A. The data in "cheese.csv" are about sales volume, price, and advertisting display activity for packages of Borden sliced "cheese." The data are taken from Rossi, Allenby, and McCulloch's textbook on *Bayesian Statistics and Marketing*. For each of 88 stores (store) in different US cities, we have repeated observations of the weekly sales volume (vol, in terms of packages sold), unit price (price), and whether the product was advertised with an in-store display during that week (disp = 1 for display). Your goal is to estimate, on a store-by-store basis, the effect of display ads on the demand curve for cheese. A standard form of a demand curve in economics is of the form $Q = \alpha P^{\beta}$, where Q is quantity demanded (i.e. sales volume), P is price, and α and β are parameters to be estimated. You'll notice that this is linear on a log-log scale,

$$\log Q = \log \alpha + \beta \log P$$

which you should feel free to assume here. Economists would refer to β as the price elasticity of demand (PED). Notice that on a log-log scale, the errors enter multiplicatively. There are several things for you to consider in analyzing this data set.

1. **The demand curve might shift (different α) and also change shape (different β) depending on whether there is a display ad or not in the store.**
2. **Different stores will have very different typical volumes, and your model should account for this.**
3. **Do different stores have different PEDs? If so, do you really want to estimate a separate, unrelated β for each store?**
4. **If there is an effect on the demand curve due to showing a display ad, does this effect differ store by store, or does it look relatively stable across stores?**
5. **Once you build the best model you can using the log-log specification, do see you any evidence of major model mis-fit?**

Propose an appropriate hierarchical model that allows you to address these issues, and use Gibbs sampling to fit your model.

Let's index stores by $i = 1, ..., n$ and observations per store $j = 1, ..., N_i$. Note that there may be a different number of observations per store. The quantity demanded at store i and observation j is denoted Q_{ij}, and similar use of indexes applies for price P and the display covariate. Using the standard form of the economic demand curve given above, we can denote our sampling model as

$$\log Q_{ij} = (\beta_0)_i + (\beta_1)_i \log P_{ij} + (\beta_2)_i \mathbb{1}(\text{disp}_{ij} = 1) + (\beta_3)_i \log P_{ij} \cdot \mathbb{1}(\text{disp}_{ij} = 1)$$

We can rewrite the above as a linear regression model:

$$y_{ij} = X_{ij}^T \beta_i + \epsilon_{ij},$$

where

$$\begin{aligned} y_{ij} &= \log Q_{ij} \\ X_{ij} &= \left(1,\ \log P_{ij},\ \mathbb{1}(\text{disp}_{ij}),\ \log P_{ij} \cdot \mathbb{1}(\text{disp}_{ij})\right)^T \\ \beta_i &= ((\beta_0)_i, (\beta_1)_i, (\beta_2)_i, (\beta_3)_i)^T \end{aligned}$$

This linear regression model can be extended to a hierarchical model with the following specification, where priors were selected to primarily induce conjugate as well as allow the data to speak for itself:

$$\begin{aligned} [y_{ij} \mid \beta_i, \sigma_i^2] &\equiv \text{Normal}\left(X_{ij}^T \beta_i, \sigma_i^2\right), \\ [\beta_i \mid \mu_\beta, \Sigma] &\equiv \text{Normal}\left(\mu_\beta, \Sigma = \text{diag}(s_1^2, \ldots, s_4^2)\right), \\ [\mu_\beta] &\propto 1, \\ [s_p^2] &\equiv \text{Inv-Ga}\left(\frac{1}{2}, \frac{1}{2}\right), \\ [\sigma_i^2] &\equiv \text{Inv-Ga}\left(\frac{a}{2}, \frac{b}{2}\right), \\ [a] &\equiv \text{Gamma}(3, 1), \\ [b] &\equiv \text{Gamma}(3, 1), \end{aligned}$$

where $p = 1, ..., P$. Here, we have four covariates, so $P = 4$. We use an improper prior for μ_β to avoid providing *a priori* information. With this model specification, we are allowing the demand curve to shift (different α) and change shape (different β) depending on whether a display ad was in effect. Inference on β_2 and β_3 will inform us regarding the marginal effect of advertising for α and β, respectively. Additionally, by varying β_0 and β_1 by store, we assume that different stores have different typical volumes and different PEDs.

To implement this hierarchical regression model in an MCMC algorithm, we need to obtain those tasty full-conditionals. The full-conditional for β_i is

$$\begin{aligned}
[\beta_i|\cdot] &\equiv \text{Normal}\left(\mu_i^*, \Sigma_i^*\right), \\
\Sigma_i^* &= \left(\frac{X_i^T X_i}{\sigma_i^2} + \Sigma^{-1}\right)^{-1}, \\
\mu_i^* &= \Sigma_i^* \left(\frac{X_i^T Y_i}{\sigma_i^2} + \Sigma^{-1}\mu_\beta\right),
\end{aligned}$$

where $X_i = [X_{i,1}^T, ..., X_{i,N_i}^T]^T$ and $Y_i = [Y_{i,1}, ..., Y_{i,N_i}]^T$.

The full-conditional for μ_β is

$$\begin{aligned}
[\mu_\beta|\cdot] &\equiv \text{Normal}\left(\bar{\beta}, \frac{1}{n}\Sigma\right), \\
\bar{\beta} &= \frac{1}{n}\sum_{i=1}^{n}\beta_i
\end{aligned}$$

The full-conditional for s_p^2 is

$$[s_p^2|\cdot] \equiv \text{Inverse-Gamma}\left(\frac{n}{2} + \frac{1}{2}, \frac{1}{2}\left(1 + \sum_{i=1}^{n}(\beta_{ip} - \mu_{\beta_p})^2\right)\right)$$

The full-conditional for σ_i^2 is

$$[\sigma_i^2|\cdot] \equiv \text{Inverse-Gamma}\left(\frac{a}{2} + \frac{N_i}{2}, \frac{1}{2}\left(b + (Y_i - X_i\beta_i)^T(Y_i - X_i\beta_i)\right)\right)$$

Note that the full-conditionals for a and b are not conjugate. Therefore, we will update these parameters with Metropolis-Hastings ratios. Consider using a Uniform$(1, 10)$ proposal for both a and b. Then, their Metropolis-Hastings ratios would look like

$$\begin{aligned}
mh_a &= \frac{\prod_{i=1}^{n}\left([\sigma_i^2|a^{(*)}, b]\right)[a^{(*)}][a^{(k-1)}|a^{(*)}]}{\prod_{i=1}^{n}\left([\sigma_i^2|a^{(k-1)}, b]\right)[a^{(k-1)}][a^{(*)}|a^{(k-1)}]}, \\
mh_b &= \frac{\prod_{i=1}^{n}\left([\sigma_i^2|b^{(*)}, a]\right)[b^{(*)}][b^{(k-1)}|b^{(*)}]}{\prod_{i=1}^{n}\left([\sigma_i^2|b^{(k-1)}, a]\right)[b^{(k-1)}][b^{(*)}|b^{(k-1)}]}
\end{aligned}$$

With these updates in hand, we can implement our MCMC algorithm and obtain the estimates for σ_i^2 and β_i.

First, let's examine the trace plots for $\mu_{\beta,p}$ and s_p^2 in Figure 1. We appear to learn very well and see good mixing. This is to be expected since we have plenty of data to learn about these eight parameters across the many sites. The trace plots are the result of a burn-in period of 3000 iterations and thinning every other iteration.

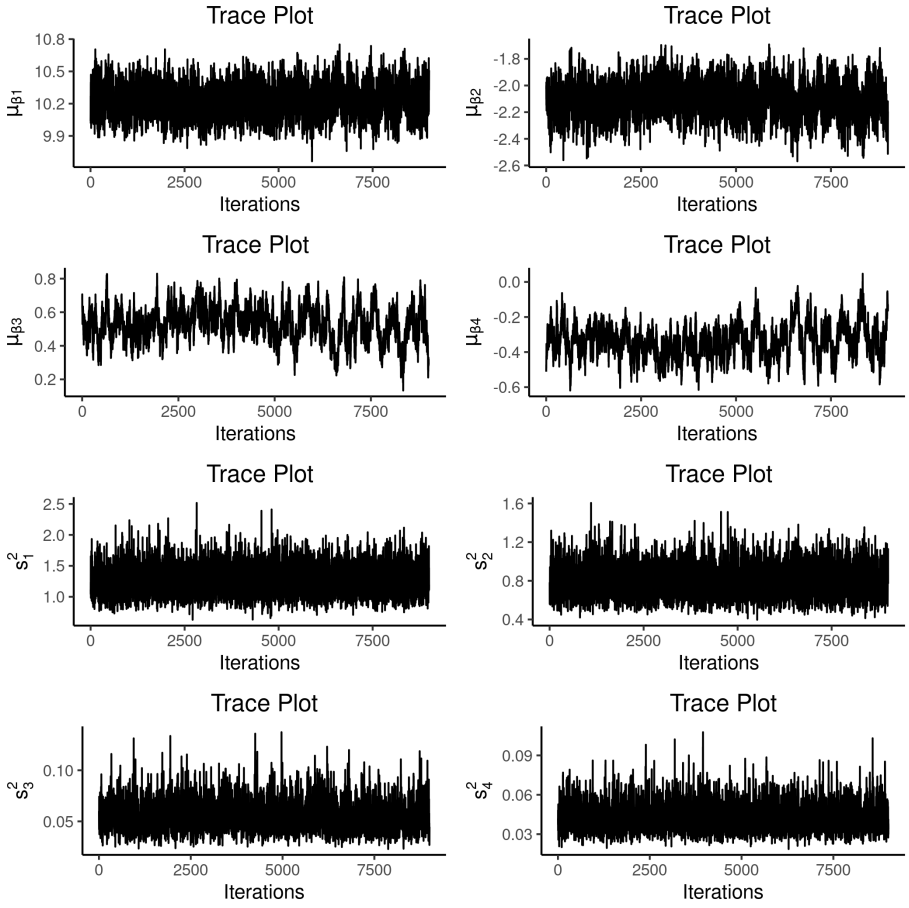

Figure 1: Trace Plots for s_p^2 and μ_β.

Now, let's take a look at a plot of σ_i^2 per store; note that the ordering does not mean anything in Figure 2. We can see that the store-level variance for volume sold varies only slightly; estimates range from 0.01 to 0.41. The relatively low variance for each store is likely due to the relatively large number of observations per store. Note that different numbers of observations (i.e. N_i) per store may lead to different values of σ_i^2. Therefore, it appears that modeling per-store variance was an appropriate move, as having a global variance parameter would not enable us to see the different in per-store variance, albeit the difference is quite small.

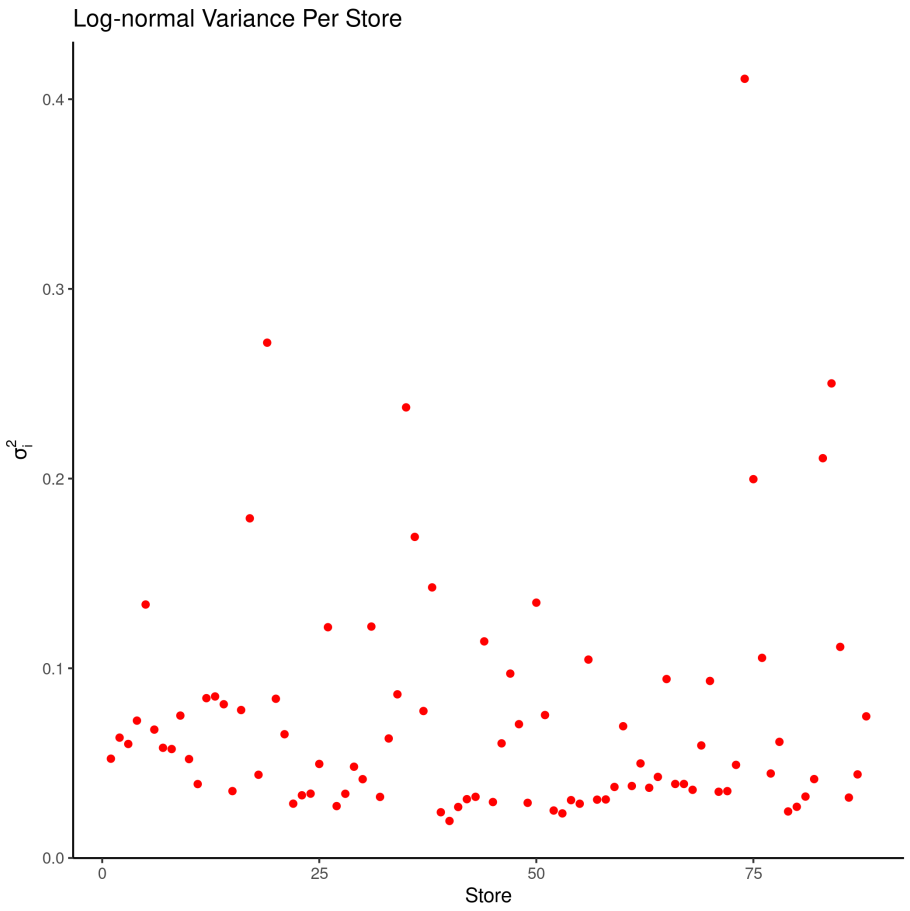

Figure 2: σ_i^2 v. Store.

Now, let's examine the histograms of the β values. In Figure 3, we obtain histograms for the β coefficients for the four parameters across all stores. We can interpret each of these parameters in the log-quantity sold space. It appears that the typical intercept for the demand curve is roughly 10 for each store. While there is some variability, the standard deviation is quite low. The marginal shift in the demand curve due to advertising is roughly 0.5 for each store, indicating that advertising leads to an increase in volume sold. The change in shape for the demand curve given a unit change in price appears to be -2 for each store, indicating that a unit increase in log-price will decrease log-volume sold by 2 units. Finally, the marginal change in shape when advertising appears to be -0.5 per store, indicating that an advertisement can not offset an increase in price on the volume sold.

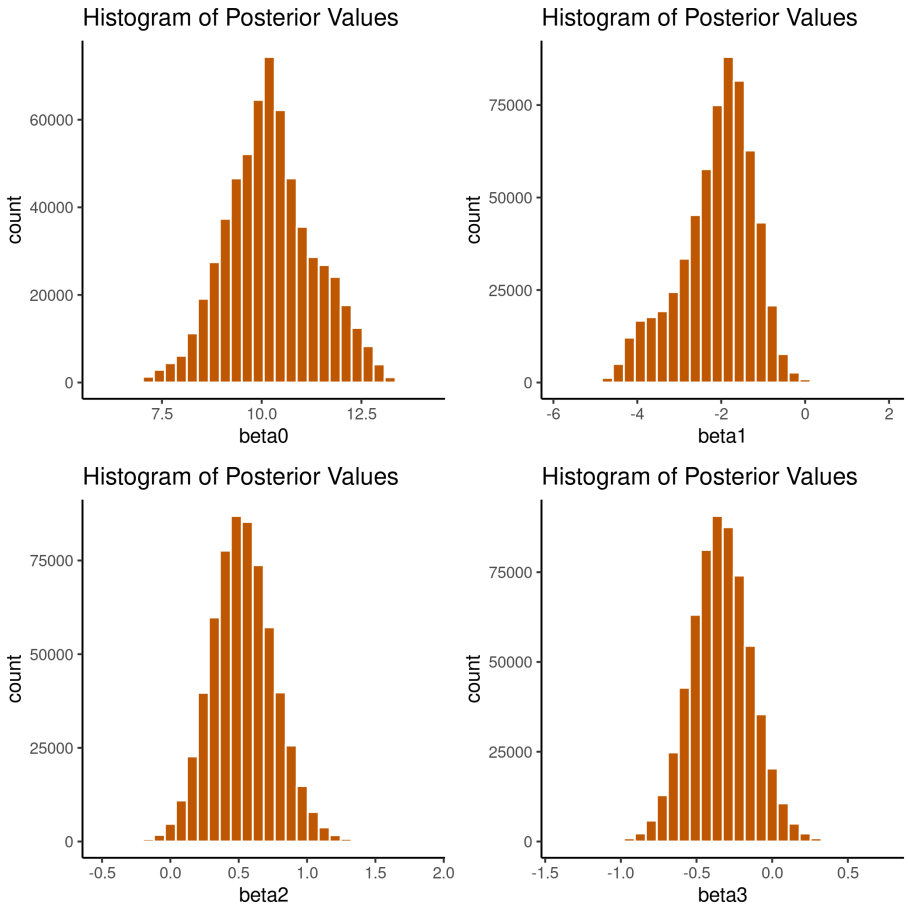

Figure 3: Histogram for β Coefficients.

Finally, let's examine the posterior means for β_2 and β_3 for each store in Figures 4 and 5, respectively. Once again, the ordering of these points does not imply any relationship. Rather, we can tell from the difference in heights of the points that each store has a different posterior estimate for both β_2 and β_3. It appears that the posterior estimates for β_2 vary slightly more than those for β_3. This indicates that the effect on the demand curve when advertising is different across each store. However, one may argue that the difference is not entirely significant. For example, we found that the intercept $\beta_0 \approx 10$, so a change of 0.4 does not greatly affect the overall intercept of the model when advertising.

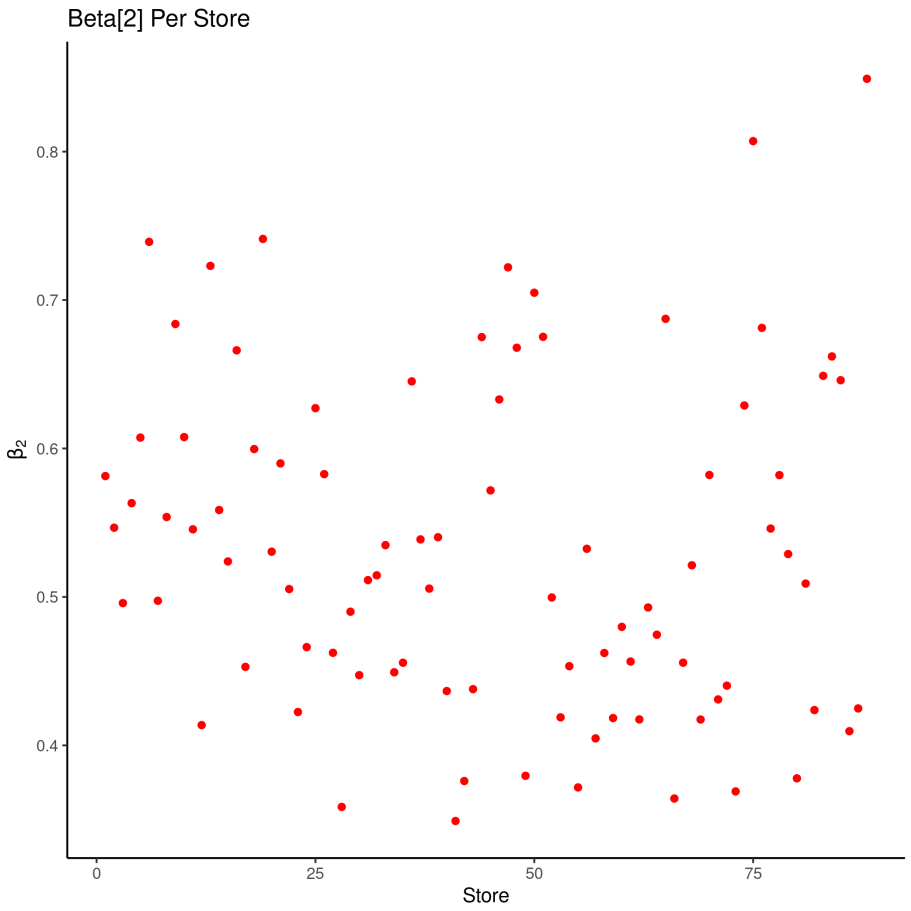

Figure 4: Estimates of β_2 per store.

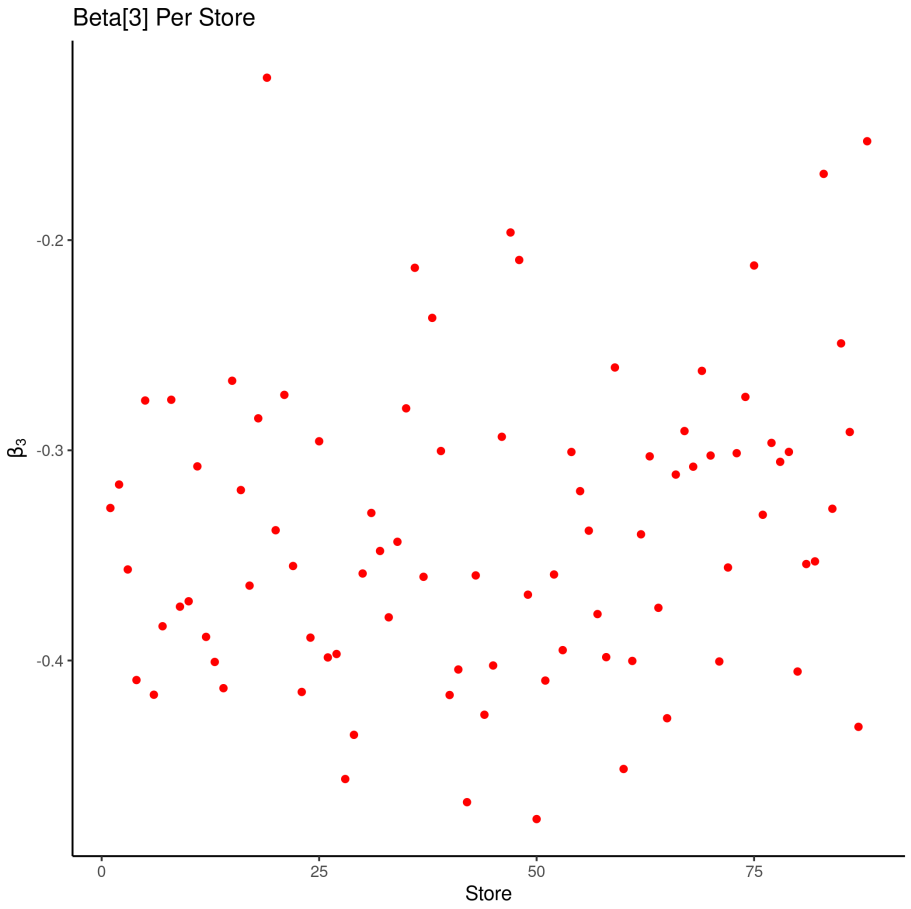

Figure 5: Estimates of β_3 per store.

Now, I address each of the five considerations given in the problem statement.

1. By the model specification, we allow the demand curve to shift and change shape depending on whether or not a display was present in the store. We do this by including $(\beta_2)_i$ and $(\beta_3)_i$ in our model.
2. We account for the difference in typical volumes between stores by having the coefficient $(\beta_0)_i$ vary depending on the store.
3. We account for the difference in PEDs by including the coefficient $(\beta_1)_i$ per store.
4. It appeared that the effect on the demand curve due to showing a display ad was very similar across stores relative to the quantities of interest.
5. When fitting this model, I saw a case of major model mis-fit in the quite terrible updating of a and b; see Figure 6. Evidently, we can obtain much better mixing and exploration than our current Metropolis-Hastings scheme. In the future, I would try different proposal distributions for a and b, as I do not suspect the prior for a or b to be the main culprit.

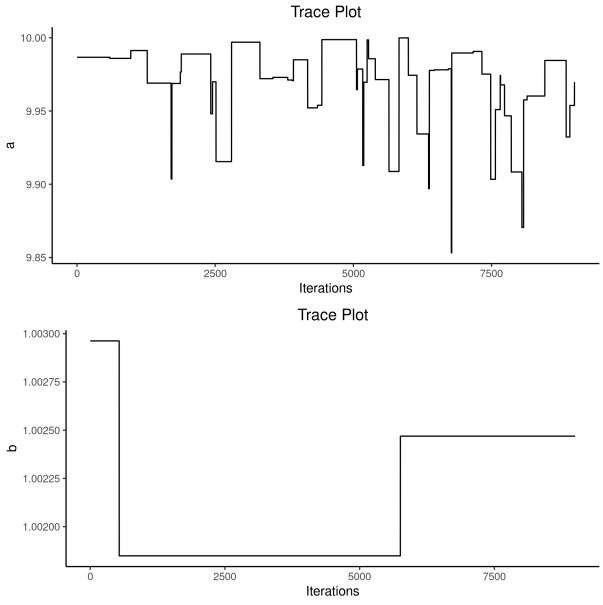

Figure 6: Trace plots for a and b.

A. In "polls.csv" you will find the results of several political polls from the 1988 U.S. presidential election. The outcome of interest is whether someone plans to vote for George Bush (senior, not junior). There are several potentially relevant demographic predictors here, including the respondent's state of residence. The goal is to understand how these relate to the probability that someone will support Bush in the election. You can imagine this information would help a great deal in poll re-weighting and aggregation (ala Nate Silver). Use Gibbs sampling, together with the Albert and Chib trick, to fit a hierarchical probit model of the following form:

$$\begin{aligned} \mathbf{Pr}(y_{ij} = 1) &= \Phi(z_{ij}) \\ z_{ij} &= \mu_i + x_{ij}^T \beta_i \,. \end{aligned}$$

Here y_{ij} is the response (Bush=1, other=0) for respondent j in state i; $\Phi(\cdot)$ is the probit link function, i.e. the CDF of the standard normal distribution; μ_i is a state-level intercept term; x_{ij} is a vector of respondent-level demographic predictors; and β_i is a vector of regression coefficients for state i.

Of the provided covariates, I decided to use age, education, race, and sex. I converted the categorical data of age and education into three indicator variables each. This decision was made so that we are able to see the marginal effect of more education or age on voting outcome. My hierarchical model is an extension of that found in Albert & Chib (1993):

$$\begin{aligned} P(Y_{ij} = 1) = \Phi(X_{ij}^T \alpha_i), \\ \alpha_i \sim \text{Normal}(\beta_i^*, B^*), \\ \beta_i^* \sim \text{Normal}(0, 10^4 \cdot I_{P+1}), \\ B^* \sim \text{Inverse-Wishart}(P+2, I_{P+1}), \end{aligned}$$

where $i = 1, ..., n$ indexes the n states in the data, $j = 1, ..., N_i$ indexes the individuals in each state, and there are P covariates and an intercept in the model. I used a multivariate normal prior on β_i^* and an inverse-Wishart on B^* to exploit conjugacy. With this hierarchical model, we get the following

conjugate posteriors:

$$
\begin{aligned}
[\alpha_i|\cdot] &\equiv \text{Normal}(\tilde{\beta}, \tilde{B}), \\
\tilde{B} &= \left[(B^*)^{-1} + X_i^T X_i\right]^{-1}, \\
\tilde{\beta} &= \tilde{B}\left[(B^*)^{-1}\beta_i^* + X_i^T Z_i\right], \\
[Z_{ij}|\cdot] &\equiv \begin{cases} \text{Normal}(X_{ij}^T\alpha_1, 1)\Big|_{[0,\infty)} & , Y_{ij=1} \\ \text{Normal}(X_{ij}^T\alpha_1, 1)\Big|_{(-\infty,0]} & , Y_{ij=0} \end{cases}, \\
[\beta_i^*|\cdot] &\equiv \text{Normal}(A^{-1}b, A^{-1}), \\
A &= (B^*)^{-1} + (10^4 \cdot I_{P+1})^{-1}, \\
b^T &= \alpha_i^T (B^*)^{-1}, \\
[B^*|\cdot] &\equiv \text{Inverse-Wishart}\left(n + P + 2, I_{P+1} + \sum_{i=1}^{n}(\alpha_i - \beta_i^*)(\alpha_i - \beta_i^*)^T\right)
\end{aligned}
$$

After removing the NA values in the provided data set, I was able to fit this model and obtain the posterior mean estimates for the coefficients in Figure 7. In this figure, we can see the intercept μ and coefficients β_2 through β_9.

Based off the way I set up my indicator functions μ_i is the mean of a non-Black, 65+ year-old, no-high-school male's propensity to vote for Bush in state i. The marginal change in voting propensity for Bush as a woman in state i is measured by β_2. The marginal change for a black American is measured by β_3. The marginal change if someone were to have a Bachelor's degree, high school, or some college is measured with β_4, β_5 and β_6, respectively. The marginal change if someone were to be 18-29, 30-44, or 45-64 are measured by β_7, β_8, and β_9, respectively.

Note that not every state is in each of the subplots in Figure 7. This is because some states did not have enough data to accurately estimate some parameters. For example, Maine did not have a complete observation for a black American; therefore, "ME" does not show up in the subplot for β_3. As one would expect, black voters tend to vote more democratic (blue) than non-black voters. This is why most of the points in the subplot for β_3 are blue.

Just for fun, I included trace plots for the beta estimates for the states of Texas (Figure 8) and New York (Figure 9). With enough data, we are able to accurately estimate the coefficients. For example, New York was a swing state with a margin of just 4.10% for Michael Dukakis. Despite the relatively tight race, we have enough points to robustly estimate the β's.

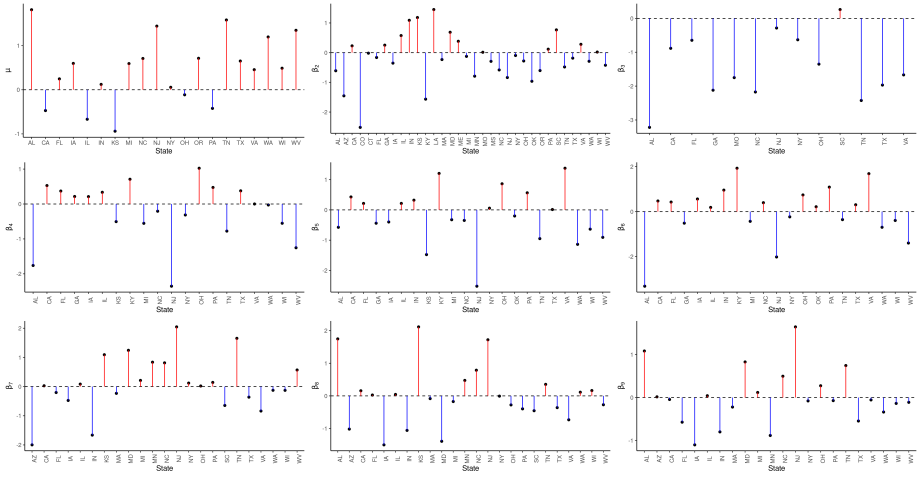

Figure 7: Posterior Mean Estimates for Coefficients.

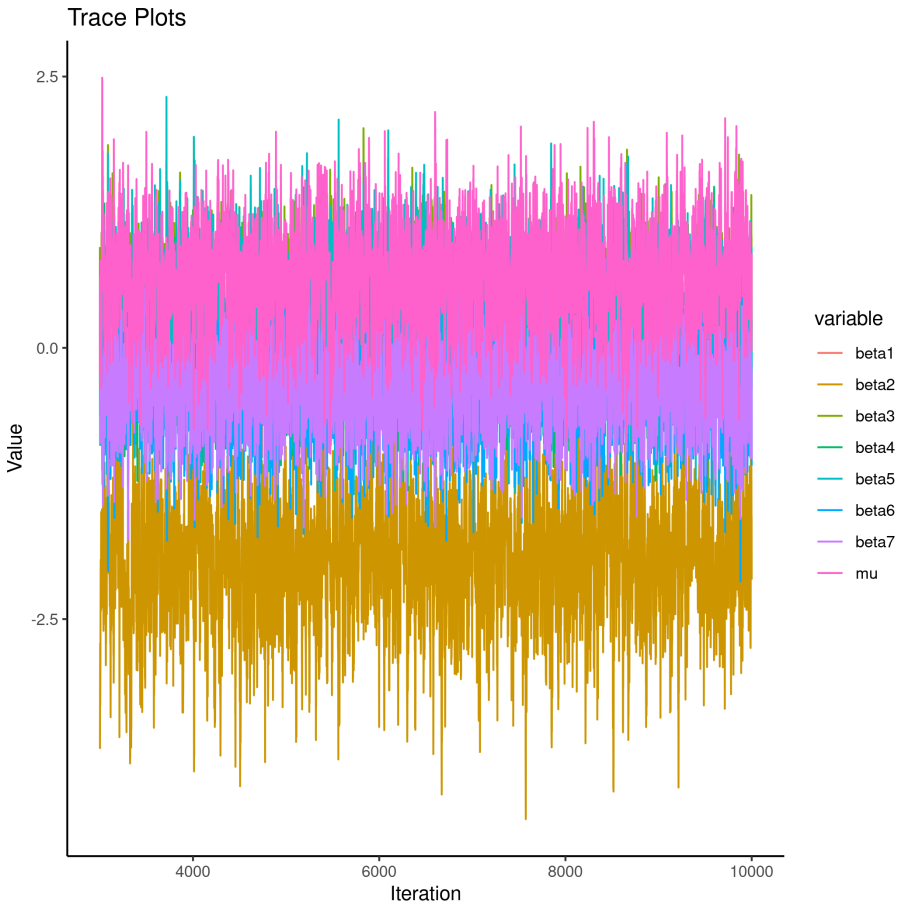

Figure 8: Estimates for Texas.

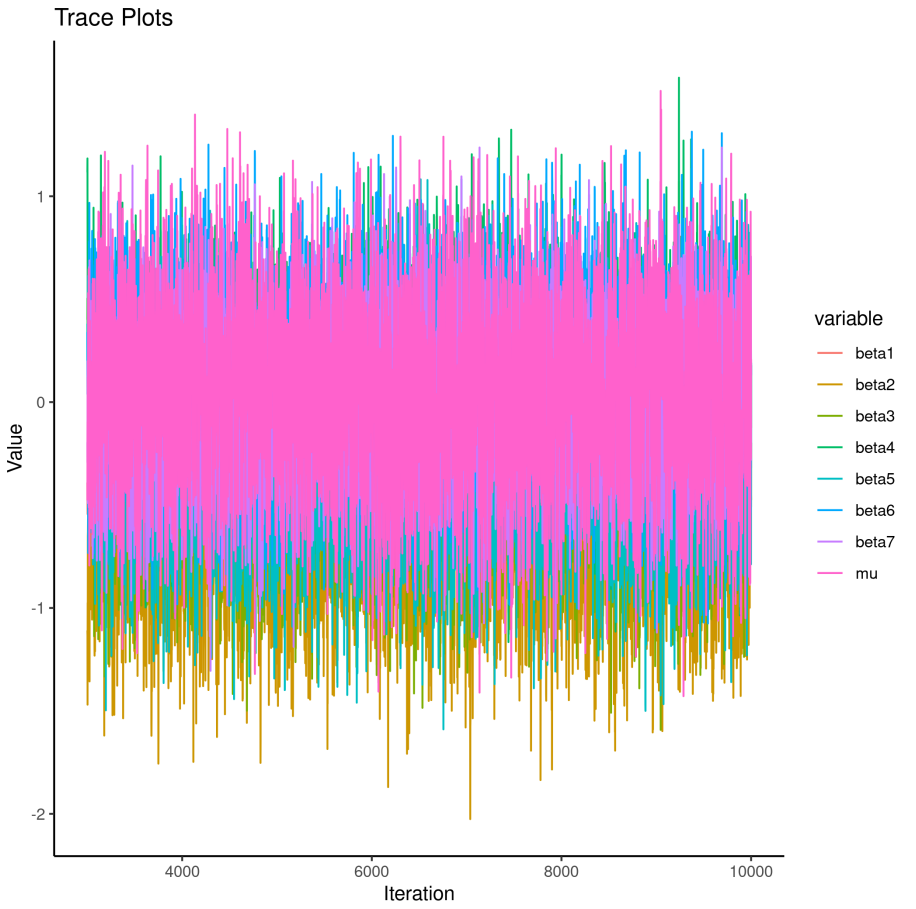

Figure 9: Estimates for New York.