Likelihood to Surpass 400 million Streaming

Taylor Swift

(Bayesian Statistics Analysis)

Abstract

According to Global Music Report 2023, released by International Federation of the Phonographic Industry, IFPI [1], Taylor Swift is charted No.1 of IFPI global recording artist of the year (2022). Inevitably, with great reputation comes with great anticipation and expectation. As the evaluation of IFPI global recording arist of the year is based on global album-equivalent units earned, which comprises of music downloads, streaming, and physical sales [2], having the knowledge of the likelihood of surpassing the established record, pertaining to the evaluation standard, would prepare her better for winning the next one, and this succinct project is to aim for that goal by focusing soely on the area of streaming.

Data Collection

As digital technology is leading the trend in this AI growing world, in order to obtain the music data in the realm of digital platforms, YouTube, iTunes, Spotify...etc., are the relevant and reliable sources to consider and cover. The total information collected have around twenty variables and nine-hundred or so records, including track names, artist names, track released timeline, streams profiles, music elements and music characteristics.

Exploratory Data Analysis

In integrating and reviewing the dataset, there are many steps to consider in order to ensure data integrity and validity. A key step is to conduct a validity check which encompasses three areas as follows:

Area1: Clarify variable and facilitate modeling (add 5 variables)
In light of Area1, there are 4 variables added to justify the analysis pertaining to Taylor Swift. These variables are artist1, artist2, artist3, and artist4, extended from the variable of artist names. In so doing, each individual artist involved has a clear variable to associate. In addition, these four variables are in the order of importance, meaning artist1 has the most important role in the song, compared to artist2. The same goes to artist 2 to artist3 and so on and so forth. As for artist5, the 5th variable added, it is for the Hierarchical modeling purpose, selecting the top 7 streaming artists in the dataset.

Area2: Verify variable types
Due to the complexity that data export process involved in some programing languages or handling, the data types do not usually line up with their claims. With the collected dataset,

there are three variables found coded as character while they should be numeric. The handling applied here is to keep the original while adding three additional variables to correct the discrepancies.

Area3: Handle Missing information
Every record is equally valuable when the information is accurate. Therefore, making every effort to retrieve the correct information is necessary. Here, variables regarding mode and key are missing for some records and the extra mile is taken to search and verify by the source and my own music training.

Modeling Process

Modeling process is like a patient being ready for doctors to examine in order to understand whether the patient's health condition is satisfactory or some treatments are to take place. Doctors have toolkits to know where and how to evaluate a patient and so as an economist or a statistian.

Postulate a model
There are many ways in defining success in music industry. One of which, if not the most relevant one, is the indicator of streaming. In examining the dataset, it is found that the average stream volume for Taylor Swift is around 390 million on average per song. Therefore, ***the objective is to know the likelihood of going beyond roughly 400 million for her next song*** so that she can either allocate resources to improve other areas or come up with strategies to strengthen her streaming market.

To achieve that, ***Hierarchical model structure is chosen with Poisson distribution as the likelihood and Gamma distribution as the prior.*** Specifically, the hyperparameters, α and ß, are governing the prior distribution for λ (lam) parameters where each λ (lam) represents the expected streams per song of an artist. Thus, α and ß control the likelihood of the means among artists. The mean of the prior distribution, Gamma, will serve as the overall mean of number of streams for all songs. The variance of this gamma distribution determines the variability among artists.

To be more precise in prediction, Top7 artists in the dataset in terms of streams volume are selected while taking the top 10 artists of IFPI report 2023[1] into the consideration as well. The list of Top7 artists include lam[1] Taylor Swift, lam[2] The Weeknd, lam[3] Bad Bunny, lam[4] Drake, lam[5] Harry Styles, lam[6] Ed Sheeran, and lam[7] SZA.

Fit the model with iterations embedded
The package used to conduct the analysis is "rjags" package in R. The first layer of the Hierarchical model is the prior, which is Gamma distribution for estimating λs. The second layer of the Hierarchical model is the likelihood, which is Poisson distribution for estimating streams. Since α and ß are the hyperparameters, leveraging the likelihood of the means among the artists, some trials and erros took place at this stage to find the optimal. To advance and

progress the modeling stage, ***three chains and 5000 iterations*** setup were initiated at this stage as well.

Verify the model

1. **Convergence**

   The first assessment of model checking is convergence. Given the traceplot (Figure1) and the autocorrelation plot (Figure2), all agree, indicating the model is converged.

(From the trace plot, the chain is stationary for each λ, the μ, and the significance as well. And with the autocorrelation, it can be concluded that the convergence is reached.)

Figure1. Trace Plots

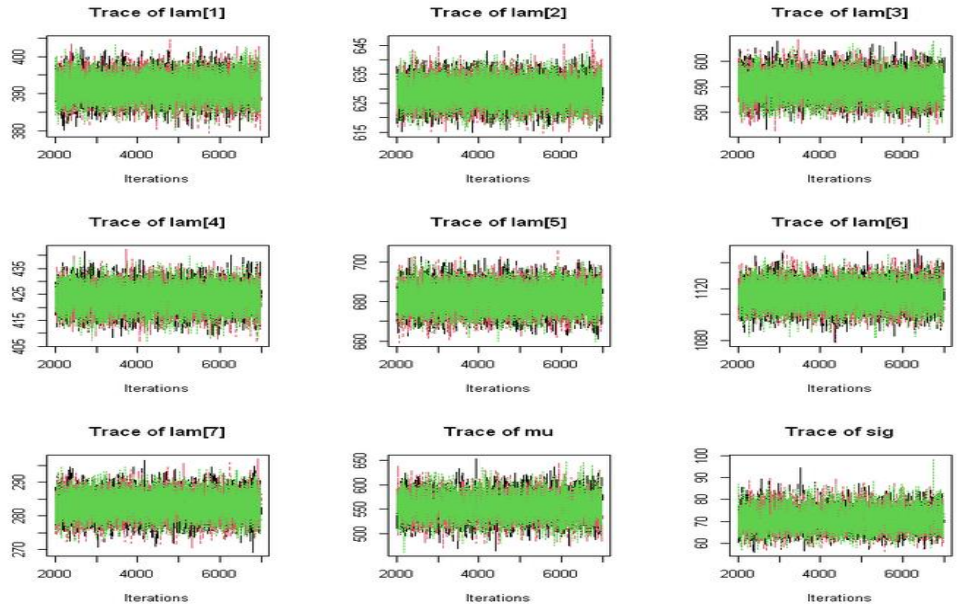
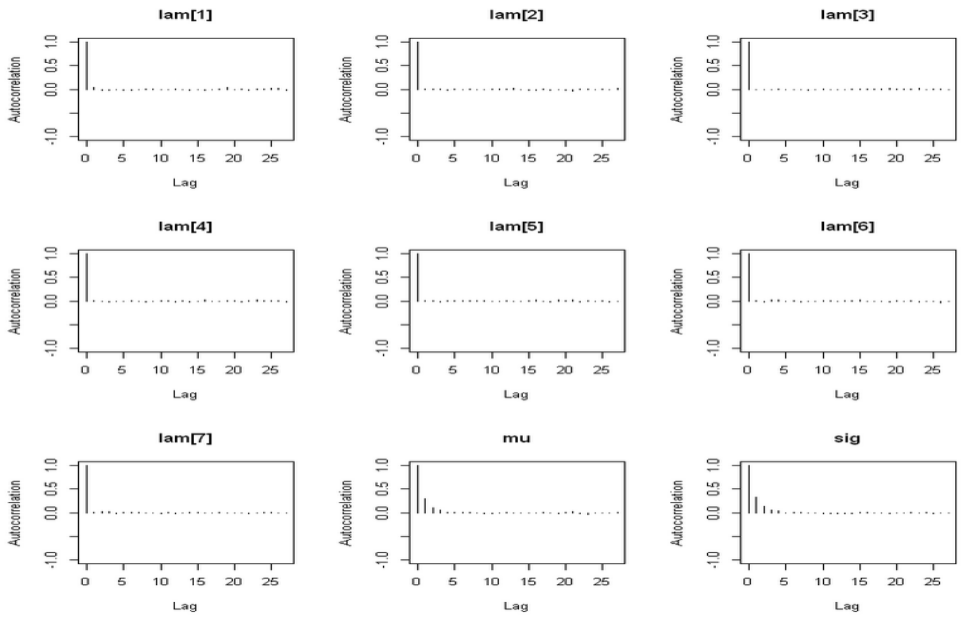


Figure2. Autocorrelation Plots



2. Residuals

In a Hierarchical model, there are two levels of residuls: the mean of the observation and the artist levels

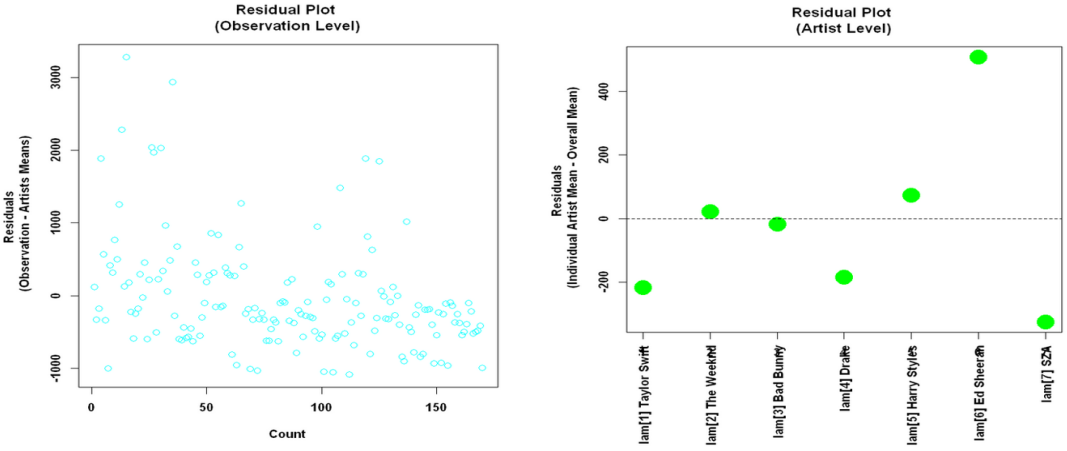
(1) The Observation Mean Level:

In Figure3, it depicts that there is no alarming violation of the model assumption

(2) The Artist Mean Level

In Figure4, it illustrates the similar message in that there is no significant violation of the model assumption

Figure3/4 Residual Plots (Observation Level vs Artist Level)



Conclusion

Putting the newly created model into action, given Taylor Swift's current average streaming status per song, what is the likelihood (posterior probability) that Taylor Swift is to surpass 400 million streaming volumes in her next song? The probability is about 32%, suggesting Taylor Swift has strong fan supports in many areas, but probability not so much in the streaming service as the result shows it is not one of her fortes compared to other top artists. Therefore, if she would like to keep up with her No1 status according to IFPI Global Recording Artist standard, encouraging her fans to listen more of her songs in the streaming platform is one way to get started.

Reference:

[1] **Global Music Report 2023 by International Federation of the Phonographic Industry (IFPI)** [https://www.ifpi.org/wp-content/uploads/2020/03/Global_Music_Report_2023_State_of_the_Industry.pdf]

[2] **Global Recording Artist of The Year from Wikipedia** [https://en.wikipedia.org/wiki/Global_Recording_Artist_of_the_Year]