Lab 2

Luis Garcia, Giorgio Soggiu, Anuradha Passan

3/21/2022

4. A Results Section

	Dependent variable:		
	popularity		
	(1)	(2)	(3)
danceability	14.256***	19.664***	19.721***
	(0.290)	(0.335)	(0.297)
explicit		10.781***	7.748***
		(0.259)	(0.218)
speechiness		−4.946***	−1.586***
		(0.208)	(0.150)
valence		−14.978***	−11.156***
		(0.224)	(0.196)
energy		10.155***	6.181***
		(0.265)	(0.232)
acousticness		−11.783***	−8.339***
		(0.180)	(0.159)
log(followers)			2.645***
			(0.014)
Constant	20.338***	25.275***	−5.879***
	(0.169)	(0.253)	(0.278)
Observations	140,874	140,874	140,302
R^2	0.017	0.172	0.349
Adjusted R^2	0.017	0.172	0.349
Residual Std. Error	17.365 (df = 140872)	15.939 (df = 140867)	14.077 (df = 140294)
F Statistic	2,404.062*** (df = 1; 140872)	4,865.927*** (df = 6; 140867)	10,723.350*** (df = 7; 140294)
Note:			*p<0.1; **p<0.05; ***p<0.01

There is statistical significance in all of the three models. The coefficient for our predictor variable of interest is consistent across models and is big enough to validate at a first stage our theory of danceability determining popularity of a song. Robust standard errors have been used in all the models.

The model (1) has a R^2 of **0.017** and an effect size for danceability of **14.256**. R^2 is low, which means there is a big part of the variability in our target variable that is not explained by the model. But since the aim of the study is to explain causal relationships rather than predicting, we will focus in the effect size of

danceability, which is big enough to back our theory. For each unitary increase in danceability, the song's popularity should increase by **14.256**. To put this in context, It is important to note that by definition, danceability ranges from 0 to 1 and popularity from 0 to 100.

The second model (2) incorporates other features we consider important to determine the popularity of a song. The coefficient of determination increases significantly to **0.172** and the effect size of danceability increases from **14.256** to **19.664**. The effect sizes for explicit, speechiness and valence are 10.781, -4.946 and -14.978 respectively, being all three statistically significant. For the remaining 2 variables, energy and acousticness add predictive power to the model but are highly correlated. Therefore, the interpretation of these coefficients (10.155 and -11.783 respectively) should take this into account.

The last model (3) adds the artist's followers into the equation, arising a previously introduced reverse causality problem. Therefore, there could be violations to the one-equation structural model. The variable is log-transformed and the model increases its R^2 to **0.349** and the effect size of danceability is **19.721**, which is a similar level than the one of model (2). The effect size of the predictors present in the model (2) vary in some variables but remains robust at least for the sign and the relative absolute value, being 7.748, -1.586 and -11.156, 6.181 and -8.339 for explicit, speechiness, valence, energy and acousticness respectively. The coefficient for the log of followers is 2.645, which is interpreted as a 1% increase in followers make 2.645 increase in the song‘s popularity. This interpretation may not represent the reality because of the violations to the one-equation structural model.

To sum up, the effect size of our variable of analysis, danceability is consistent to any of the feature additions respect the base model (1), which means that there is a causal relationship between danceability and popularity. As a result, increasing danceability of a song can produce an increase on the song's popularity. The coefficients of determination increase significantly with the addition of features, wut we have to reject the use of the third model in favor of the model (2), because of a reverse causality problem with the artist's followers and the song's popularity.

Then, we will select the following model:

$$popularity = \beta_0 + \beta_1 danceability + \beta_2 explicit + \beta_3 speechiness + \beta_4 valence + \beta_5 energy + \beta_6 acousticness$$

5. Limitations of your Model

Possible limitations of the model are the fulfillment of the large sample assumptions (IID and unique BLP) and limitations to the structural model, such as omitted variable bias and violations to the one-equation structural model.

5.a Large sample model assumptions

The assumptions for the large sample model are:

1. I.I.D. data (Independent and Identically Distributed)
2. Unique BLP existence

Also, we need to have a large sample size so that we can rely on asymptotics, which is the case for our dataset.

1. I.I.D. data (Independent and Identically Distributed)

The dataset once cleaned of rows containing NA's, contains 140874 songs. Are songs independent of each other such as knowing one song does not provide any information on any other songs?

We can wonder how the songs were selected by Spotify. We can suppose that a logical process has been applied to select songs. We can strongly doubt that songs were randomly selected by Spotify.

From a technical point of view, data can be considered clustered as artists can produce several songs. Songs from similar artists might share musical characteristics making them not independent of each other.

Aggregating data by artist_id reveals that half of the artists (15985 over 33215) have one single song in the dataset, nevertheless, the other half have more than 1 song. Can be pointed out some extreme cases such

as id_artists (3meJIgRw7YleJrmbpbJK6S ->name?) that has 1140 songs on Spotify. Nevertheless, a song might not be identifiable only focusing on technical considerations difficult to perceive by listeners. Each song has its DNA and thus its own Identity. Song's unicity balances the previous arguments and contributes to making data IID enough to perform our study.

2. Unique BLP existence

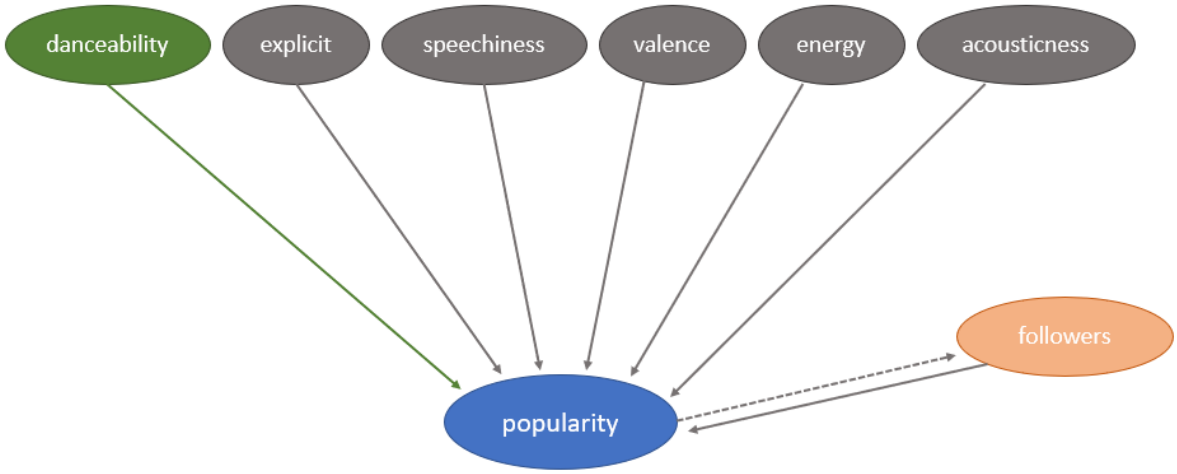
A BLP exists when the covariance between predictors and the covariance between predictors and target variable is finite. If a BLP exists, it is unique when there is no perfect multicolinearity between features. In other words, any feature can be expressed as linear combination of other features.

By looking at how the features are created (based on musical elements, tempo, BPM, decibels, etc), it is difficult to think there could be perfect multicolinearities. There are no fat tails on the distributions of features and any of the models dropped any feature or raised any warning on multicolinearity. Also, although the correlation plot could only show perfect colinearity between two features, none of the pairs present evidence of it.

There is evidence enough to assure this assumption is met.

5b. Structural limitations of your model

The causal diagram considering all the analyzed variables is shown in the figure below:
*Note the selected model (2) does not include the relationship between followers and popularity.



Omitted variable bias

We have identified a potential omitted variable that may have a significant effect on a song's popularity, which is the promoting power that the record company is putting into the song. We consider this feature has a positive effect size on the song's popularity and therefore, the coefficients may be biased, meaning this that the coefficients of the true structural model may differ from the coefficients of our model.

Since we do not have the information we can only discuss about the sign and strength of the bias. For danceability, the Omitted Variable Bias is $\tilde{\beta_1} - \beta_1 = \beta_{prom}\delta_1$. We could expect β_{prom}, which is the coefficient for the omitted variable in the original model, to be positive. δ_1 is the effect size of danceability in a model with the original features regressing the unknown variable, the promoting power, for which we can not form an opinion about direction and size. Therefore, we can only state that the variable promoting power may modify the coefficients.

Reverse causality

We consider that there could be a reverse pathway in the structural model that can violate the one-equation structural model. The artist's number of followers can be a cause for a song's popularity but also creating popular songs can increase the number of followers. For this reason, we selected model (2) over (3).

6. Conclusion