Simple and Multiple Regression Analysis

Statistical Learning Models - Project 1

Pedro Nunes, PK - pedronunes0028@gmail.com

October 2023

1 Introduction

Regression analysis is the statistical method used to determine the structure of a relationship between two variables (Simple Regression) and three or more (Multiple Regression). On this paper, a study of both types of regressions will be exposed. All the code done can be found in `https://github.com/Viperxyzzz/Statistical-Learning-Models`

2 Simple Regression

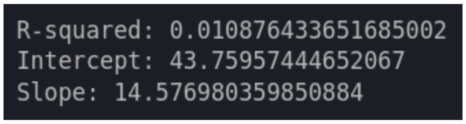
2.1 Dataset

For this project, the chosen dataset was the Song Popularity Dataset, which can be accessed at the following link: `https://www.kaggle.com/datasets/yasserh/song-popularity-dataset/data`. The dataset represents various songs with features such as popularity, danceability, acousticness, liveness, etc. It had 18835 entries with 15 columns of data.

Two variables were chosen, popularity and danceability. My assumption was that if a song has good danceability, then it must be popular. Thus, the independent variable is danceability, and the dependent variable is popularity.

2.2 Regression Analysis

A simple regression model was fit, and we got the following values:



```
R-squared: 0.010876433651685002
Intercept: 43.75957444652067
Slope: 14.576980359850884
```

Figure 1: Results of the regression model

The R-Squared value is really low, which means that only about 1.09% of the variance in the dependent variable can be explained by the linear relationship with the independent variable. This means that our model doesn't do a good job in this situation.

The Intercept value is about 43.76, which means that when the danceability approaches 0, the popularity of the song is 43.76.

The Slope is around 14.58, which means that for each unit of danceability, the popularity increases around 14.58 units. The positive value means that when the independent variable increases, so does the dependent variable.

2.3 Outlier Analysis

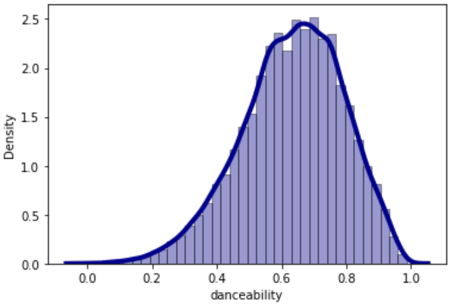
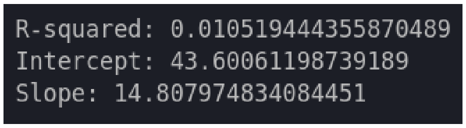


Figure 2: Danceability

Our danceability follows a normal distribution. For removing outliers an interquartile range was used (IQR). IQR is the range between the first and the third quartiles namely Q1 and Q3: IQR = Q3 − Q1. The data points which fall below Q1 − 1.5 IQR or above Q3 + 1.5 IQR are outliers. At first we had around 18835 entries. After removing the outliers we were left with 18702 entries. A linear regression was then calculated, and a new analysis was done.



```
R-squared: 0.010519444355870489
Intercept: 43.60061198739189
Slope: 14.807974834084451
```

Figure 3: Results after outlier removal

Astonishingly, our results were even worse, with a lower value of R-Squared. Thus, no more outlier removal was carried on.

2.4 Homoskedasticity Check

Homoskedacity means that the variance of the residuals is constant across all levels of the independent variable. To get a solid understanding we plot the residuals against the predicted values.

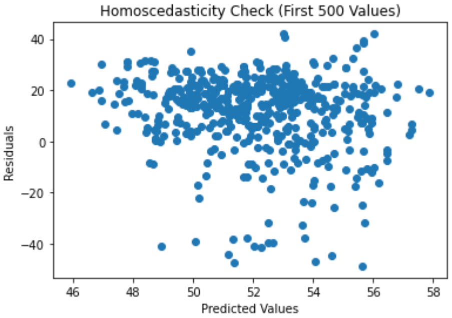


Figure 4: Homoskedacity graphic

With this graphic we can understand that it is taking a cone shape, which means that it isn't respecting homoskedacity. Rather it indicates heteroscedasticity, since the size of the error term differs across values of an independent variable instead of being a constant.

2.5 Independence via Autocorrelation

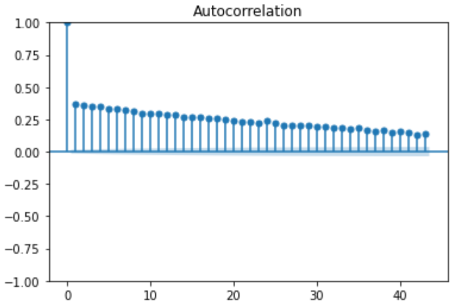


Figure 5: Autocorrelation by lags graph

The ACF values are consistently around 0.25, which implies that there is a moderate positive autocorrelation between the 'song_popularity' variable and its past values at various lags.

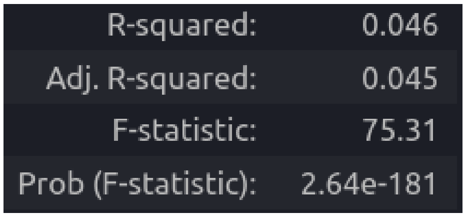
This means that the variable is not independent, but rather shows some temporal dependencies.

3 Multiple Regression

3.1 More Features

For a multiple regression, more features were chosen. In this case - duration, acousticness, danceability, energy, instrumentalness, key, liveness, loudness, speechiness, tempo, time_signature and audio_valence. Each of them with 18835 entries, thus meeting the 15 * p criteria.

3.2 Multiple Regression Analysis

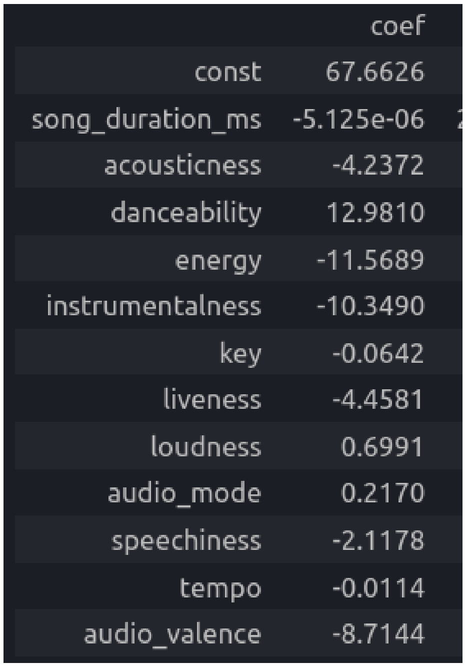


R-squared:	0.046
Adj. R-squared:	0.045
F-statistic:	75.31
Prob (F-statistic):	2.64e-181

Figure 6: Multiple Regression Results

For the multiple regression, the R-Squared value is about 0.046, whereas the adjusted is 0.045. When compared to the simple regression, our values are 4 times greater. However, they are still really really low, which means our model is failing to do a good job. However, the f-statistic values has a low p-value, meaning it's statistically significant. Which means that at least there is a relationship between the independent and dependent variables.

3.3 Parameters Interpretation



	coef
const	67.6626
song_duration_ms	-5.125e-06
acousticness	-4.2372
danceability	12.9810
energy	-11.5689
instrumentalness	-10.3490
key	-0.0642
liveness	-4.4581
loudness	0.6991
audio_mode	0.2170
speechiness	-2.1178
tempo	-0.0114
audio_valence	-8.7144

Figure 7: Independent variables coefficients

By looking at the coefficients, we can conclude that the ones with the higher influence are danceability, energy and instrumentalness. When danceability increases one unit, the popularity increases by 12.98. On the other hand, when energy and instrumentalness increase by one unit, the popularity decreases, by 10.35 and 11.57, respectively.

4 Conclusions

In conclusion, the regression analysis carried on in this study, both simple and multiple, showed relatively low R-Squared values, thus meaning that the chosen independent variables had low explanatory power for predicting a song popularity.

Due to the complexity of the problem, more sophisticated models are required to produce better results.