Analyzing Song Popularity with Multiple Linear Regression

Carlotta Gherardi, UID: 505812614

1 Introduction

To leverage the continuous growth of the music industry, it can be useful to better understand what musical features of a song increase its popularity. The dataset used for this analysis contains information about songs from the streaming service Spotify. Specifically, the raw dataset contains 28,356 observations (each represents a song), and 23 variables. Ultimately, the aim of this project is to investigate what musical characteristics of a song make it more popular by fitting a linear regression model to the data, using 'track_popularity' as the response variable.

2 Data Cleaning

The raw dataset included several duplicate 'track_name' and 'track_artist' pairs, such as the same song released in multiple albums. Since the goal of this project is to analyze which musical attributes of a song increase the song's popularity (and 'track_album' is not considered such by this paper) duplicates were removed by retaining only the most popular version of each song. Additionally, all variables unrelated to the musical composition of the song were removed. The 'track_artist' variable is a covariate, as some singers may be better known leading to their songs being more popular regardless of their musical structure, but it's a categorical variable with too many categories to be included in the model. So, to account for the possible confounding effect, a 'top_artist' binary variable was created. This variable takes a value of 1 if the mean 'track_popularity' for a song's artist is in the top 10 percent. Also, to control computational cost, the size of the dataset was reduced by keeping only the most recent, and presumably relevant observations (December 2019 onwards). This cleaned dataset had no missing values, 1580 observations, and 19 variables.

3 Descriptive Statistics

Table 1: Statistics for continuous variables

	Track Pop.	Danceability	Energy	Key	Loudness	Speechiness
Min	0.00	0.1730	0.0526	0.000	-26.207	0.02450
Mean	45.53	0.6769	0.6713	5.443	-6.908	0.11530
Median	45.00	0.6970	0.6870	6.000	-6.404	0.06975
Max	98.00	0.9720	0.9940	11.000	-0.247	0.68200

	Acousticness	Instrumentalness	Liveness	Valence	Tempo	Duration ms
Min	0.0000022	0.000000	0.02820	0.0335	57.5	65000
Mean	0.2102998	0.110078	0.18495	0.4683	122.4	195762
Median	0.1125000	0.000019	0.12500	0.4625	123.8	189994
Max	0.9720000	0.962000	0.97000	0.9810	204.1	515703

From Figure 1 we can see the response variable is approximately normally distributed. Figure 2 highlights that most predictors show weak correlations with 'track_popularity', although those with

Table 2: Statistics for binary variables

Category	Mode	Top Artist
0	756	1397
1	824	183

Table 3: Statistics for categorical variables

Playlist Genre	Count
edm	356
latin	258
pop	288
r&b	150
rap	444
rock	84

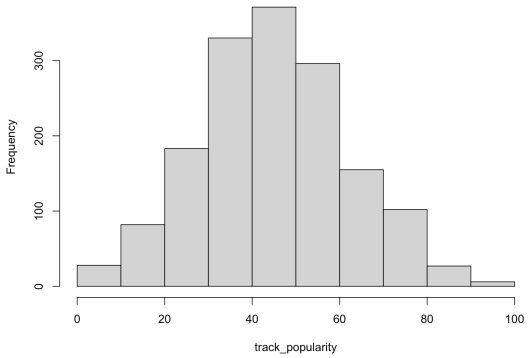

Figure 1: Histogram of track popularity

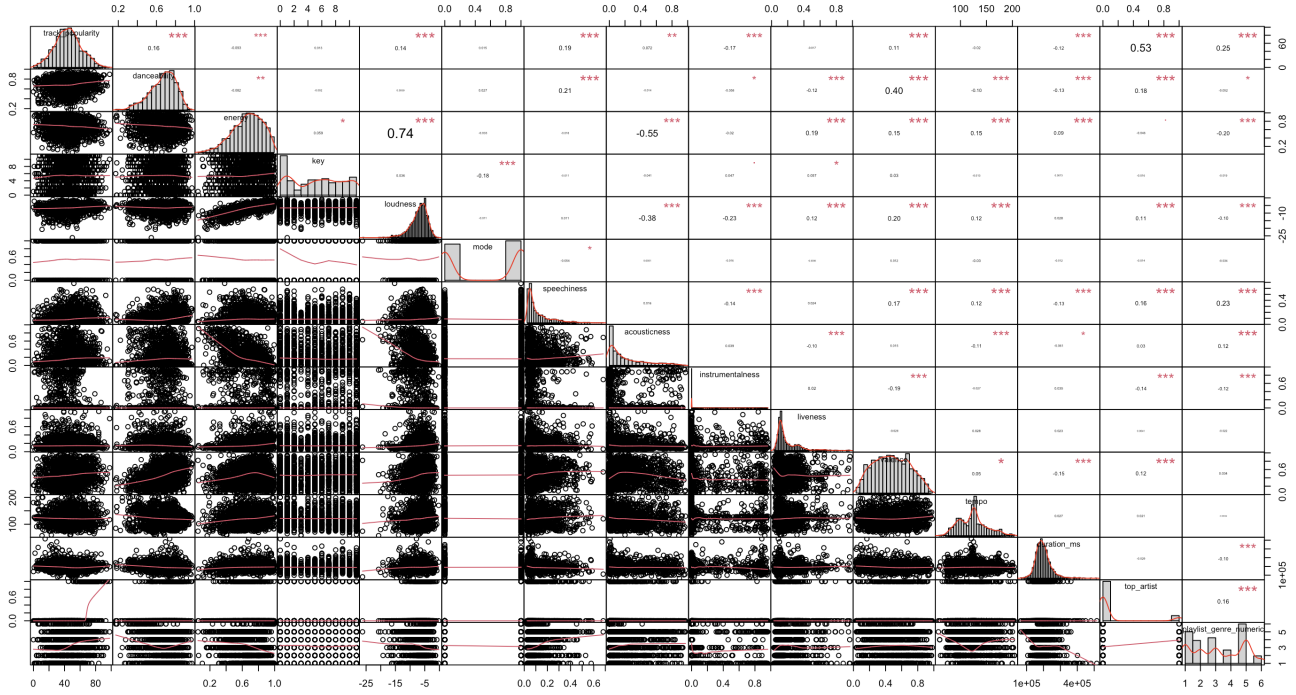

Figure 2: Correlation plot

'top_artist' and 'playlist_genre' appear stronger. Some predictors seem more highly correlated, such as 'energy' and 'loudness', but all correlation coefficients are below the 0.8 threshold typically indicative of multicollinearity. Both Figure 2 and Table 1 suggest several predictors, like 'speechiness', 'loudness', and 'duration_ms' are skewed, so the model may not capture the the full effects of these variables accurately. Also, only 11.58% of observations are from top artists, and these tend to have higher popularity scores, so it may lead the model to overstate the effect of a song being from a top artist. There is some imbalance in the 'playlist_genre' categories too, with edm being the most represented genre, and hence possibly adding bias to the model.

4 Model Diagnostics

A baseline model containing all 19 predictors was fit to the cleaned dataset. The diagnostic plots in Figure 3 indicate the baseline model lacked validity. The Residuals vs Fitted plot appears to support the assumption of linearity and that the mean of the error terms is 0, as the scatter of the points makes the red line mostly straight at 0. However, this plot, like the Scale-Location plot, shows a funnel-shaped pattern that indicates non-constant variance of the error terms. The Normal Q-Q plot indicates that the normality assumption is met as most points lie on the straight line with slope of 1.

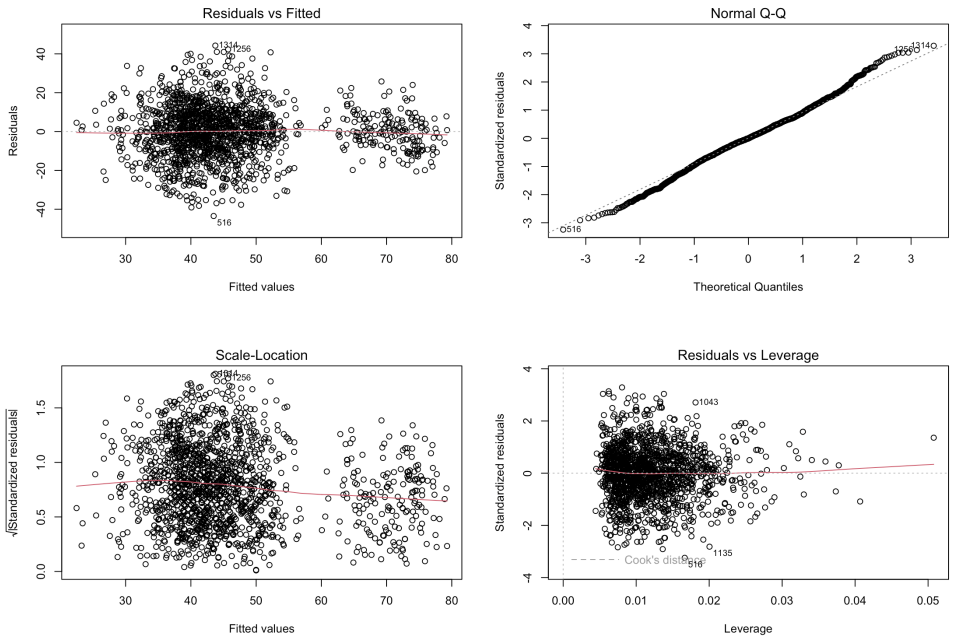

Figure 3: Baseline model diagnostic plots

Some deviations in the tails suggest that outliers may be unduly influencing the model. In fact, in the Residuals vs Leverage plot various points have a standardized residual outside the -2 to 2 range, and high leverages greater than $2 * (p + 1)/n \approx 0.024$: hence, likely a large cook's distance as well.

Outliers were investigated using standardized residuals and cook's distance, and these observations were mostly rare cases of extremely popular and unpopular songs. Given that the focus of this project is to understand the typical factors that influence track popularity rather than predict popularity, and since removing the outliers would still leave 1453 observations, the outliers were removed from the data. Moreover, to try to improve the model's inability to meet the homoscedasticity assumption, a box cox transformation was applied to the response variable ($\Lambda = 1.2$). Additionally, none of the predictor variables had a Variance Inflation Factor (VIF) greater than 5, so we assumed there was no multicollinearity that could cause coefficients to be poorly predicted. Regardless, to improve the model further, all-subset variable selection was employed, using Bayes Information Criteria (BIC) to determine the subset of predictors, as the subsets suggested by other metrics seemed to cause overfitting. The final model included 7 predictors, and had a BIC of -966.3887.

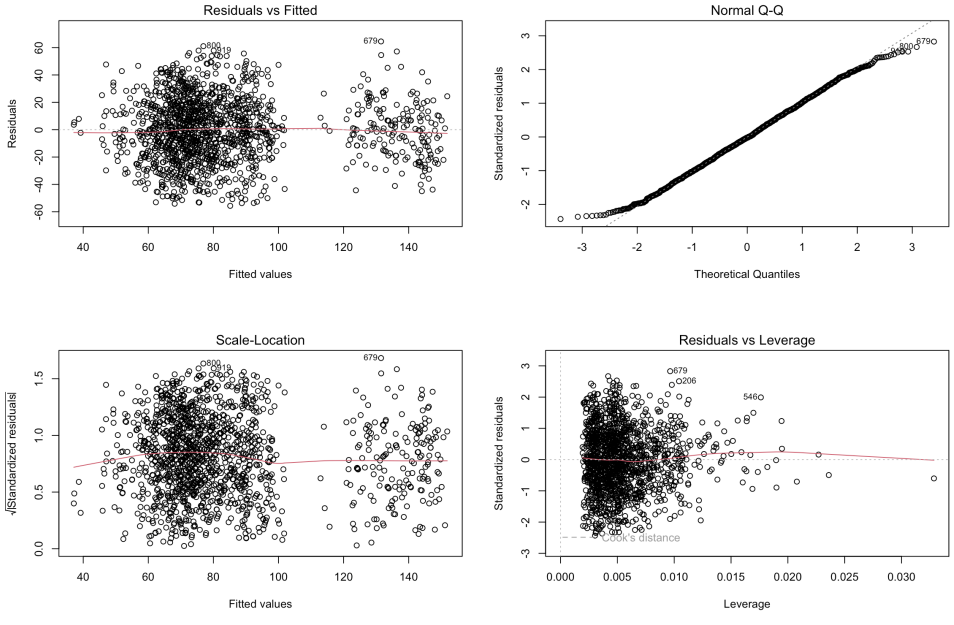

Figure 4: Final model diagnostic plots

Figure 4 shows visible improvement in the model's validity. Particularly, in comparison to the baseline model, the final model's Residuals vs Fitted, and Scale-Location plots no longer show an obvious funnel-shaped pattern. Instead, variance in the residuals seems to remain more constant across different values in the response variable, and there is increased random scatter about the horizontal line at 0, making the red lines straighter as well: all suggesting the homoscedasticity assumption is more satisfied. Also, in the Normal Q-Q plot points lie closer to the straight line, especially in the tails, implying improvement in the normality of the residuals too. The Residual vs Leverage plot shows fewer points with high standardized residuals and leverage, indicating a lower chance of outliers excessively affecting the model. Lastly, to facilitate conclusions on the final model, the continuous predictors were scaled.

5 Inference

The following is the final model:

$$\begin{aligned}\text{popularity\_transformed} = 67.8419 &- (6.6547 \times \text{energy}) + (9.1527 \times \text{loudness}) - (2.4990 \\ &\times \text{duration\_ms}) + (52.2254 \times \text{top\_artist}) + (6.8497 \\ &\times \text{playlist\_genrelatin}) + (5.6592 \times \text{playlist\_genrepop}) + (20.5283 \\ &\times \text{playlist\_genrerap}) + \epsilon\end{aligned}$$

A t-test was run for each coefficient of the model. In other words, for each β_i with $i \in \{1, 2, \ldots, 7\}$:

$$\begin{aligned}&\text{Null hypothesis } (H_0) : \beta_i = 0 \\ &\text{Alternative hypothesis } (H_1) : \beta_i \neq 0\end{aligned}$$

Table 4 shows all the t-tests resulted in p-values smaller than 0.05, so we rejected the null hypotheses,

Table 4: t-test results

Term	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	67.8419	1.0489	64.676	< 2e-16 ***
energy	-6.6547	0.9536	-6.979	4.53e-12 ***
loudness	9.1527	0.9295	9.847	< 2e-16 ***
duration_ms	-2.4990	0.6137	-4.072	4.91e-05 ***
top_artist	52.2254	1.9060	27.401	< 2e-16 ***
playlist_genrelatin	6.8497	1.8316	3.740	0.000191 ***
playlist_genrepop	5.6592	1.7802	3.179	0.001510 **
playlist_genrerap	20.5283	1.6068	12.776	< 2e-16 ***
Adjusted R-squared			0.5036	

and concluded that every predictor, when all other predictors are kept constant, significantly impacts 'popularity_transformed'. Moreover, an Adjusted R-squared of 0.5036 means that the variation in the predictors explains about 50.36% of the variation in 'popularity_transformed', which is a moderate model fit. ANOVA was also run on the model. This tested the following hypothesis (sequentially):

$$\begin{aligned}&\text{Null hypothesis } (H_0) : \beta_1 = \beta_2 = \ldots = \beta_7 = 0 \\ &\text{Alternative hypothesis } (H_1) : \text{At least one } \beta_i \neq 0 \text{ for } i \in \{1, 2, \ldots, 7\}\end{aligned}$$

The last p-value in Table 5 (for 'playlist_genrerap') is associated with the overall F-test, meaning the hypothesis shown above including all the predictors. Since this p-value is much smaller than 0.05, we can strongly reject the null hypothesis, and conclude that at least one of the predictors impacts the response variable 'popularity_transformed'. In other words, the proposed model is a substantially better fit than an intercept only model. However, the p-value for 'playlist_genrelatin' is

Table 5: ANOVA F-test results

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
energy	1	12564	12564	23.9365	1.07e-06 ***
loudness	1	178645	178645	340.3446	< 2e-16 ***
duration_ms	1	23299	23299	44.3885	3.817e-11 ***
top_artist	1	472995	472995	901.1264	< 2e-16 ***
playlist_genrelatin	1	579	579	1.1031	0.29376
playlist_genrepop	1	3092	3092	5.8912	0.01534 *
playlist_genrerap	1	85674	85674	163.2213	< 2e-16 ***
Residuals	1445	758471	525		

greater than 0.05, suggesting that, after accounting for the variation explained by 'energy', 'loudness', 'duration_ms', and 'top_artist', including the 'playlist_genrelatin' predictor in the model does not significantly increase the model's explanatory power. An added variable plot seemed to confirm this, but a partial F-test was conducted to test a reduced model without 'playlist_genrelatin' against the full model. In other words:

$$\text{Null hypothesis } (H_0) : \text{The reduced model is sufficient to explain the response variable}$$
$$\text{Alternative hypothesis } (H_1) : \text{The full model explains the response variable significantly better}$$

The p-value in Table 6 is smaller than 0.05 so we reject the null hypothesis and keep the full model.

Table 6: Partial F-test results

Model	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
Model 1: Reduced Model	1446	765812				
Model 2: Full Model	1445	758471	1	7341.2	13.986	0.0001914 ***

6 Conclusion

The final model reveals that a song's baseline 'transformed_popularity', with all numeric predictors at zero and the genre set to the reference, edm, is approximately 67.8419. Also, it shows that a one-unit increase in a song's scaled energy or duration (assuming all else is kept constant) decreases its popularity by about 6.6547 and 2.4990 units respectively, whereas songs louder by one-unit see an increase of about 9.1527 units in popularity. This suggests 'loudness' is the most influential of these predictors given its greater coefficient, while 'duration_ms' impacts popularity the least. The model also highlights the significance of the 'top_artist' predictor, as a song by a top artist has, on average, a 'transformed_popularity' of about 52.2254 units higher. Regarding genre, the model predicts that, holding other factors constant, latin, pop, and rap songs are more popular than edm songs, with rap being the most popular as the popularity of rap songs is expected to be 20.5283 units higher than that of edm songs.

Finally, the model suggests that louder, less energetic, and shorter songs by top artists are more popular. Latin, pop, or especially rap songs tend to outperform edm songs in popularity. However, since the descriptive statistics had shown some skewness in the 'loudness' and 'duration_ms' predictors, as well as some imbalance in the 'top_artist' and 'playlist_genre' categories, more analysis, perhaps with a more balanced dataset, should be done.

7 References

30000 Spotify Songs. https://www.kaggle.com/datasets/joebeachcapital/30000-spotify-songs.