DEPARTMENT OF MATHEMATICS AND APPLIED MATHEMATICS
UNIVERSITY OF CRETE

MEM264-Applied Statistics

Project report

Prediction on worldwide movie revenue using multiple linear regression

---

Advisor: Dr. Sofia Triantafyllou

Students: Chalkidis Theodoros csd4198
Manos Georgios csd4333
Zarkos Christos csd4516

HERAKLION, JUNE 2023

Contents

List of Figures

List of Tables

1 Introduction

The goal of this project is to predict the worldwide gross revenue of movies using information in the data we get for some movies from The movie database. To do that we will use multiple linear regression. Initially, we will use only 4 explanatory variables (proposed by the project assignment). Consequently, we will try to expand our model by adding explanatory variables and then dump variables that are not really useful to our model. At the same time we will answer some key questions about the model that are also asked by the project assignment.

2 Variable selection and analysis

In this section we are going to use the 4 explanatory variables proposed by the project assignment. We will to compute the correlation coefficient between each explanatory variable and the response variable. Also, we will visualize their relationships using scatterplots and lastly identify the most important of the 4 explanatory variables.

2.1 Correlation coefficients

In this section we will show the calculation results of the correlation coefficients between each explanatory variable and the response variable.

Explanatory variable	**Correlation coefficient (R)**
budget	0.753
is_english	0.142
runtime	0.216
popularity	0.461

Table 2.1: Correlation coefficients table

From the above results, we notice that the explanatory variable with the best results is **budget**, so this is if we could only use only one variable we would use **budget** since due to its higher correlation coefficient we can conclude that it is the variable with the strongest linear association with the explanatory variable.

2.2 Scatterplots

In this section we will show the scatterplots between each explanatory variable and the response variable. We expect the scatterplot between budget and revenue to be the one looking the most linear since they got the highest correlation.

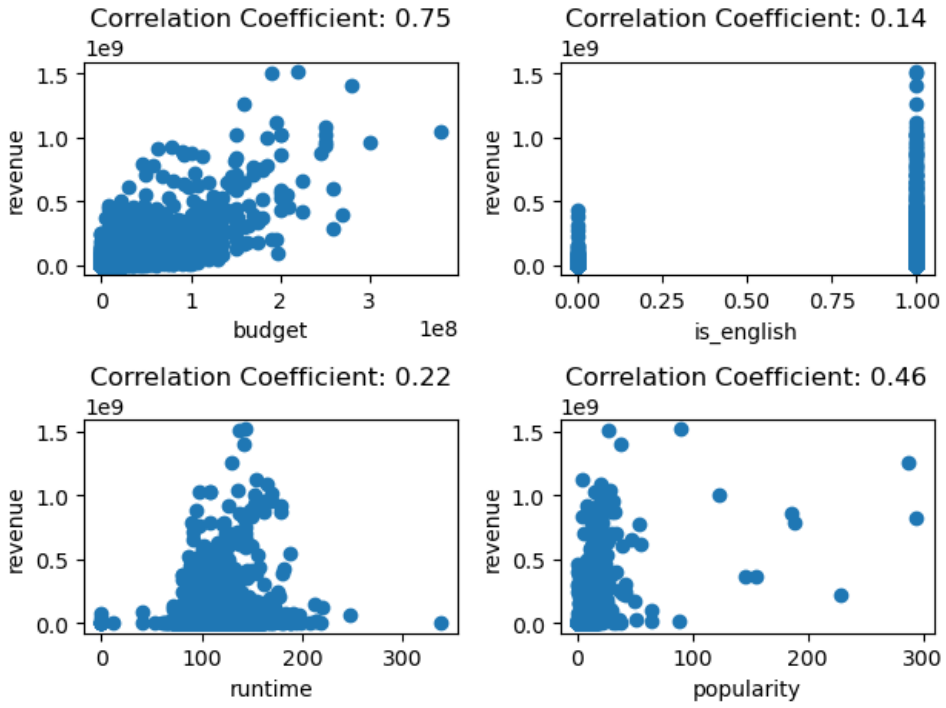

Figure 2.1: Scatterplots between the explanatory variables and the response variable

3 Regression and results

3.1 Data Issues

The dataset appears to be noisy. There are many missing values (i.e. NaN values), as well as out of bounds values (e.g. budget = 0) on many samples. For simplicity purposes, we handle NaN values by throwing away the respective sample, and we don't apply any integrity constraints or use any imputation methods.

Dropping all samples that have NaN values even results in throwing almost 2800 samples out of 3000. However, in order for our feature construction to work properly but also make sure we throw the least samples possible, we drop them after selecting the subset of variables we want to apply Multiple Regression on, resulting in keeping 2991 samples out of 3000.

3.2 Description of the analysis

For our regression task, we constructed 7 new variables, and included 4 original ones. A detailed description of the data is found in List 3.2. We assessed each movie based on how many famous actors it includes, how many famous directors it involved, how many famous production companies were involved and also based on its primary (1st) genre type. To do so, we created 3 separate lists of famous actors, directors and production companies, and also a map of genres and scores based on how much revenue the respective genre type usually generates (i.e. we know that thrillers usually make less revenue than action movies, based on internet surveys). To avoid any potential bias, we used internet resources and surveys to classify each genre to a respective score (1-5), with 5 responding to the genres that usually generate the most revenue.

- *budget*: Budget of a movie in dollars. 0 values mean unknown (original)
- *runtime*: Total runtime of a movie in minutes (original)
- *popularity*: Popularity of the movie in float (original)
- *is_english*: 1 if the movie's language is English, else 0 (custom)
- *genre_score*: An integer value (1-5) corresponding to the movie's primary genre category (custom)
- *is_holiday*: 1 if the movie's release month was between May and August or November and December, where usually holidays take place and cinemas are more crowded (custom)
- *movie_age*: We know that really old movies (i.e. <1988) didn't generate as much revenue in dollars as the ones today due to currency inflation changes (custom)
- *actor_score*: The number of famous actors the movie includes (custom)
- *director_score*: The number of famous directors the movie involved (custom)
- *company_score*: The number of famous production companies the movie included (custom)
- *revenue*: Total revenue earned by a movie in dollars (original target variable)

All of the aforementioned features are constructed logically based on other observations. We tried to avoid inducing any bias or correlation between samples (i.e. how much revenue the first movie of a director produced compared to a second one). We could also potentially try to use kernels and predict on a different space to better model the linear relationship of a variable with the target variable (e.g. if budget had a multinomial correlation or if popularity had an exponential correlation to the revenue variable, we could probably try constructing squared or exponential features).

With the aforementioned features, we constructed a linear regression model using statsmodel api. The model with 10 predictors and 2991 samples resulted in an R^2 score of 0.619, but since its a multiple linear regression model, we also computed the R^2_{adj} score of 0.618. The scores are also included in the model summary as provided by the package in Figure 3.1.

OLS Regression Results

Dep. Variable:	revenue	R-squared:	0.620
Model:	OLS	Adj. R-squared:	0.618
Method:	Least Squares	F-statistic:	485.6
Date:	Sat, 17 Jun 2023	Prob (F-statistic):	0.00
Time:	21:19:05	Log-Likelihood:	-58851.
No. Observations:	2991	AIC:	1.177e+05
Df Residuals:	2980	BIC:	1.178e+05
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.759e+07	1.05e+07	-2.638	0.008	-4.81e+07	-7.09e+06
budget	2.4519	0.054	45.498	0.000	2.346	2.558
is_english	-7.266e+06	4.83e+06	-1.503	0.133	-1.67e+07	2.21e+06
runtime	1.001e+05	7.5e+04	1.334	0.182	-4.7e+04	2.47e+05
popularity	2.606e+06	1.38e+05	18.917	0.000	2.34e+06	2.88e+06
genre_score	1.67e+05	1.29e+06	0.129	0.897	-2.36e+06	2.7e+06
is_holiday	8.984e+06	3.19e+06	2.818	0.005	2.73e+06	1.52e+07
movie_age	2.056e+05	1.1e+05	1.873	0.061	-9613.063	4.21e+05
actor_score	-3.175e+06	3.91e+06	-0.811	0.417	-1.08e+07	4.5e+06
director_score	3.713e+07	8.2e+06	4.528	0.000	2.11e+07	5.32e+07
company_score	4.338e+06	3.03e+06	1.431	0.153	-1.61e+06	1.03e+07

Omnibus:	2028.965	Durbin-Watson:	2.026
Prob(Omnibus):	0.000	Jarque-Bera (JB):	63313.653
Skew:	2.774	Prob(JB):	0.00
Kurtosis:	24.846	Cond. No.	3.02e+08

Figure 3.1: Ordinary Least Squares regression summary

Note that the Condition Number (Cond. No.) is large. The summary mentions that this might indicate that there are strong multicollinearity or other numerical problems, but given the noise in our data (e.g. budget values of 0) this is something we would expect. The const variable mentioned first corresponds to the intercept of the model. In section 3.5, we select the most significant variables to understand which ones are the most important ones for our model.

3.3 Linear regression conditions

In this section we will check if the conditions for multiple linear regression are satisfied by our model.

3.3.1 Nearly normal residuals

To check if our residuals are (nearly) normally distributed we will plot a histogram of our residuals.

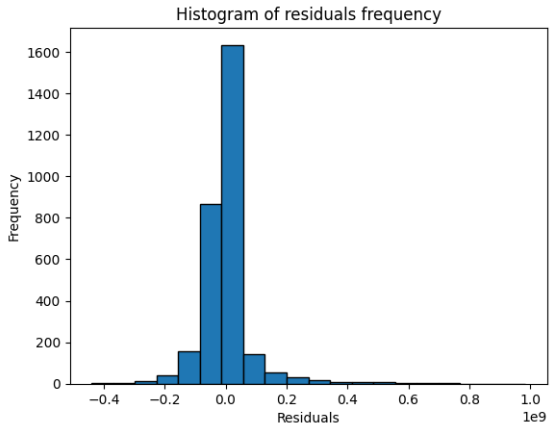

Figure 3.2: Histogram of the frequency of the residuals

It is clear from the histogram that our residuals are (nearly) normally distributed.

3.3.2 Constant variability in residuals

To check for constant variability we will do a scatterplot of the residuals and the absolute value of residuals vs. the predicted values of our model.

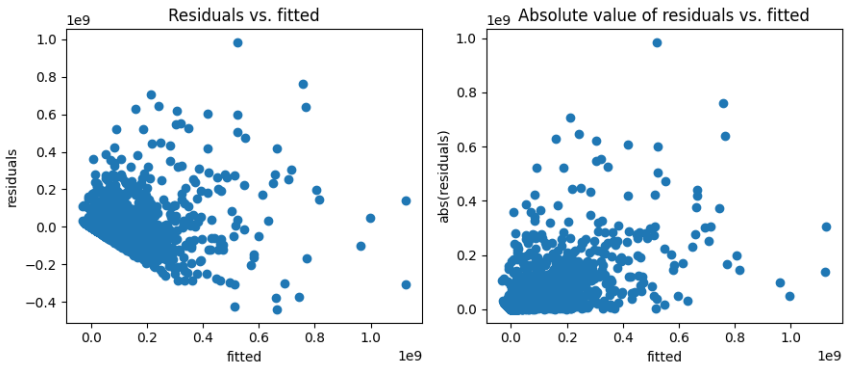

Figure 3.3: Scatterplots of the residuals and abs(residuals) with the predictions

In this scatterplots, we can see that our residuals are mostly in a range of 0.4e9 units within our predicted value. We clearly have some outliers, in the positive range but given the size of our sample, we would say these outliers are not significant and thus the condition is met.

3.3.3 Independence of residuals

Since our samples are independent we can also conclude that our residuals are independent too.

3.3.4 Each numerical variable linearly related to the outcome

To check for linearity we will plot a scatterplot of the residuals vs. each numerical explanatory variable. We only have 3 such variables in our model (budget, runtime and popularity).

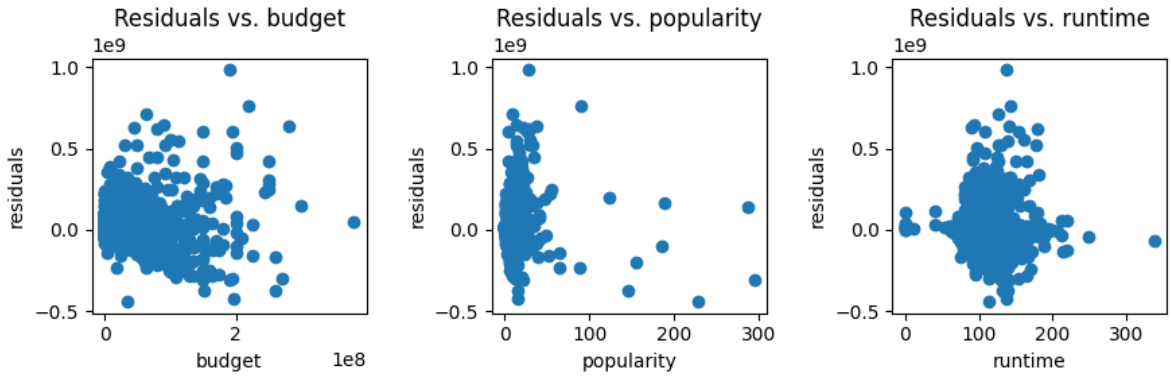

Figure 3.4: Scatterplots of the residuals vs. each numerical explanatory variable

3.4 Analysis of the most predictive variable

As we can see from our model summary the slope of our most predictive variable in exercise 1, which is budget, is 2.4519. The interpration of this slope is that if everything else is held constant, an increase to the budget of the movie by 1 dollar would lead the movie revenue increasing by 2.4519 dollars on average.

3.5 Most significant variables

In this section, we implemented both forward selection and backward elimination algorithms to select the most significant variables for predicting movie revenue. We tried

both of them separately, as well as the forward-backward selection algorithm. Both followed the p-value approach as when we care about understanding which variables are statistically significant predictors of the response, the p-value approach is preferred. In our case, we deploy a statistical (conditional) independence test using causal-learn library to compute a p-value and decide which variables to select in each case. We tried Fisher-z test, Chi-Square test, Kernel-based conditional independence (KCI) test and G-Square test. However, while Fisher-z was really fast, KCI required quite some time to finish and produce the exact same results. Chi-Square and G-Square were only viable in forward selection, resulting in only 2 features, while it required way too much memory for backward selection thus deeming them impossible to run in our case.

Using Fisher-z test, all 3 aforementioned approaches resulted in the following features being selected: budget, popularity, is_holiday, movie_age, director_score. The resulting model, as seen on the summary in Figure 3.5, has decreased R-squared and adjusted R-squared scores compared to the original one. This is expected as we selected features based on p-values and not the ones that increase R-squared metrics. Also note that the Cond. No. is still large.

OLS Regression Results

Dep. Variable:	revenue	R-squared:	0.619
Model:	OLS	Adj. R-squared:	0.618
Method:	Least Squares	F-statistic:	969.2
Date:	Sat, 17 Jun 2023	Prob (F-statistic):	0.00
Time:	21:44:43	Log-Likelihood:	-58854.
No. Observations:	2991	AIC:	1.177e+05
Df Residuals:	2985	BIC:	1.178e+05
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.213e+07	3.48e+06	-6.359	0.000	-2.9e+07	-1.53e+07
budget	2.4733	0.047	53.101	0.000	2.382	2.565
popularity	2.607e+06	1.37e+05	19.024	0.000	2.34e+06	2.88e+06
is_holiday	9.614e+06	3.18e+06	3.026	0.003	3.38e+06	1.58e+07
movie_age	2.189e+05	1.05e+05	2.089	0.037	1.34e+04	4.24e+05
director_score	3.757e+07	8.15e+06	4.612	0.000	2.16e+07	5.35e+07

Omnibus:	2031.611	Durbin-Watson:	2.023
Prob(Omnibus):	0.000	Jarque-Bera (JB):	63463.260
Skew:	2.780	Prob(JB):	0.00
Kurtosis:	24.871	Cond. No.	2.27e+08

Figure 3.5: Ordinary Least Squares regression summary after variable selection