Applied Statistics

IFIGENEIA TSIFLIDOU (MATH 5736)

In this project, we will try to predict the box office revenue (in dollars) of movies. The data include 3000 Movies in The movie database. The data include information like the movie's budget, language, genre, and cast. Some of are numerical random variables (e.g., budget), some are categorical (e.g., genre) and some information is in text format. Your goal is to predict the worldwide gross revenue (variable name: revenue) using the information in the data. A detailed description of the data can be found in the last page of this document.

Exercise 1

We will first consider only the following explanatory variables:

- *budget*
- *binary variable denoting if the movie is english or not*
- *running time*
- *popularity*

For each of the numerical explanatory variables, compute the correlation coefficient and visualize the relationship of the variable with the response variable with a scatter plot. If you could only use one variable to predict revenue, which one would you use?

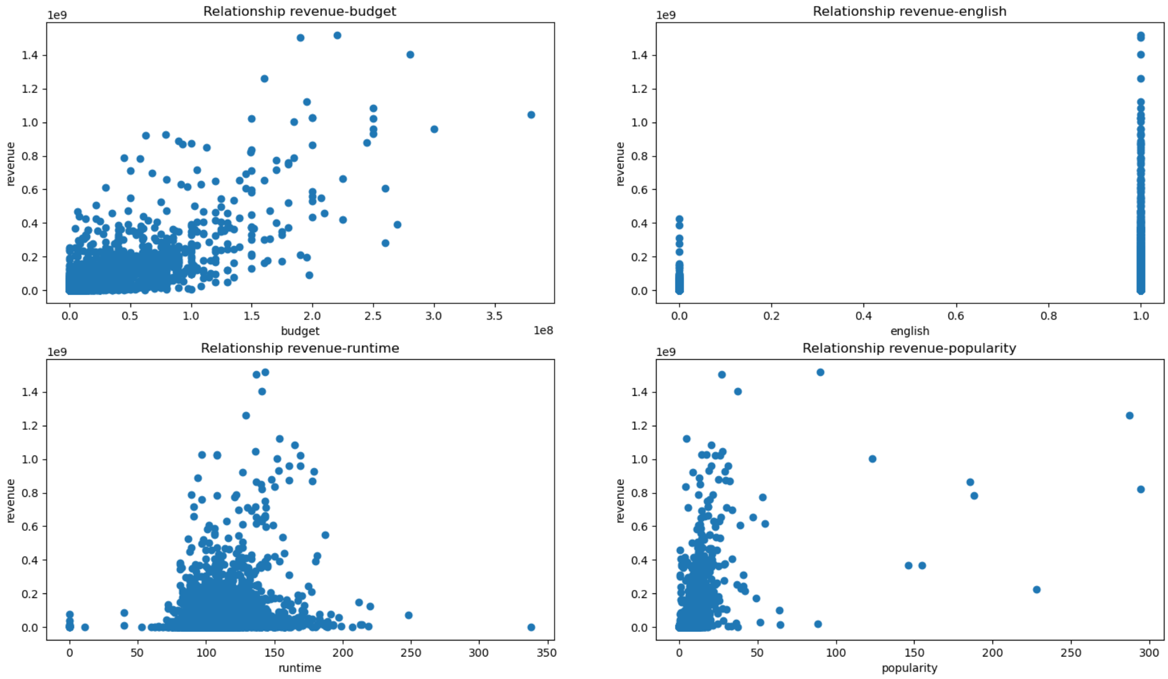
To provide revenue, we will use the correlation coefficient between the response variable (revenue) and the four explanatory variables: budget, english, runtime, popularity. The binary variable english was created by comparing the original variable original_language to the value english.

A correlation coefficient close to 0 indicates that there is little to no dependency between the two variables. In this case, the variable does not provide useful information for prediction. On the other hand, a value close to the extremes, +1 (for positive correlation) or -1 (for negative correlation), means that when one variable changes, the other tends to change in the same or opposite direction, respectively. Practically, this can give us valuable insights for our prediction.

After calculating the correlation coefficients for the pairs revenue-budget, revenue-english, revenue-runtime, and revenue-popularity, we observe that the most interesting value is that of revenue-budget, with the value of ~0.75. If we had to choose only one variable, we would select the variable ***budget***.

```
The correlation coefficient between revenue and budget is: 0.7529645103815285
The correlation coefficient between revenue and english is: 0.14212987285400097
The correlation coefficient between revenue and runtime is: 0.21638013018147245
The correlation coefficient between revenue and popularity is: 0.4614602896736137
```

We can visually confirm this choice by plotting each pair in separate scatter plots. For the revenue-budget pair, we observe a relatively linear relationship.



Exercise 2

You will now use multiple regression to predict movie revenue. You can use just the variables in Exercise 1 above, or you can construct additional features: For example, you could construct a binary variable encoding if Brad Pitt is in the cast, or a numerical variable that encodes the number of female actors in the film.

a) Briefly describe your analysis (no more than 2 paragraphs). Did you include additional variables, and, if so, which one? What is the R^2 of the model?

In addition to the four variables used in Exercise 1, we now explore whether other variables could be equally useful for predicting revenue by examining their correlation coefficients. Since revenue is the target variable, it cannot be used as a predictor. Most of the remaining data in the original dataset is either in text form (e.g. titles, descriptions) or sorted in JSON format. From this data, we derived two additional variables:

- The number of **male actors** in each movie (based on the original *cast* variable)
- The number of **production companies** involved in each movie (based on the original *production_companies* variable)

As in the previous exercise, we calculate the correlation coefficients between *revenue* and each of these new variables:

```
Correlation between revenue and men: 0.3725865209264919
Correlation between revenue and companies: 0.15569953980317414
```

We then use the sklearn library to build a multiple linear regression model to predict revenue. Based on the correlation values, we select the four most promising predictor variables: *budget, popularity, men,* and *companies*. Using the .fit method, the model calculates the optimal slopes and intercept for the regression line.

The $\boldsymbol{R^2\ \ coefficient}$, which ranges from 0 to 1, indicates how well the model fits the data. The $\boldsymbol{R^2}$ value obtained is:

```
R-squared: 0.38467963775962166
```

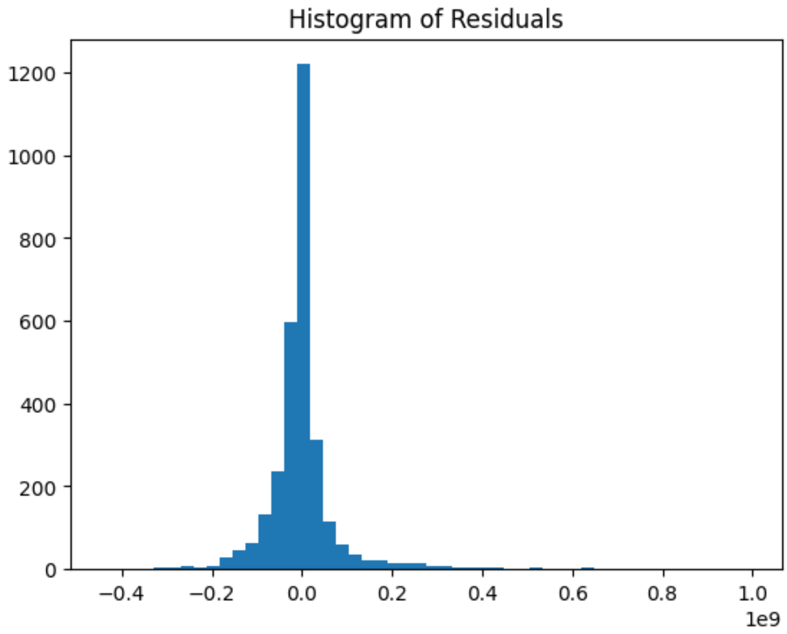
b) Are the conditions for the multiple linear regression model satisfied? Explain your answer

The multiple linear regression model is defined by the following equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Where β_0 is the intercept, and $\beta_1, \ldots, \beta_p$ are the slopes of the variables. The assumptions that must be met for a linear regression model to be valid are the following:

1. The residuals should follow a roughly normal distribution. This can be verified using a histogram and by visually inspecting the shape of the distribution — which is the case here.



2. The residuals should have constant variance.
3. The residuals should be independent.
4. Each predictor variable should have a linear relationship with the dependent variable.

c) *What is the slope associated with the variable you identified as most predictive in Exercise 1 above?*

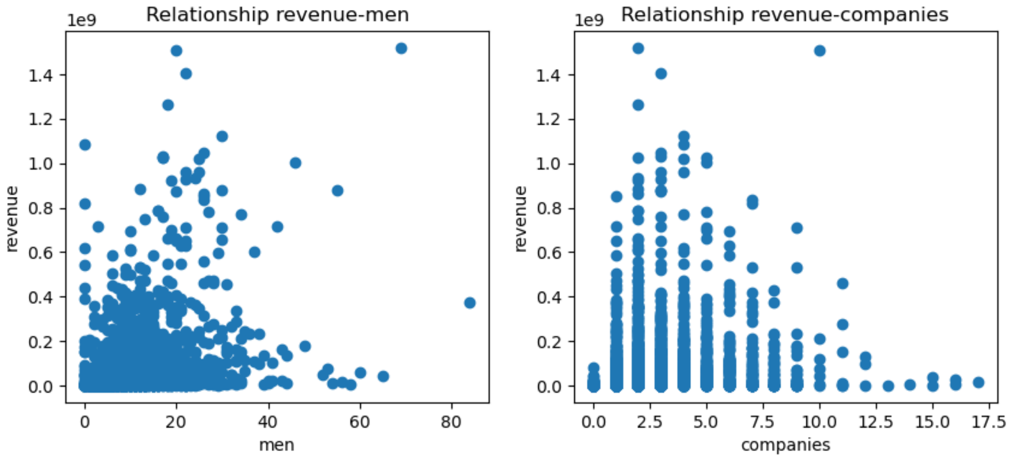
The slopes calculated from the multiple linear regression for each of the four variables are:

```
Slope: [ 2.48287008e+00  2.58367158e+06  9.96693078e+05 -3.97996983e+06]
```

The most predictive variable, *budget*, has a slope of ~2.48. This relatively small value is due to the large scale (~10^9) of *budget-revenue*, in contrast to the other variables. The slope of each variable represents how much the dependent variable **Y** (*revenue*) changes on average when that specific variable $\boldsymbol{X_i}$ increases by one unit, keeping all other variables constant.

d) *Which variables are significant for predicting movie revenue?*

As mentioned in Exercise 1 and in part *a)*, the usefulness of a variable depends on its correlation coefficient with the target variable (*revenue*). The further the coefficient is from 0-whether positive or negative- the more informative the variable is for predicting revenue. Overall, the most important variables, ranked in descending order, are budget, *popularity, men, companies*. Starting with *budget* and gradually adding the other variables to the model, we observed that the $\boldsymbol{R^2}$ value improves at each step, although with diminishing returns. That is, each additional variable contributes less to the model's explanatory power than the previous one.


Relationship revenue-men
1e9
revenue
men
Relationship revenue-companies
1e9
revenue
companies