ISYE 6414 Regression Analysis

Project Report

Movie Box Office Prediction

Submitted on

Dec 6 2021

Submitted by

Alex Yang, 108, [zyang363@gatech.ed](mailto:zyang363@gatech.edu)u

Dheerachote Korpajarasoontorn, 260, dheerachote.k@gatech.edu

Maxime Alos, 201, malos3@gatech.edu

**Abstract**

In view of the considerable amounts of money involved in the production of movies, it is critical to have an estimate of how much money a new movie is likely to make when deciding whether or not to film it. However, predicting the box office of a movie before its release or even its production is often a difficult and imprecise endeavour. Moreover, movie producers need to understand what factors will influence the revenue the most to efficiently make commercial movies. This project aims to give an estimate of what a movie’s box office is going to be. To do this we have implemented 4 different regression models: simple linear regression (OLS), weighted least squares regression (WLS), LASSO regression and Box Cox transformation followed by OLS. With the simple linear regression, we reached the highest model accuracy with an R-squared value of 0.6966 on the test data set.

**Introduction**

Movie production, like any other business, has the objective to make as much profit as possible with each movie. In 2020, the US theatrical and entertainment industry was worth $32 billion [1]. Some movies can have a box office in the billions of dollars such as Avatar which has the world record lifetime gross revenue of $2,847,246,203. Considering how large the sums of money at stake can be, it is critical for movie producers to be able to predict how much money a new movie idea is likely to be worth. Therefore, predicting a movie’s box office is a crucial part of decision-making for movie producers when considering whether or not to go through with a new movie idea.

When confronted with a new movie, the reaction of an audience can vary wildly based on highly variable and unpredictable factors such as temporary trends, critics’ reviews and more. This makes it extremely difficult to predict what factors will contribute to its success, and how much the casts and directors will influence the movie’s performance. Our team is curious about this and expects that this project will allow us to understand the movie industry. The question we will solve is how we can predict movie box office with regression models and which predictors are most important.

We will start by describing what data sources that we have used for this project and how we have assembled them to work as a coherent whole. We will then go on to explaining what our approach to this problem was, going into the details of how and why we have decided to implement the 4 following models: simple linear regression (OLS), weighted least squares regression (WLS), LASSO regression and Box Cox transformation followed by OLS. Finally, we will present our analysis of the results given by each of those models and go on to draw conclusions about this problem from our results.

**Problem statement and Data sources**

The goal of this project is to accurately predict the box office of a movie. To do this we implemented multiple models involving regression to try and predict what the revenue of a movie is going to be from different predictors.

In order to do this, we will use two different data sets. The first data set was extracted from The Movie DataBase (TMDB) and can be found on Kaggle under the name “TMDB Box Office Prediction”[2]. This dataset has 7398 entries for movies and 22 features for each of those movies. The train data set has 3000 data points and the test data set is the remaining 4398 entries. Features include both quantitative information such as the budget, length, and popularity of the movie as well as descriptive information such as the cast list, genre, title, and languages the movie was translated into.

The second dataset is made up of salary information about the 1,000 most famous directors in the world. It includes their name, their ranking, number of movies produced, and average salary on average on all of their produced movies. The two datasets have been joined on the movie director’s name.

**Methodology**

The exploratory data analysis was conducted to understand the correlation and missing data. With an intuition of simplifying the model, we started the feature selection with the domain knowledge of the movie industry. By examining each column, we decided that some columns are similar to the identification of the movies, such as the homepage weblinks, taglines, overviews, id, and etc. We first dropped the columns: "homepage", "imdb\_id", "original\_title", "overview", "poster\_path", "tagline", "title", “genre”, and "Keywords". All of these are descriptional text information that may not be useful in our project. For some other parameters, we decided that a binary variable would be helpful such as “belongs\_to\_collection”. We also converted other JSON strings into binary such as English binary columns from original languages, and into numerical values such as spoken languages.

For the “Cast” and “Crew” columns, we decided that this information would have a significant influence on the movie's box office so we will need feature engineering on these two features. We found additional datasets that summarize the top casts’ and top directors’ total revenue to replace the text information. However, for the cast columns, there are multiple casts in one movie and had made the problem extremely sophisticated, and as we fitted the engineered cast total revenue columns into the stepwise feature selection steps, we did not observe a significant improvement in the accuracy and R squared value. Therefore, we decided not to include cast information. On the other hand, a similar approach was done to the director columns and we did observe an obvious change in the accuracy using aggregated director revenues.

For the original data set, we observed that several columns have more than 50% of the missing entries and most of these columns are the text information. We dropped the rows with the missing values after removing the unwanted columns, and we dropped about 4.2% of the data points from the 3,000 data points train set.

During the exploratory data analysis, we further investigated the correlation and multicollinearity of the parameters as well as their repartition. Figure 1 below displays how correlated each predictor is to the others.



Figure 1. Multicollinearity of the parameters.

We noticed that most predictors have low correlation levels to each other except for the budget, director revenue, and revenue. Revenue designates the box office of the movie and is our response variable. Therefore, we know that both the movie’s budget and the revenue of its director will be strong indicators of what the movie’s box office is expected to be, making them important predictors.

The data cleaning and modeling was initially done on python through Google Colab where the collaboration of coding is enabled. Data cleaning required a lot of merging or tables and manipulation of the data, so we decided that python would be a convenient tool to start with. Nonetheless, the regression model constructed in Python is not explicitly displayed. In other words, we cannot visualize the model parameter directly in Python. So we changed tools during the modeling stage to switch to R. From R, we successfully export a model with numerical representation of the coefficients that we can present in the slideshows and perform further prediction with individual data points that we obtained from running movies in the theater.

We started the modeling with the simple linear regression (ordinary least square method). As we converted all the parameters to numerical values, it is making sense to use the linear regression to describe the data set. Based on the exploratory data analysis conclusions, we decide to choose the following seven parameters for modeling: budget, popularity, runtime, spoken\_languages, collection binary, English binary and Director’s revenue. The reason for choosing these parameters is based on the stepwise feature selections and multicollinearity.

However, it is also observed that the skewness level of the data set is very high. For example, the money value for budget and revenue did not exhibit a normal distribution as we saw in the exploratory data analysis stage. Therefore, a transformation of the data set may be helpful in resolving the heteroscedasticity in the residual distributions and deviation of the QQ-plot from the linear model. Then we performed box-cox transformation to the Y data which is the revenue. This improved the residual distribution on the residual plot and the QQ-plot is closer to the expectation of 45 degrees.

While we were initially using stepwise feature selection, we wanted to improve the efficiency of the selection, and then we used LASSO regression on the data set to perform feature selection.

Due to the observation of the heteroscedasticity of the residuals distribution, it is believed that a Breusch-Pagen test may be needed to determine if further model trials should be implemented. After the BP test, it is believed that the weighted least square method would be proper to further improve the model accuracy. Then, we decided to train a weighted least square regression model for the data sets.

**Analysis**

* **Exploratory Data Analysis**

Through EDA, we observed that the most important parameter that influences the movie revenue is the movie budget, and also the engineered parameter Director Revenue, which depicts a director’s influence on the box office for a movie, is also important. We experimented with different combinations during stepwise feature selection and observed that fitting these two parameters alone contributed to the most of the accuracy while other chosen parameters slightly increased the accuracy. Yet, this assumption is not validated as the overfitting could be a concern for the modeling.

* **Regression Analysis**

First, we train a linear regression model with our preprocessed data feeding in all of the features into the model (Model 1). The resulting model achieves a 69.7% Adjusted R Squared Value and has the p-value of the overall F-test less than 0.05 meaning that at least one variable that we include in this model has statistically significant relationships with revenue. The training set RMSE is 74,346,123 and test set RMSE is 79,894,875 From the separated t-test, we found that budget, popularity, runtime, collection, and director revenue are the variables that constitute the relationship. Detailed model results can be found in Appendix 2.

However, as shown in the residual plots of Appendix 2, the residual of this model has an unconstant variance. Moreover, from the QQ plot, the error is normally distributed. Given the skewness of the data, we also believed that scaling the data would improve the model accuracy. Thus, we try fixing this problem using Box-Cox transformation. The 95% confidence interval of the lambda is 0.24 - 0.27 as shown in Appendix 3, so we transform our response variable using a lambda of 0.25 then train linear regression again.

The residual of this new model behaves better; its distribution is a lot closer to normal as depicted from the QQ plot in Appendix 4. However, the Adjusted R-Squared dropped to 58.7%. And all of the features show significant relationships with revenue at a 5% significance level. The training and test set RMSE are 330,240,119 and 369,890,219, respectively. (Note that the RMSE is scaled back to the original revenue unit so that it is comparable to the RMSE of the other model.)

Next, we consider using Weighted Least Square to solve our heteroscedasticity problem. We performed the Breusch-Pagen test and received a p-value of less than 0.05 confirming that Weighted Least Square might be appropriate here. However, after reiterating WLS 1,000 times, the model doesn’t converge.

As the second part of the feature selection steps, the LASSO regression did not perform well as expected. It didn’t drop any variables. This may be because all features have a significant relationship with our response as shown by the result of the Regression model using transformed revenue previously. Thus, the resulting Lasso model was exactly the same as model 1 and did not improve the accuracy nor decrease the error. Therefore, it is not recommended in this project.

From all the models that we’ve tried, we decided to use the model that is trained using untransformed revenue data to predict the revenue of a new movie and exhibit it in the presentation. And from the model’s coefficients, we can interpret that the variable that has the most impact on the revenue is the budget used in creating the movie.

Eventually, we wanted to test the final selection model’s accuracy. We chose a new movie that is not included in the data set and handpicked the associated parameters that we used in the linear regression models. For example, one of the movies we chose was 007 (2021), with the given features that we collected from TMDB and IMDB, we ran the regression prediction to get a predicted revenue of 691,483,170, while the true revenue is 734,563,795 As of Nov 25. 94% accuracy was achieved in this case. We believed that there is a random effect that has been included in this prediction but since we have also tested our regression model on test data set that includes more than 4,000 data points and reached ~70% of R squared value, it may be concluded that the model we chose is good enough to make a decent prediction for movie revenue.

**Conclusion**

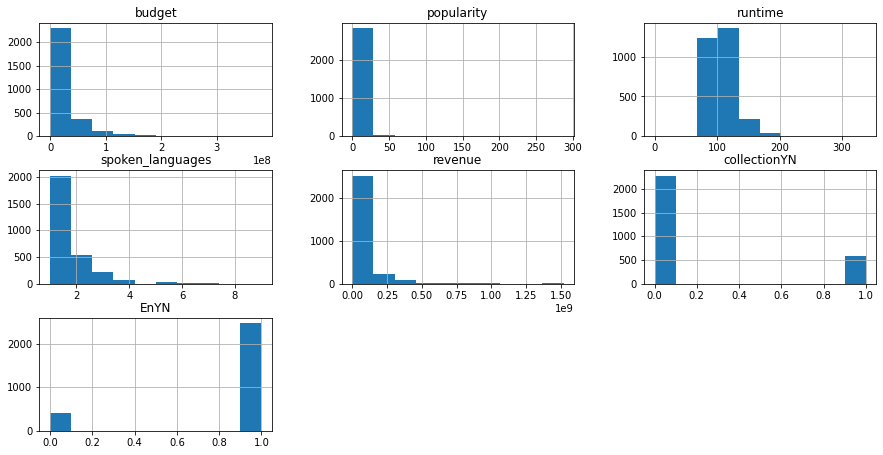
In conclusion, with the linear regression model, we reached the prediction R squared value of 70%. While other models may provide better residual plots and QQ-plot shape, we believe that the linear regression model is the best model to be used in this project due to both accuracy and simplicity. The most important step in the project is the data cleaning and feature engineering as 70% of the time in this project was spent on this step. This is realized by a deep exploratory data analysis, where the data was fully understood. Extensive domain knowledge is needed when making the data integration type of feature engineering decisions. A better feature selection would be helpful in further improving the model’s simplicity and explainability. Future studies should be done to further investigate the implementation of other models in this data set, as some are failing in the projects including weighted least square and Box-Cox transformation.

**Lesson Learned**

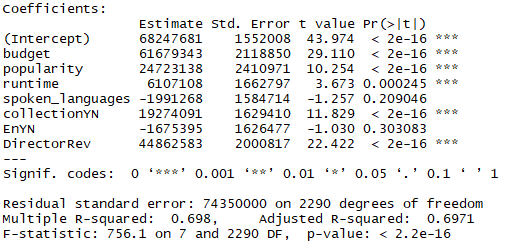
The lesson we learned from this project is to apply different regression models to describe and predict based on the given data set. With all the techniques that we learned in class, it is necessary to run experiments to compare the modeling results. Interpretation of the results is also necessary to reach a valid conclusion about a target model. Many criteria are available to evaluate a model and one should not baldly rely on one to make an arbitrary evaluation, because the random effect and overfitting may influence the actual quality of the model and affect the final judgement. In this class, we learned various tools in regression that we can apply to different statistical problems. The real-world problem that followed each theoretical topic in the lecture was extremely helpful in projecting the concepts into daily scenarios. We especially enjoyed the application problem solving part of the class where we combined gained theoretical knowledge to conduct a free-topic project to evaluate the learning.

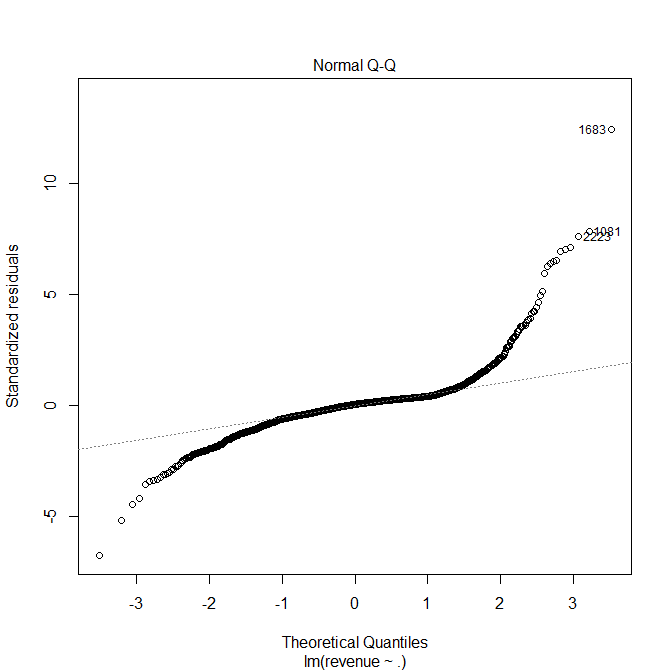
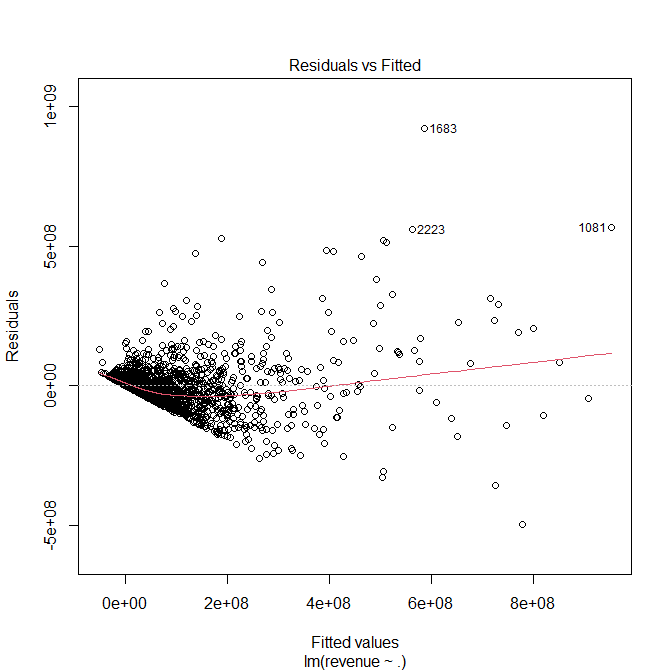
**Appendixes**

Appendix 1: Data repartition

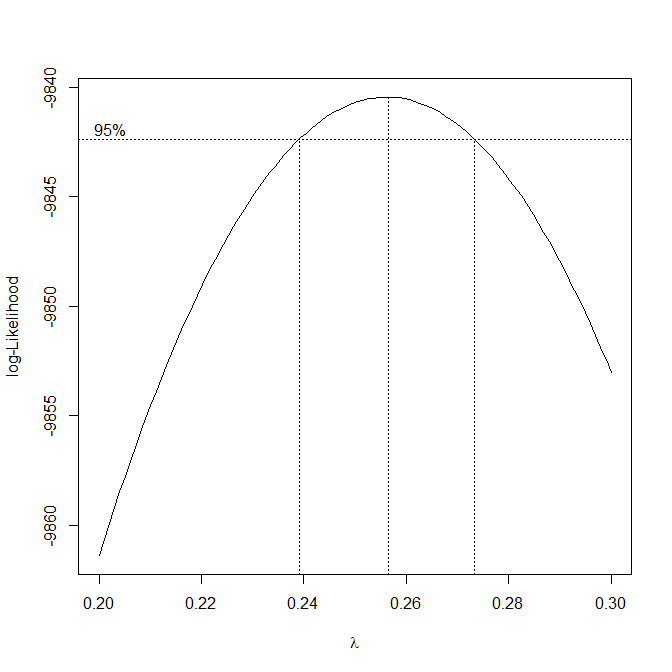
****

Appendix 2: Model 1 result and residual

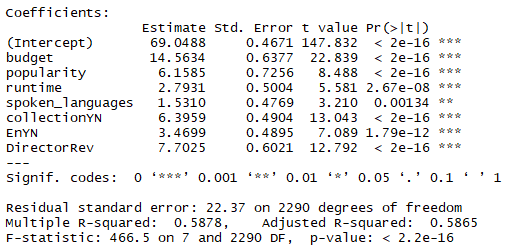


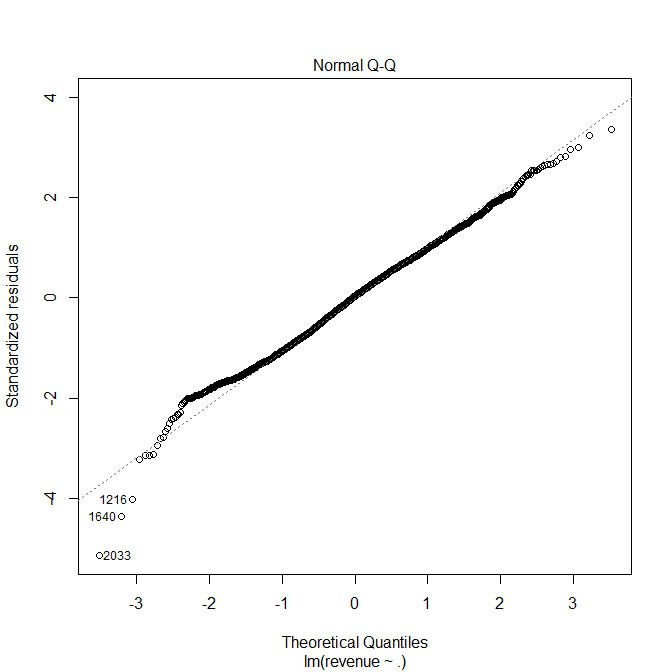
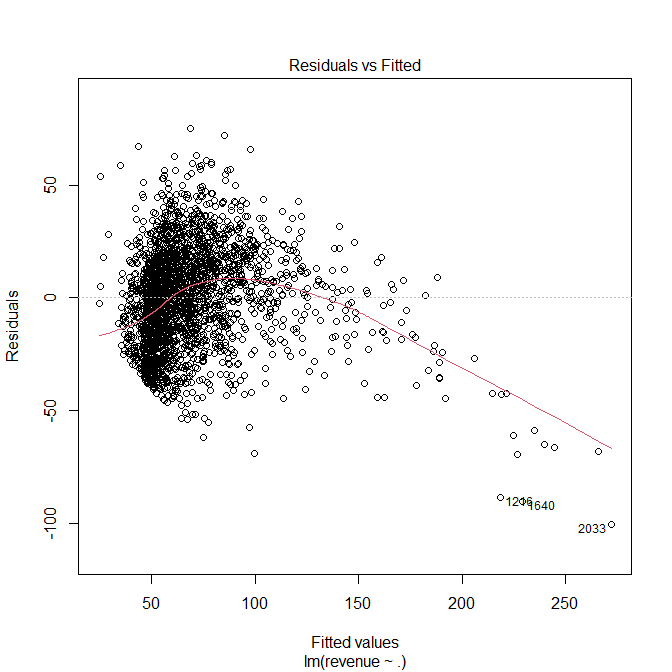


Appendix 3: Box Cox Transformation



Appendix 4: Model 2 result and residual





**Reference**

[1]B. Adgate, “The impact covid-19 had on the entertainment industry in 2020,” Forbes, 13-Apr-2021. [Online]. Available: https://www.forbes.com/sites/bradadgate/2021/04/13/the-impact-covid-19-had-on-the-entertainment-industry-in-2020/?sh=50934978250f. [Accessed: 02-Dec-2021].

[2]“TMDB box office prediction,” Kaggle. [Online]. Available: *https://www.kaggle.com/c/tmdb-box-office-prediction/data?select=test.csv. [Accessed: 02-Dec-2021].*