What Attributes of Movies are Associated with Domestic Gross earning?

Takumi Horiba, Arul Howard, Saumya Jain, and Dinghang Xie

Introduction

Research Question and Motivation

There are several attributes of movies such as ratings on critic websites that may be associated with how much money a movie earns. We would like to investigate what factors are significantly associated with the domestic gross of a movie, so that we can explain the relationship in data. Domestic gross of a movie is a critical measure for production companies and film industries, and it is of our interest to explore what factors are associated with this value because production companies would like to produce movies that can bring profits to them. After this analysis, we may be able to infer the relationships in our data. For example, high ratings or budget may or may not be associated with high domestic gross. In our following statistical analysis, we investigate relationships between log domestic gross of a movie and other attributes of a movie.

Description of the Dataset

Of the available 35 variables, we removed some variables based on if they had too many missing values, were composite and could be calculated using other variables. The following shows response and explanatory variables in our dataset. Note that in our model building, we use log domestic gross as a response variable for better fitting, but in our dataset it is originally stored as described below.

Response Variable

- Domestic gross (Quantitative | In Million $): Total Gross earned In North America (USA and Canada) in millions of dollars (USD).

Explanatory Variable

- Year (Categorical | Year Type): The release year of the movie.
- Rotten Tomatoes critics (Quantitative | In 100 Score): Movie ratings by critics on Rotten Tomatoes.
- Metacritic critics (Quantitative | In 100 Score): Movie ratings by critics on Metacritic.
- Rotten Tomatoes Audience (Quantitative | In 100 Score): Movie ratings by audience on Rotten Tomatoes.
- Metacritic Audience (Quantitative | In 100 Score): Movie ratings by audience on Metacritic.
- Primary Genre (Categorical | Genre Type): The broad genre of the movie.
- Budget (Quantitative | In Million $): Total budget in millions of dollars (USD)
- Oscar Winners (Categorical | Oscar Win or Not): indicates if a movie won an Oscar by the value “Oscar winner”, otherwise the movie did not win an Oscar.

- Distributor (Categorical | Distributor Name): The film production and distribution company associated with the film
- IMDb Rating (Quantitative | In 10 Score): The IMDb rating of the movie.

We do not use following variables: Film, Foreign Gross, Worldwide Gross, Percentage of Gross earned abroad, Release Date. Notably, considering the realistic connections and statistical interpretation, foreign and worldwide gross, percentage of gross earned abroad are highly correlated with response variables. It is not our interest to explain the movie's performance on the domestic market by domestic gross given worldwide or foreign gross. We exclude these variables out of the analysis to better answer our research question.

Data Cleaning

Accessing from the original 5 years of Hollywood data, we process the quantitative data in our dataframe in following criteria: (1) Data has missing a value or “N/A” value stored, which indicates data is not accepted for numerical operation. (2) Data contains “0” or extreme value stored, which is apparently recognized as outliers. (3) Data has meaningless symbols, such as “-” and “.”, which is considered as invalid input of data. Processing the original dataframe, 121 valid inputs of data which are mainly distributed over the year of 2020 and 2022 of Hollywood data remain.

Based on the dataframe after filtering, we operate on uniforming the units in each variable. For categorical variables, we convert it to factors for further operation, using as.factor().

Analysis

Data Visualization

We split the explanatory variable into quantitative and categorical parts for visualization. For quantitative variables, we choose the scatter plot for observing the distribution pattern and heatmap for better interpretation of correlations. For categorical variables, we employ the boxplot for comparison of types of individuals and reveal its categorical features.

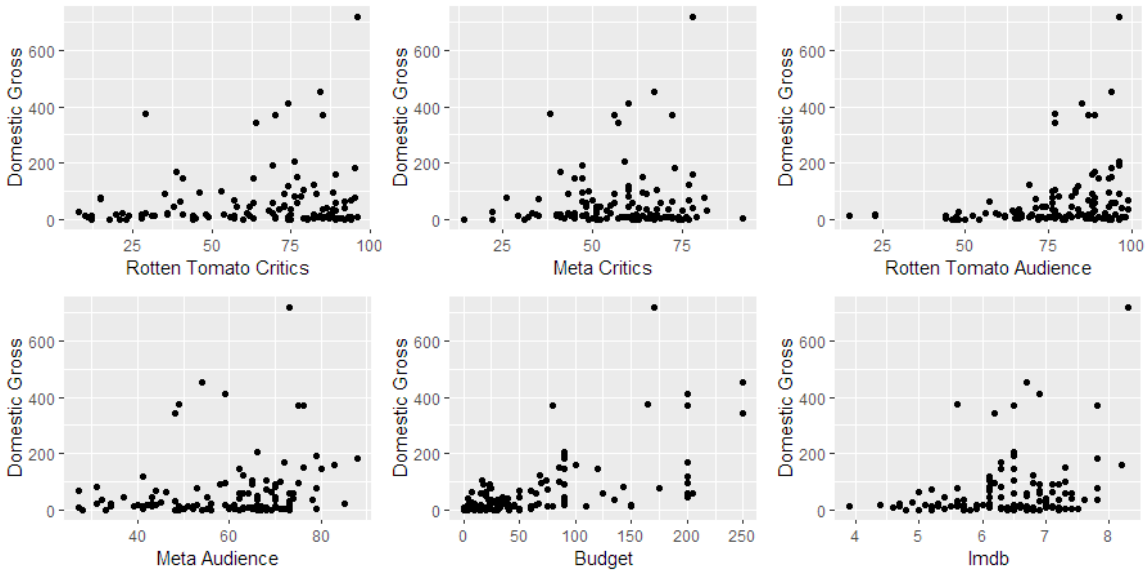

Figure 1 - Tables of Scatter Plot Between Domestic Gross and Quantitative Variables

If we focus on the relationship between each quantitative variable with domestic gross, no obvious linearity appears, although we can observe the trend of positive correlation. Generally, variance in domestic gross increases as each variable increases, hence we would have to employ some transformation. Further transformation or higher order terms may apply for improving the relation and better fitting the regression model.

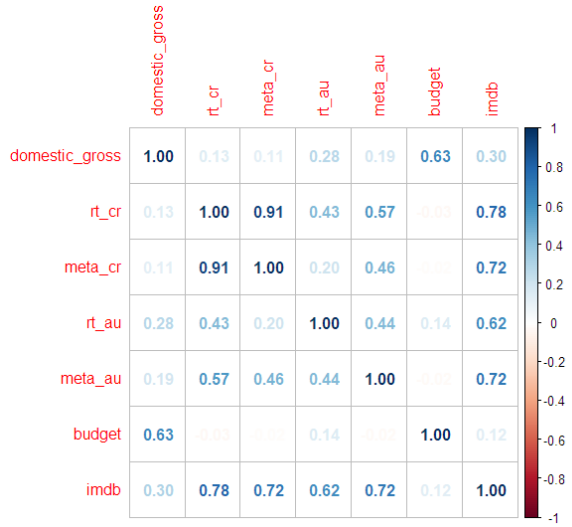

Figure 2 - Heatmap of Correlation Between Quantitative Variables

Progressing from the table of the plots, we demonstrate the correlation of each quantitative variable. Generally, explanatory variables are not strongly correlated with the domestic gross. However, we can observe that there exist some highly correlated variable pairs, which provides the indication for the problem of collinearity for the following model comparison and selection.

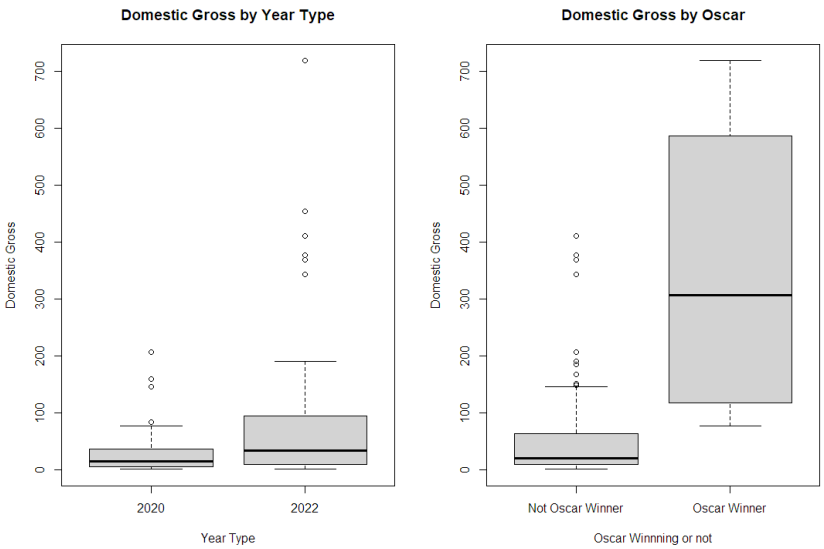

Figure 3 - Tables of Boxplot Between Domestic Gross and Year Type & Oscar

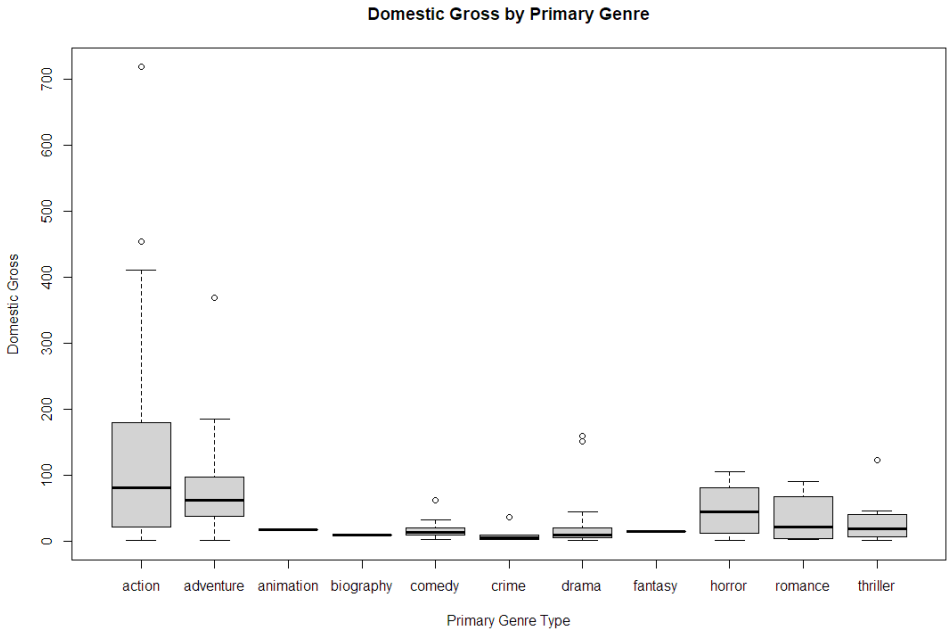

Figure 4 - Boxplot Between Domestic Gross and Primary Genre

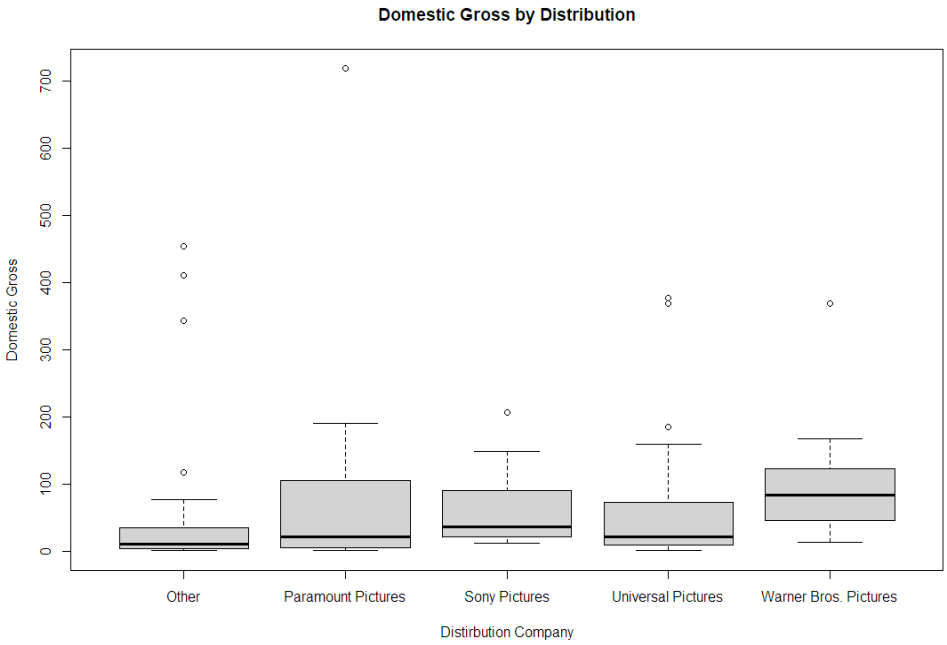

Figure 5 - Boxplot Between Domestic Gross and Distribution

Having explored the features of variables above, we turn our attention to the four categorical variables. Each boxplot precisely visualizes the data features for the corresponding variable. For instance, the year 2022 has a wider upper whisker than the other year, which conveys the market-heat of the movie industry in different years. Some particular kinds of genre and distribution companies are popular among the others. We choose the five distributors with highest mean domestic gross as individual categories and regard other distributors as a single category of "Other" to reduce the number of dummy variables in our model. Originally it contained more than 30 distributors including those that do not distribute movies very frequently. We have checked that the selected five distributors distribute movies multiple times a year. Oscar winners have commonly higher distributed domestic gross, and its median value is apparently greater than the non-oscar movies. These features possibly implies their close relation with domestic gross and further research will be addressed in the model selection.

Model Selection and Diagnostics

We used the regsubsets function from leaps library to select the best model for each number of parameters to conduct exhaustive search for the best model for each number of parameters. Based on the Cp plot, the model with 7 or 8 parameters is a strong candidate for our explanatory model because their Cp values are close to the number of parameters. Since it is not sensible to drop some of the dummy variables in one categorical variable, we keep a categorical variable if any of the corresponding dummy variables are included by the result of regsubsets.

The model with 8 parameters selected, which we call Model 1, contains the following variables: year, primary genre, metacritics critics score, budget, imdb, and distributor, with no interactions nor higher order terms. Note that our response variable is log transformed domestic gross for reasons which we discuss later.

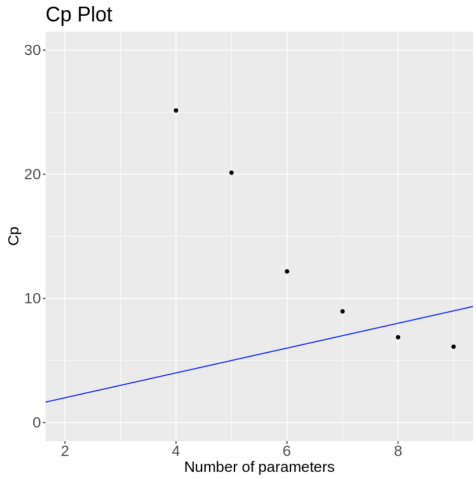

Figure 6 - Cp Plot

In Model 1, year is not a significant covariate because it has p-value of bigger than 0.05, which suggests excluding it from the model may give us a better model for explaining the statistically significant association. In the Model 2 in which year is excluded, we observe that the meta_cr variable does not have a significant coefficient any more, which is a plausible result because meta_cr score and imdb are both ratings with relatively high correlation in between, as we explored in preliminary data analysis. In this case, by excluding either of these correlated explanatory variables, we could reduce the chance of having collinearity problems in our model. Additionally, we observe that many of the dummy variables for primary genre are not significant, except for one identified by the regsubsets() function, namely the dummy variable for the horror genre. Since we keep 10 dummy variables for only one significant dummy variable, it is worth checking the model without this categorical variable.

With these considerations in mind, we tried a few more models and we obtained Model 6 which contains three variables: budget, imdb, and distributor. Budget and imdb have significant coefficients and many of dummy variables corresponding to the distributor variable also do. Other models we have tried had similar performance but some contained non-significant coefficients. Compared to original models Model 1 or 2, Model 6 has smaller adjusted R^2, but it does have much lower BIC and this model does not have coefficients that are not statistically significant except for some of the dummy variables. Since our research question is to find explanatory variables that have associations that are statistically significant, we believe that Model 6 serves our purpose better. Additional benefit of Model 6 is the smaller size of model compared to Model 1 or 2 that have 10 more parameters for dummy variables in the primary genre. This is most strongly reflected in the score of BIC which penalizes larger models more severely than AIC. The resulting residual plots for those three models are very similar and do not show strong patterns except that there are few points in the bottom of the plots that are far from the majority of points. This suggests that while the fitting of these models could be mostly appropriate from a point of view of residuals, there are possibly some outliers in the data. The QQ-plots for all three models show some patterns, which suggests that underlying distribution is right skewed, and thus not normal. With a relatively large number of observations of 121, the normal assumption in our dataset for our linear model is likely to be violated. This is why we applied log transformation on the response variable to reduce the effect of this issue and reduce the variability in variance of the response variable, however, we could not ensure that normality condition is not violated with the chosen dataset with log transformation.
We have tried several transformations including square root, however, we observed that log transformation performed better than others while we acknowledge that it is not the perfect transformation given the QQ-plot that suggests right skewed underlying distribution. Right-skewed distribution is not

suitable for normality assumption of the model, but it is plausible for domestic gross to have such a distribution because most movies do not perform exceptionally well in a sense that few blockbusters do. With original response variable or square root transformation of response variable, we were able to obtain significant coefficients for some of the explanatory variables in our models we tried, however these models have residual plots that show some patterns, hence we concluded that such transformations are not preferable choices. Hence, we determined that log transformation is better than other transformations we have tried, and thus tried model building using the log transformed response variable.

	Comments	Adjusted R^2	AIC	BIC
Model 1		0.559	363.050	421.762
Model 2	Model 1 without year	0.548	365.220	421.136
Model 6	Includes budget, imdb, and distributor.	0.511	365.213	390.375

Table 1 - Models and their Performances by Statistical Metrics

In the residual plot for Model 6, we suspect some data points in the bottom of the plots may be considered as outliers. We used leverage and cook's distance to investigate this further. With the definition in our course, there are 11 data points with high leverage values, however, there are no data points with cook's distance greater than 1. Removing a few points with lowest residuals was shown to improve the residual plot of the new model. However, merely low performance of movies does not give a strong justification to remove them as outliers. They are farther from the general trend in data than observations but they are the actual movies that did not perform well. From model comparison, we found that the existence of these points does not change the model vastly. Hence, we think of them as valid observations and thus we keep them in the data. Therefore, while some points in our residual plot may appear to be outliers, we would not remove any points in this analysis.

Figure 7 - Residual and QQ plot for Model 6.

Model Interpretation

We choose Model 6 for its statistically significant coefficients and much smaller size of model of data with small loss in adjusted R^2. The model can be written as:

$$log(Y) = -0.955 + 0.0144x_1 + 0.650x_2 + 1.39x_3 + 0.900x_4 + 0.222x_5 + 0.531x_6 + 0.432x_7 + \epsilon,$$

where Y is the domestic gross earnings. x_1 is the budget and x_7 is the IMDB rating of the movie. ϵ denotes the error term of this linear regression model which is assumed to be normally distributed. The variables x_2 to x_6 are dummy variables corresponding to the distributor of the movie. The baseline is "Other" distributors that significantly perform worse than top 5 distributors in the dataset. x_2 is the dummy variable which takes 1 if a movie is distributed by Paramount Pictures and 0 otherwise. Similarly, x_3 corresponds to movie distributed by Sony Pictures, x_4 corresponds to movie distributed by Universal Pictures, x_5 corresponds movie distributed by Walt Disney Studios Motion Pictures, x_6 corresponds to movie distributed by Warner Bros. Pictures.

All the coefficients for the chosen model are positive indicating that they have a positive influence on the domestic gross earnings. For coefficient of budget, if all dummy variables are zero and other factors are held constant, for unit increase in budget, the expected value of natural log of domestic gross increases by 0.0144. Similar interpretation applies to IMDB scores. For distributors of movies, if other factors are held constant, and if the movie is distributed by Paramount pictures, then the expected value of log domestic gross increases by 0.650 compared to when the movie is distributed by companies that are classified in "Other". Again, similar interpretations can be drawn for other dummy variables. We observe that higher budget and/or IMDB scores are associated with higher gross, which is plausible because it makes sense for higher rated movies to have higher earnings. With a higher budget, it is reasonable to believe that movies can attract people by having popular actors and spectacular visuals, and thus the positive coefficients in both IMDB and budget are reasonable. The dummy variables for distributors are all positive which seems feasible since we selected some distributors with high mean domestic gross. It is reasonable to have significant coefficients for distribution companies because they have influence on the production and marketing of movies, which are both related to domestic gross. The most significant explanatory variable to the model is the budget (p-value: 8.42e-10) followed by the dummy variable for Sony Pictures (p-value: 4.25e-5). The significant coefficients provide strong evidence against our null hypothesis that there is no linear dependence between these two explanatory variables and the gross earnings.

Conclusion

From our analysis, we have found that there are statistically significant relationships between domestic gross of a movie and some factors, namely, IMDB score, budget, and distributor of the movie by building a model that has statistically significant coefficients for these variables. We considered 10 explanatory variables and their interactions and higher order terms of them if applicable but we have arrived at the model that is simple and understandable with three explanatory variables involved. We have found that higher ratings and budgets are associated with the success of movies as a business measured by domestic gross, which is what we expected and understandable by common sense. This analysis could lead to more sophisticated statistical analysis using more careful examination of underlying distribution and better treatments of categorical features and missing values, which requires a higher level of statistical knowledge and maturity.