Random Effect Priors in Binary Longitudinal Models

Juna, Ucheoma, Nathan

Research Problem

Sufficient quantity of sleep is very important for overall health, and yet many people find it difficult to sleep enough despite their better efforts. Scientific literature published by the (CDC) suggests at least seven (7) hours is necessary for adults to remain healthy.

Data

- Generated by Fitbit device’s of 36 individuals
- Measurements taken over span of four months

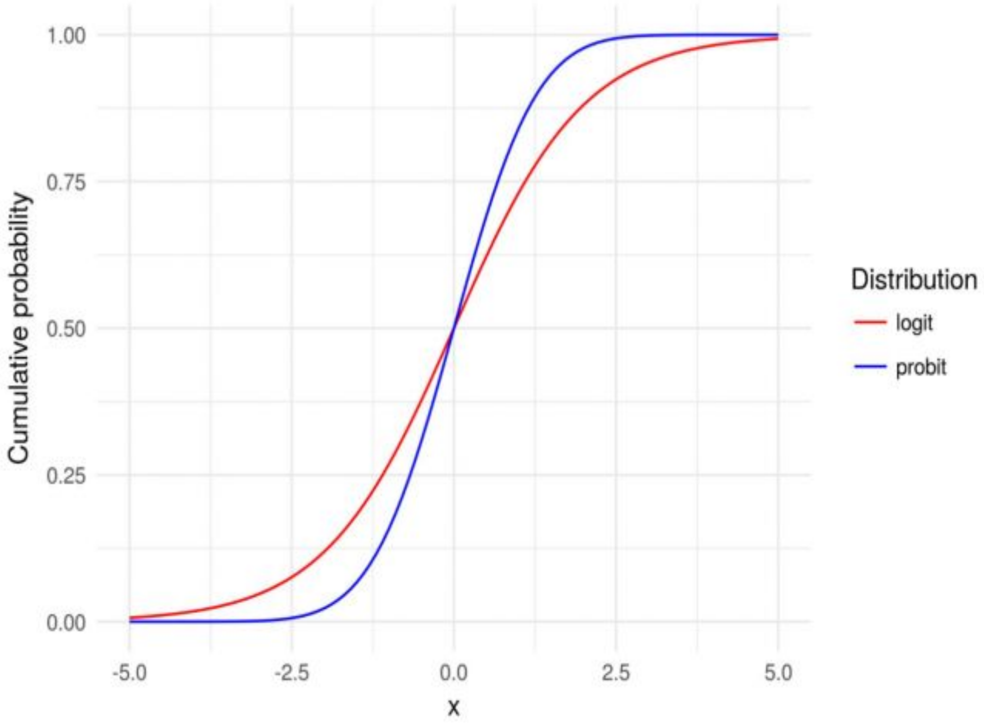
sleep_duration_hr	steps	nightly_temperature	stress_score	daily_temperature_variation	lightly_active_minutes	moderately_active_minutes	very_active_minutes
8.683333333333334	8833.0	34.1376871401	78.0	-1.7883249002998594	149.0	24.0	33.0
9.133333333333333	9727.0	33.7945437956	80.0	-2.4627093726523013	132.0	25.0	31.0
9.333333333333334	8253.0	34.6110107335	84.0	-2.385801017132203	112.0	27.0	31.0

Motivation

- Using the bayesian approach we incorporate what we believe about the data to inform our model
- Hierarchical model is applied to capture the impact random effects have on the response variable

Model Specification

- Uses Generalized Linear Regression with a Probit Link Function
- Introduces latent response variables to account for the binary aspect of the output
- Uses a flat prior for beta, for simplicity
- Applies a hierarchical approach to account for random effects



Methods

Represent binary random effect longitudinal data using a latent variable that is equivalent to using the probit link function in the GLM framework.

Latent Variable Model

$$y_i = \begin{cases} 0 & \text{if } z_i \leq 0 \\ 1 & \text{if } z_i > 0 \end{cases}$$

$$z_i \sim \mathcal{N}(x_i'\beta, 1)$$

Random Intercept Model

$$y_{ij} = \beta_0 + x_{ij}\beta_1 + u_i + \epsilon_{ij}$$

$$u_i \sim \mathcal{N}(0, \sigma_u^2)$$

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma_e^2)$$

Model Development: Likelihood

Consider

- y_i vector of outcomes for subject i
- u_i vector of subject-specific intercept for subject i
- X_i matrix of covariates for subject i
- β vector of fixed-effect coefficients

$$y_i \sim Bern(\theta_i) \text{ where } \theta_i = \mathbb{P}(y_i = 1 | X_i, \beta, u_i) = \Phi(X_i\beta + u_i)$$

$$y_i = \begin{cases} 0 & \text{if } z_i \leq 0 \\ 1 & \text{if } z_i > 0 \end{cases} \quad \text{and} \quad z_i \sim \mathcal{N}(X_i\beta + u_i, 1)$$

Model Development: Likelihood

$$\mathcal{L}(Y|Z, X, \beta, U) \propto \prod_{i=1}^{n} \{1(z_i > 0)1(y_i = 1) + 1(z_i \leq 0)1(y_i = 0)\}$$
$$\times \exp\left[-\frac{1}{2}(z_i - (X_i\beta + u_i))^2\right] \qquad (1)$$

Model Development: Posterior

Model relies on the following assumptions

- Flat prior on β → $\pi(\beta) \propto 1$
- Normal prior on U → $\pi(\tau_u^2) \propto (\tau_u^2)^{-1}$
- Non-informative prior τ^2_u → $U \sim MVN(0, (1/\tau_u^2)I_{n\times n})$
- Designs matrix D

$$p(\beta, Z, U, \tau_u^2 | Y, X) \propto \mathcal{L}(Y|Z, X, \beta, U)\pi(\beta)\pi(U|\tau_u^2)\pi(\tau_u^2)$$

$$= (\tau_u^2)^{n/2-1} \exp\left[-\frac{1}{2}\|Z - (X\beta + DU)\|_2^2 - \frac{\tau_u^2}{2}\|U\|_2^2\right]$$

$$\left[\prod_{i=1}^{n}\{1(z_i > 0)1(y_i = 1) + 1(z_i \leq 0)1(y_i = 0)\}\right]$$

Random Intercept: Conditional Posteriors

$$
\begin{aligned}
p(U|\beta, \tau_u^2, Z, Y, X) &\propto \exp\left[-\frac{1}{2}\|Z - (X\beta + DU)\|_2^2 - \frac{\tau_u^2}{2}\|U\|_2^2\right] \\
&= \exp\left[-\frac{1}{2}\|(Z - X\beta) - DU)\|_2^2 - \frac{\tau_u^2}{2}\|U\|_2^2\right] \\
&= \exp\left[-\frac{1}{2}\left((Z - X\beta) - DU)'((Z - X\beta) - DU) + \tau_u^2 U'U\right)\right] \\
&= \exp\left[-\frac{1}{2}\left((Z - X\beta)'(Z - X\beta) - 2(DU)'(Z - X\beta) + (DU)'DU + \tau_u^2 U'U\right)\right] \\
&\propto \exp\left[-\frac{1}{2}\left((DU)'DU + \tau_u^2 U'U - 2(DU)'(Z - X\beta)\right)\right] \\
&= \exp\left[-\frac{1}{2}\left(U'D'DU + \tau_u^2 U'U - 2U'D'(Z - X\beta)\right)\right] \\
&= \exp\left[-\frac{1}{2}\left(U'D'DU + \tau_u^2 U'U - 2U'D'(Z - X\beta)\right)\right] \\
&= \exp\left[-\frac{1}{2}\left(U'(D'D + \tau_u^2 \boldsymbol{I}_{n\times n})U - 2U'D'(Z - X\beta)\right)\right] \\
&= \exp\left[-\frac{1}{2}\left(U'(D'D + \tau_u^2 \boldsymbol{I}_{n\times n})U - 2U'(D'D + \tau_u^2 \boldsymbol{I}_{n\times n})\right.\right. \\
&\qquad \left.\left. \times (D'D + \tau_u^2 \boldsymbol{I}_{n\times n})^{-1} D'(Z - X\beta)\right)\right]
\end{aligned}
$$

Gibbs Sampler

Our Gibbs Sampler iterates through draws and updates

- $\beta^{(b)}|U^{(b-1)},(\tau_u^2)^{(b-1)},Z^{(b-1)},Y,X \sim MVN\left[(X'X)^{-1}X'(Z^{(b-1)}-DU^{(b-1)}),(X'X)^{-1}\right]$
- $$U^{(b)}|\beta^{(b-1)},(\tau_u^2)^{(b-1)},Z^{(b-1)},Y,X \sim MVN\Big[(D'D+(\tau_u^2)^{(b-1)}\boldsymbol{I}_{n\times n})^{-1}D'(Z^{(b-1)}-X\beta^{(b-1)}), (D'D+(\tau_u^2)^{(b-1)}\boldsymbol{I}_{n\times n})^{-1}\Big]$$
- $(\tau_u^2)^{(b)}|\beta^{(b-1)},U^{(b-1)},Z^{(b-1)},Y,X \sim Gamma\left(\frac{n}{2},\frac{1}{2}\|U^{(b-1)}\|_2^2\right)$
- For all i st $y_i=0$, update $z_i^{(b)}|\beta^{(b-1)},u_i^{(b-1)},y_i=0,X_i \sim \mathcal{N}\left[X_i\beta^{(b-1)}+u_i^{(b-1)},1\right]1(z_i\leq 0)$
- For all i st $y_i=1$, update $z_i^{(b)}|\beta^{(b-1)},u_i^{(b-1)},y_i=1,X_i \sim \mathcal{N}\left[X_i\beta^{(b-1)}+u_i^{(b-1)},1\right]1(z_i> 0)$

Assessing Model Convergence

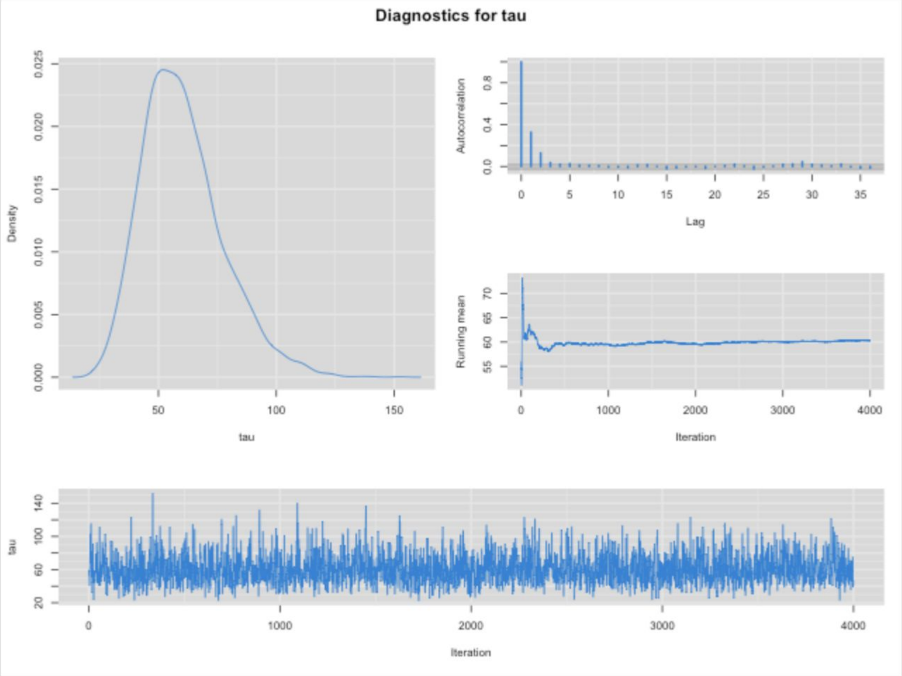
To determine convergence of our Gibbs sampler we considered:

- Visual Inspection of MCMC Plots outputs (trace, trailing mean, autocorrelation plots)
- Geweke Diagnostic

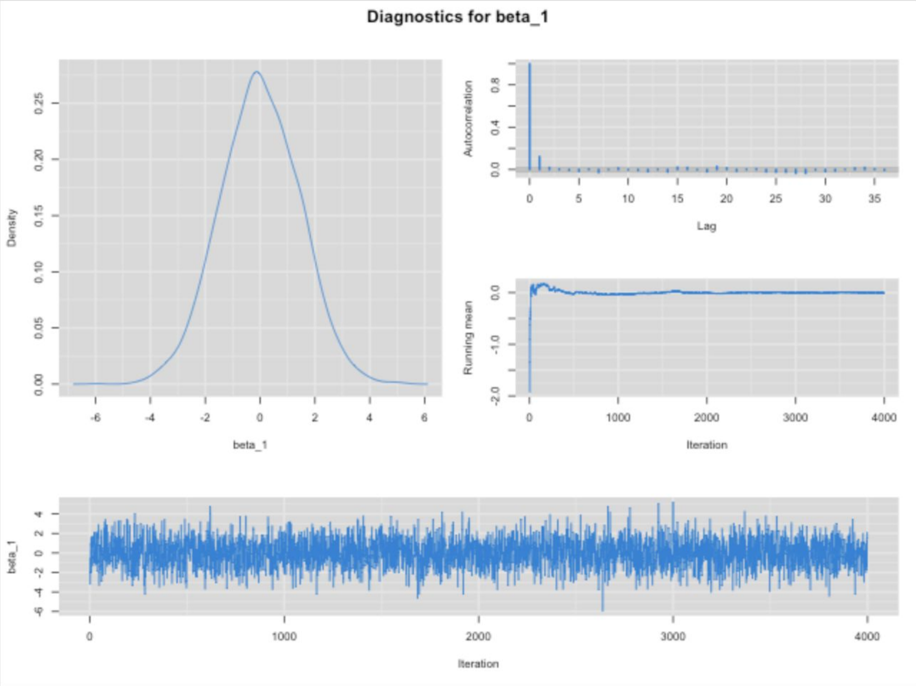
We did not consider an acceptance rate because the algorithm did not include an M-H step

Convergence Plots

Plots for tau



Plots for beta

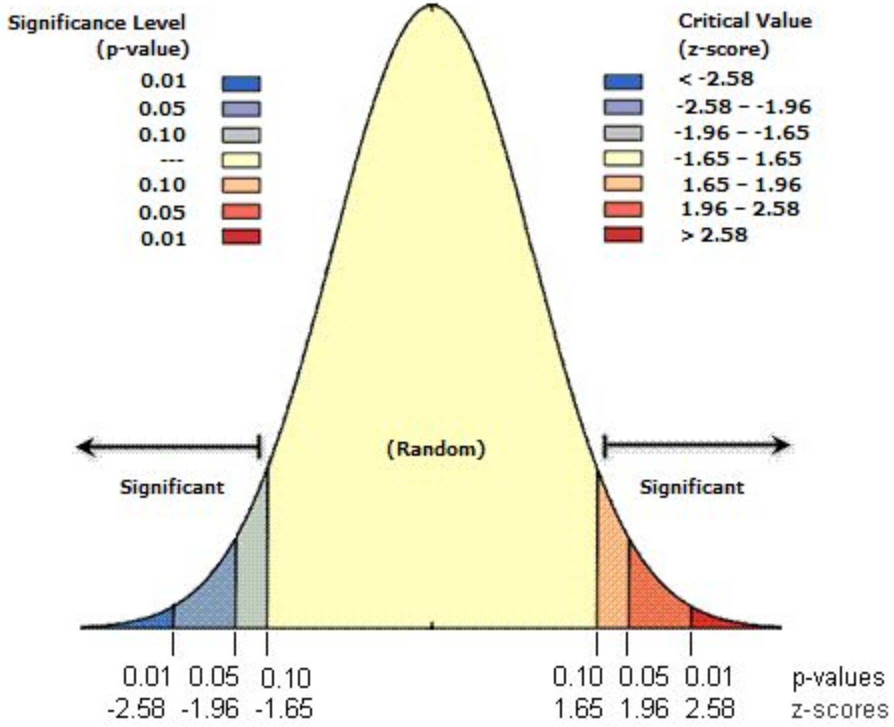


Geweke Diagnostic

Recall: This performs something like a classical stats T-test to compare the means of two samples. In this case, the samples are the first 10% and latter 50%. The output is compared to z-score critical values depending on the chosen significance level.

Almost all our cases “pass the test”

Image



Geweke Diagnostic Table

Our beta coefficients look to converge using this metric, but two of the random intercepts do not.

Below we see Geweke diagnostic for the β coefficients, demonstrating the convergence (full table in Appendix).

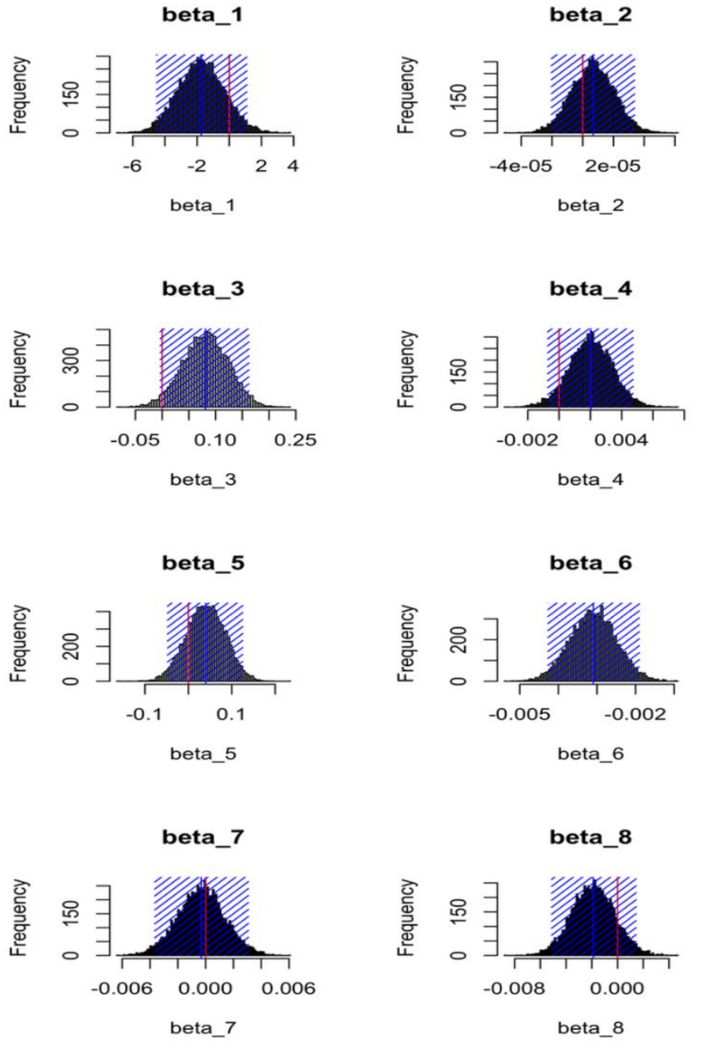
β_1	0.514926
β_2	-0.474728
β_3	-0.576305
β_4	0.970695
β_5	-0.431691
β_6	1.404739
β_7	0.702277
β_8	-0.834102

Below we call out random intercepts with absolute value Geweke diagnostics greater than 2.58.

u_6	-2.949920
u_{33}	-2.638370

Coefficient Results

We produced coefficients for our variables. As seen at right, many of them may not be significantly different than zero, except for the 3rd and 6th plots at right corresponding to nightly temperature and lightly active minutes



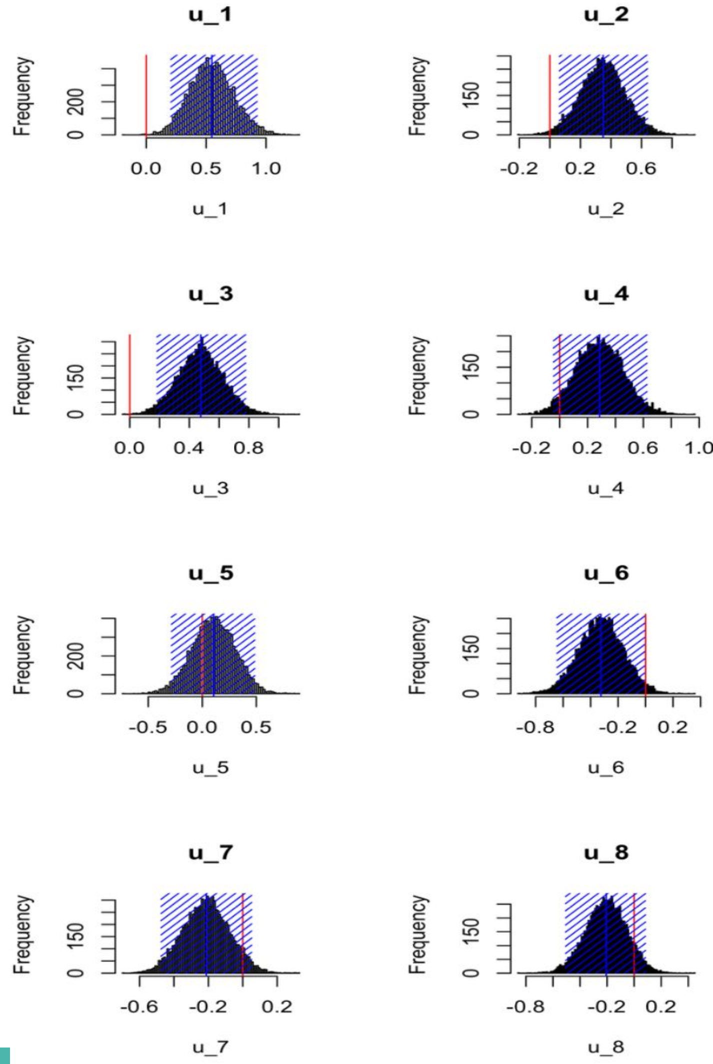
Random Intercept Results

We generate 36 random intercepts, some sample distributions shown at right, each for an individual in the sample from the data.

We see individual 1, 2, 3 show higher probability of sleeping a full night.

Individual 6 shows lower probability.

Other individuals may be zero effect.



Discussion of Limitations and Future Work

One goal for the group was to incorporate a Lasso regularization in the Bayesian Regression model. This would replace the non-informative prior on the Betas with a Laplace prior. We see that some of our predictors could possibly be removed.

We would like to extend the analysis to conducting Bayesian inference on new data, or by implementing a test/train split. This would allow us to assess predictive capabilities in addition to the model's interpretation capabilities.

Regularization

As discussed in class, we could implement these alternate priors in our model to achieve the regularization methods. (source)

6.2 Shrinkage Methods 227

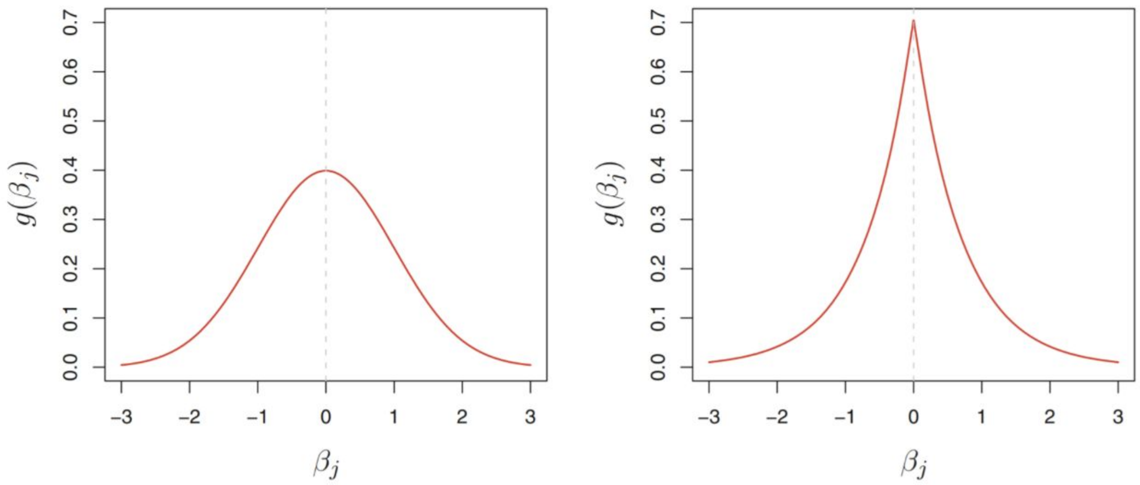


FIGURE 6.11. Left: *Ridge regression is the posterior mode for* β *under a Gaussian prior.* Right: *The lasso is the posterior mode for* β *under a double-exponential prior.*