Predicting Movie Ratings With Regression Model Based on Budget, Gross Revenue, and Runtime

Lior Levy Meruk, Mary McSweeney, Henry Johnson, Ainsley Strang, and Evan Sang

Department of Statistics and Data Science, University of California, Los Angeles

Stats 101A: Introduction to Data Analysis and Regression, Lecture 1

Professor Maria Cha

March 17, 2024

Introduction

Our research goal when approaching this project was to help current film students at UCLA and in the greater Los Angeles area better understand what factors influence movie success. For our purposes, we defined movie success as a high audience score. For this research, we reviewed our original data, transformed our variables, and then tested and validated different models to find the best predictors and finally, interpreted our results. In order to create our model, we used data from IMDB that reported factors such as budget, director, movie name, gross, rating, runtime, and what we wanted to predict, user rating (score). We only ended up using the numerical data, giving us four variables budget, runtime, gross, and score. Originally, there were 6820 variables in the data set, but to minimize the conflicting variable of time, we decided to focus our research on the year 2001, giving us 200 movies to work with. Before working with the data set, it was necessary to clean our observations. All rows that were missing any data were deleted (n = 32) with row-wise deletion, leaving 168 observations. Additionally, our monetary numbers of budget and gross were divided by a billion to make our model on the same relative scale and for easier interpretability. Another thing to note was that our score variables were a rating out of 10.

Data Description

Summary statistics for each of our variables (Score, Budget, Gross, and Runtime) are included in Figure 1. After these modifications, the general relationships between variables in our data were relatively linear, but with room for improvement (Figure 2)

Mean:

score	budget	gross	runtime
6.34397590	0.03883313	0.09289490	107.25301205

Standard Deviation:

score	budget	gross	runtime
0.96655896	0.03143674	0.14427589	20.33587269

Summary:

core	budget	gross	runtime	
Min.	2.300000	0.00010000	0.000080631	81.000
1st Qu.	5.800000	0.01500000	0.016158239	93.000
Median	6.400000	0.03000000	0.038577035	101.500
Mean	6.343976	0.03883313	0.092894899	107.253
3rd Qu.	6.975000	0.05600000	0.099609142	119.000
Max.	8.800000	0.14000000	1.006968171	210.000

Figure 1: Summary Statistics

.

Initial Scatterplots:

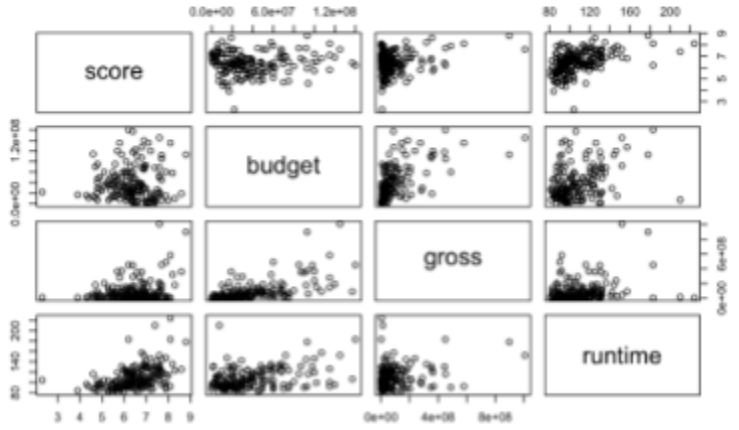

Figure 2: Scatter plots between non transformed variables

Methods

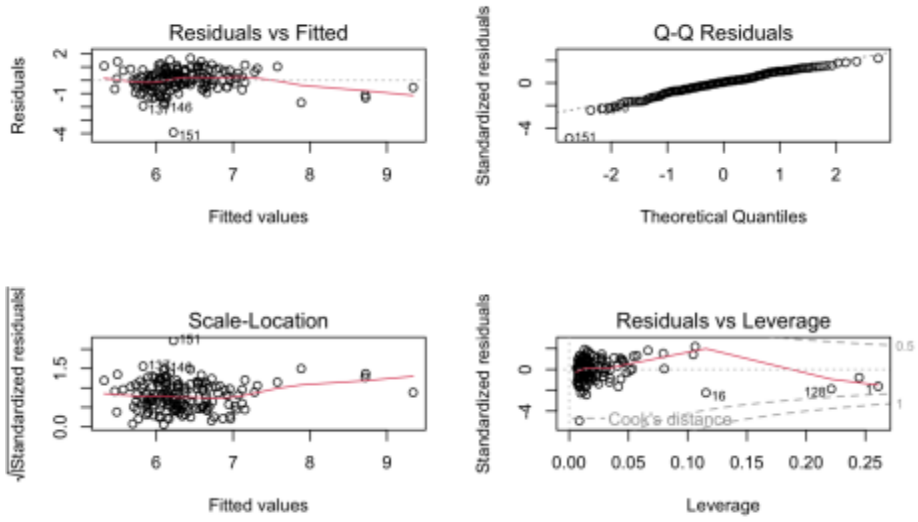

Figure 4: Untransformed Diagnostic Plots

The initial model appeared to be significant from the summary statistics (Figure 3). Every predictor had a small p-value and the overall F-test for the model was found to be significant as well. Our adjusted R^2 was 0.3272, implying that 32.7% of the variance in score was accounted for by budget, gross, and runtime.

Upon inspection of the four diagnostic plots for this model (Figure 4), it is obvious that some of the model assumptions are violated. The residual plots do not show constant variance around zero and there are some potential bad leverage points. This implies that the assumption of constant variance in the residuals

```
Call:
lm(formula = score ~ budget + gross + runtime)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9243 -0.4411  0.0526  0.5306  1.6546

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.134388   0.344458  12.003  < 2e-16 ***
budget      -9.920431   2.528044  -3.924 0.000128 ***
gross        2.451998   0.552143   4.441 1.65e-05 ***
runtime      0.022070   0.003335   6.617 5.11e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8001 on 162 degrees of freedom
Multiple R-squared:  0.3272,     Adjusted R-squared:  0.3147
F-statistic: 26.26 on 3 and 162 DF,  p-value: 6.771e-14
```

Figure 3: Summary of Untransformed Linear Model

is violated, meaning inference based on our model will be invalid. In addition, the Q-Q plot shows our data is mostly normal, but could be improved upon as well. To account for the violation of the linear model assumptions, we considered transforming our variables next.

With our initial scatterplot between our non-transformed variables, we first decided to fit a full, non-transformed linear model with budget, gross, and runtime variables predicting score. We then analyzed model fit and assessed any violations of the linear model assumptions using residual plots, standardized residual plots, Q-Q plots, and leverages. Based on our results, we then considered transforming our variables to account for any violations in our assumptions, as well as fitting a model with those transformed variables. We then evaluated any evidence of multicollinearity and conducted variable selection using all possible subsets, forwards, and backwards selection procedures to find a final model.

To try and lessen the violations of the assumptions and improve validity, we applied a Box-Cox transformation to our model. The results from the Box-Cox transformation suggested a power of 2 for score, 0.33 for budget, 0.18 for gross, and -2 for runtime. The output also suggested rejecting the log-log transformation as well as no transformation(Figure 5).

Similarly, this model showed all predictors were significant and the overall F-test was significant.

The diagnostic plots of the transformed model show a significant improvement, especially in the residual plots, we also notice a decreased number of outliers and leverage points (Figure 7).

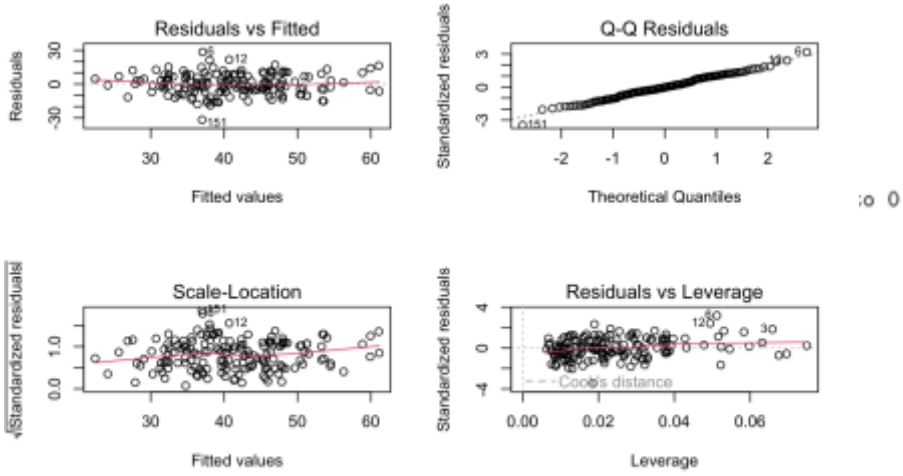

Figure 7: Transformed Model Residual Plots

```
##
## Call:
## lm(formula = t_score ~ t_budget + t_gross + t_runtime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.713  -5.575  -0.319   6.920  28.538
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.999e+01  4.552e+00  13.180  < 2e-16 ***
## t_budget    -6.105e+01  1.011e+01  -6.037 1.03e-08 ***
## t_gross      3.669e+01  6.746e+00   5.438 1.95e-07 ***
## t_runtime   -2.239e+05  2.660e+04  -8.419 1.92e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.228 on 162 degrees of freedom
## Multiple R-squared:  0.4152, Adjusted R-squared:  0.4043
## F-statistic: 38.33 on 3 and 162 DF,  p-value: < 2.2e-16
```

Figure 6: Transformed Model Statisitics

We continued with variable selection to ensure no variables were highly correlated with one another, this would give us misleading results which would make our model weaker. The results from the VIF test in R suggested no signs of multicollinearity (Figure 8).

```
##  t_budget   t_gross t_runtime
##  1.938831  1.940866  1.112464
```

Figure 8: VIF

After exploring all possible subsets, we found the most optimal models to be the following at each number of predictors:

One predictor: $Score^2 \sim Runtime^{-2}$
Two predictors: $Score^2 \sim Budget^{0.33} + Runtime^{-2}$
Three predictors: $Score^2 \sim Budget^{0.33} + Gross^{0.17} + Runtime^{-2}$

After using the method of determining all possible subsets, we concluded that based on all measures, the full transformed model would be the most accurate. (Figure 9 and 10).

```
## Selection Algorithm: exhaustive
##          t_budget t_gross t_runtime
## 1  ( 1 ) " "      " "     "*"
## 2  ( 1 ) "*"      " "     "*"
## 3  ( 1 ) "*"      "*"     "*"
```

Figure 9: Model Selection Prediction

```
##   Size     Radj2      AIC     AICc      BIC
## 1    1 0.2656625 774.5293 774.6756 780.7533
## 2    2 0.2999124 767.5853 767.8307 776.9213
## 3    3 0.4043372 741.7497 742.1200 754.1976
```

R^2, AIC, AICc, and BIC all suggest the model with 3 predictors.

Figure 10: Model Selection Verification

Final Model:

$$Score^{2} = 59.99 - 61.05 \cdot Budget^{0.33} + 36.69 \cdot Gross^{0.17} - (2.239E^{5}) \cdot Runtime^{-2}$$

The score has a positive relationship with gross, and runtime, but a negative relationship with budget.

Discussion

The goal of our project was to make a model that would help those interested in movie success such as movie sites and UCLA film students predict a movie's success. Using data from IMDB about movies from 2001, a time for movies that many people (reported to be very nostalgic and happy). While it is difficult to interpret the results because of the way our model was transformed, it makes sense that movies that made more money would have a higher score as people would be more likely to spend money on it. Having a positive relationship with runtime could also make sense, as longer movies are more typically adult movies, and adults are more likely to be reporting on these websites. However, the negative relationship with budget seems counterintuitive, as one would think that movies with more money put into them would be of higher quality and therefore people would like better. Something we could do to improve this model and account for this confusing relationship and our low R^2 value would be to add more prediction variables. For example, in one study, *Movie Success Prediction Using Data Mining,* researchers use more categorical variables such as actors and movie ratings (G, PG, R, etc.) in order to create their model. Additionally, research papers such as *Movie Success Prediction using Machine Learning Algorithms and their Comparison* use features such as number of tickets sold and number of audience in their prediction model. Alongside the improvement in using additional variables, it would also be helpful to increase the number of observations we used. We could do this by picking a specific decade rather than just a single year.

Works Cited

J. Ahmad, P. Duraisamy, A. Yousef and B. Buckles, "Movie success prediction using data mining," 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, India, 2017, pp. 1-4, doi: 10.1109/ICCCNT.2017.8204173. keywords: {Motion pictures;Data mining;Correlation;Mathematical model;Investment;Data models;Predictive models;movie success;data mining;movies;attributes},

R. Dhir and A. Raj, "Movie Success Prediction using Machine Learning Algorithms and their Comparison," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, 2018, pp. 385-390, doi: 10.1109/ICSCCC.2018.8703320. keywords: {Motion pictures;Correlation;Machine learning algorithms;Prediction algorithms;Machine learning;Computational modeling;Social networking (online);Box office gross;Data Mining;Machine learning;Movie success;Movie;Predictive analytics;Critical review;Rating},