Beyoncé*

Tam Ly

April 14, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

Table of contents

[*] Code and data are available at: https://github.com/atn-ly/beyonce

1 Introduction

2 Data

2.1 Source and Methodology

The dataset used in this paper was created using data gathered from Spotify (**spotify?**) and Billboard (**Billboard?**). The data was collected for the purpose of this paper to investigate the number of Spotify streams based on weeks spent on the *Billboard Hot 100* for Beyoncé songs. Song titles, album titles, and the number of Spotify streams were collected from Spotify (**spotify?**), while weeks spent on the *Billboard Hot 100* were collected from Billboard (**Billboard?**). There were no similar datasets available that could have been used.

We analyzed the data in R (R Core Team 2023) using the following packages: `arrow` (**arrow?**), `ggplot2` (**ggplot2?**), `here` (**here?**), `janitor` (**janitor?**), `kable` (**kable?**), `knitr` (**knitr?**), `marginaleffects` (**marginaleffects?**), `modelsummary` (**modelsummary?**), `rstanarm` (Goodrich et al. 2022), `scales` (**scales?**), and `tidyverse` (**tidyverse?**).

2.2 Variables

There are 4 variables in this dataset:

1. `song` which represents the song titles,
2. `album` which represents the album titles,
3. `spotify_streams` which represents the number of Spotify streams, and
4. `wks_on_chart` which represents the number of weeks spent on the *Billboard Hot 100*.

Table 1 shows a sample of the dataset with all the variables and the first 10 out of 140 observations.

Table 1: Sample of the cleaned dataset with the first 10 observations

Song	Album	Spotify Streams	Weeks on Chart
Crazy in Love	Dangerously in Love	1141641440	27
Naughty Girl	Dangerously in Love	197200179	22
Baby Boy	Dangerously in Love	309099095	29
Hip Hop Star	Dangerously in Love	6773025	0
Be With You	Dangerously in Love	14676517	0
Me, Myself and I	Dangerously in Love	167639029	24

Song	Album	Spotify Streams	Weeks on Chart
Yes	Dangerously in Love	14646805	0
Signs	Dangerously in Love	15041276	0
Speechless	Dangerously in Love	17300089	0
That's How You Like It	Dangerously in Love	13007089	0

Figure 1 summarizes the data with all 140 observations. We see a positive correlation between the number of Spotify streams and weeks spent on the *Billboard Hot 100*.

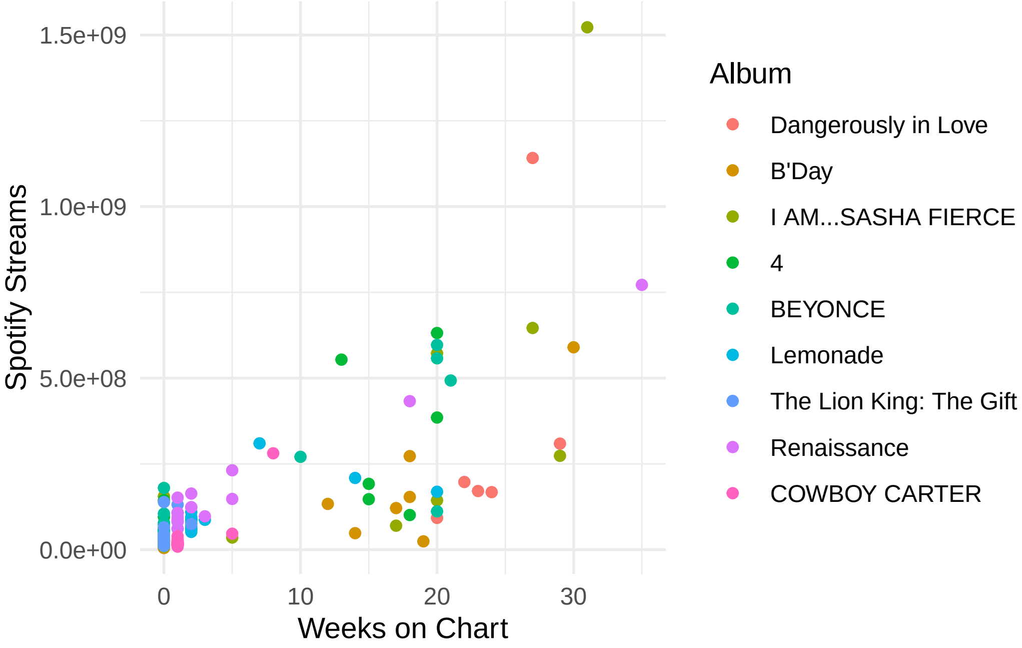


Figure 1: Relationship between the number of Spotify streams and weeks spent on the *Billboard Hot 100* for Beyoncé songs

2.3 Measurement

3 Model

The goal of our modeling strategy is to predict the number of Spotify streams for a Beyoncé song based on the number of weeks it spent on the *Billboard Hot 100*. We used a negative binomial regression model in a Bayesian framework. Negative binomial regression is a type of generalized linear model that is useful for modeling count data.

3.1 Model set-up

The model that we are interested in is:

$$
\begin{aligned}
y_i|\mu_i, r &\sim \text{NegBinom}(\mu_i, r) && (1)\\
\log(\mu_i) &= \alpha + \beta \times \text{Number of weeks}_i && (2)\\
\alpha &\sim \text{Normal}(0, 2.5) && (3)\\
\beta &\sim \text{Normal}(0, 2.5) && (4)\\
& && (5)
\end{aligned}
$$

Where:

- y_i is the outcome variable, representing the number of Spotify streams for song i,
- μ_i is a parameter for the negative binomial distribution, representing the probability of success in a single trial,
- r is a parameter for the negative binomial distribution, representing the number of successes,
- Number of weeks$_i$ is the predictor variable, representing the number of weeks spent on the Billboard Hot 100 for song i,
- α is a parameter, representing the intercept with a specified prior probability distribution that is Normal with a mean of 18 and standard deviation of 2.5, and
- β is a parameter, representing the slope coefficient with a default prior probability distribution that is Normal with a mean of 0 and standard deviation of 2.5.

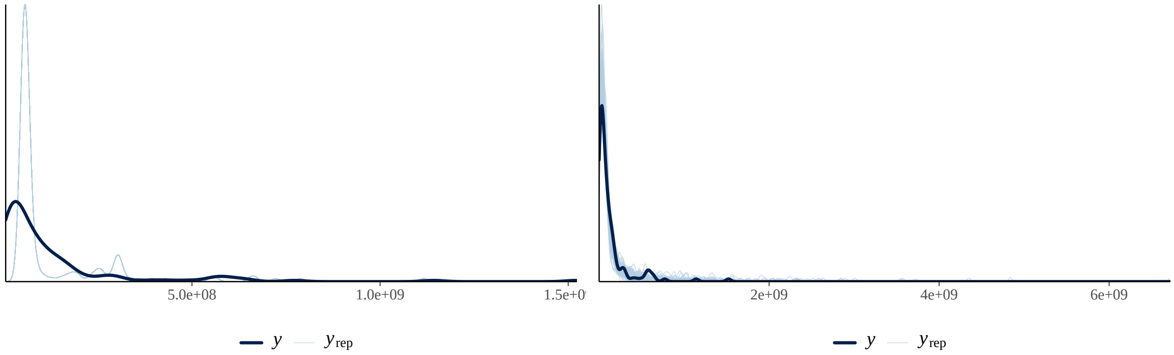
3.2 Model justification

We expect a positive relationship between the number of Spotify streams and weeks spent on the *Billboard Hot 100* based on the positive correlation in the graph that we observed in Section 2.

Negative binomial regression operates under several assumptions. It assumes linearity between the outcome and predictor variables, independence of observations, and no multicollinearity.

We considered Poisson regression as an alternative model since it is also used for count data. However, one of the restrictions with Poisson regression is that it assumes equal mean and variance. Negative binomial regression relaxes this assumption to allow for over-dispersion. We fitted both and compared them using posterior predictive checks in Figure 2. We see that the negative binomial approach does a better job of fitting the data.

We implemented additional model checking and diagnostic issues. Details and graphs can be found in Section A.



(a) Poisson model

(b) Negative binomial model

Figure 2: Comparing posterior prediction checks for Poisson and negative binomial models

4 Results

5 Discussion

5.1 First discussion point

5.2 Second discussion point

5.3 Third discussion point

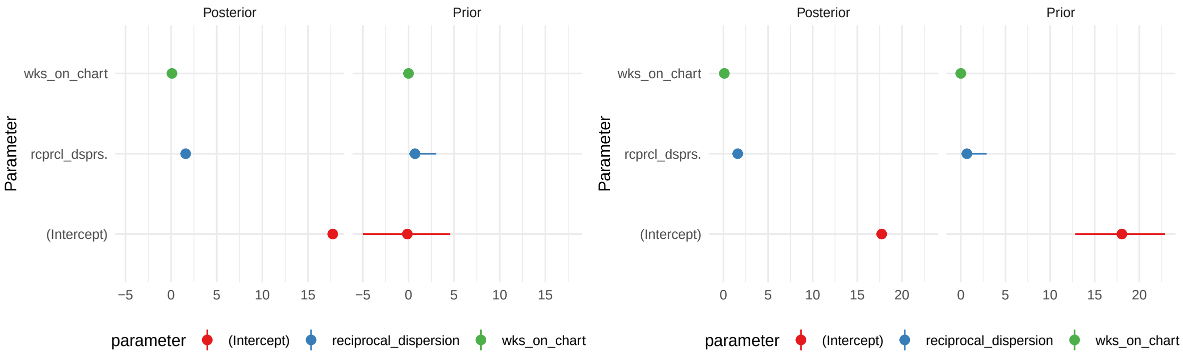
5.4 Weaknesses and next steps

Appendix

A Model details

A.1 Posterior predictive check

We compare the posterior with the prior. Figure 3a suggests that there is an issue with the default prior we specified for α. Our re-specified model in Figure 3b shows that it does a better job of fitting the data.



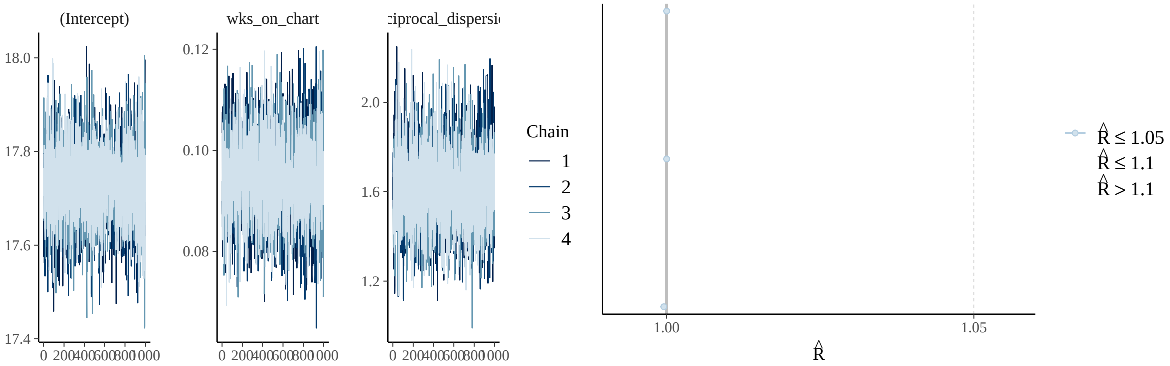
(a) Default priors

(b) Re-specified priors

Figure 3: Comparing the posterior with the prior

A.2 Diagnostics

Figure 4a is a trace plot and it shows that there are no horizontal lines that appear to bounce around and have a nice overlap between the chains. Figure 4b is a Rhat plot and it shows that everything is close to 1 and no more than 1.1. This suggests that there are no problems in both.



(a) Trace plot

(b) Rhat plot

Figure 4: Checking the convergence of the MCMC algorithm

References

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.