Predicting Movie Revenue using Linear Regression

Kenneth Gutierrez King

December 13, 2025

Throughout the last century, the movie and television industry has developed into one of the largest and most lucrative industries in the United States, and has since been a staple in American culture and society. Given this large opportunity for success, it is important that movie studio executives understand which factors contribute the most to a movie's overall revenue and success. In order to investigate what these factors may be, this project focuses on predicting movie revenue using linear regression, with the goal of predicting whether increasing a movie's production budget is justified and under what circumstances.

To begin, this project analyzes a dataset containing statistics for 120 movies, including each movie's production date, genres, runtime in minutes, director, the director's birth year, viewer rating on a scale of 1 to 10, production budget, and gross revenue. Before analyzing the data, data cleaning addressed production dates recorded with two-digit years, which were converted to four-digit years to ensure correct centuries. One movie (An Everlasting Piece) with a $4 million budget but only $75,000 gross, was removed from the dataset as there was likely a typo. Additionally, budgets and gross revenues were adjusted for inflation using consumer price index data from the U.S. Bureau of Labor Statistics. When adjusting for inflation a release year column was created to simplify the analysis and avoid using the full date.

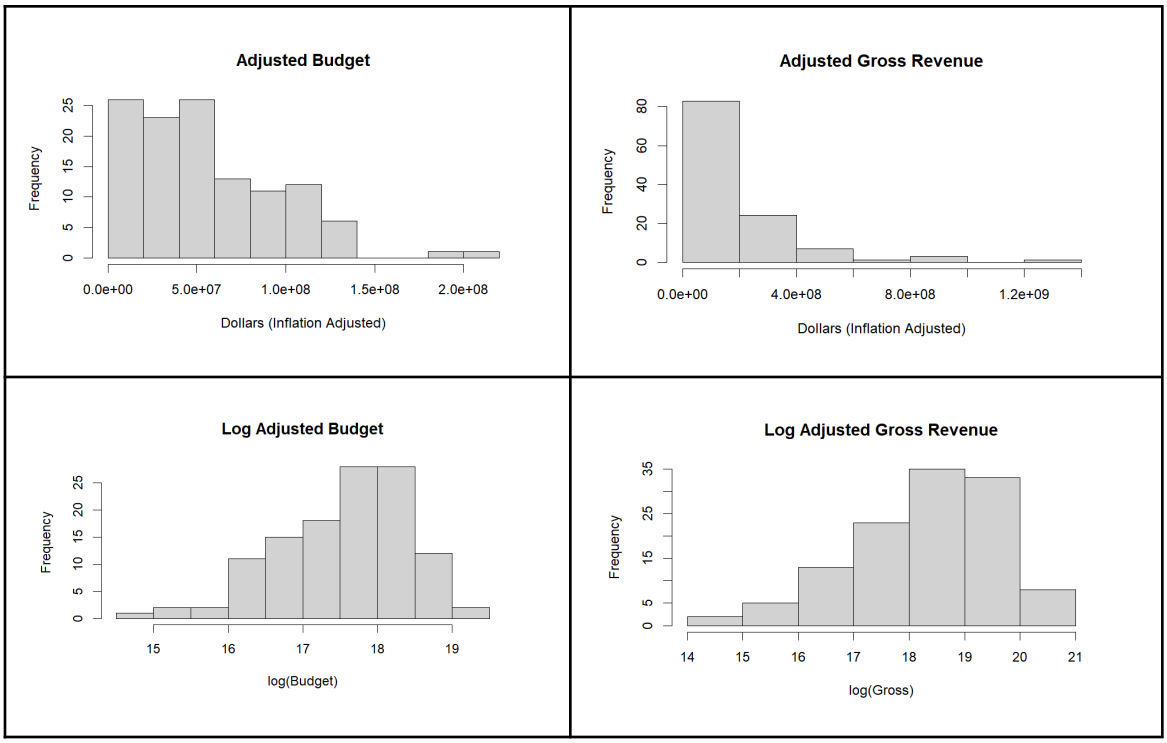
After cleaning the dataset, exploratory data analysis was conducted, starting with the categorical variables genre and director. Each movie in the dataset had at least one genre and could have up to three, which were filtered alphabetically. The dataset contained 20 unique genres, with the most frequent being Drama (71 movies), Comedy (41 movies), and Action (35 movies). To ensure that genre could fit in a model without overparameterization, movies were manually grouped into 6 similar categories. Movies with multiple genres were assigned subjectively with intent of best-fit, which could introduce some bias into the model.

Group 1: Action/ Adventure /Sci-Fi	**Group 2:** Fantasy/Horror/ Thriller/Mystery	**Group 3:** Comedy	**Group 4:** Drama/ Romance	**Group 5:** Crime/ Biography/ History/War/ Western	**Group 6:** Family/ Animation/ Music/Musical/ Sport
31 movies	16 movies	27 movies	12 movies	12 movies	21 movies

Genre groups and number of movies per group

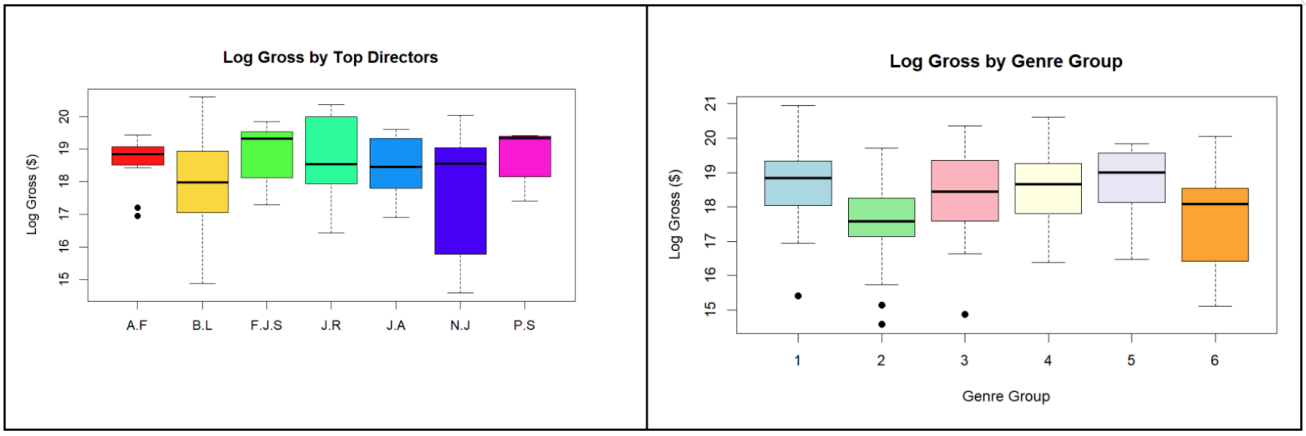
For the director variable, there were 30 unique directors, but only 7 had more than five films in the dataset. To avoid overparameterization and confounding with other variables, a dummy variable was created indicating if they are a top director, as well as dummy variables for each top director.

The continuous data being evaluated in this project was the movie's budget, gross profit, runtime in minutes, average movie rating, release year, and a newly made director's age variable created from the given director birth year information. Some key findings from the basic summary statistics were that production budgets and gross revenues (adjusted for inflation) vary widely, with budgets ranging from about $2 million to $204 million and gross revenues from around $2 million to $1.27 billion. Movie runtimes and viewer ratings are more consistent, averaging 116 minutes and 6.6 out of 10 respectively. Movies in the dataset were produced from 1947 to 2019 (which lead to accounting for inflation), and directors' ages at the time of each film range from 31 to 77, with an average of 48. After plotting histograms for each continuous variable, histograms for runtime, average movie rating, and director's age all appeared approximately normally distributed. The distribution for release year was left-skewed, however because inflation adjustments were made, no transformations were needed. The adjusted gross revenue and adjusted budget variables were right-skewed, so a logarithmic transformation was applied to these variables to reduce skewness, stabilize variance, and satisfy the needed assumptions to perform linear regression.



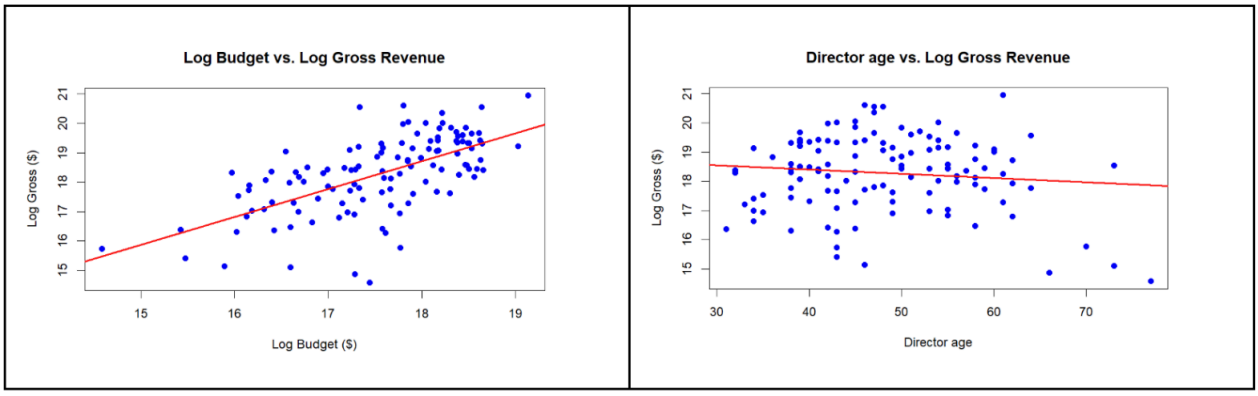
Budget and gross revenue before and after log transformations

Bivariate relationships between the response variable, log gross profit, and the categorical predictors were examined using box plots and ANOVA tests. After comparing top directors against non-top directors, the box plots produced were nearly identical, and the ANOVA test ($p=0.416$) showed no statistically significant difference in log gross revenue between the two groups. Furthermore, after comparing the 7 directors deemed as top, the box plots had some slight variation, however the ANOVA test ($p=0.526$) also showed no statistically significant difference in log gross revenue between the top directors. Due to these results, it can be inferred that the director will have no significance in the model. Looking at genre groups, the box plots showed some differences in log gross profit, with Groups 2 and 6 standing out. The ANOVA test confirmed that these differences are statistically significant, indicating that the mean log gross revenue does vary across genre groups ($p = 0.0123$). This suggests that genre should be considered as a predictor in the model.



Box plots for log gross revenue by top directors (left), and by genre group (right)

Bivariate relationships between the response variable, log gross profit, and the continuous predictors adjusted budget, movie average rating, director age, runtime, and release year were examined using scatterplots and correlation tests. For each predictor, a scatterplot with a simple regression line was created against log gross profit. Additionally, correlation values were calculated for each predictor, and a scatterplot matrix was produced to visualize trends among all continuous variables. The visual and correlation results indicated that log budget and log gross revenue had a strong positive relationship (0.631). Average movie rating showed a moderate positive correlation with log gross profit (0.369). Director age had a very weak correlation (-0.108), runtime showed a modest positive correlation (0.324), and release year had a weak correlation (-0.151) with log gross profit. From this, it can be inferred that director's age and release year will not be helpful for building a model. The scatterplot matrix shows that there is a positive relationship between runtime and movie average rating, but this correlation is not strong enough to influence model selection.



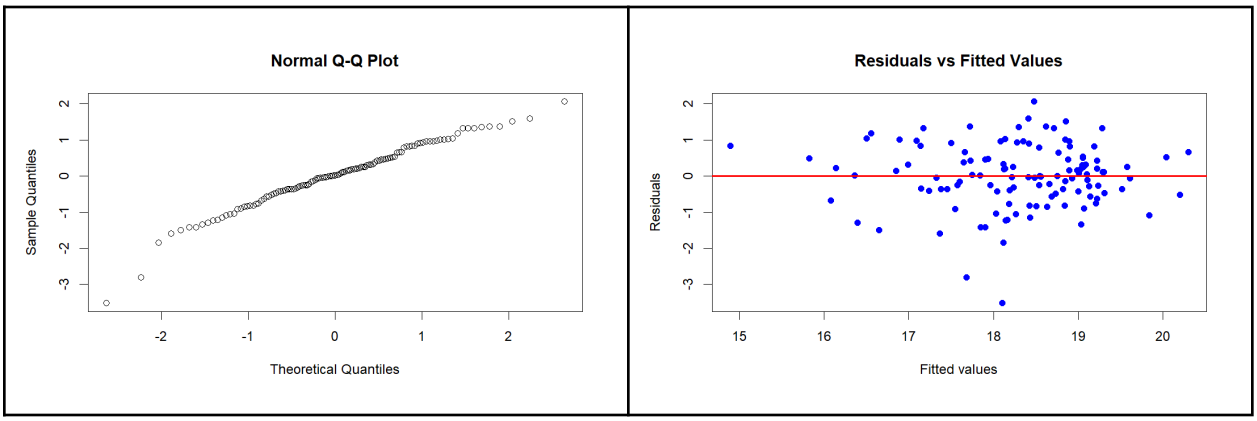
Scatterplots of log gross profit vs. log budget (strong positive) and director age (weak).

To find the best possible linear regression model, subset selection using the R function regsubsets() was performed including log budget, average movie rating, genre group, runtime, and release year as potential predictors. Log budget and average rating consistently appeared in the best models, while genre groups occasionally improved fit, particularly for Group 6 (family themed genre). Runtime and release year were therefore no longer considered. Two models were then compared, with the first model including log budget and average rating, and the second model adding an indicator for genre Group 6. The second model has a lower AIC (-21.05 vs -15.04) and all predictors were statistically significant. After further testing, no other predictors or interactions were found to be statistically significant, making our optimal model:

$$log\ gross_{adj} = \beta_0 + \beta_1(log\ budget_{adj}) + \beta_2(average\ movie\ rating) + \beta_3(genre\ group\ 6) + \varepsilon$$

The final model predicts log-adjusted gross revenue using log-adjusted budget, average movie rating, and whether a movie is in Genre Group 6. (β_0 = -0.68) is the intercept, and is the baseline log gross revenue when all predictors are 0. (β_1= 0.91) represents the expected increase in log gross revenue for a one-unit increase in log-adjusted budget, holding other factors constant. (β_2= 0.47) represents the expected increase in log gross revenue for a one-unit increase in average movie rating, holding other factors constant. (β_3= -0.61) represents the expected decrease in log gross revenue for movies in Genre Group 6 compared to other genres, controlling for budget and rating. The final model explains about 52% of the variation in log-adjusted gross revenue, has a residual standard error of 0.90 on 115 degrees of freedom, and is highly statistically significant overall ($p < 0.001$). Diagnostics of the final model were also evaluated. The Q-Q plot of the residuals indicates that they are approximately normally distributed,

Additionally, the residuals versus fitted values plot shows no clear patterns, suggesting that the assumptions of linearity and constant variance are satisfied.



Q-Q plot (left) and Residuals vs Fitted Values (right) confirming the model assumptions of normality and constant variance.

Based on the final linear regression model, log-adjusted budget has a strong positive effect on log-adjusted gross revenue, as it had the highest beta coefficient, and was extremely significant ($p<0.001$). While this model strongly suggests that increasing a movie's budget is generally associated with higher expected revenue, this estimate may be biased because our model only contains 3 predictors, and is only considering a few genres that were objectively grouped. To get a better prediction if increased budget is justified, the model should consider data regarding franchise status, cast popularity, other genres, and other factors that can influence a movie's success beyond budget and rating that could better predict the effect of budget on revenue.

Using the final model, it can be predicted that a new comedy movie produced on June 5, 2013 with a $10 million budget (adjusted for inflation), an average rating of 7 out of 10, a runtime of 2 hours, and directed by Antoine Fuqua is predicted to generate approximately $34.6 million in gross revenue. The 95% prediction interval ranges from about $5.6 million to $212 million. This model only takes into consideration the budget, average rating, and genre if it is a family or music movie, however a reasonable prediction of gross revenue is still made.

Reference

U.S. Bureau of Labor Statistics. (2025). Consumer Price Index for All Urban Consumers. Bls.gov; U.S. Bureau of Labor Statistics. https://data.bls.gov/pdq/SurveyOutputServlet