Dane Rasmussen STAT 4000 Final Project

**Dataset Description:**

For my dataset, I created a web scraper to get the production budget, gross domestic box-office earnings, and other data on the **top 300 domestic box-office earners** (From <https://www.the-numbers.com/> ).

Research Questions:

* Can we predict the domestic box-office earnings based on the movie’s budget
* Will we see any differences in box-office earnings based on whether the movie is animated, a sequel, or a remake?
* Do certain genres of movies make more money than other genres?

Variables collected:

* Rank – a rank of 25 indicates that that movie made the 25th most money at the domestic box office
* Title
* Gross domestic box office earnings
* Opening Weekend – how much money the movie made in its opening weekend
* Legs – measure of how the movie performed in the long run (Total Domestic Gross / Opening Weekend Gross)
* Domestic Share – number from 1 – 100 indicating how much money the movie made in the domestic box-office as opposed to the international box-office
* Production budget
* Inflation adjusted Domestic Box office earnings
* Number of theaters showing the movie on its opening weekend
* Max number of theaters showing the movie
* Animated – 1 if the movie was animated 0 if the movie was not
* Sequel – 1 if the movie is a sequel 0 if the movie is not
* Remake – 1 if the movie is a remake 0 if the movie is not
* Domestic ROI – return on investment based on the production budget and domestic box-office earnings of a movie ((Domestic Box-office earnings / Production Budget) – 1)
* Genre – categorical variable
  + Sci-fi
  + Superhero
  + Action
  + Adventure
  + Comedy
  + Fantasy
  + Biopic
  + Horror
  + Romance
  + Christmas / Holiday
  + War
  + Western
  + Concert

Some notes about the data

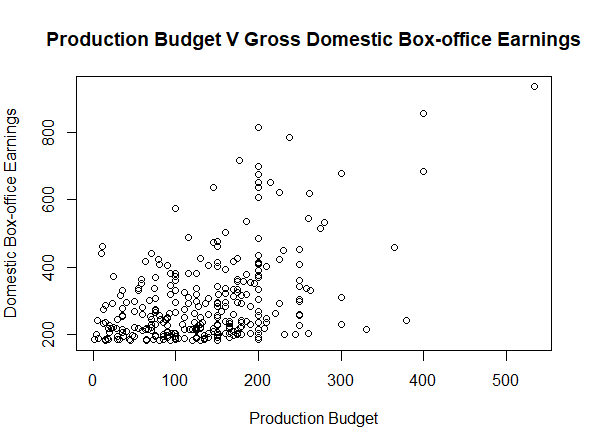
The website gives the caution that production budgets are not completely accurate since studios can use accounting tricks to inflate or deflate the budget for movies.

Additionally, the indicator variables (animated, sequel, remake, genre) were all hand jammed by me. This may have introduced unintentional bias.

Lastly, since these are the top 300 highest grossing movies in the domestic box-office, it could be argued that the model I choose is limited to a small subset of movies and lack any predictive power for movies that are not in the top 300.

**Initial plots:**

One assumption for fitting a Least squares regression against data is linearity. To check this, we can look at a scatter plot for our data.

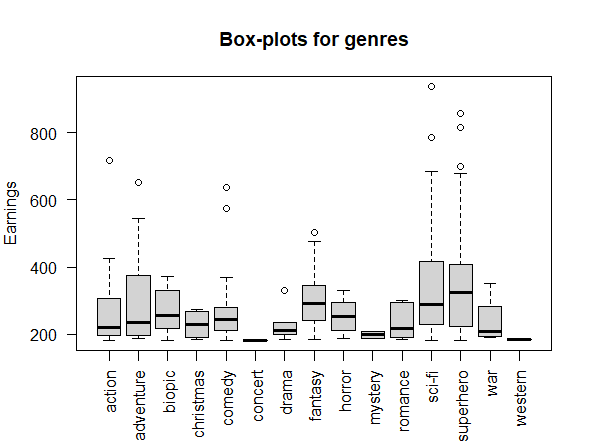


Given the plot and correlation coefficient of .4651331, there is weak evidence of a linear relationship between production budget and domestic box-office earnings.

Chart, scatter chart

AI-generated content may be incorrect.

Based off genre, there are no obvious trends as far as difference in means for the genres goes. Box-plots will confirm or deny this



Given the box-plot, I would say that there is clear evidence of skewed data. This makes sense considering we are taking the data from the top performing movies in the domestic market. The top 50 or so movies boast domestic earnings from 900 million to 400 million. After that, the earnings converge around 300-200 million and end up slowly dropping to about 180 million.

I started my analysis by trying to fit models with domestic box office earnings as response variables, but after trying many models with different transformations, I landed on the following model. I will check assumptions, outlier/influential points, and give confidence/prediction intervals for coefficients.

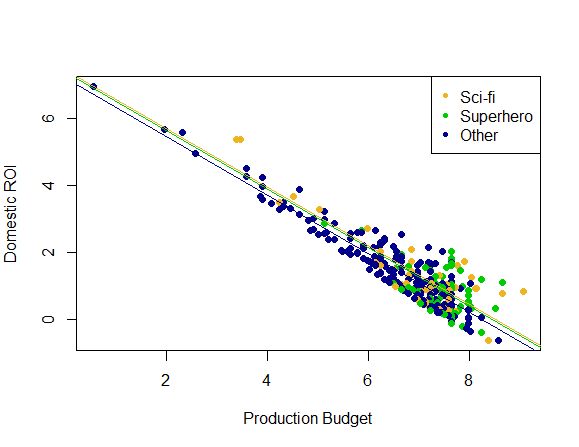
**Updated Model**:

Log(Domestic ROI + 1) = B0 + B1 \* Log(Production Budget) + B1 \* Sci-fi + B2 \* Superhero

Both Domestic ROI and Production are transformed by taking the natural log. Additionally, to get rid of negative values in Domestic ROI, I add 1 to each value.

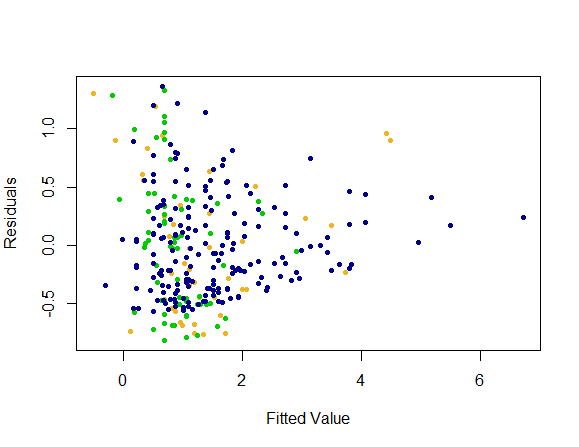
Whether the movie was a sci-fi movie, superhero movie, or neither seemed to be the only factors that contributed to predicting the Domestic Return on Investment.

Scatter plot:



From this scatter plot, I see that if the movie is sci-fi or a superhero movie does result in a higher ROI. Additionally, I see no evidence for interaction between genre and Production budget.

Residuals plot:



There is clear evidence of non-constant variance. As the fitted value increases, we see a clear decrease in the variance of the residuals.

My first idea before fitting a weighted least squares model was to identify influential points.

Leverage:

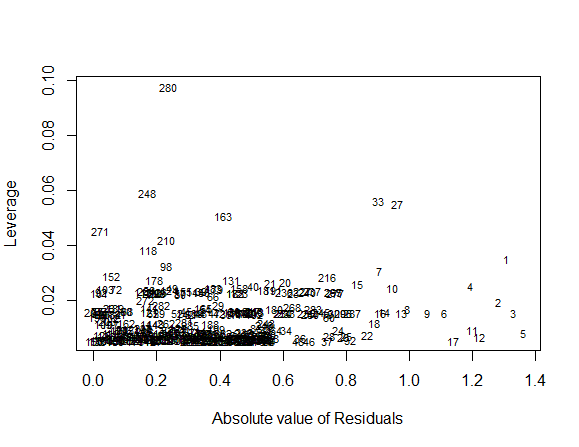
Chart, scatter chart

AI-generated content may be incorrect.

We can Identify 280 (Snow White and the seven dwarves), 27 (Star Wars Episode IV: A New Hope), 33 (E.T. Extra Terrestrial), 248 (Gone With the Wind), 163 (My Big Fat Greek Wedding), 271 (Grease), 210 (Back to the Future), 98 (Star Wars Episode VI: Return of the Jedi), and 1 (Star Wars Episode VII: The Force Awakens) as data points with peculiar X values.

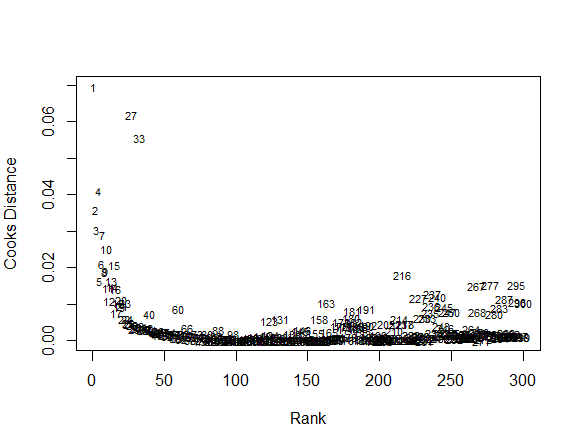
Influence:

Residuals V Leverage:



33 (E.T.) .and 27 (A New Hope) stand out particularly when plotted against residuals.

Cooks Distance:



Cooks Distance Singles out 1, 27, and 33. We can also consider looking at 4, 2, 3, 7, and 10. Considering these all have high ranks, we can assume that they have high positive residuals. 280 and 248 (Snow White and Gone with the Wind respectively) are hard to pick out in this graph.

I am going to remove points 1 (The Force Awakens, 2 (Avengers: Endgame), 3(Spider-man: No Way Home), 4 (Avatar), 7 (Avatar: The Way of Water), 27 (A New Hope), 33 (E.T), 248 (Gone with the Wind), 280 (Snow White). Even if our model can predict how well Gone with the Wind and Snow white would do, I doubt how accurate the data is for these movies considering how long ago they were made. This will probably not be enough to solve our problem with non-constant variance, so I will make a Weighted Least Squares Model using the same variables on the edited data

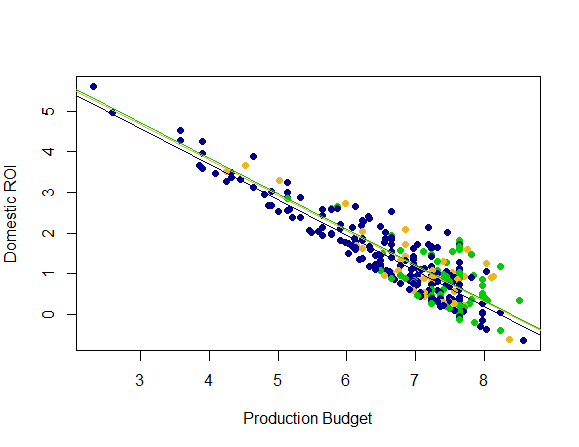
**Final Model Using Weighted Least Squares:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | T value | P value | Lower bound | Upper bound |
| Intercept | 7.44626 | 0.10619 | 70.123 | <2e-16 | 7.2372945 | 7.6552792 |
| (Production\_budget) | -0.91652 | 0.01833 | -49.988 | <2e-16 | -0.952 | -0.8804 |
| Superhero | 0.12917 | 0.07199 | 1.794 | 0.0739 | -0.0125 | 0.2709 |
| Sci-fi | 0.11408 | 0.07177 | 1.590 | 0.113 | -0.027 | 0.2534 |

Residual Standard error: 0.4495 R-squared: .803 adj R-squared: .8009

By removing those high influence points (which had quite a few sci-fi movies, and 2 Superhero movies), we now see that using the sci-fi genre or the superhero genre as indicator variables is no longer significant at the 0.05 level. It is noteworthy that the sci-fi and superhero genre had so many outliers, which could inform future producers about their unprecedented potential in the market.

Linearity:



The Assumption of linearity is validated here. There is a strong linear relationship between the base 2 log of Domestic ROI and the base 2 log of Production Budget + 1.

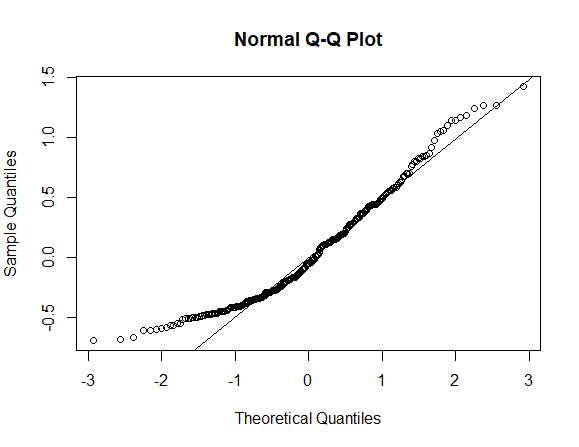
Constant Variance

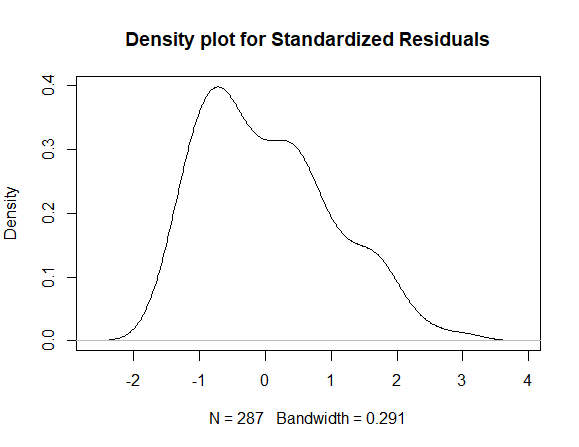
Chart, scatter chart

AI-generated content may be incorrect.

The weight that I used is . This gives a favorable weight to larger fitted values and gives a less favorable weight to values fitted close to 0.

Normality of Residuals





Based on the Density plot and QQ plot, I would say that our data is slightly right skewed. This should be expected since I have picked my data from the very best performing movies in the domestic market. To remedy this, I could remove more data points with high ranks, which tend to outperform predictions.

**Independence of Residuals**:

I could not get the data for the years for each of the movies. If I could, I would observe the residuals across time to see if there is any evidence of autoregression.

**Model Interpretation:**

The interpretation for the intercept is that if you made a movie that neither a superhero nor a sci-fi movie, and you made it with a budget of 0, you would expect to see a log base 2 ROI + 1 of 7.44626. This means you could expect an ROI of around 127. Our confidence interval gives a range of

With a confidence interval for B1 of [-0.952, -0.8804] we can be 95% confident that as the production budget doubles, the Base 2 log of the ROI plus one will decrease anywhere between .952 and .8804. A decrease in 1 of the bases 2 log of ROI would mean that your ROI is halved. With our Coefficient, we would not quite see a halving in the ROI, but it would be close.

This essentially means that as the production budget increases, you start to see diminishing returns on Domestic ROI. It should come as no surprise that the more money you pay to make a movie, the harder it is to make that money back.

For our estimate of the indicator variable for superhero movies, we got a value of .129. That means that compared to a movie that is not superhero or a sci-fi movie, we could expect to see domestic ROI increase by a factor of . Our confidence interval tells us that we may see an increase by a factor of [] This interval includes 2 to the power of 0, which would mean an increase by a factor of 1, which would mean 0 change in ROI.

Our indicatory coefficient for the sci-fi genre has a similar interpretation to the indicator for the superhero genre. In this case our interval results in a factor increase in ROI that ranges from to . Likewise, this interval includes 2 to the power of 0, which results in no change.