Afrobeats榜单持续时间的贝叶斯负二项回归

Cabansag, Ian Jacob 和 Ntegeka, Paul

arXiv:2601.01391v1 [eess.AS] 4 Jan 2026

1 摘要

Afrobeats曲目在流媒体平台上争夺听众的注意力，在这些平台上，排行榜曝光度可以直接转化为收入和文化影响力。本项目使用尼日利亚2024年Spotify Top 200每日排行榜，研究合作与榜单持续时间之间的关系。每首曲目均以其在该年度进入Top 200的天数，以及在尼日利亚市场的年度总播放量进行概括。研究拟合了一个贝叶斯负二项回归模型，其中响应变量为上榜天数，预测变量包括合作指标（独唱曲目与多艺人合作曲目）以及年度总播放量的对数。该模型能够恰当地处理过度离散的计数数据，并允许在控制整体受欢迎程度后，直接解释合作效应。后验推断通过PyMC使用NUTS完成，并报告后验速率比、合作具有促进作用的后验概率以及后验预测检验。结果表明，在以总播放量为条件的情况下，合作曲目的上榜天数略少于可比的独唱曲目，估计速率比约为$\exp(\beta_{\text{collab}}) \approx 0.93$（95%区间为 $[0.88, 0.99]$），后验概率为 $P(\beta_{\text{collab}} > 0 \mid \text{数据}) \approx 0.007$。

2 引言

Afrobeats 在全球的崛起，部分得益于流媒体平台：编辑精选播放列表和算法推荐提高了热门歌曲的曝光度。对于艺人和唱片公司而言，衡量成功的一个重要指标是榜单持续时间：一首曲目在某个国家的 Top 200 排行榜上停留多久。上榜时间越长，意味着听众兴趣持续越久，也意味着获得更多曝光机会。

艺术家和行业观察者普遍认为，合作（例如在一首曲目中邀请另一位艺术家参与）有助于歌曲在更长时间内保持热度。然而，合作也与其他因素相关，例如营销预算和艺术家知名度。简单比较独唱曲目和合作曲目的平均上榜天数，会将这些影响混杂在一起。

本项目旨在通过严谨的统计框架，量化合作对榜单持续时间的影响。研究使用2024年尼日利亚市场的每日Spotify Top–200数据，重点关注排行榜中呈现的Afrobeats及相关流派。核心问题是：

> 在控制总体流行度（年度总播放量）后，合作曲目在尼日利亚 *Top 200* 排行榜上的上榜时间是否比独唱曲目更长？

从方法论角度看，本项目不采用简单的共轭贝叶斯 A/B 检验，而是使用具有负二项似然的贝叶斯广义线性模型（GLM）。这一选择源于响应变量（上榜天数）属于计数数据，且存在严重过度离散。模型使用马尔可夫链蒙特卡罗（MCMC）进行估计，并通过 PyMC 实现 NUTS 算法，从而得到参数和派生量的完整后验分布。

数据部分介绍数据集和预处理。贝叶斯模型部分阐述负二项回归的形式及先验选择。计算部分概述计算设置和收敛诊断。结果部分报告后验汇总和预测检验，讨论部分则对研究发现及其局限性作实质性解读，并作总结。

3 数据

3.1 数据来源和结构

该数据集由尼日利亚市场在2024年1月1日至2024年12月31日期间的 Spotify Top–200 每日快照组成。每行对应某一天的一首轨迹，并包含以下变量等信息：

- date：排行榜条目的日历日期。
- uri：Spotify 曲目的唯一标识符。
- rank：该日期的榜单排名（1–200）。
- 轨迹_名称：曲目标题。
- 艺人_姓名：以逗号分隔的艺人姓名列表。

表1：轨迹层面变量的汇总统计量（尼日利亚 SpotifyTop200，2024年）。

变量	Mean	SD	Min	Max
上榜天数 n_i	54.7	81.9	1.0	365.0
总播放量 S_i	1,908,473	3,532,364	10,132	29,251,917
对数 S_i	12.735	2.100	9.223	17.191
合作率 $P(x_{i,\ \text{collab}} = 1)$		0.456		

- 播放量：该轨迹在该日期于尼日利亚的播放次数。

数据来自 Spotify 官方网站上的尼日利亚地区每日排行榜 [3]:

https://charts.spotify.com/charts/view/regional-ng-daily/latest

原始数据集包含 $N = 72,995$ 条每日曲目级观测数据，代表 $n_{\text{tracks}} = 1,335$ 首不同曲目。日期范围覆盖全年，从2024年– 01–01至2024年–12–31。

3.2 曲目级汇总

回归分析在曲目级进行，而不是日期级。对于每首曲目 i（由唯一的统一资源标识符识别），构建以下汇总变量：

- n_i：2024年该曲目 i 出现在尼日利亚Top–200排行榜上的不同日期数量。这称为上榜天数。
- S_i：该曲目 i 在其上榜的所有日期中累计获得的播放量总数，即 $S_i = \sum_t$ 播放量$_{it}$。
- $x_{i,}$ 合作：用于表示该曲目 i 是否为合作曲目（1）或独唱曲目（0）的指标。

合作指标根据艺 人_名称字段，使用一种简单的文本启发式方法得出。如果艺人字段包含逗号（多位艺人）、与号“&”，或“feat”“ft.”或“ x ”等表示客串艺人的标记，则将曲目归类为合作曲目。通过对随机抽取的部分曲目进行人工检查，发现该启发式方法与检查结果具有较高一致性。

为确保回归的数值稳定性和可解释性，使用经过对数转换的总播放量，

$$x_{i,\text{logStreams}} = \log S_i,$$

作为预测变量。

表 1 汇报了这 $n_{\text{tracks}} = 1{,}335$ 首曲目主要变量的汇总统计量。

探索性图表（ n_i的直方图、按合作状态划分的 n_i 箱线图，以及 n_i与 log S_i的散点图）显示，响应变量呈明显右偏，且相较于泊松分布存在过度离散；此外，合作状态和总播放量都与上榜天数相关。这些特征促使我们在下一节采用负二项回归模型。

4 贝叶斯模型

4.1 似然：负二项回归

令 n_i 表示轨迹 i 在 2024年登上 Top200 排行榜的天数。这些是具有显著过度离散特征的非负计数：$\text{Var}(n_i)$ 远大于 $\mathbb{E}[n_i]$。负二项分布为此类数据提供了灵活的参数族。

对于每条轨迹 i，模型假设

$$n_i \mid \mu_i, \alpha \sim \text{NegBin}(\mu_i, \alpha), \tag{1}$$

其中， $\mu_i > 0$ 是均值， $\alpha > 0$ 是离散参数。在这种参数化方式下，

$$\mathbb{E}[n_i \mid \mu_i, \alpha] = \mu_i, \qquad \text{Var}(n_i \mid \mu_i, \alpha) = \mu_i + \frac{\mu_i^2}{\alpha}.$$

为了将 μ_i 与协变量联系起来，我们使用包含两个预测变量的对数链接函数：合作状态和总播放量对数。设$x_{i1} = x_{i,\text{collab}}$ 和$x_{i2} = x_{i,\text{logStreams}}$。线性预测量为

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}, \tag{2}$$

均值则通过以下关系与之关联：

$$\mu_i = \exp(\eta_i). \tag{3}$$

系数 β_1 表示在总播放量固定时，作为合作曲目所产生的对数乘法效应：

$$\exp(\beta_1) = \frac{\text{Expected days on chart for collab track}}{\text{Expected days on chart for solo track}} \quad \text{at fixed } S_i.$$

同样， β_2 描述上榜天数如何随总播放量变化。

4.2 先验

回归系数和离散参数采用弱信息先验。系数（$\beta_0, \beta_1,$ β_2）采用相互独立的正态先验

均值为零，标准差为 2：

$$\beta_0 \sim \mathcal{N}(0, 2^2), \tag{4}$$

$$\beta_1 \sim \mathcal{N}(0, 2^2), \tag{5}$$

$$\beta_2 \sim \mathcal{N}(0, 2^2). \tag{6}$$

这些先验允许存在相当大的变异（例如，先验上 $\exp(\beta_j)$ 以较高概率大致位于 e^{-4} 和 e^{4} 之间），同时仍能对极端系数进行温和的正则化。

离散参数 α 被限制为正值，并采用半正态先验：

$$\alpha \sim \text{HalfNormal}(2). \tag{7}$$

该先验倾向于采用适中的过度离散值，同时允许数据揭示额外泊松变异的程度。

4.3 后验分布和派生量

令 $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$，并用 $\theta = (\boldsymbol{\beta}, \alpha)$ 表示所有参数。给定数据 $\mathbf{n} = (n_1, \ldots, n_{n_{\text{tracks}}})$ 和协变量 $\mathbf{x}$，后验密度为

$$p(\theta \mid \mathbf{n}, \mathbf{x}) \propto \left[\prod_{i=1}^{n_{\text{tracks}}} p\big(n_i \mid \mu_i(\theta), \alpha\big) \right] p(\boldsymbol{\beta}) p(\alpha). \tag{8}$$

有两个派生量尤其值得关注：

（1）合作比率比

$$R_{\text{collab}} = \exp(\beta_1),$$

在固定 s_i 的情况下，该指标衡量合作曲目预计会在排行榜上停留的天数是独唱曲目的多少倍（或少多少倍）。

（2）合作有帮助的后验概率：

$$P(\beta_1 > 0 \mid \mathbf{n}, \mathbf{x}),$$

该指标可作为支持正向合作效应的贝叶斯度量。

n_i 的后验预测分布也会被计算出来，并与观测分布进行比较，作为模型检验。

5 计算

后验推断使用 Python 中的 PyMC 库（4.0.1 版）完成，并结合 ArviZ（0.21.0 版）进行诊断和可视化。模型采用负二项似然以及均值–离散度参数化和对数链接函数。采样使用无转向采样器（NUTS）进行。

运行两条链，每条链进行 2，000 次预热迭代和 2，000 次预热后抽样，共得到 4，000 个后验样本。目标接受率设为 0。9，以促使步长较小并确保采样稳定。

使用 ArviZ 报告的 $\hat{R}$ 诊断指标和有效样本量评估收敛情况。所有参数的 $\hat{R}$ 值都非常接近 1。00，有效样本量则达到数千，表明混合效果良好。β_1、β_2和 α 的轨迹图均未显示非平稳性或多峰性的迹象。

使用 PyMC 后验预测采样器（pm.sample_post erior_predictive）生成后验预测样本，为每条曲目生成 n_i^{rep} 重复抽样。

同时还考虑了一个协变量完全相同的泊松回归模型。后验预测诊断显示，泊松模型低估了方差，且无法再现持久性很强的曲目所呈现的厚右尾；相比之下，负二项模型的拟合效果显著更好。这支持了上文所述的贝叶斯建模框架。

6 结果

6.1 后验汇总

表 2 报告了由 ArviZ 得出的回归系数和离散参数的后验均值及 94% 最高密度区间（HDI）。

表 2：模型参数的后验汇总（均值和 94% 最高密度区间）。

参数	Mean	标准差	3% 最高密度区间	97% 最高密度区间
β_0 (截距)	−8.126	0.122	−8.354	−7.904
β_1 (合作)	−0.074	0.030	−0.129	−0.014
β_2 (流媒体播放量对数)	0.859	0.009	0.843	0.876
α (离散度)	5.037	0.237	4.621	5.503

总播放量对数的系数 β_2 显著为正，这表明总播放量较高的曲目往往会在排行榜上停留更久，这符合预期。我们主要关注的量是 β_1，即合作的影响。β_1 的后验分布中心位于 -0.074，94% 的最高密度区间为 -0.129 至 -0.014，完全低于零。

根据 β_1 的后验分布，合作率比

$$R_{\text{collab}} = \exp(\beta_1)$$

后验分布中位数约为 $R_{\text{collab}} \approx 0.929$，95% 可信区间为 $[0.879, 0.986]$。效应为正的后验概率极低：

$$P(\beta_1 > 0 \mid \mathbf{n}, \mathbf{x}) \approx 0.007.$$

也就是说，在控制年度总播放量后，合作曲目预计比可比的独唱曲目在 Top–200 榜单上的上榜天数略少，而且有强有力的贝叶斯证据表明合作效应并非正向。

独唱与合作曲目的预期上榜天数

为了提供更直观的概述，我们来看一首处于对数总播放量中位数水平的“典型”曲目，即 $\tilde{x}_{\text{logStreams}} \approx$12。71。对于每一次后验抽样（$\beta_0, \beta_1, \beta_2$），在这一参考值下，独唱曲目和合作曲目的预期上榜天数分别为

$$\mu^{(s)}_{\text{solo}} = \exp\left(\beta_0^{(s)} + \beta_2^{(s)} \tilde{x}_{\text{logStreams}}\right),$$
$$\mu^{(s)}_{\text{collab}} = \exp\left(\beta_0^{(s)} + \beta_1^{(s)} + \beta_2^{(s)} \tilde{x}_{\text{logStreams}}\right).$$

根据后验均值和标准差采用对数正态近似，代表性独唱曲目的预期上榜天数约为 $\mu_{\text{solo}} \approx 16.3$（95% 区间约为 $[11.8, 22.7]$），而代表性合作曲目的相应数值为 $\mu_{\text{collab}} \approx 15.2$（95% 区间约为 $[10.9, 21.2]$）。这些数值进一步印证了这样的解读：在总体流行度保持不变时，合作与榜单持续时间小幅但不可忽视的缩短相关。

后验预测检验

后验预测检验将 n_i 的经验分布与从后验预测分布中抽取的复制计数 n_i^{rep} 进行比较。图1展示了一项具有代表性的检验，其中观测值 n_i 的直方图（或核密度估计）与后验预测复制值的分布叠加显示。

负二项模型相当准确地再现了上榜天数的偏斜分布，包括那些持续上榜时间极长的曲目所形成的长右尾，不过它略微低估了持续上榜时间最长的热门歌曲的频率（例如，上榜天数超过

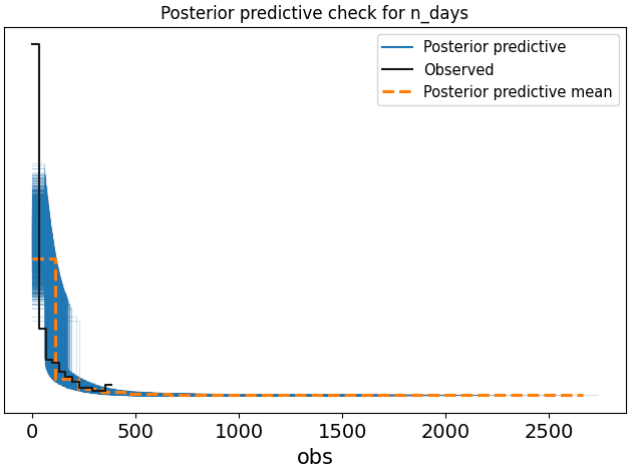

图1：上榜天数的后验预测检验。实线表示 n_i的经验密度，而阴影区域显示后验预测复制值的分布 n_i^{rep}。

300 天的曲目）。按合作状态分层的后验预测图（未显示）表明，该模型也捕捉到了独唱曲目与合作曲目之间的分布差异。

讨论

本项目应用贝叶斯负二项回归模型，研究2024年尼日利亚 Spotify Top–200 中 Afrobeats 曲目的榜单持续时间。通过将每日榜单数据汇总为曲目级计数，并将上榜天数建模为合作状态和总播放量对数的函数，该分析超越了简单的 A/B比较，在明确控制总体流行度的同时量化了合作效应。

最主要的实质性发现是，在固定年度总播放量的情况下，合作与上榜天数方面的轻微劣势相关。后验速率比估计值约为0.93，这意味着在总播放量相同的情况下，合作曲目预计比独唱曲目少上榜约7%的天数。合作具有有益效应的后验概率仅约为0.7%，表明贝叶斯证据强烈反对“合作本质上会延长上榜时长”这一观点。

在方法论上，本项目展示了完整的贝叶斯广义线性模型（GLM）工作流程：构建恰当的似然函数、设定弱信息先验，

使用 PyMC 进行马尔可夫链蒙特卡罗拟合，解读后验效应大小和概率，并进行后验预测模型检验。尤其是，采用负二项似然对于处理上榜天数的过度离散至关重要；泊松模型会低估变异性，预测拟合效果也明显较差。

需要指出几项局限。首先，合作指标基于对艺人姓名的启发式解析，因此可能会误判某些曲目。其次，模型以年度总播放量为条件，而年度总播放量本身是播放列表收录、营销和推荐算法等因素共同作用的复杂结果。更进一步的分层模型可以尝试联合建模播放量与上榜天数，或引入艺人层面的随机效应。第三，分析仅限于单个年份和单一市场；将研究扩展到多个年份或国家，或许能揭示合作效应是否会随时间和地域而变化。

尽管存在这些局限，该模型清楚地展示了如何将贝叶斯回归应用于实际流媒体数据，以回答一个具体的实质性问题。未来的研究可以在此框架基础上，加入流派信息、唱片公司指标或艺人网络特征，从而更好地理解合作影响榜单表现的机制。

参考文献

[1] Andrew Gelman、John B. Carlin、Hal S. Stern、David B. Dunson、Aki Vehtari 和 Donald B. Rubin。2013年。贝叶斯数据分析 （第3版）。CRC出版社。[2] Osvaldo Martin。2021年。使用 *Python*进行贝叶斯分析 （第2版）。Packt Publishing。[3] Spotify。2024年。Spotify排行榜：尼日利亚每日前200名。https://charts.spotify.com/charts/view/regional-ng-daily/latest 访问日期：2026年1月4日。

一段 PyMC 模型代码

以下 Python 代码片段展示了分析中使用的核心 PyMC 模型。

```
使用 pm.Model() 作为 nb_模型: beta0 =pm.Normal
```