Bayesian Statistics Final Report

Julia Blake & Kevin Lotharp

Due: Thursday, May 9, 2024

**Research Question**

Are there certain attributes that contribute significantly to a song’s popularity? If so, could we use these attributes to predict how popular a song will become? In an era with music streaming platforms dominating the landscape, the availability of vast amounts of data offers unprecedented opportunities for researchers and industry professionals to delve deeper into understanding what makes a song popular. If successful, these models have the potential to benefit music producers and artists seeking to write and record music with the best chance of becoming the next big hit.

**Dataset Description**

We used a dataset titled, “Spotify - All Time Top 2000s Mega Dataset”. This dataset contained audio statistics of the top 2000 tracks on Spotify, released from 1956 to 2019. The various attributes, correlations, and histograms can be seen in Figures 1 and 2. All attribute columns were standardized to aid in the interpretability of results. *Popularity*, the variable of interest, showed a weak correlation with the other attributes in the dataset. However, certain attributes (e.g., *Energy, Loudness,* and *Acousticness*) displayed higher levels of correlation. Although this is a dataset of the top 2000 songs, we can see that the Popularity variable has a distribution covering most of the sample space (0,100). Additionally, we constructed two variables from the data, *Years Since Release* (coded as “yrs\_since\_release”) and *Era*. *Years Since Release* calculated the difference in years between a song’s release date and the year 2019. *Era* was a categorical variable that sorted each song into one of 11 time periods between 1956 and 2019. To account for sparsity in the data, songs recorded before 1970 were grouped together, and all other groups were defined in five-year chunks.

Fig. 1 Fig. 2

A diagram of different types of data

Description automatically generatedA chart of different types of data

Description automatically generated with medium confidence

**Methods**

Our project sought to answer three main questions:

1. What impact do our chosen attributes have on a song’s popularity?
2. Are there latent variables that can further explain a song’s popularity?
3. Does the era in which a song is released have an impact on a song’s popularity?

To answer these questions, we employed a panel of regression models in R. Initially, we established an analytical reference point by performing an OLS regression with the following formula:

We compared the results of this model with those of a Bayesian linear regression to answer Question 1. Next, we used frequentist and Bayesian methods to perform a factor analysis for Question 2. Lastly, we used Bayesian multilevel modeling to answer Question 3.

**Results**

Bayesian Linear Regression

From performing OLS linear regression, we found that the strongest coefficient estimates were *Loudness, Speechiness, Years Since Release*, and *Danceability*. The value was 0.13, suggesting that our model could explain a small but significant portion of variance in values of *Popularity*. The results of this analysis are listed in Table 1.

Table 1  
A screenshot of a computer

Description automatically generated

We performed Bayesian Linear Regression (BLR) twice, first with default priors determined by the *brms* package and then with custom priors determined from our limited knowledge of the data. Our initial run of the model consisted of:

* 4 chains, each with 2000 iterations
* Warmup: 1000
* Thin: 1
* Priors: default

From Table 2 below, it is clear that our estimates were extremely close to the OLS estimates. Further, the Rhat values for all parameters equaled 1.00, and the Effective Sample Size (ESS) was greater than 10% of iterations. We also plotted the trace plots for each chain to ensure convergence (Figure 3).

Table 2

A screenshot of a computer

Description automatically generated

Fig. 3 (Sample of Convergence Checks)

A screenshot of a graph

Description automatically generated

Further analyzing our model, Figure 4 shows the posterior predictive check - comparing the observed outcome of *Popularity* (y) to simulated datasets (yrep) from the posterior predictive distribution. These lines closely tracked with each other.

Fig. 4 Table 3

|  |  |
| --- | --- |
| **Attribute** | **Prior Distribution** |
| *BPM* | Normal(120,5) |
| *Energy* | Normal(50,5) |
| *Danceability* | Normal(50,5) |
| *Loudness\_db* | Normal(-10,2) |
| *Liveness* | Normal(20,5) |
| *Valence* | Normal(50,5) |
| *Duration* | Normal(120,20) |
| *Acousticness* | Normal(20,5) |
| *Speechiness* | Normal(10,5) |

A graph of a line

Description automatically generated

Our BLR with different priors was analyzed in a similar manner. Table 3 lists the priors we used, and Table 4 lists the results. These priors were constructed by using our novice experience in the music space. For example, we knew that 120bpm is very common. However, we used the internet when we had no knowledge for certain variables like *loudness\_db*.

Table 4

A screenshot of a computer

Description automatically generated

We used the Widely Applicable Information Criterion (WAIC) to compare the two BLR models and also to estimate the effective number of parameters needed to adjust for overfitting. WAIC is an extension of the Akaike Information Criterion (AIC) and produces more fully Bayesian results. There was no meaningful change in the estimates between the models, and this model’s trace plots also confirmed convergence. However, the WAIC comparison between models led us to conclude that the initial model had greater predictive power due to its higher Expected Log Predictive Density and smaller WAIC.

Factor Analysis

This step in the project utilized both frequentist and Bayesian methods. Firstly, we used factor analysis to probe the dataset for underlying latent structures. The scree plot of eigenvalues (Figure 5) suggested a potential five-factor solution, with the first five eigenvalues (2.67, 1.59, 1.16, 1.08, 1.03) exceeding one. However, the scree plot did not display a clearly defined elbow point, indicating ambiguity in determining the appropriate number of factors to retain.

Fig. 5  
A graph with numbers and lines

Description automatically generated

The loadings from the factor analysis revealed several patterns. The first factor was primarily related to *Energy, Loudness\_db,* and *Acousticness*, while the second factor appeared to represent *Danceability* and third factor *Popularity*. The fourth factor seemed to reflect a combination of *Loudness\_db, Years Since Released,* and *Valence*. Lastly, the fifth factor related to *Duration* and *Valence* (see Table 5)*.*

However, the model’s low fit statistic (0.0497) indicated that the identified factors may not adequately capture the underlying structure of the data. This suggests that factors beyond the attributes examined in this analysis, such as lyrics, artist fame, or emotional attachment to the songs, may play significant roles in determining song popularity.

Table 5

A screenshot of a computer

Description automatically generated

To further explore the latent structure of the data, we conducted Bayesian Confirmatory Factor Analysis (BCFA) using the *blavaan* package. The model specified five latent factors and we assigned attributes from the dataset to each of the factors:  
***Vivacity*** *– Energy, Loudness\_db, BPM;* ***Wordiness*** *– Acousticness, Speechiness;* ***Happy*** *– Danceability, Valence;* ***Rawness*** *– Liveness, Duration; and* ***Age*** – *Years Since Release, Popularity.* Unfortunately, the results of this analysis were inconclusive due to several warnings, including divergent transitions, high Rhat values, and low effective sample sizes. These issues evidenced problems with model convergence and reliability of parameter estimates. Further refinement of the model or exploration of alternative modeling approaches may be necessary to obtain more robust results.

Multilevel Models

We used Bayesian multilevel modeling from the *brms* package in R to fit three types of multilevel models to investigate differences in song popularity across different eras. From plotting OLS regression lines across the various era clusters (Figure 6), we had reason to believe that *Popularity* might vary across time periods. The values for the 11 regression models ranged from 0.09 to 0.4.

Fig. 6

A chart of different colored dots

Description automatically generated

#### *Unconditional Means Model*

The first model we fitted was the unconditional means model, which estimates the average popularity across different eras without considering any predictor variables. The output revealed an adjusted intraclass correlation coefficient (ICC) of 0.043, indicating that 4.3% of the total variance in popularity is attributed to differences between eras. This model provides a baseline for comparing the predictive performance of more complex models.

#### *Random Intercept Model*

Next, we constructed a random intercept model. In this model, we introduced the predictor variables used in earlier regression models to predict popularity, while allowing for random intercepts for each era. The output showed that the adjusted ICC decreased slightly to 0.014, suggesting that the inclusion of predictor variables explained a portion of the between-era variance in popularity.

#### *Random Intercept and Slope Model*

Finally, we fitted a more complex model by adding random slopes for the predictor variables, allowing the effects of predictors to vary across eras. This model provided insights into whether the relationship between our predictor variables and *Popularity* varied both within eras and across different eras. The associated ICC for this model was 0.16.

**Conclusion**

*What impact do our chosen attributes have on a song’s popularity?*

From our regression analysis, *Loudness\_db, Speechiness*, and *Age*, demonstrate notable influence on a song's popularity. While our models explained only a modest portion of the variance in *Popularity*, they underscored the importance of these attributes in shaping a song’s appeal to listeners.

*Are there latent variables that can further explain a song’s popularity?*

Our exploration into latent variables through factor analysis proved to be largely inconclusive. While we identified several factors related to attributes such as *Energy, Loudness,* and *Danceability*, the low fit statistic suggested that these factors may not fully capture the underlying structure influencing a song’s popularity. This indicates the presence of additional latent variables beyond the attributes examined in our analysis, warranting further investigation into factors such as lyrical content, artist reputation, and emotional resonance with listeners.

*Does the era in which a song is released have an impact on a song’s popularity?*

Our multilevel modeling analysis explored the effects of different periods in modern music history on song popularity. While the effects are small, it appears that there is enough between-group variation across musical eras to conclude that the musical period in which a song is recorded has an effect on the song’s popularity.

**Appendix**

> stancode(brm\_model)  
 // generated with brms 2.21.0  
 functions {  
 }  
 data {  
 int<lower=1> N; // total number of observations  
 vector[N] Y; // response variable  
 int<lower=1> K; // number of population-level effects  
 matrix[N, K] X; // population-level design matrix  
 int<lower=1> Kc; // number of population-level effects after centering  
 int prior\_only; // should the likelihood be ignored?  
 }  
 transformed data {  
 matrix[N, Kc] Xc; // centered version of X without an intercept  
 vector[Kc] means\_X; // column means of X before centering  
 for (i in 2:K) {  
 means\_X[i - 1] = mean(X[, i]);  
 Xc[, i - 1] = X[, i] - means\_X[i - 1];  
 }  
 }  
 parameters {  
 vector[Kc] b; // regression coefficients  
 real Intercept; // temporary intercept for centered predictors  
 real<lower=0> sigma; // dispersion parameter  
 }  
 transformed parameters {  
 real lprior = 0; // prior contributions to the log posterior  
 lprior += student\_t\_lpdf(Intercept | 3, 62, 14.8);  
 lprior += student\_t\_lpdf(sigma | 3, 0, 14.8)  
 - 1 \* student\_t\_lccdf(0 | 3, 0, 14.8);  
 }  
 model {  
 // likelihood including constants  
 if (!prior\_only) {  
 target += normal\_id\_glm\_lpdf(Y | Xc, Intercept, b, sigma);  
 }  
 // priors including constants  
 target += lprior;  
 }  
 generated quantities {  
 // actual population-level intercept  
 real b\_Intercept = Intercept - dot\_product(means\_X, b);  
 }