Random Effect Priors in Binary Longitudinal Models

Nathan Bick
Georgetown University
Washington, DC

Juna Luzi
Georgetown University
Washington, DC

Ucheoma Ukah
Georgetown University
Washington, DC

1 Introduction

Sufficient quantity of sleep is very important for overall health, and yet many people find it difficult to sleep enough despite their better efforts. There is scientific literature (Centers for Disease Control and Prevention, accessed May 12 2023) that suggests at least seven (7) hours is necessary for adults to remain healthy. Using Fitbit data collected from 36 participants over the span of several months, we will explore the impact predictors such as steps taken, stress, and duration of activity have on an individual's nightly hours of sleep. These predictors will be assessed at daily increments to produce a binary response variable defined as True when the individual does sleep at least seven hours and False otherwise.

When conducting a longitudinal study, such as this, we introduce variability not only between participants but also within the measurements of each individual participant. Treating a longitudinal study with several participants as a fixed effect model would neglect to capture the impact of the individual's variability on the results. A more robust approach would be to develop a model that incorporates both the within group variability as well as the between group variability. In this paper we attempt to capture this by developing a binary linear regression model with random effects in the Bayesian context.

2 Methods

To streamline the proposed approach, we adopt the methodology developed by Albert and Chib (Albert and Chib, 1993), who present a Bayesian framework for the analysis of binary and polychotomous response data through data augmentation. Specifically, they incorporate latent variables that conform to a truncated normal distribution and develop a Gibbs sampling algorithm for the estimation of the posterior distribution of β. We leverage this methodology to model our binomial response variable of whether an individual achieves at least seven hours of sleep per day.

Consider y_i to be a vector of outcomes for subject i and has the form $y_i' = [y_{i1}, ..., y_{im_i}]$ where $y_{ij} \in 0, 1$ for $j = 1, ..., m_i$.[1] Let $X_i \in \mathbb{R}^{m_i \times p+1}$ be a matrix of covariates for subject i, $u_i \in \mathbb{R}^{m_i \times 1}$ denote the subject-specific intercept, and $\beta \in \mathbb{R}^{p+1 \times 1}$ represent the fix-effect coefficients. Define $y_i \sim Bern(\theta_i)$ where $\Phi^{-1}(\theta_i) = X_i\beta + u_i$, and $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of the standard normal distribution. Then, $\theta_i = \mathbb{P}(y_i = 1 | X_i, \beta, u_i) = \Phi(X_i\beta + u_i)$. Subsequently, the probit regression model is equivalent to

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

where $z_1, ..., z_n$ are independent latent variables.[2] Subsequently, utilizing the approach outlined by Albert and Chib (Albert and Chib, 1993), the likelihood takes the form of the product of:

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

Our model relies on the following assumptions: a flat prior is assumed for β, thus $\pi(\beta) \propto 1$, and a normal prior is assumed for u_i, specifically $U \sim MVN(0, (1/\tau_u^2)\boldsymbol{I}_{n\times n})$, with a non-informative prior on τ_u^2, $\pi(\tau_u^2) \propto (\tau_u^2)^{-1}$. We define the full posterior distribution[3] to be

$$p(\beta, Z, U, \tau_u^2 | Y, X) = (\tau_u^2)^{n/2-1} \exp\left[-\frac{1}{2}\|Z - (X\beta + DU)\|_2^2 - \frac{\tau_u^2}{2}\|U\|_2^2\right] \left[\prod_{i=1}^{n} \{1(z_i > 0)1(y_i = 1) + 1(z_i \leq 0)1(y_i = 0)\}\right] \tag{2}$$

By utilizing a normally distributed prior on the random intercept, we are able to operate on recognizable conditional distributions, thereby facilitating the implementation of a Gibbs sampler.

Thus, our Gibbs Sampler iterates through draws and updates

1. $\beta^{(b)}|rest \sim MVN\left[(X'X)^{-1}X'(Z^{(b-1)} - DU^{(b-1)}), (X'X)^{-1}\right]$
2. $U^{(b)}|rest \sim MVN\left[(D'D + (\tau_u^2)^{(b-1)}\boldsymbol{I}_{n\times n})^{-1}D'(Z^{(b-1)} - X\beta^{(b-1)}), (D'D + (\tau_u^2)^{(b-1)}\boldsymbol{I}_{n\times n})^{-1}\right]$
3. $(\tau_u^2)^{(b)}|rest \sim Gamma\left(\frac{n}{2}, \frac{1}{2}\|U^{(b-1)}\|_2^2\right)$

[1] Our dataset is unbalanced; therefore, we allow the number of measurement occurrences m to vary by subject.

[2] The latent variables z_i, which are introduced for computational convenience, do not affect the model's outcome.

[3] Full mathematical derivations can be found in the appendix.

4. For all i such that $y_i = 0$, update
$z_i^{(b)}|rest \sim \mathcal{N}\left[X_i\beta^{(b-1)} + u_i^{(b-1)}, 1\right] 1(z_i \leq 0)$

5. For all i such that $y_i = 1$, update
$z_i^{(b)}|rest \sim \mathcal{N}\left[X_i\beta^{(b-1)} + u_i^{(b-1)}, 1\right] 1(z_i > 0)$

The data utilized in this model is the LifeSnaps dataset (Yfantidou et al., 2022), which was obtained from Zenodo[4]. The LifeSnaps dataset comprises self-tracking observations collected via Fitbit watches over a four-month period from 71 individuals, with varying observation counts per individual. The dataset comprises more than 35 different data types, recorded at various granularities ranging from seconds to daily intervals, which total over 71 million rows of data. However, in this study, we focus solely on a subset of daily variables that were deemed to be intuitively related to sleep. After filtering and joining the raw data, our analysis data includes 1,766 daily observations of 8 variables across 36 individuals. The variables include steps, nightly temperature, stress score, temperature variation, lightly active minutes, moderately active minutes, very active minutes, and sleep duration in hours. We calculate the binary sleep response variable.

3 Results

Given our samples for the parameters β, u, and τ, we can inspect the distributions and draw interpretations to understand the relationships in the data:

- By inspecting the mean value of the sampled distribution, we can interpret relationship between the predictor and the response in the case of the β, and the baseline response probability in the case of the u

- By inspecting whether and (and if so, where) the distribution includes zero, we can interpret the significance level of the coefficient. Furthermore, we can use the feature of Bayesian analysis where we can provide degrees of credibility.

Let us consider the β parameters. We included the following variables in our analysis: steps, nightly temperature, stress score, daily temperature variation, lightly active minutes, moderately active minutes, and very active minutes. The coefficients are presented in this order throughout, and also include the overall intercept term.

We see that steps, nightly temperature, stress score, and daily temperature have positive relationships, while the lightly active minutes, moderately active minutes, and very active minutes have negative relationships with the response. There substantial reason to question the significance of the β_7 and β_8, while $\beta_2, \beta_3, \beta_4, \beta_5$ have less reason. For example, for the parameter of steps, we see $P(\beta_2 > 0) = 0.6872$.

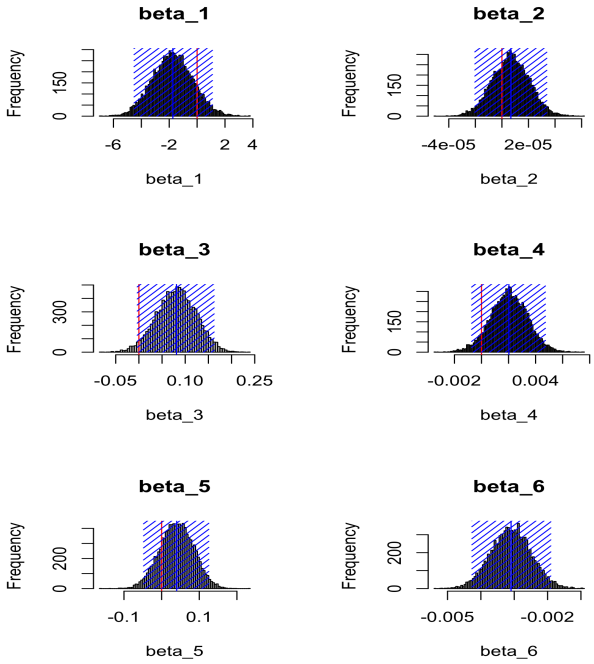
Below is a summary table of the distributions for each β parameter.

[4]Zenodo is an open research repository operated by CERN that provides a general-purpose platform for sharing research outputs

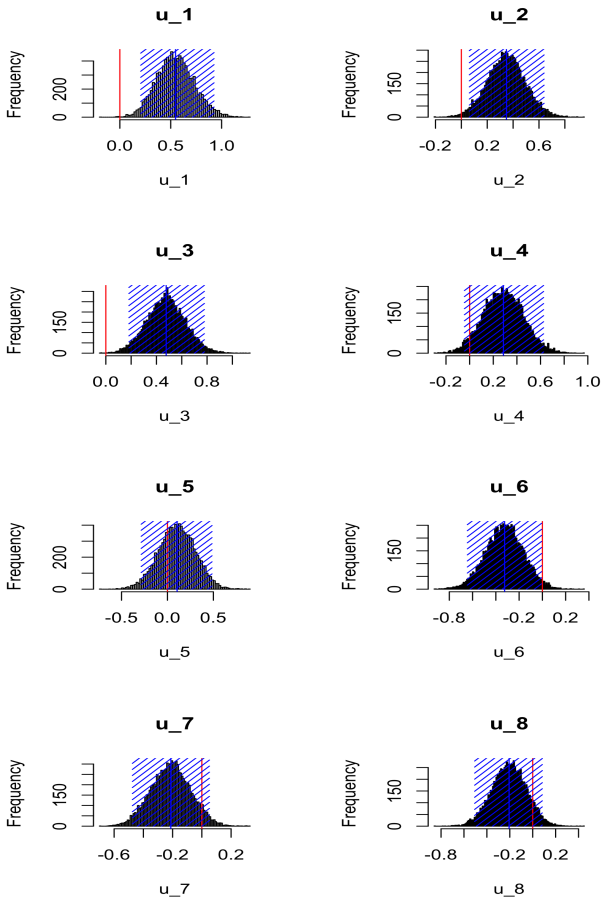
Value	β_1	β_2	β_3	β_4
2.5%	-4.605	-2.017e-05	-0.00225	-0.00068
50%	-1.747	6.722e-06	0.0814	0.002000
97.5%	1.0568	3.381e-05	0.165	0.00468
Mean	-1.745	6.803e-06	0.0813	0.00199
Value	β_5	β_6	β_7	β_8
2.5%	-0.04665	-0.00426	-0.00367	-0.00518
50%	0.04019	-0.00309	-0.000313	-0.00191
97.5%	0.1274	-0.00192	0.00309	0.00137
Mean	0.0401	-0.00309	-0.000309	-0.00191

These results suggest that the strongest relationships are demonstrated by the variables of nightly temperature (higher temperature suggests higher likelihood of full night sleep) and lightly active minutes (lower number of minutes suggests higher likelihood of sleep) .

We present plots for the sample distributions of β parameters. Each features a vertical red line for $x = 0$ and a shaded region between the 2.5-percentile and 97.5-percentiles to aid the reader in interpretation.



We also present a selection of plots for the distributions of the random intercepts. For each individual, our model successfully creates an individual-specific random intercept capturing individual-specific effects, 36 in total. As we consider the examples depicted below, we see that individuals 1, 2, and 3 have significant intercepts higher than zero, meaning that they are more likely to sleep seven hours. Other individuals depicted, such as 4, 5, 7, and 8 are not credibly different than zero. Individual 6, however, is shown to have a lower probability due to the distribution being negative. The remaining distribution plots are in the Appendix.



Although this will be discussed further in the Discussion section, we have reason to believe that a regularization technique may be well suited to this scenario, as evidenced by the credible interval for several of the parameters including zero.

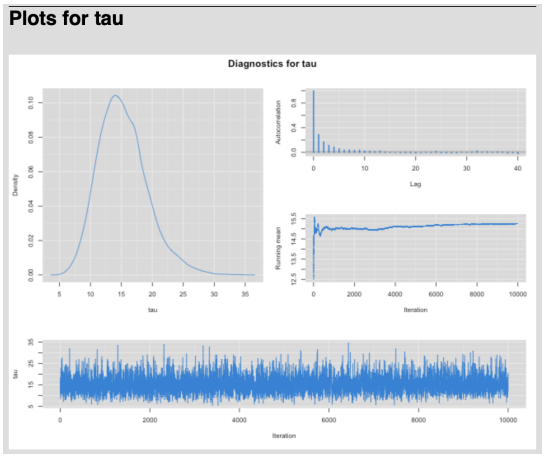
It is important to consider the convergence results fo the Gibbs sampler. We find that our computing methods show convergence for our parameters of interest, so the resulting samples are in fact draws from the full posterior. In our Gibbs Sampler, we used 20,000 total iterations including 10,000 burn-in/warm-up iterations. To investigate the convergence of our sampler, we used two primary methods:

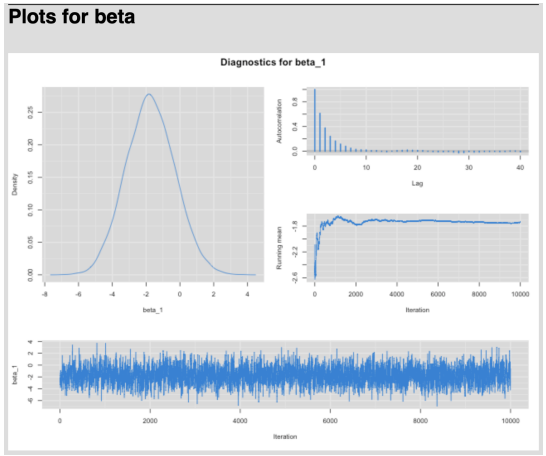
- standard MCMC plots, including for autocorrelation, running mean, and trace plots.
- Geweke Diagnostic

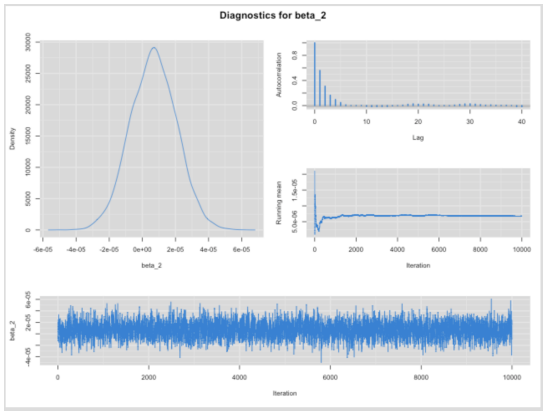
Because our conditional posterior distributions were recognizable and did not require any M-H step, we do not report such convergence statistics as acceptance rate.

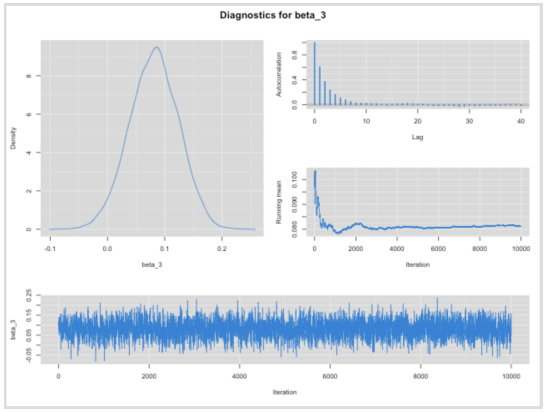
We present example MCMC plot outputs below, which we used for visual inspection. We include a selection of plots for the τ, β, and u parameters.

We first consider the β parameters, and we see that the trace plots show no trend across iterations; the autocorrelation plots go to zero reasonably quickly; and the running mean plots more or less stabilize as well for all parameters.



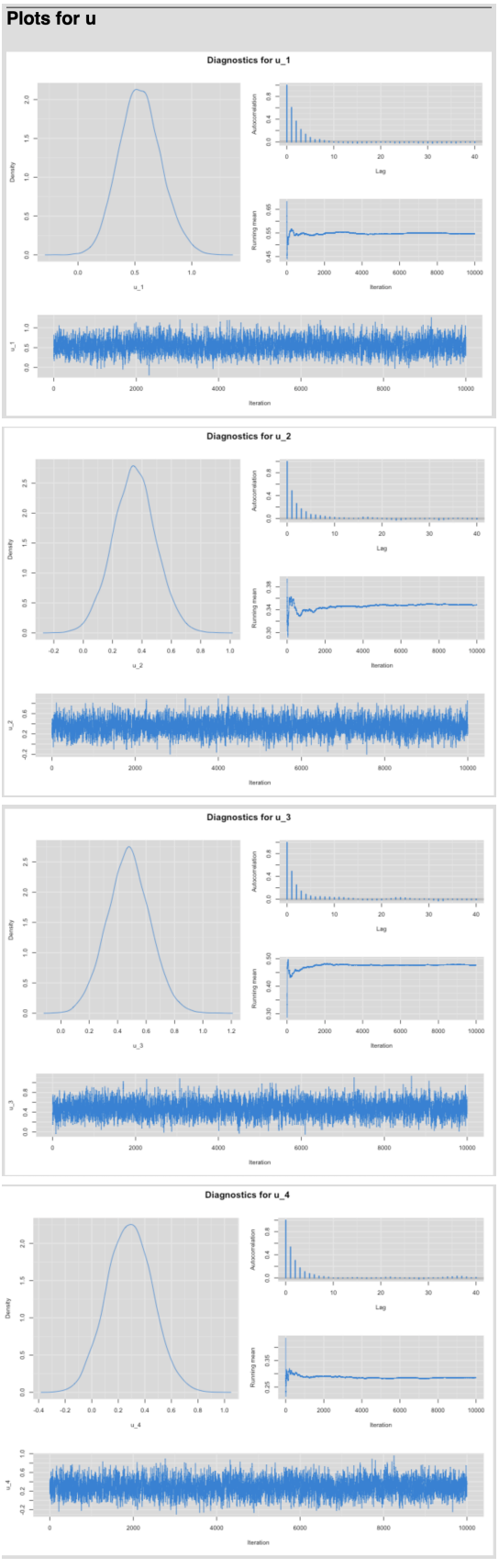






For the random intercepts, our sampler achieves overall convergence, but for a small subset of the intercept parameters there are certain diagnostics to carefully consider and may require additional scrutiny. None of the u parameters display

telltale signs of non-convergence such as late jumps in the running mean or drift in the trace plot, but a few display slower than optimal reduction in autocorrelation or very slight drift in trailing mean. Below we present some examples from among the first plots of the individual random intercepts:



The remainder MCMC plots are available in the supplement at the end of the paper.

We also considered the Geweke diagnostic to assess the convergence. This diagnostic takes the first 10 percent and last 50 percent of the Markov chain and performs a test akin to the classical statistical t-test for the equality of the means. The output is a z-score which we can test against the conventional critical values to numerically determine if there is a convergence issue. We compared these to the 2.58 or 0.01 critical value and found that some of the u's diagnostics are significant at that 0.01 level. We see that all the betas "pass the test", while u_6, and u_33 do not pass the test.

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

4 Discussion

The use of longitudinal binary model with random intercept is a good fit for the problem we investigated. Our model adequately captured the individual-specific effects via the random intercept and the relationship between our predictors and response. However, we noticed that some of the predictor coefficients may not be credibly different from zero, suggesting that the variables available in our data set may not be the most useful for predicting the full-night-sleep binary. On the other hand, we see a range of outcomes for the random intercepts, which makes intuitive sense given individual variability, including in the variable number of observations.

Our implementation of the Gibbs sampler to our model converged on the whole as shown by the plots in the supplement to the paper and the Geweke diagnostics.

In future exetensions to the above analysis, we would like to investigate alternate choices of priors for the regression β coefficients. The choice of non-informative prior could be improved by choosing the Laplace as prior. This would lead to a Lasso regularization regression, which would help with removing unnecessary predictors.

Additionally, we would like to spend more time on implementing Bayesian inference on a test dataset. While it is indeed useful to create a model for interpretation purposes as we have done, it is also valuable to both test the model on new data and to put it to use in prediction. This would require either a test/train split for our data, or to find additional data.

References

J. Albert and S. Chib. 1993. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88:669–679.

Centers for Disease Control and Prevention. accessed May 12, 2023. How much sleep do i need? `https://www.cdc.gov/sleep/about_sleep/how_much_sleep.html`.

Sofia Yfantidou, Christina Karagianni, Stefanos Efstathiou, Athena Vakali, Joao Palotti, Dimitrios Panteleimon Giakatos, Thomas Marchioro, Andrei Kazlouski, Elena Ferrari, and Šarūnas Girdzijauskas. 2022. LifeSnaps: A 4-month multi-modal dataset capturing unobtrusive snapshots of our lives in the wild.

A Model Derivations

The complete posterior distribution of our model is given by

$$
\begin{aligned}
p(\beta, Z, U, \tau_u^2 | Y, X) &\propto \mathcal{L}(Y|Z, X, \beta, U)\pi(\beta)\pi(U|\tau_u^2)\pi(\tau_u^2) \\
&\propto \left[\prod_{i=1}^{n} \{1(z_i > 0)1(y_i = 1) \right. \\
&\quad \left. +1(z_i \leq 0)1(y_i = 0)\} \right] \\
&\quad \times \exp\left[-\frac{1}{2}\|Z - (X\beta + DU)\|_2^2\right] \\
&\quad \times (\tau_u^2)^{n/2} \exp\left[-\frac{\tau_u^2}{2}\|U\|_2^2\right] (\tau_u^2)^{-1} \\
&= \left[\prod_{i=1}^{n} \{1(z_i > 0)1(y_i = 1) \right. \\
&\quad \left. +1(z_i \leq 0)1(y_i = 0)\} \right] \\
&\quad \times (\tau_u^2)^{n/2-1} \exp\left[-\frac{1}{2}\|Z - (X\beta + DU)\|_2^2 \right. \\
&\quad \left. -\frac{\tau_u^2}{2}\|U\|_2^2\right]
\end{aligned}
$$

The complete conditional distribution of the β parameter in our model is

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

We recognize the distribution as the kernel of a normal distribution with mean $\mu_\beta = (X'X)^{-1}X'(Z - DU)$ and variance $v_\beta = (X'X)^{-1}$

The complete conditional distribution of the U parameter

in our model is

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

We recognize the distribution as the kernel of a multivariate normal distribution with parameters $(D'D+\tau_u^2\boldsymbol{I}_{n\times n})^{-1}D'(Z-X\beta)$ and $(D'D+\tau_u^2\boldsymbol{I}_{n\times n})^{-1}$.

The complete conditional distribution of the τ_u^2 parameter in our model is

$$p(\tau_u^2|\beta,UZ,Y,X)\propto(\tau_u^2)^{n/2-1}\exp\left[-\frac{\tau_u^2}{2}\|U\|_2^2\right]$$

We recognize the distribution as the kernel of the Gamma distribution with parameters $\frac{n}{2}$ and $\frac{1}{2}\|U\|_2^2$.

The conditional distribution for each z_i on $y_i=0$ is given by

$$p(z_i|\beta,u_i,y_i=0,X_i)\propto\exp\left[-\frac{1}{2}\left(z_i-(X_i\beta+u_i)\right)^2\right]1(z_i\le 0)$$

We recognize the distribution as the kernel of a normal distribution with mean $X_i\beta+u_i$ and variance 1 truncated below zero. Similarly, the conditional distribution for each z_i when $y_i=1$ is

$$p(z_i|\beta,u_i,y_i=1,X_i)\propto\exp\left[-\frac{1}{2}\left(z_i-(X_i\beta+u_i)\right)^2\right]1(z_i>0)$$

Following the previous result, we recognize the distribution as the kernel of a normal distribution with mean $X_i\beta+u_i$ and variance 1 truncated above zero.

B Tables

Below we see the full table of distribution summaries for the u parameters.

Percentile	u					
	u_1	u_2	u_3	u_4	u_5	u_6
2.5%	0.207	0.0638	0.185	-0.0421	-0.286	-0.6437
50%	0.544	0.345	0.475	0.288	0.104	-0.320
97.5%	0.922	0.634	0.778	0.630	0.4925	-0.0070
mean	0.548	0.345	0.477	0.288	0.105	-0.322

	u_7	u_8	u_9	u_10	u_11	u_12
2.5%	-0.476	-0.504	-0.136	-0.688	-0.300	-0.180
50%	-0.211	-0.203	0.175	-0.357	0.0845	0.118
97.5%	0.0503	0.0862	0.505	-0.0343	0.482	0.419
mean	-0.211	-0.205	0.177	-0.358	0.0861	0.118

	u_13	u_14	u_15	u_16	u_17	u_18
2.5%	-0.924	-0.430	-0.657	-0.219	-0.543	-0.549
50%	-0.600	0.0369	-0.3267	0.0880	-0.0438	-0.263
97.5%	-0.300	0.5057	-0.01350	0.4043	0.443	0.0119
mean	-0.604	0.0373	-0.329	0.0902	-0.0469	-0.265

	u_19	u_20	u_21	u_22	u_23	u_24
2.5%	-0.161	-0.344	0.348	0.240	-1.141	-0.448
50%	0.228	-0.0746	0.661	0.588	-0.800	-0.155
97.5%	0.651	0.200	1.004	0.985	-0.482	0.128
mean	0.232	-0.0739	0.665	0.595	-0.802	-0.157

	u_25	u_26	u_27	u_28	u_29	u_30
2.5%	-0.593	-0.3715	0.0844	0.0811	-0.8962	-0.808
50%	-0.225	0.0982	0.435	0.475	-0.404	-0.454
97.5%	0.130	0.595	0.798	0.922	0.0204	-0.119
mean	-0.227	0.101	0.437	0.483	-0.413	-0.457

	u_31	u_32	u_33	u_34	u_35	u_36
2.5%	-0.371	-0.473	-0.642	-0.698	0.182	-0.153
50%	-0.0687	-0.164	-0.3318	-0.3548	0.522	0.136
97.5%	0.238	0.151	-0.037	-0.0317	0.890	0.431
mean	-0.0686	-0.163	-0.334	-0.356	0.526	0.136

Below we see the full table of Geweke Diagnostic values:

Parameter	Value
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
u_1	-0.287257
u_2	-0.975786
u_3	-1.276105
u_4	-0.200308
u_5	-0.515895
u_6	-2.949920
u_7	-1.585263
u_8	-0.353173
u_9	-0.800180
u_{10}	-1.786157
u_{11}	0.001445
u_{12}	-1.628287
u_{13}	-1.704308
u_{14}	-1.029297
u_{15}	-2.230273
u_{16}	0.608960
u_{17}	0.328686
u_{18}	-0.216783
u_{19}	-0.635291
u_{20}	-0.785723
u_{21}	1.408848
u_{22}	1.166730
u_{23}	-1.814024
u_{24}	-1.776982
u_{25}	-0.095651
u_{26}	0.600251
u_{27}	-0.442609
u_{28}	1.114402
u_{29}	-1.391743
u_{30}	-1.483992
u_{31}	-0.142445
u_{32}	-1.168170
u_{33}	-2.638370
u_{34}	-0.680126
u_{35}	0.824166
u_{36}	-0.252140
τ	-1.875053

C Replication Code

Our replication code can be found in the group's GitHub repository at the following link: `https://github.com/BickieSmalls/math640-final-proj` and it is included in the file $\texttt{bayesian}_{c}omputing.R$.