Applied Statistics

Programming Assignment

Kontogeorgis Nikolaos,
Papageridis Vasileios,
Rentzelas Athanasios-Panagiotis

June 22, 2023

Introduction

This report analyzes data from *The Movie Database* to predict box office revenue. The main goal is to identify key factors impacting movie income and develop a predictive model for informed decision-making in the film industry. The dataset includes information on 3.000 movies, covering variables like budget, language, genre, cast, running time, and popularity. Two exercises are conducted: an exploratory analysis to determine the most influential variable and a multiple regression analysis incorporating additional features like well-known actors, directors, and cast gender composition. The report evaluates regression assumptions and emphasizes the significance of understanding factors driving box office success. Stakeholders can leverage the insights to optimize budget allocation, marketing strategies, and talent selection.

Excercise 1: Exploratory Analysis

In this exercise, we focus on exploring the relationship between individual explanatory variables and revenue. The variables considered are budget, a binary variable denoting whether the movie is in English or not, running time, and popularity. For each numerical explanatory variable, we compute the correlation coefficient and visualize the relationship with revenue using scatter plots. Based on these analyses, we determine which variable exhibits the strongest association with revenue and would be the most reliable predictor if only one variable could be used.

Results: It seems that budget is the explanatory variable most strongly associated with the revenue, since it has the highest correlation coefficient and we can see a trend in the scatter plot.

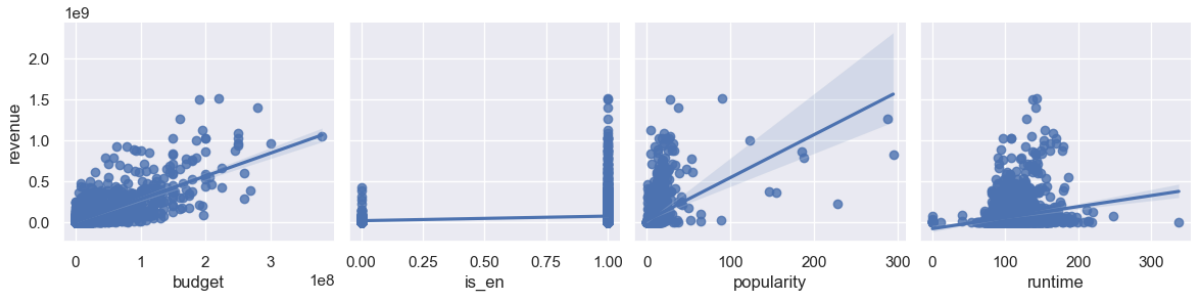

Figure 1: Correlation Coefficient

Exercise 2: Multiple Regression Analysis

In this exercise, we employ multiple regression analysis to predict movie revenue. We begin by incorporating the variables from Exercise 1 and subsequently introduce additional features if deemed relevant. These additional variables include:

- "has_actor": A binary variable indicating the presence of a well-known actor in the cast.
- "$has\_director$": A binary variable indicating the involvement of a renowned director.
- "$female\_percentage$": A numerical variable representing the percentage of women in the cast.

a)

We used the same variables as before, and added the 2 "binary" variables has_actor and $has\_director$, which refer to well-known actors and directors respectively. As well as the variable $female\_percentage$, which refers to the percentage of women in the cast of the film. The R^2 value of our model is 0.616, meaning that the variables included explain approximately 61.6% of the variability in movie revenue and adjusted R^2 value is 0.615 accounting the number of predictors in the model.

b)

We evaluate the conditions required for the multiple linear regression model to be valid, including linearity, normality, homoscedasticity, and independence.

Linearity: For confirming the linearity assumption, we will look at the plot of the actual vs predicted revenue:

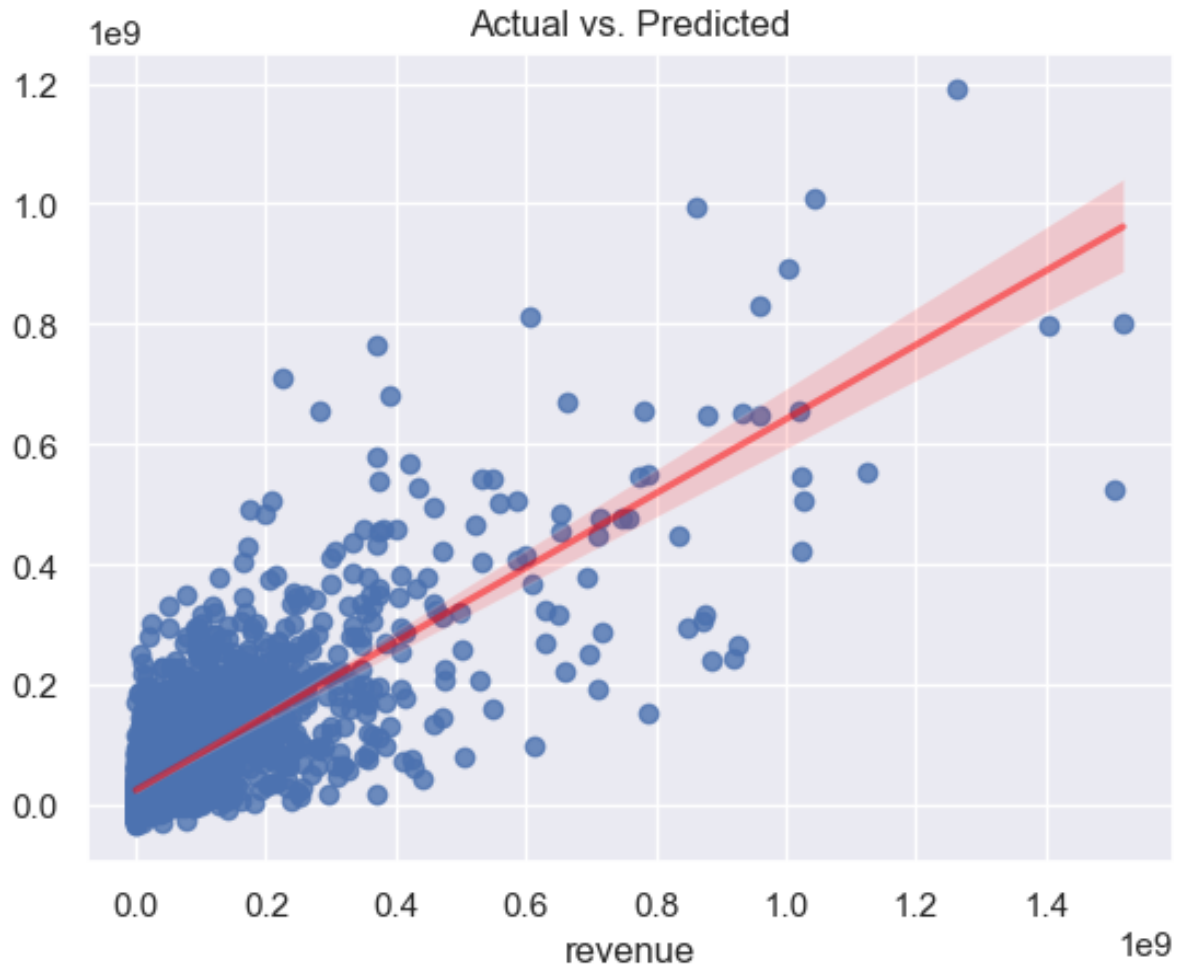

Figure 2: Scatter Plot of Predicted and Actual Revenues

Since there appears to be a linear relationship, we conclude that the assumption of linearity is satisfied.

Normality Observing the distribution of our residuals in the plot below, we notice that it is nearly normal so the normality condition is also satisfied.

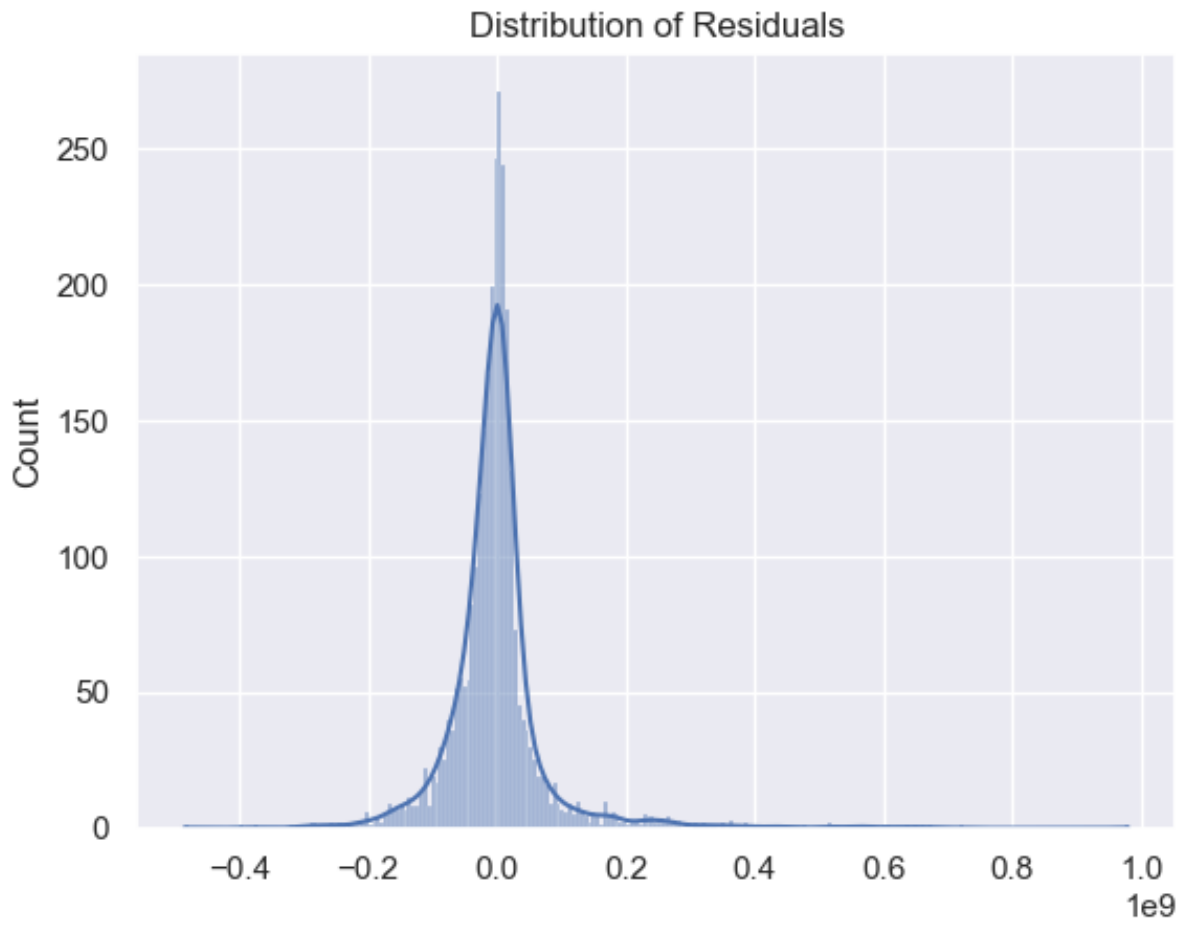

Figure 3: Distribution of Residuals

Homoscedasticity: We can see the residuals have constant variance, so the homoscedasticity assumption is met.

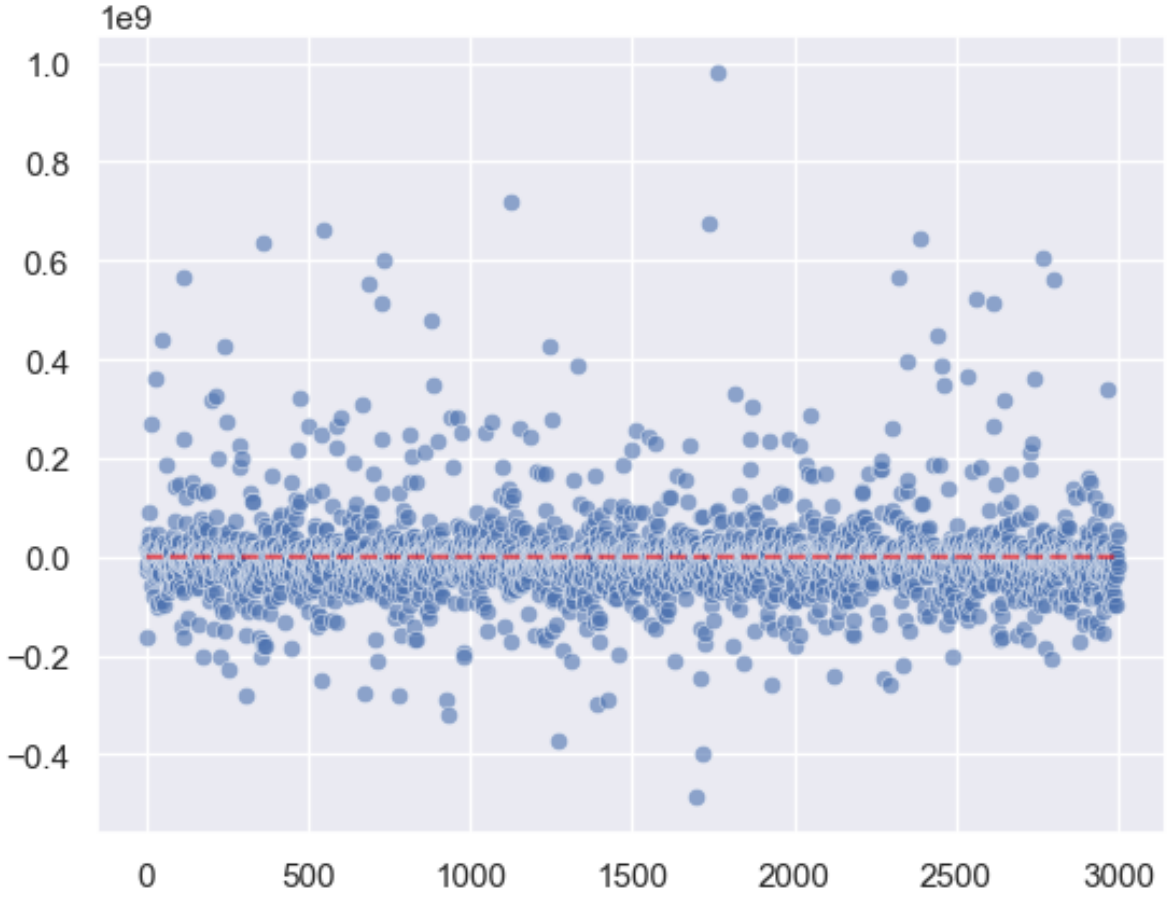

Figure 4: Residuals Plot

Independence: As for independence, we are going to have a look at the correlation matrix and the VIF values for our features:

	budget	female_pct	has_actor	has_director	popularity	runtime
budget	1.000000	-0.100811	0.158802	0.160743	0.344388	0.226170
female_pct	-0.100811	1.000000	-0.001863	-0.055663	-0.013052	-0.054402
has_actor	0.158802	-0.001863	1.000000	0.036478	0.095703	0.077155
has_director	0.160743	-0.055663	0.036478	1.000000	0.083465	0.081436
popularity	0.344388	-0.013052	0.095703	0.083465	1.000000	0.127282
runtime	0.226170	-0.054402	0.077155	0.081436	0.127282	1.000000

Figure 5: Correlation Matrix

	feature	VIF
0	budget	1.688575
1	female_pct	3.649285
2	has_actor	1.054551
3	has_director	1.045620
4	popularity	1.816764
5	runtime	4.616885

Figure 6: VIF values

We can see that there are no features with very high correlation, and that no VIF value is greater than 5 (even though runtime comes close), so we can conclude that the features are independent.

c)

The slope associated with the most predictive variable identified in Exercise 1, i.e., budget, is 2.4127. This implies that, on average, a one-unit increase in budget will result in a 2.4127 unit increase in revenue, assuming all other variables are held constant (Ceteris Paribus).

d)

The most important variables for predicting the revenue of a film (based on p-values) are:

- Budget
- Popularity
- Whether the film is directed by one of the well-known directors we have defined

Conclusion

In conclusion, this analysis used data from $The\ Movie\ Database$ to predict box office revenue for movies. The **strongest** association with revenue was found for the budget variable. A multiple regression model incorporating additional features achieved an R^2 value of 0.616, explaining approximately 61.6% of revenue variation. The assumptions of multiple linear regression were met. Budget, popularity, and the involvement of renowned directors were identified as significant predictors of movie revenue. These findings emphasize the importance of budget allocation, popularity, and industry professionals in driving movie success, providing valuable insights for stakeholders to optimize financial outcomes.