Bayesian Statistics Final Report (4-5pgs)

Julia Blake & Kevin Lotharp

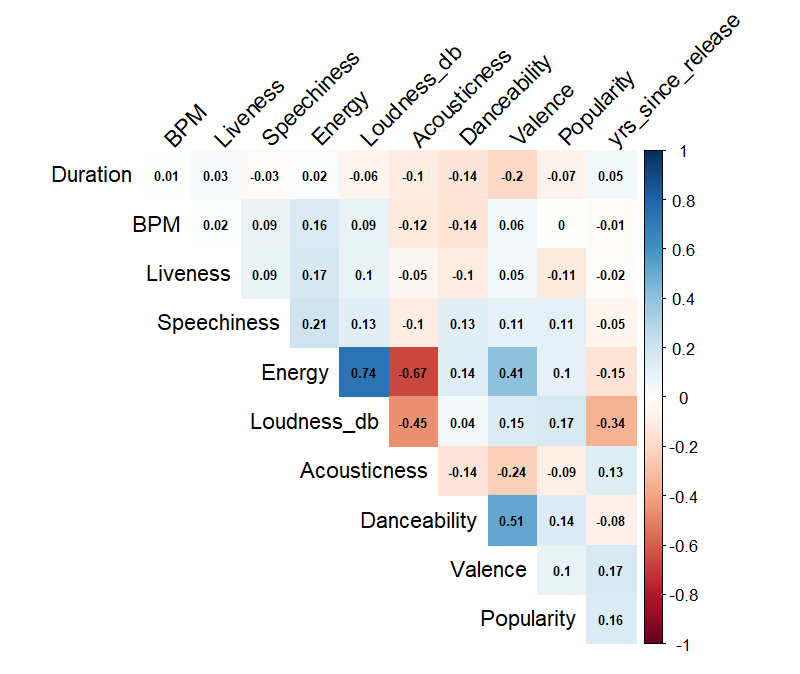
Due: Thursday May 9, 2024

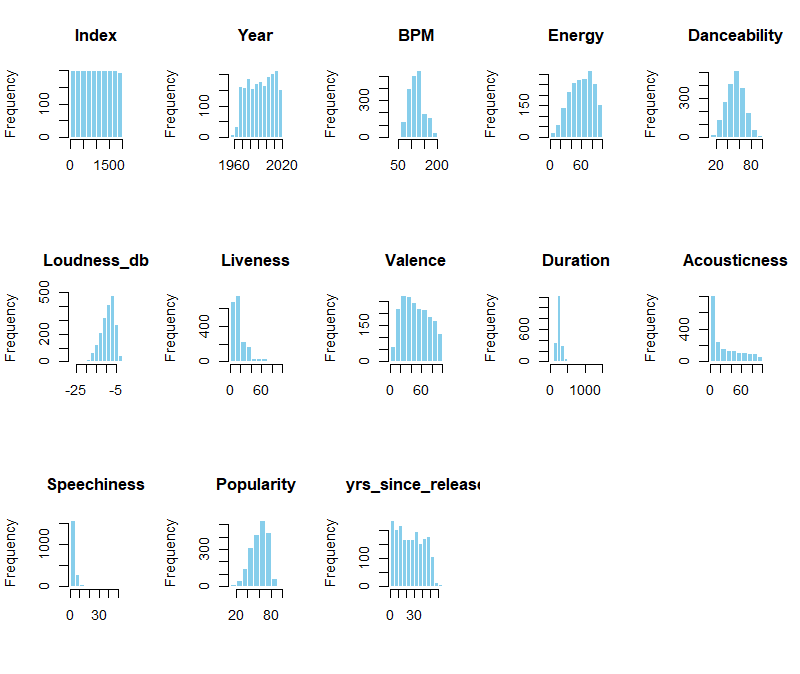
**Research Question**

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 contains audio statistics of the top 2000 tracks on Spotify, released from 1956 to 2019. The various attributes correlations and histograms can be seen below. Notably, Energy, Loudness and Acousticness are the most correlated with each other. Moreover, Popularity, the variable of interest does not appear strongly correlated with any attributes at face level.





Additionally, we can see below that Popularity seems to have declined on average over time. Meaning, that older songs are more popular. However, we can see a couple explanations for this:

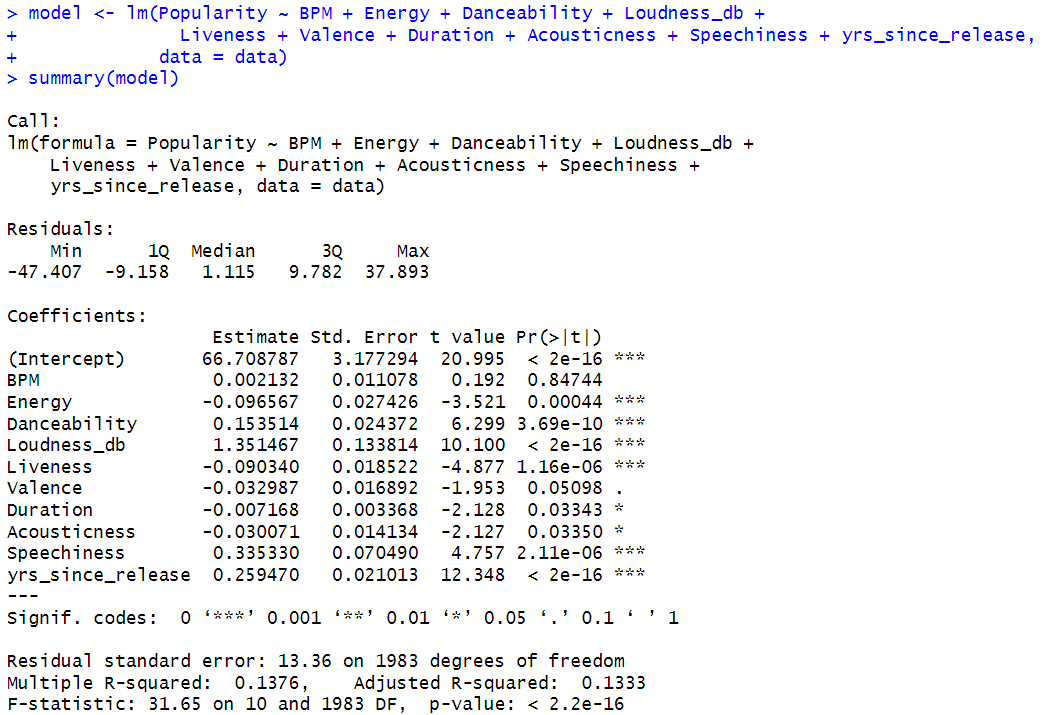
– Maybe older songs have had more time to become popular, allowing for more “plays/streams” or for songs to have resurgences if not popular at release, or,

– Sample size could be at play. It appears there are much fewer older songs in this dataset as compared to newer songs. This has the potential to skew relative impacts.

**Analysis**

First, lets run our traditional frequentist OLS model to have a good baseline for comparison for our Bayesian Linear Regression Model. We used the below equation as our model.

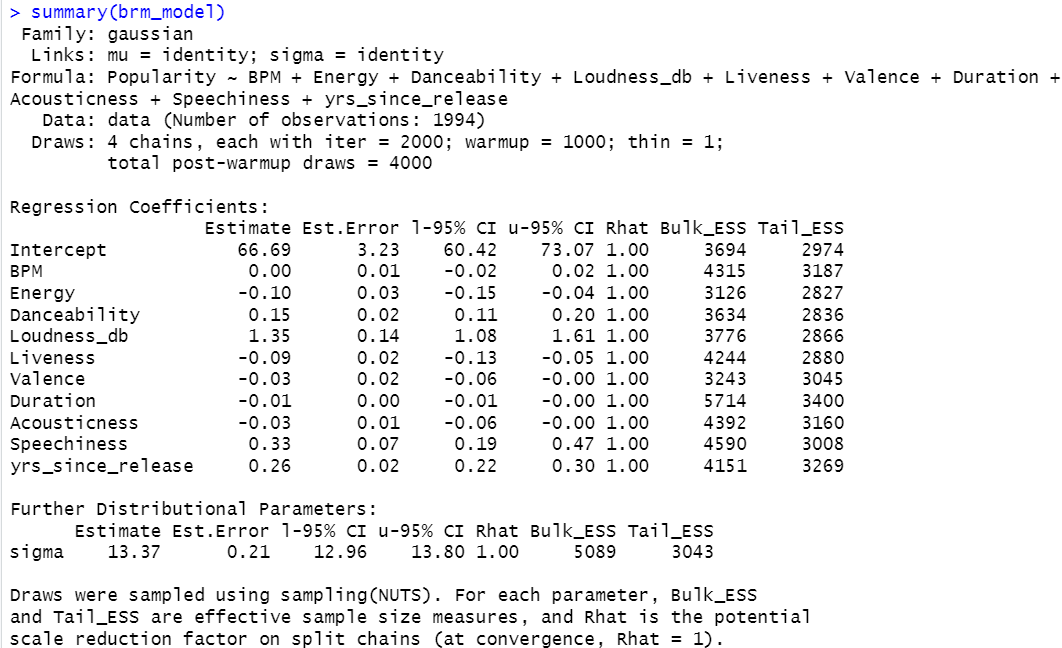
Our coefficient results are as follows. We can see that the strongest coefficient estimates are the intercept, loudness, speechiness, yrs\_since\_release, and danceability. The R-squared value is 0.13.



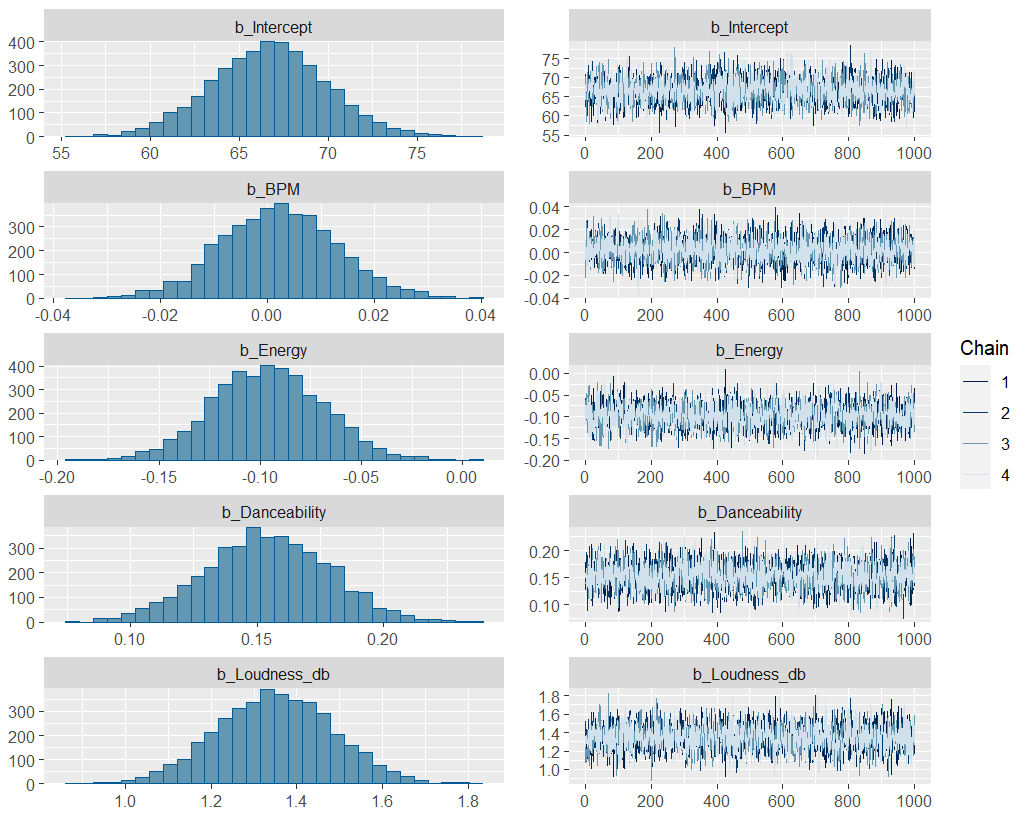
Now we will run our Bayesian Regression Model with the following set up:

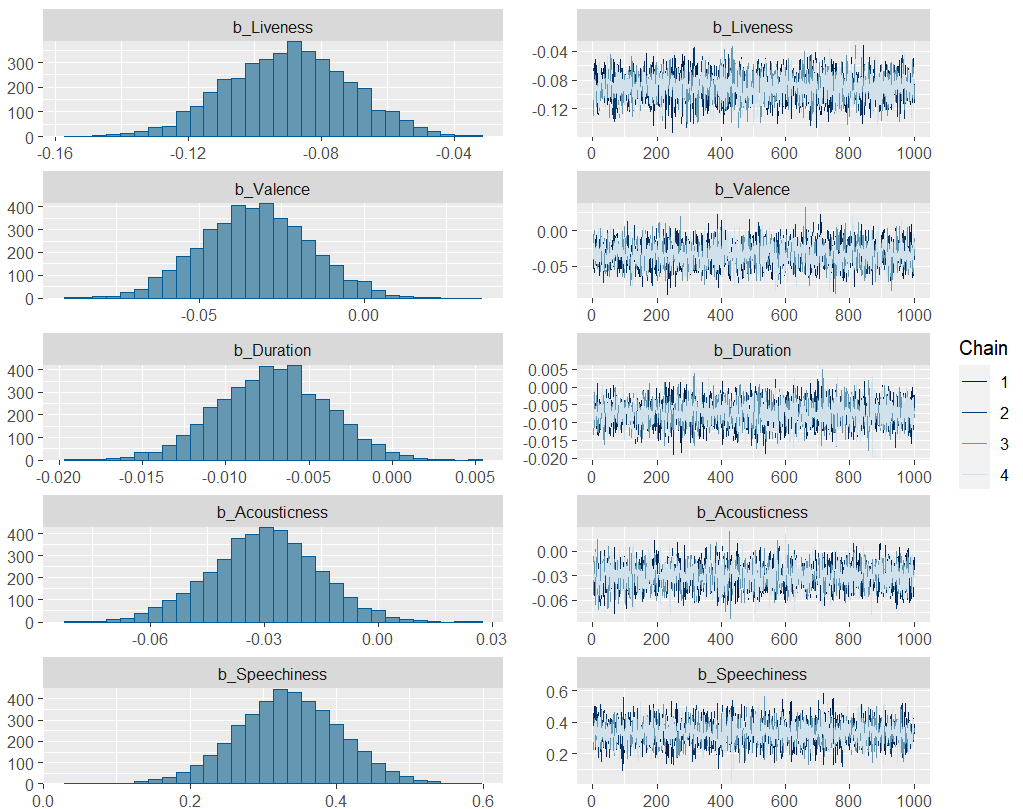
* 4 chains, each with 2000 iterations
* Warmup: 1000
* Thin: 1
* Priors: default

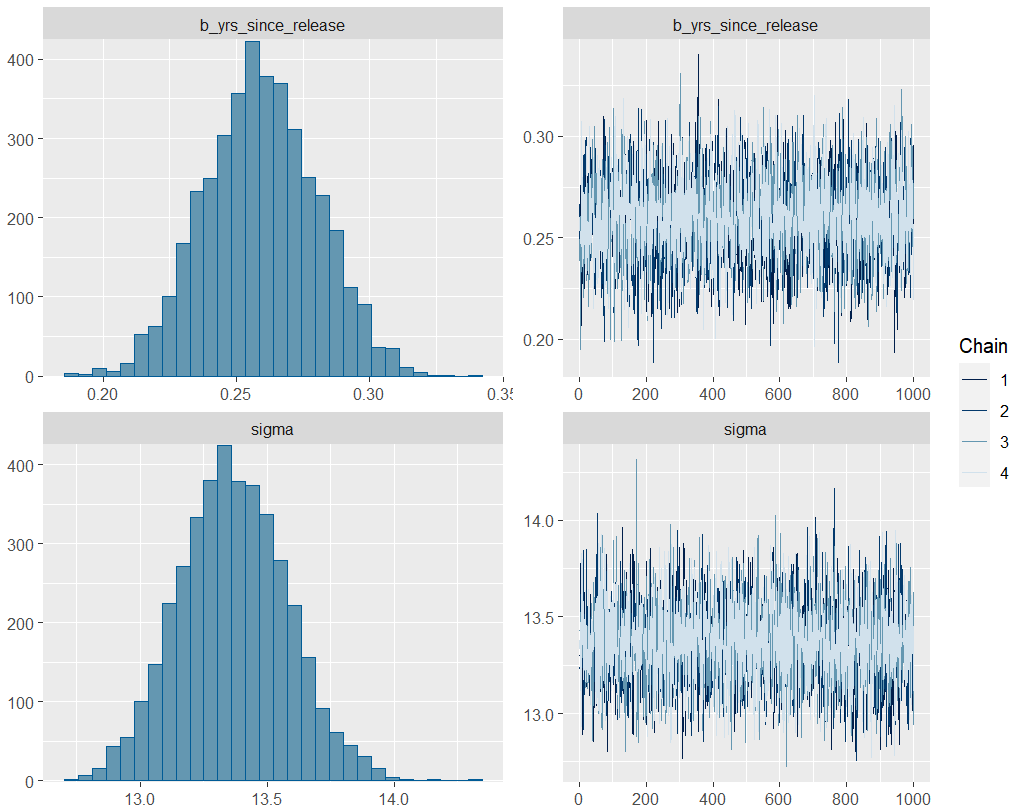
We can see below that our estimates are extremely close to the OLS estimates. Further, we can see that our Rhat equals 0 for all estimates and Effective Sample Size (ESS) is greater than 10% of iterations.



For further comfort that our chains converge, we plotted the trace plots and can see that the chains converged for each variable.

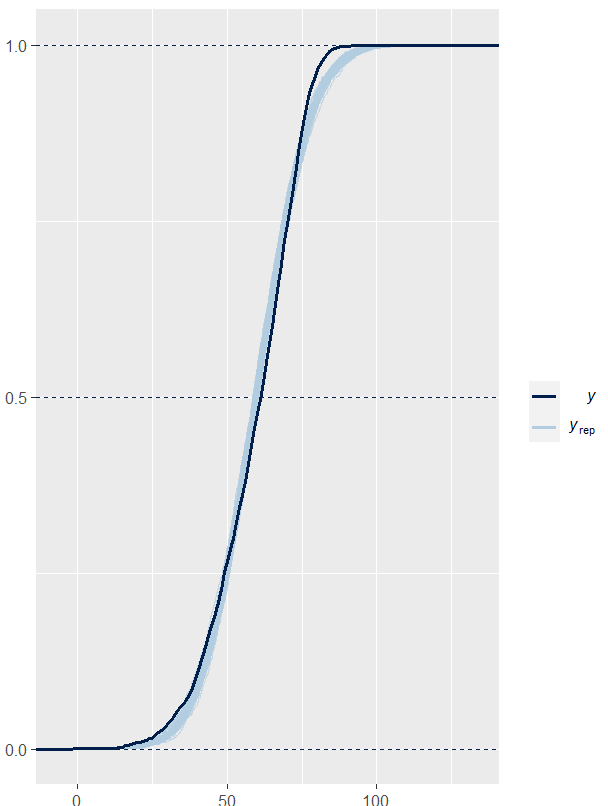




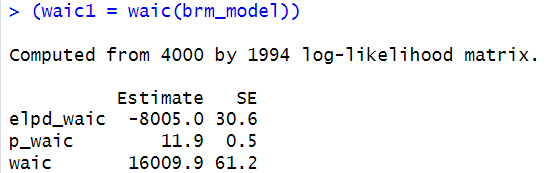


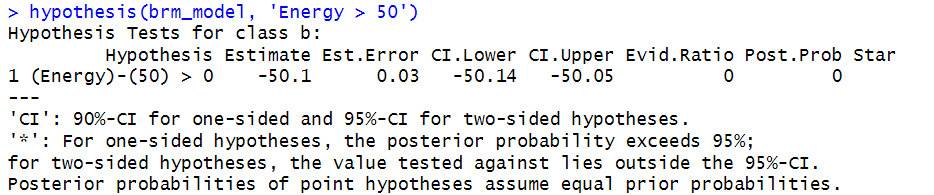
Let’s do some further analysis into our model.

The below plot shows us the posterior predictive check - comparing the observed outcome variable y to simulated datasets yrep from the posterior predictive distribution. We can see it follows the line closely until the turns.

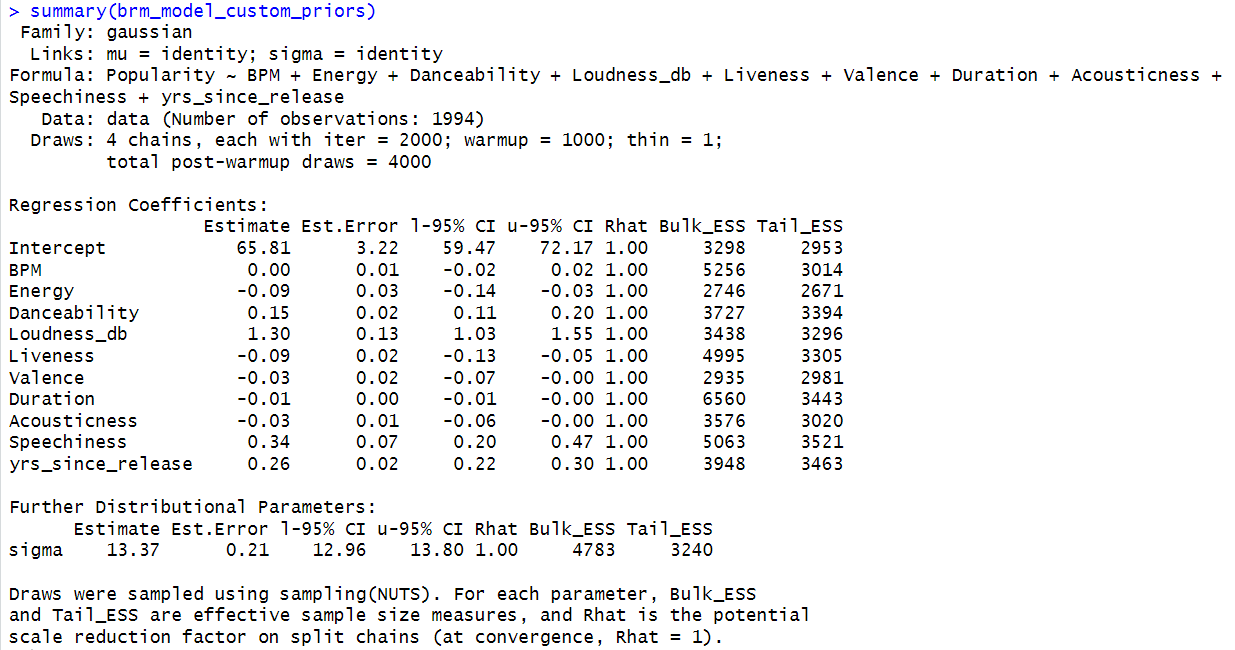


Next, lets look at the Widely Applicable Information Criterion (WAIC), WAIC is an extension of the Akaike Information Criterion (AIC) that is more fully Bayesian . The WAIC estimates the effective number of parameters to adjust for overfitting.

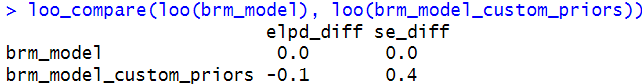




Now, lets try our model but with different prior distributions on the predictors.



Comparing the WAIC of this model with the original bayesian regression model with default priors we can see that the initial model is better at predicting; the Expected Log Predictive Density (ELPD) is higher



**Conclusion**

As a result of our regression analysis, we can see that the predictors Loudness, Speechiness and Age of the release have the greatest impact on a song’s popularity, albeit small coefficients. Thus, we can conclude that audio attributes do not necessarily affect how popular a song becomes, rather it is more likely that external factors/ideas such as the emotion behind lyrics, media, or popularity of the particular artists would have a greater impact.

**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);  
 }