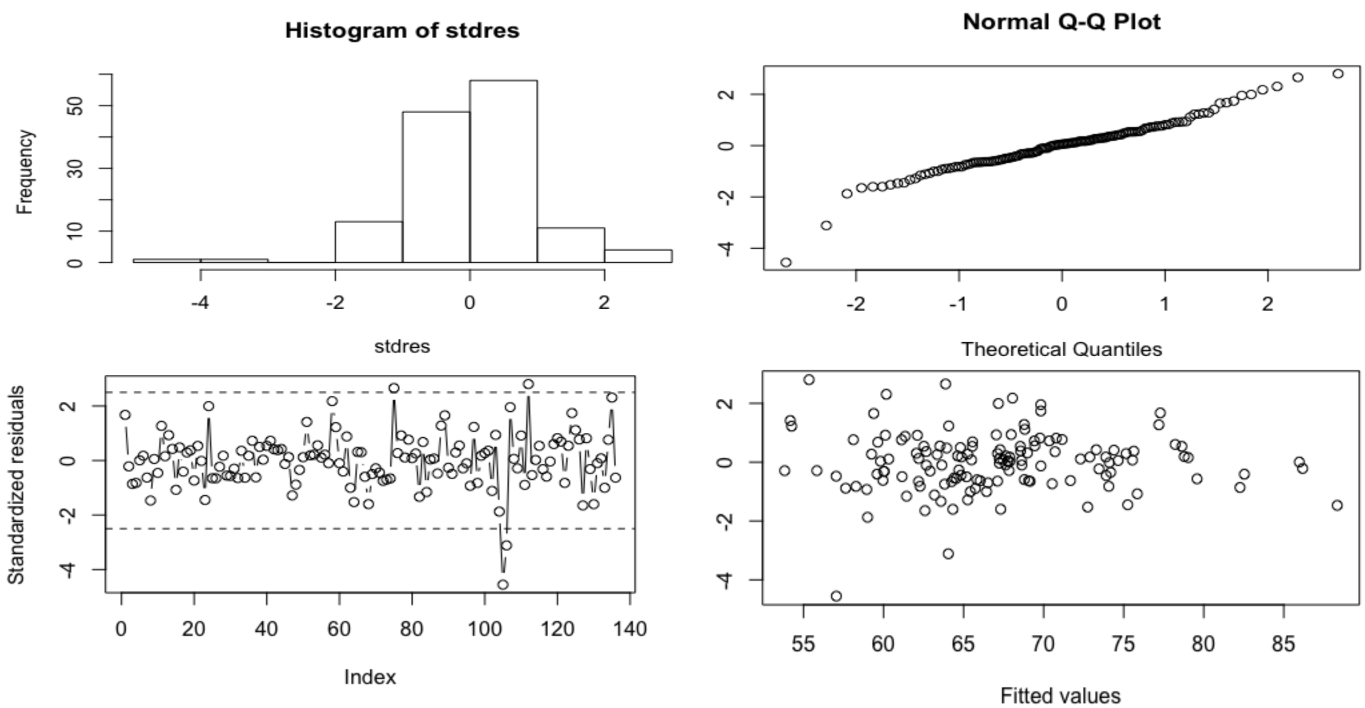
The regression is quite strong with an adjusted R-squared of .5095. Somewhat shockingly, it seems that given the other information, a movie’s gross income adds virtually no predictive power to the model. This suggests that a blockbuster, that brought in a lot of money is not necessarily the best choice when looking for a film most people would love. Similarly, given the other values, the number of votes a film receives seems to also contribute insignificantly to the predictive power of the model. The coefficient for **Runtime** shows that given the other 3 predictors are held constant, a one-minute increase in runtime is associated with an estimated expected increase in the movie’s IMDb score of .05 points. The coefficient for **Metascore** says that given the other 3 predictors are held constant, a one-point increase is associated with an expected increase in IMDb score of .33 points. The value for the residual standard error implies that a rough 95% prediction interval for a movies IMDb score using this model is ±12.964.

To measure multicollinearity, I used the variance inflation factor (VIF), which assesses how much the variance of an estimated regression coefficient increases if the predictors are correlated. Each of these VIF values corresponds to an estimate of how much we think the variance of its predictor’s β hat, the variance of its slope coefficient, has been inflated relative to what the variance would have been if our data had perfectly uncorrelated predictors. My VIF numbers are relatively close to 1, so collinearity doesn’t seem to be a concern at this point. Apparently, the number of votes a movie gets, its gross income, its runtime and its perception by critics do not necessarily go together.

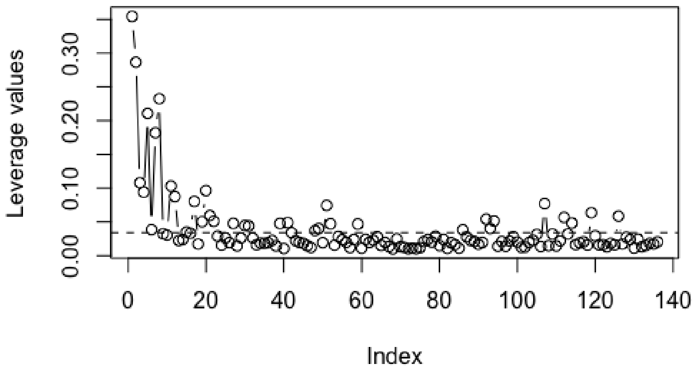
VIF VALUES

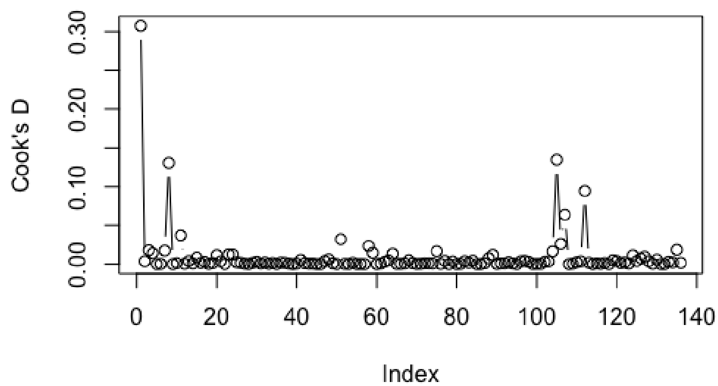
Votes	Gross	Runtime	Metascore
1.402295	1.247410	1.596317	1.196928

Looking now at the relevant residual plots, the standardized residual versus fitted values plot, the normal plot of the residuals and the standardized residuals versus index show the movie “Cats” to be a clear outlier. The movie has a standardized residual of much higher than ±2.5, meaning that we expect it to occur far less than 1% by random chance in the sample. The movie faced a slightly better critic than public reception, but in general underperformed massively. With an IMDb score of 28, the smallest one in the data set, the model is unable to predict it. The runtime was close to the typical 115minute-long film, running for 110 minutes. The main problem is that the intercept is on its own higher than the IMDb score received. The normal plot shows the some heavy-tailedness, with the right upper end of the normality plot going above the hypothetical straight and the left lower end going below it. The residuals versus fitted values plot also shows some non-constant variance.



To explore the potential existence of leverage points or influential points, I looked at Leverage values and Cook's distances; We can recognize here that a few points seem to have high leverage values and Cook’s distance, especially the observation at index 0, therefore it is important to explore the implications of removing them on our model.





We should now consider our options to simplifying the model. A reasonable first step would be to omit the Gross Income as a predictor, or the number of Votes, or both. Instead I decided to first consider a few linear restrictions on the model. I wanted to see see if only the sum of all the predictors might matter. I created the variable:

$$\boxed{\text{Total} = \text{Metascore} + \text{Gross Income} + \text{number of votes} + \text{runtime}}$$

The null hypothesis in this case would be that the simpler model that follows, that ther predictors added up are adequate and so one predictor is needed rather than 4. The alternative hypothesis would then be the full model.

The full model is:

$$\boxed{\text{IMDb score} = \beta_0 + \beta_1 * \text{Metascore} + \beta_2 * \text{Gross Income} + \beta_3 * \text{number of votes} + \beta_4 * \text{runtime} + \varepsilon}$$

The restricted (subset) model is

$$\boxed{\text{IMDb score} = \beta_0 + \beta_1 * (\text{Metascore} + \text{Gross Income} + \text{number of votes} + \text{runtime}) + \varepsilon}$$

The second model is basically a simplification of the first with $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \gamma_1$

```
Linear hypothesis test

Hypothesis:
- Runtime  + Metascore = 0
Votes - Runtime = 0
Gross - Runtime = 0

Model 1: restricted model
Model 2: IMDB ~ Votes + Gross + Runtime + Metascore

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    134 8991.8
2    131 5503.8  3      3488 27.673 6.295e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

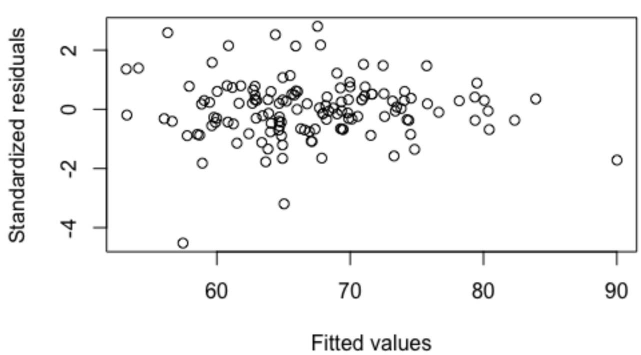
To explore the validity of this formula, I used a partial F-Test. Not surprisingly, the F value of 27.673, with the appropriate degrees of freedom corresponds to a p-value of of $6.295 * 10^{-14}$. Therefore there is really strong evidence against the null hypothesis and we accept the alternative. As we expected, the simplified model is not sufficient, both because the predictors have different predictive strength as explored above, but also because they are on a different scale.

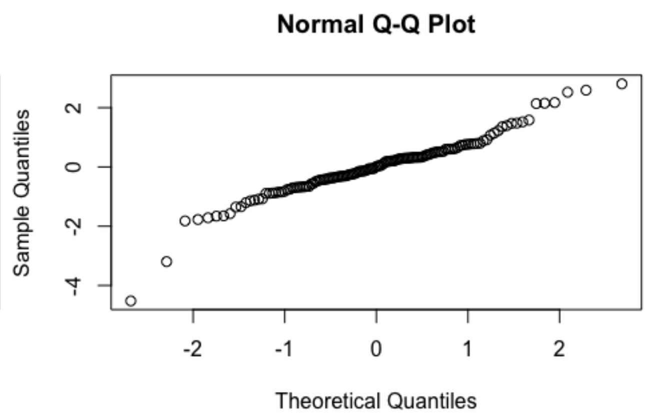
Eliminating the least effective predictor, gross income adjusted R-square actually increases to 51.07%. The regression:

```
lm(formula = IMDB ~ Metascore + Runtime + Votes)

Residuals:
     Min       1Q   Median       3Q      Max 
-28.8543  -3.6416   0.5729   3.3124  17.7323 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.911e+01  3.804e+00  10.281  < 2e-16 ***
Metascore   3.362e-01  3.733e-02   9.007 2.06e-15 ***
Runtime     5.951e-02  3.423e-02   1.739   0.0844 .  
Votes       1.643e-05  6.381e-06   2.575   0.0111 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.474 on 132 degrees of freedom
Multiple R-squared:  0.5216,	Adjusted R-squared:  0.5107 
F-statistic: 47.96 on 3 and 132 DF,  p-value: < 2.2e-16
```

Regression Equation

$$\text{IMDb score} = 39.11 + 1.643(10^{-5})Votes + 0.05951\,Runtime + .3362Metascore$$





I also tried removing the second least effective predictor, number of Votes. Adjusted R square falls to 48.887%

```
lm(formula = IMDB ~ Metascore + Runtime + Gross)

Residuals:
     Min       1Q   Median       3Q      Max 
-29.4594  -3.3356   0.0789   3.3609  18.4483 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.609e+01  3.781e+00   9.544  < 2e-16 ***
Metascore   3.425e-01  3.862e-02   8.866 4.55e-15 ***
Runtime     9.229e-02  3.269e-02   2.823   0.0055 ** 
Gross       9.794e-06  1.188e-05   0.824   0.4112    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.618 on 132 degrees of freedom
Multiple R-squared:  0.5001,    Adjusted R-squared:  0.4887 
F-statistic: 44.02 on 3 and 132 DF,  p-value: < 2.2e-16
```

Regression Equation

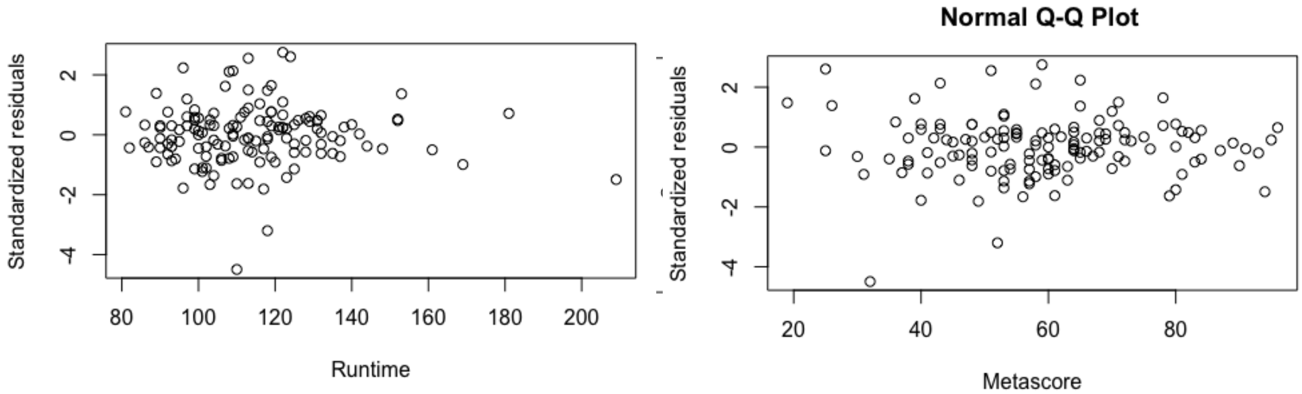
IMDb score $= 36.09 + 9.794(10^{-6})Gross + 0.09229\ Runtime + .3435 Metascore$

Now removing both the number of votes and the gross income, adjusted R-squared is smaller than the original case and the first simplified model, but higher than the second simplified model. This model however only relies on 2 predictors. The regression follows:

```
lm(formula = IMDB ~ Metascore + Runtime)

Residuals:
     Min       1Q   Median       3Q      Max 
-29.2598  -3.4911   0.1319   3.2398  18.1035 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 34.94969    3.51651   9.939  < 2e-16 ***
Metascore    0.34893    0.03777   9.238 5.28e-16 ***
Runtime      0.10131    0.03077   3.293  0.00127 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.61 on 133 degrees of freedom
Multiple R-squared:  0.4975,    Adjusted R-squared:   0.49 
F-statistic: 65.84 on 2 and 133 DF,  p-value: < 2.2e-16
```

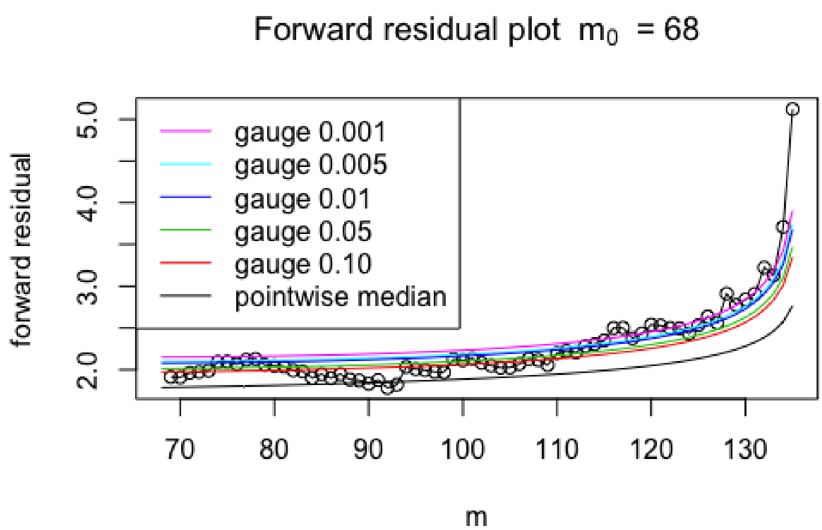
Regression Equation

IMDb score $= 34.94969 + +.10131\ Runtime + .34893 Metascore$



Looking at the standardized residuals against the individual predictors, we can see significantly stronger inconstant variance in the runtime predictor.

A reasonable next step now would be to try to run the regression, having removed the outlier:



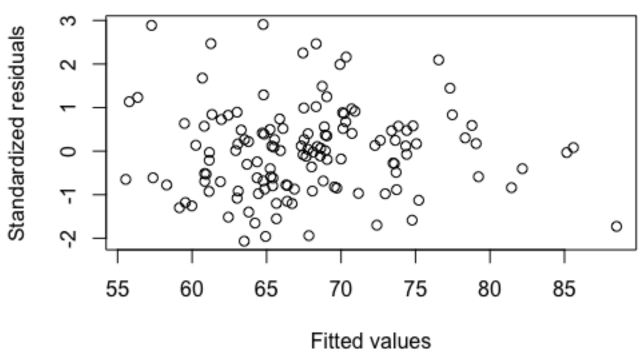
I used the package ForwardSearch to apply some outlier identification methods that are designed to avoid masking and swamping and to help me correctly identify the outlier. The system pointed out element [105] to be an outlier, which is indeed the movie "Cats". The resulting regression is as follows;

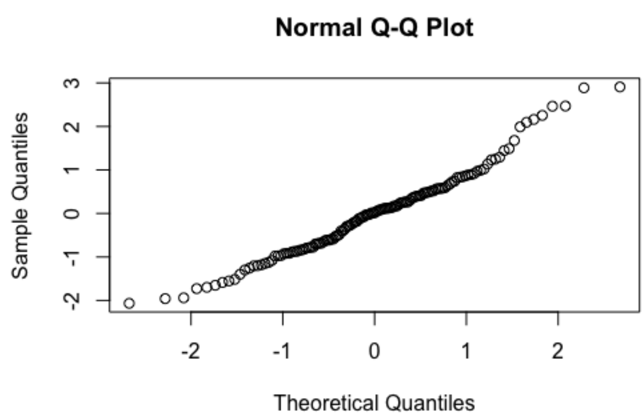
The regression is even stronger now with an R squared of 54.36%, increasing by more than 3%. The residual versus fitted values plot and normal plot show the same issues, signs of non-constant variance and heavy-tailedness.

```
Call:
lm(formula = IMDB ~ Metascore + Runtime + Votes + Gross)

Residuals:
     Min       1Q   Median       3Q      Max
-12.8800  -4.1284   0.2517   3.0516  16.4476

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.157e+01  3.576e+00  11.625  < 2e-16 ***
Metascore   2.977e-01  3.403e-02   8.747 1.07e-14 ***
Runtime     5.883e-02  3.177e-02   1.852   0.0664 .
Votes       1.490e-05  5.658e-06   2.633   0.0095 **
Gross       1.235e-05  1.028e-05   1.201   0.2319
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.71 on 128 degrees of freedom
Multiple R-squared:  0.5574,    Adjusted R-squared:  0.5436
F-statistic:  40.3 on 4 and 128 DF,  p-value: < 2.2e-16
```

Regression Equation

$$\text{IMDb score} = 41.57 + 1.49(10^{-5})Votes + 1.235(10^{-5})Gross + 0.05883\,Runtime + 0.2977Metascore$$





Removing now the Gross Income predictor, adjusted R square reaches 54.2%:
Regression Equation

$$\text{IMDb score} = 40.12 + 1.49(10^{-5})Votes + 0.07023\ Runtime + .3061 Metascore$$

```
lm(formula = IMDB ~ Metascore + Runtime + Votes)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.5201  -3.9319   0.2143   3.0039  16.7013 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.012e+01  3.372e+00  11.899  < 2e-16 ***
Metascore   3.061e-01  3.336e-02   9.176  9.3e-16 ***
Runtime     7.023e-02  3.037e-02   2.313  0.02233 *  
Votes       1.490e-05  5.667e-06   2.629  0.00962 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.72 on 129 degrees of freedom
Multiple R-squared:  0.5524,	Adjusted R-squared:  0.542 
F-statistic: 53.07 on 3 and 129 DF,  p-value: < 2.2e-16
```

Removing the number of Votes instead, Adjusted R squared falls to 52.26%:
Regression Equation

$$\text{IMDb score} = 37.81 + 1.235(10^{-5})Gross + 0.09706\ Runtime + .3088 Metascore$$

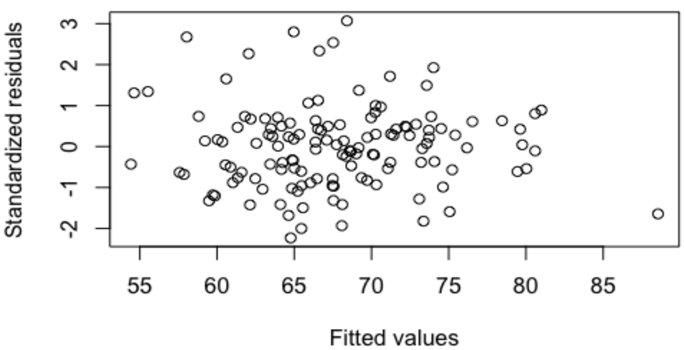
```
lm(formula = IMDB ~ Metascore + Runtime + Gross)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.8179  -3.8965   0.0304   2.9858  18.1320 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.781e+01  3.352e+00  11.278  < 2e-16 ***
Metascore   3.088e-01  3.454e-02   8.939  3.5e-15 ***
Runtime     9.706e-02  2.890e-02   3.358  0.00103 ** 
Gross       1.235e-05  1.052e-05   1.174  0.24250    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.84 on 129 degrees of freedom
Multiple R-squared:  0.5334,	Adjusted R-squared:  0.5226 
F-statistic: 49.16 on 3 and 129 DF,  p-value: < 2.2e-16
```

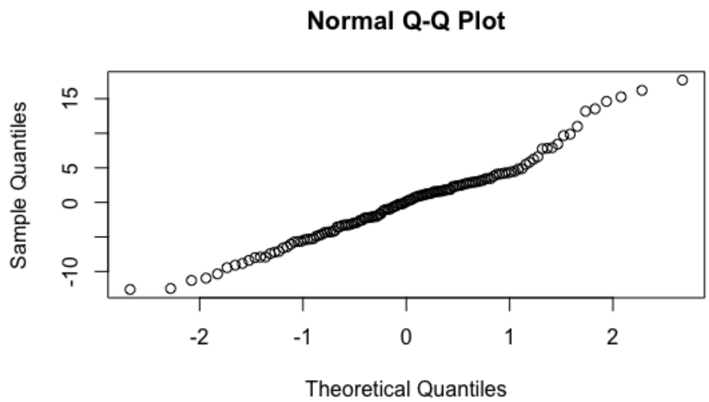
Removing now both predictors, the adjusted R squared goes to52.12%:

Regression Equation

$$\text{IMDb score} = 36.35803 + 0.10846\,Runtime + .31719 Metascore$$

```
lm(formula = IMDB ~ Metascore + Runtime)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.5905  -3.5476   0.1352   2.9477  17.6953 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 36.35803    3.12155  11.647  < 2e-16 ***
Metascore    0.31719    0.03384   9.374 2.88e-16 ***
Runtime      0.10846    0.02726   3.979 0.000114 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.848 on 130 degrees of freedom
Multiple R-squared:  0.5284,	Adjusted R-squared:  0.5212 
F-statistic: 72.84 on 2 and 130 DF,  p-value: < 2.2e-16
```





We can get the highest adjusted R-squared by using the whole model with 4 predictors, at 54.36%. However, to decide the best model out of the different options, I decided to compare the Cp values of the models. Here are the Cp values of all the best possible models for each number of predictors, for the model without the outliers. I decided to use the Cp metric and so the best model for our data seems to be a 3-predictor model, predictors being Metascore, Runtime and number of Votes.

```
> leaps(cbind(Metascore,Gross,Runtime,Votes), IMDB, nbest=2)
$which
      1     2     3     4
1  TRUE FALSE FALSE FALSE
1 FALSE FALSE  TRUE FALSE
2  TRUE FALSE FALSE  TRUE
2  TRUE FALSE  TRUE FALSE
3  TRUE FALSE  TRUE  TRUE
3  TRUE  TRUE FALSE  TRUE
4  TRUE  TRUE  TRUE  TRUE

$label
[1] "(Intercept)" "1"           "2"           "3"           "4"

$size
[1] 2 2 3 3 4 4 5

$Cp
[1] 23.920457 98.797127  7.985726  9.497982  4.625376  6.355413  5.000000
```

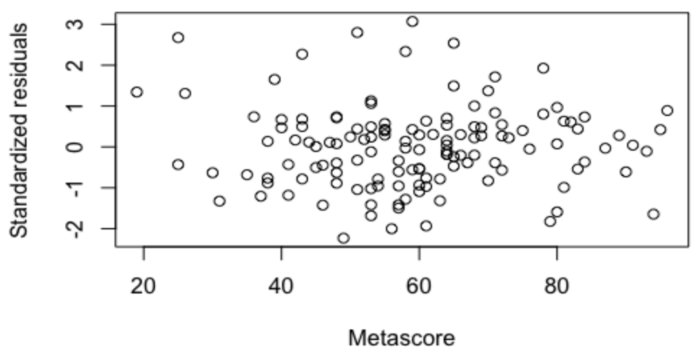
Therefore, the best regression equation is:

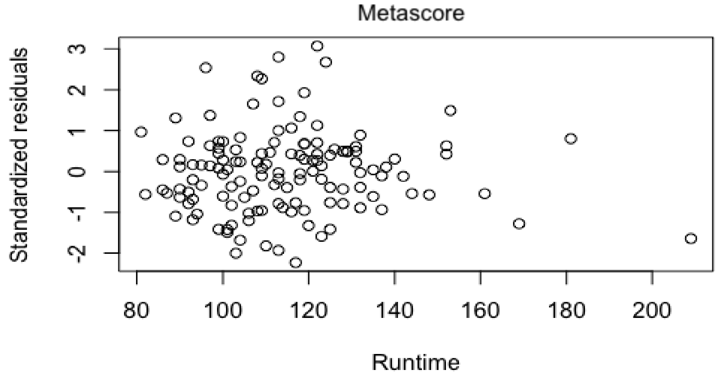
Regression Equation

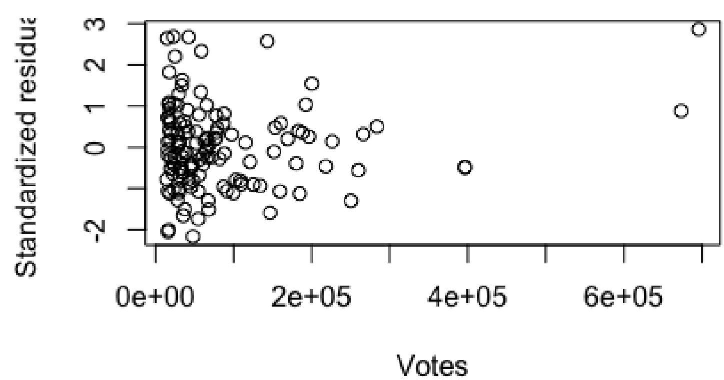
$$\text{IMDb score} = 40.12 + 1.49(10^{-5})Votes + 0.07023\,Runtime + .3061 Metascore$$

With an R squared of 54.2%.

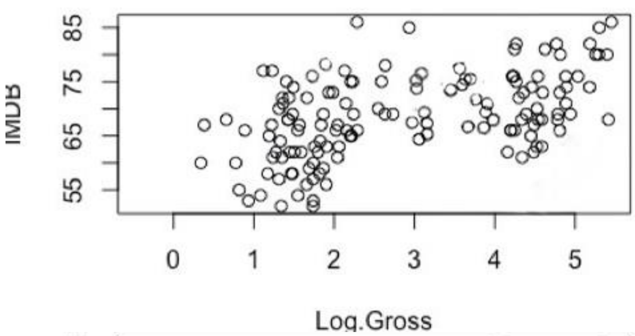
Below are plots of the standardized residuals against each of the predictors.

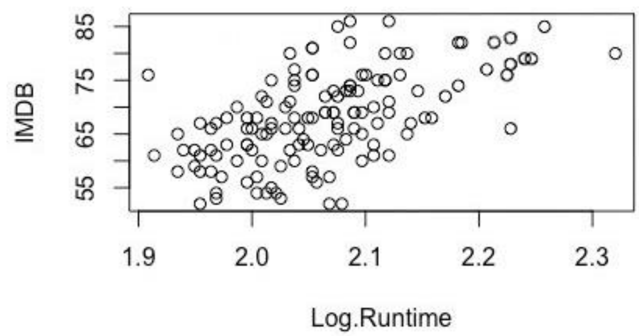


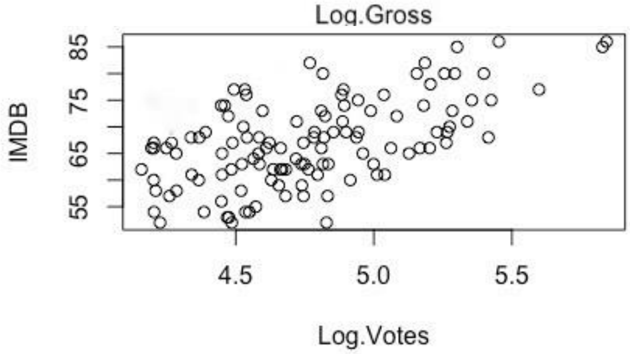




Since 3 out of the 4 predictors used on the model seemed right tailed, I decided to also explore the potential for modeling them using a semi-logarithmic model. These are the three variables logged plotted against the target variable:





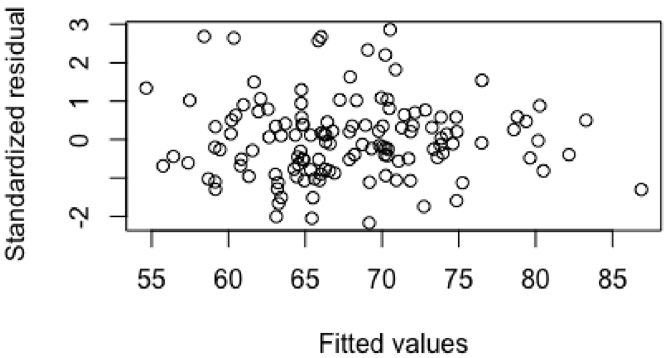


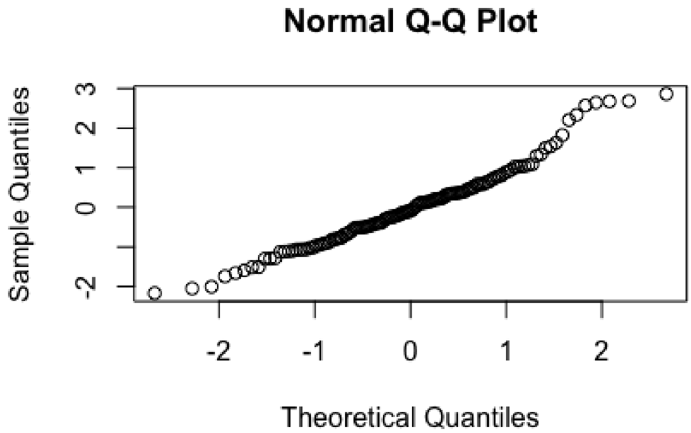
We can see there is a clear linear relationship between the log of a movie's votes and runtime versus their IMDB score. There is linear relationship with the log of the Gross income, but it is somewhat weaker.
An initial regression shows:

```
lm(formula = IMDB ~ Metascore + Log.Gross + Log.Runtime + Log.Votes)

Residuals:
    Min      1Q  Median      3Q     Max 
-12.166  -3.748  -0.483   2.790  15.481 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -7.26500   14.94319  -0.486   0.6277    
Metascore    0.29722    0.03328   8.932 3.86e-15 ***
Log.Gross    0.60019    0.32063   1.872   0.0635 .  
Log.Runtime 21.81282    8.71497   2.503   0.0136 *  
Log.Votes    2.28797    1.53994   1.486   0.1398    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.655 on 128 degrees of freedom
Multiple R-squared:  0.5473,    Adjusted R-squared:  0.5332 
F-statistic: 38.69 on 4 and 128 DF,  p-value: < 2.2e-16
```

Regression Equation

$$\text{IMDb score} \quad = -7.265 + 2.28797 * logVotes + .6 * logGross \\ +21.81282 * logRuntime + 0.29722Metascore$$





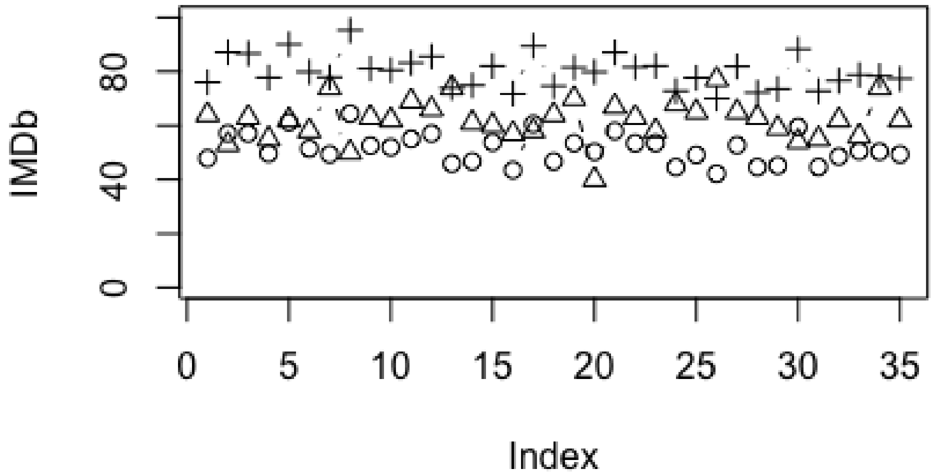
I investigated the “best model” by using a best subsets regression. I looked at Mallows Cp and choose to minimize Cp, which recommended the full four-variable model. That gives us an R squared of 54.7%. The normal plot shows the some heavy-tailedness, with the right upper end of the normality plot going above the hypothetical straight and the left lower end going below it. The residuals versus fitted values plot also shows some non-constant variance.

```
> leaps(cbind(Metascore,Log.Gross,Log.Runtime,Log.Votes),IMDB,nbest=2)
$which
      1     2     3     4
1  TRUE FALSE FALSE FALSE
1 FALSE FALSE  TRUE FALSE
2  TRUE FALSE  TRUE FALSE
2  TRUE FALSE FALSE  TRUE
3  TRUE  TRUE  TRUE FALSE
3  TRUE FALSE  TRUE  TRUE
4  TRUE  TRUE  TRUE  TRUE

$label
[1] "(Intercept)" "1"           "2"           "3"           "4"

$size
[1] 2 2 3 3 4 4 5

$Cp
[1] 21.770222 94.829720  6.207804 13.397734  5.207452  6.504017  5.000000

$r2
[1] 0.4667845 0.2084016 0.5288960 0.5034680 0.5395070 0.5349216 0.5473139

$adjr2
[1] 0.4627142 0.2023589 0.5216482 0.4958290 0.5287979 0.5241058 0.5331675
```

The applicable regression equation therefore is the one above. This equation tells us that given that Gross income, Runtime and Metascore are held fixed, a one unit increase in the logged number of votes is associated with a 2.28797 unit increase in the IMDb score of a movie. Or multiplying number of votes by 10 is associated with an expected 2.28797 unit increase IMDb score, holding all else constant. Similarly multiplying the Gross income by 10, is associated with an expected .6 unit increase IMDb score, holding all else constant. On the contrary, a 1 point increase in the Metascore, is associated with an expected .29722 unit increase IMDb score, holding all else constant. The regression is somewhat significant with 53.32% of the variability in IMDb being accounted for by the predictors. A prediction interval for the IMDb score is ±2ˆσ ≈ 11.31 that is 95% of the time the logged inbound tourism is known within ±11.31. The normal plot shows the some heavy-tailedness, with the right upper end of the normality plot going above the hypothetical straight and the left lower end going below it. The residuals versus fitted values plot also shows some non-constant variance.

One way that this model might be used is to predict other movies' IMDb scores. For that purpose, I decided to web-scrape the next 50 most voted movies for 2019 and apply this model to them:

	Title	IMDB score	fit	lower predicted limit	upper predicted limit
1	Angry Birds 2: Η Ταινία	64	61.92510111	47.87998241	75.97021981
2	Αντίστροφη Μέτρηση	53	72.09049477	57.09523686	87.08575268
3	Guns Akimbo	63	71.8695473	57.07019805	86.66889655
4	Ο Παραλίας	55	63.70387767	49.63874844	77.7690069
5	Κάποιος Υπέροχος	62	75.64797983	61.2210604	90.07489926
6	Κάνε να Χιονίσει	58	65.66397119	51.48107598	79.8468664
7	Honey Boy	74	63.60915624	49.34515963	77.87315285
8	47 Meters Down: Uncaged	50	79.83073139	64.42097872	95.24048407
9	Μετά την καταστροφή	63	66.79997664	52.53589627	81.064057
10	Color Out of Space	62	66.23365936	51.96464053	80.50267818
11	Blinded by the Light	69	69.12205104	55.04606525	83.19803683
12	Ένας Καλός Ψεύτης	66	71.31169218	57.05894608	85.56443829
13	Ένας αληθινός φίλος	74	59.93058317	45.8844072	73.97675913
14	Κινητό, αγάπη μου (2019)	61	60.80342638	46.70426152	74.90259124
15	The Kid Who Would Be King	60	67.91088085	53.76813053	82.05363117
16	Η Τρύπα	57	57.60986352	43.4722425	71.74748454
17	Jay and Silent Bob Reboot	58	75.02718795	60.58188983	89.47248608
18	Harriet	64	60.65589598	46.70515976	74.6066322
19	Queen & Slim	70	67.48519019	53.49643195	81.47394842
20	Τραύματα	40	65.04019586	50.26713831	79.81325341
21	400 Μίλια Αγάπης	67	72.50565531	57.98971305	87.02159757
22	Haunt	63	67.52681747	53.46713842	81.58649651
23	Unplanned	58	67.8103936	53.53554681	82.08524038
24	Brittany Runs a Marathon	68	58.70704007	44.60791926	72.80616088
25	Πού Χάθηκες, Μπερναντέτ	65	63.39642682	49.11994953	77.67290411
26	Τα Γεράκια της Νύχτας	77	56.08760001	42.18090122	69.9942988
27	3 Δευτερόλεπτα	65	67.30187668	52.72501693	81.87873643
28	Η Λαίδη και ο Αλήτης	63	58.39520019	44.60890149	72.18149889
29	Προάγγελμα Θανάτου	59	59.33530471	45.15085864	73.51975078
30	Χελς Κίτσεν: Οι βασίλισσες του εγκλήματος	54	73.84997005	59.46514105	88.23479905
31	The Intruder	55	58.71424755	44.64631406	72.78218105
32	Η καρδερίνα	62	62.5704273	48.51223502	76.62861958
33	Point Blank: Αντίστροφη Μέτρηση	56	64.6304615	50.51969169	78.74123131
34	The Last Black Man in San Francisco	74	64.37954615	50.34499487	78.41409743
35	Noelle	62	63.30950597	49.31099558	77.30801636



On the above graph, the actual IMDb score of each movie is shown as a triangle, the upper limit is shown as a cross and the lower as a circle. The model does a relatively good job predicting the intervals with only 3 values outside of the prediction intervals. We would expect at least 1 out of 20 values to be outside of the bounds since these are 95% prediction intervals. The three movies that are found outside of the prediction intervals are “Countdown “, “47 meters down” and “Traumas” all of which underperformed compared to our model.

Discussion and Conclusion

Although a single regression model predicting the IMDb score of a movie based on critic reviews, number of votes on IMDb, Runtime and Gross Income was successfully generated, we cannot infer causality for the variables given that this is an observational study in which the variables were not controlled but simply recorded. Prediction is possible but with somewhat limited precision.

There were also flaws in the data that were beyond the control of this study. Out of the 150 films for which IMDb scores were collected, only 136 films had meta-scores. This presented a challenge in using stepwise regression to create a linear model. This deletion of rows with missing values, reduced the sample size and affected the informativeness of the regression. To get around this problem, only the 136 films for which both an IMDb and a meta-score was available were used in stepwise regression.

In order to create a possibly more accurate multiple linear regression, further research on other factors that could affect the popularity of a movie should investigated. Data on word-of-mouth interaction, marketing budgets, production values and distributors, though hard to collect or retrieve could offer additional insights. Ideally, further research may also use a larger sample size in order to mitigate the effect of deletion due to missing observations.

Appendices

Appendix A.

In [1]:
```
urls = ["https://www.imdb.com/search/title/?title_type=feature&year=2019-01-01,2019-12-31&sort=num_votes,desc",'h
import requests
r = requests.get(urls[0])
r1 =requests.get(urls[1])
r2 = requests.get(urls[2])
r.status_code
r1.status_code
r2.status_code
```

Out[1]: 200

In [2]:
```
html = r.text
html1 = r1.text
html2 = r2.text
import bs4
soup = bs4.BeautifulSoup(html, 'html.parser')
soup1 = bs4.BeautifulSoup(html1, 'html.parser')
soup2 = bs4.BeautifulSoup(html2, 'html.parser')
type(soup)
soup.title.text.strip()
```

Out[2]: 'Feature Film,\nReleased between 2019-01-01 and 2019-12-31\n(Sorted by Number of Votes Descending) - IMDb'

In [3]:
```
movies1 = soup.find_all('div', class_="lister-item mode-advanced")
movies2 = soup1.find_all('div', class_="lister-item mode-advanced")
movies3 = soup2.find_all('div', class_="lister-item mode-advanced")
movies=[]
for movie in movies1:
    movies.append(movie)
for movie in movies2:
    movies.append(movie)
for movie in movies3:
    movies.append(movie)
len(movies)
```

Out[3]: 150

In [4]:
```
tags=[]
for per_movie in movies:
    collection = per_movie.findAll("img")
    for img in collection:
        if 'alt' in img.attrs:
            if img.attrs['alt'] not in tags:
                tags.append(img.attrs['alt'])
len(tags)
```

Out[4]: 150

ı [121]:
```
metascores=[]
for per_movie in movies:
    meta_score = per_movie.find('div', class_="inline-block ratings-metascore")
    if meta_score!=None:
        if meta_score.find('span', class_="metascore mixed") != None:
            metascores.append(meta_score.find('span', class_="metascore mixed").text.strip())
        elif meta_score.find('span', class_="metascore favorable") != None:
            metascores.append(meta_score.find('span', class_="metascore favorable").text.strip())
        elif meta_score.find('span', class_="metascore unfavorable") != None:
            metascores.append(meta_score.find('span', class_="metascore unfavorable").text.strip())
        else: metascores.append(None)
    else: metascores.append(None)
len(metascores)
```

ıt[121]: 150

ı [122]:
```
movie_ratings=[]
for per_movie in movies:
    rating= per_movie.find('strong').text
    rating = float(rating)/.1
    movie_ratings.append(rating)
len(movie_ratings)
```

ıt[122]: 150

In [123]:
```
movie_genres=[]
for per_movie in movies:
    genre= per_movie.find('span', class_="genre").text.strip()
    movie_genres.append(genre)
len(movie_genres)
```

Out[123]: 150

In [124]:
```
runtimes=[]
for per_movie in movies:
    runtime= per_movie.find('span',class_="runtime")
    runtime = int(runtime.text.strip()[0:3])
    runtimes.append(runtime)
len(runtimes)
```

Out[124]: 150

In [125]:
```
How_PG =[]
for per_movie in movies:
    PG= per_movie.find('span',class_="certificate")
    if PG!= None:
        How_PG.append(PG.text.strip())
    else: How_PG.append(None)
len(How_PG)
```

Out[125]: 150

In [5]:
```
gross_l = []
for per_movie in movies:
    gross= per_movie.find('p',class_="sort-num_votes-visible").find_all('span')[-1].text.strip()
    gross=gross.replace('M',"").replace('$',"")
    gross_l.append(gross)
len(gross_l)
```

Out[5]: 150

```
In [ ]: import xlsxwriter
        workbook = xlsxwriter.Workbook('2019_Movie_Data.xlsx')
        worksheet = workbook.add_worksheet()

        headings = ["Movie Title", "IMDB rating", "Metascore", "Genres","Runtime","PG rating","Number of Votes",
                    "Gross Income"]
        row=0
        column=0
        for heading in headings :
            worksheet.write(row, column, heading)
            column += 1
        #titles
        row=1
        column=0
        for title in tags :
            worksheet.write(row, column, title)
            row += 1
        #IMDB Rating
        row=1
        column=1
        for IMDB_rating in movie_ratings :
            worksheet.write(row, column, IMDB_rating)
            row += 1
        #Metascores
        row=1
        column=2
        for score in metascores :
            worksheet.write(row, column, score)
            row += 1
        #Genres
        row=1
        column=3
        for genres in movie_genres :
            worksheet.write(row, column, genres)
            row += 1
        #Runtimes
        row=1
        column=4
        for time in runtimes :
            worksheet.write(row, column, time)
            row += 1
        #How_PG
        row=1
        column=5
        for pg in How_PG :
            worksheet.write(row, column, pg)
            row += 1
        #n_votes
        row=1
        column=6
        for n in n_votes :
            worksheet.write(row, column, n)
            row += 1
        #gross_l
        row=1
        column=7
        for income in gross_l :
            worksheet.write(row, column, income)
            row += 1
        workbook.close()
```

Appendix B.

Movie Title	IMDB rating	Metascore	Genres	Runtime	PG rating	Number of V	Gross
Avengers: En	85	78	Action, Adve	181	PG-13	671,859	858.37
Captain Mar	69	64	Action, Adve	123	PG-13	395,782	426.83
Once Upon a	77	83	Comedy, Dra	161	R	393,816	135.37
Parasite	86	96	Comedy, Dra	132	R	275,863	275,863
Spider-Man:	75	69	Action, Adve	129	PG-13	265,380	388.53
Star Wars: T	68	53	Action, Adve	142	PG-13	258,853	258,853
The Irishmar	80	94	Biography, C	209	R	248,286	248,286
John Wick: C	75	73	Action, Crime	131	R	225,687	171.02
Shazam!	71	71	Action, Adve	132	PG-13	217,593	140.37
1917	85	78	Drama, War	119	R	195,788	195,788
Alita: Battle	73	53	Action, Adve	122	PG-13	191,688	85.71
Knives Out	80	82	Comedy, Crir	131	PG-13	190,357	190,357
Aladdin	70	53	Adventure, F	128	PG	187,524	354.87
Us	69	81	Horror, Myst	116	R	184,057	175.01
Glass	67	43	Drama, Sci-F	129	PG-13	182,566	111.04
Marriage Sto	80	93	Comedy, Dra	137	R	178,107	178,107
The Lion King	69	55	Animation, A	118	PG	168,348	540.08
Toy Story 4	79	84	Animation, A	100	G	159,291	433.03
It Chapter Tw	66	58	Drama, Fant	169	R	158,707	193.77
El Camino: A	74	72	Action, Crime	122	TV-MA	151,119	151,119
Ford v Ferrar	82	81	Action, Biogr	152	PG-13	149,989	149,989
Ad Astra	66	80	Adventure, D	123	PG-13	145,817	35.4
Jojo Rabbit	80	58	Comedy, Dra	108	PG-13	138,375	0.35
Fast & Furiou	65	60	Action, Adve	137	PG-13	133,752	165.55
X-Men: Dark	58	43	Action, Adve	113	PG-13	125,072	65.85
Midsommar	72	72	Drama, Horr	148	R	119,793	27.33
Pokémon De	66	53	Action, Adve	104	PG	114,965	144.11
Godzilla: Kin	61	48	Action, Adve	132	PG-13	109,195	110.5
Uncut Gems	76	90	Crime, Dram	135	R	105,619	105,619
6 Undergrou	61	41	Action, Adve	128	R	102,138	102,138
Terminator:	63	54	Action, Adve	128	R	98,507	98,507
Rocketman	73	69	Biography, D	121	R	97,021	96.37
Triple Fronti	65	61	Action, Adve	125	R	91,153	91,153
How to Train	75	71	Animation, A	104	PG	87,533	160.8
Men in Black	56	38	Action, Adve	114	PG-13	86,588	79.8
Zombieland:	68	55	Action, Come	99	R	85,595	26.8
Jumanji: The	69	58	Action, Adve	123	PG-13	84,551	84,551
Murder Myst	60	38	Action, Come	97	PG-13	81,977	81,977
Yesterday	69	55	Comedy, Fan	116	PG-13	78,982	73.29
Doctor Sleep	74	59	Drama, Fant	152	R	76,792	76,792
The Lighthou	77	83	Drama, Fant	109	R	76,346	0.43
The Two Pop	76	75	Comedy, Dra	125	PG-13	75,651	75,651
Frozen II	71	64	Animation, A	103	PG	75,463	75,463
Long Shot	69	67	Comedy, Ror	125	R	71,162	30.32
Escape Room	63	48	Action, Adve	99	PG-13	68,218	57.01
Pet Sematar	57	57	Horror, Myst	101	R	67,806	54.72
Hellboy	52	31	Action, Adve	120	R	67,188	21.9
Booksmart	72	84	Comedy	102	R	66,162	22.68
Ready or Not	68	64	Comedy, Hor	95	R	65,378	26.74
Polar	63	19	Action, Crime	118	TV-MA	65,307	65,307
Extremely W	66	52	Biography, C	110	R	64,380	64,380
The King	73	62	Biography, D	140	R	64,231	64,231
Little Wome	80	91	Drama, Rom	135	PG	64,156	64,156
Brightburn	61	44	Drama, Horr	90	R	62,257	17.3
The Highway	69	58	Biography, C	132	R	60,823	60,823
I Am Mother	68	64	Drama, Myst	113	TV-14	60,485	60,485
Klaus	82	65	Animation, A	96	PG	58,244	58,244
Rambo: Last	62	26	Action, Adve	89	R	57,594	18.87
Dumbo	63	51	Adventure, F	112	PG	56,004	114.77
Gemini Man	57	38	Action, Dram	117	PG-13	55,107	20.55
Maleficent: I	67	43	Adventure, F	119	PG	54,973	36.95
Isn't It Roma	59	60	Comedy, Fan	89	PG-13	54,771	48.79
Hustlers	64	79	Comedy, Crir	110	R	54,077	80.55
Fighting with	71	68	Biography, C	108	PG-13	52,258	22.96
Angel Has Fa	64	45	Action, Thrill	121	R	52,009	67.16
Cold Pursuit	62	57	Action, Crime	119	R	47,872	32.14
Velvet Buzzs	57	61	Horror, Myst	113	R	47,827	47,827
Crawl	62	60	Drama, Horr	87	R	46,384	39.01
The Lego Mo	66	65	Animation, A	107	PG	45,875	105.81
Happy Death	62	57	Comedy, Hor	100	PG-13	45,765	28.05
Annabelle Co	59	53	Horror, Myst	106	R	45,108	73.65
Scary Stories	62	61	Horror, Myst	108	PG-13	42,929	62.74
Captive State	60	54	Drama, Horr	109	PG-13	42,476	5.96
Good Boys	67	60	Adventure, C	90	R	41,670	69.06
Anna	66	40	Action, Thrill	119	R	40,540	7.74
The Gentlem	81	51	Action, Come	113	R	39,714	39,714
Dolemite Is I	73	76	Biography, C	118	R	39,280	39,280
Fractured	63	36	Mystery, Thri	99	TV-MA	38,380	38,380
Always Be M	68	64	Comedy, Ror	101	PG-13	38,362	38,362
The Secret Li	65	55	Animation, A	86	PG	38,102	158.14
The Dead Do	55	53	Comedy, Fan	104	R	37,188	6.56
Shaft	64	40	Action, Come	111	R	36,514	21.36
In the Tall Gr	54	46	Drama, Horr	101	TV-MA	35,306	35,306
Pain and Glo	76	87	Drama	113	R	34,224	34,224
The Hustle	54	35	Comedy, Crir	93	PG-13	34,105	35.42
The Peanut E	77	70	Adventure, C	97	PG-13	33,863	13.12
The Dirt	70	39	Biography, C	107	TV-MA	33,640	33,640
The Laundror	63	57	Comedy, Crir	95	R	33,131	33,131
Child's Play	58	48	Horror, Sci-Fi	90	R	32,935	29.21
Bombshell	68	64	Biography, D	109	R	32,889	32,889
The Farewell	77	89	Comedy, Dra	100	PG	30,749	16.88
The Silence	52	25	Drama, Horr	90	PG-13	30,520	30,520
In the Shado	62	48	Crime, Myste	115	TV-MA	30,389	30,389
Midway	67	47	Action, Dram	138	PG-13	29,869	29,869
The Curse of	53	41	Horror, Myst	93	R	29,665	54.73
Five Feet Ap	72	53	Drama, Rom	116	PG-13	29,600	45.73
Serenity	53	37	Drama, Myst	106	R	29,349	8.55
Stuber	61	42	Action, Come	93	R	28,090	22.37
A Beautiful D	74	80	Biography, D	109	PG	28,017	28,017
Last Christm	65	50	Comedy, Dra	103	PG-13	27,900	27,900
Ma	56	53	Horror, Myst	99	R	27,895	45.37
Downton Abl	74	64	Drama, Rom	122	PG	27,835	31.03
Gully Boy	82	65	Drama, Musi	153	Not Rated	24,826	5.57
Judy	69	66	Biography, D	118	PG-13	24,121	24,121
After	54	30	Drama, Rom	105	PG-13	24,098	12.14
Tolkien	68	48	Biography, D	112	PG-13	23,227	4.54
The Wanderi	60	57	Action, Dram	125	TV-MA	23,179	2.19
Between Tw	61	59	Comedy	82	TV-MA	21,851	21,851
Motherless E	69	60	Crime, Dram	144	R	21,452	21,452
The Report	72	66	Biography, C	119	R	19,841	19,841
The Prodigy	58	45	Horror, Thrill	92	R	19,243	14.86
Late Night	65	70	Comedy, Dra	102	R	19,107	15.5
The Art of Se	67	65	Comedy, Crir	104	R	18,389	2.41
Richard Jewe	75	68	Biography, C	131	R	18,162	18,162
Close	57	51	Action, Dram	94	TV-MA	18,161	18,161
Abominable	70	61	Animation, A	97	PG	18,068	20.61
Portrait of a	82	95	Drama, Rom	122	R	17,586	17,586
Togo	81	71	Adventure, B	113	PG	17,309	17,309
The Boy Who	76	68	Drama	113	TV-PG	17,096	17,096
21 Bridges	66	51	Action, Crime	99	R	17,033	17,033
What Men W	52	49	Comedy, Fan	117	R	16,766	54.61
I Lost My Bo	76	80	Animation, D	81	TV-MA	16,702	16,702
The Addams	58	46	Animation, C	86	PG	15,989	30.3
Wine Countr	54	56	Comedy	103	R	15,985	15,985
Dora and the	60	63	Adventure, F	102	PG	15,850	54.89
Missing Link	67	68	Animation, A	93	PG	15,840	16.65
The Aeronau	66	60	Action, Adve	100	PG-13	15,786	15,786
A Rainy Day	66	48	Comedy, Ror	92	PG-13	15,503	15,503
Dark Waters	76	72	Biography, D	126	PG-13	15,307	15,307
Someone Gr	62	63	Comedy, Ror	92	R	14,377	14,377