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

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

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 a, b, μ_β, and s_p^2 in Figures 1 and 2. Note the rather low-acceptance rate for a compared to b. This may be because our proposal distribution for a (i.e. the Uniform(0,10) distribution) does not enable us to efficiently learn about a. However, this model appears to learn about b relatively well. Therefore, the proposed model is less than ideal for learning about a. For μ_β and s_p^2, 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 four parameters across the many sites.

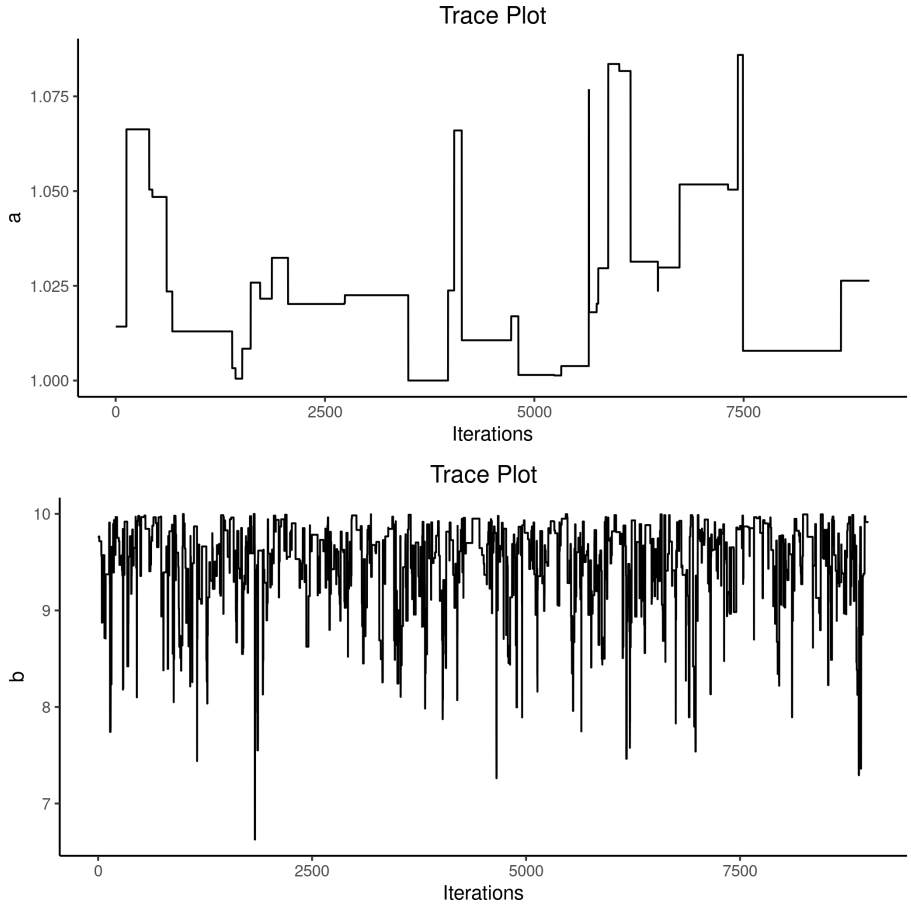

Figure 1: Trace Plots for a and b.

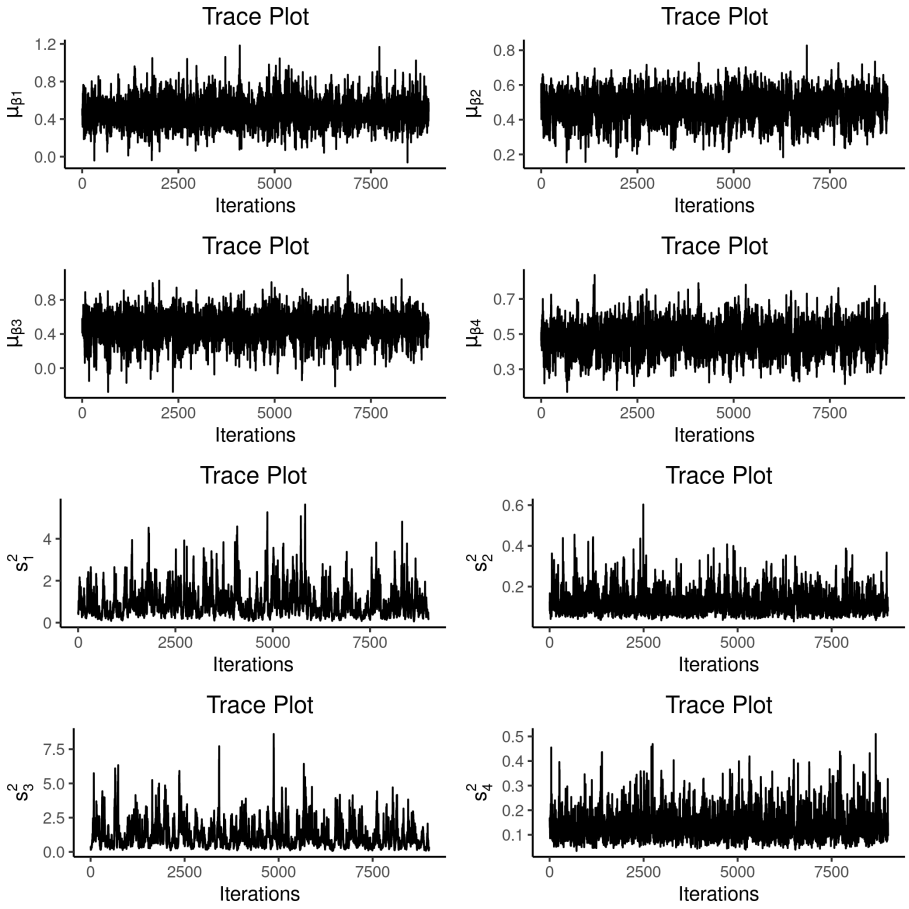

Figure 2: Trace Plots for s_p^2 and μ_β.

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 3. We can see that the store-level variance for volume sold varies greatly; estimates range from 20 to 320. This is likely due to the different number of observations (i.e. N_i) per store. Therefore, it appears that modeling per-store variance was an appropriate move, as having a global variance parameter would not enable us to learn adequately in the model.

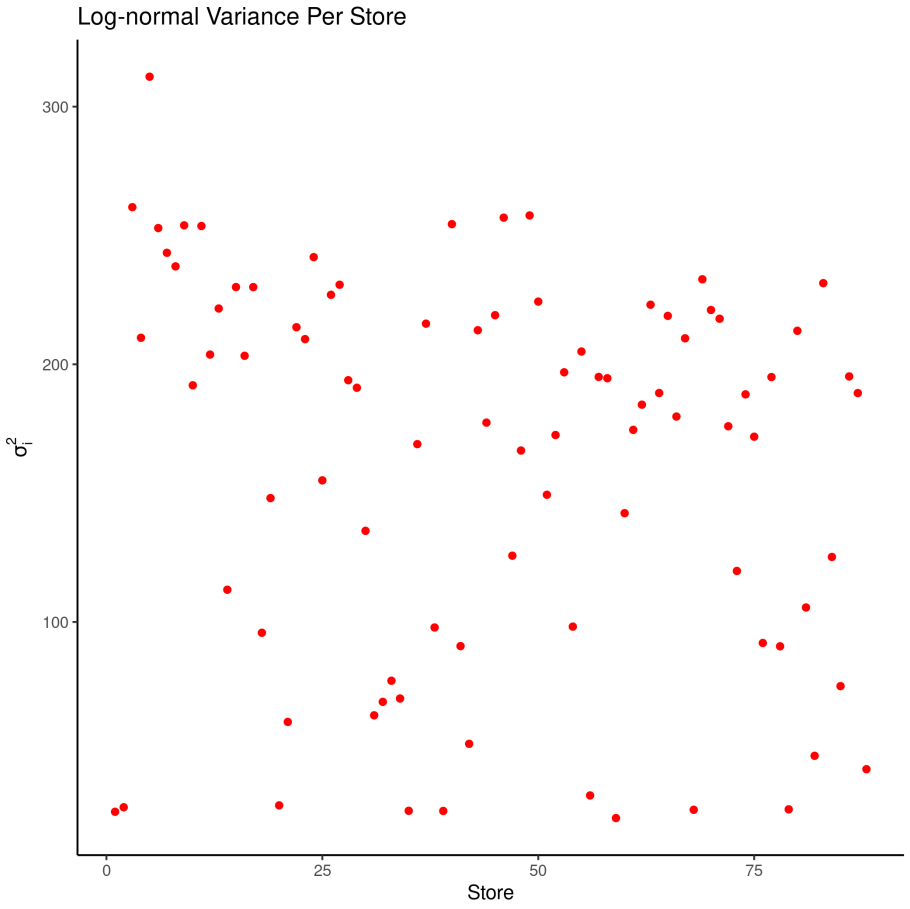

Figure 3: σ_i^2 v. Store.

Finally, let's examine the histograms of the β values. In Figure 4, we obtain histograms for the β coefficients for the four parameters across all stores. We do this because it appears that the β values do not significantly differ from store to store. 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 0.1 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 0.9 for each store. The change in shape for the demand curve given a unit change in price appears to be 0.05 for each store. Finally, the marginal change in shape when advertising appears to be 0.3 per store.

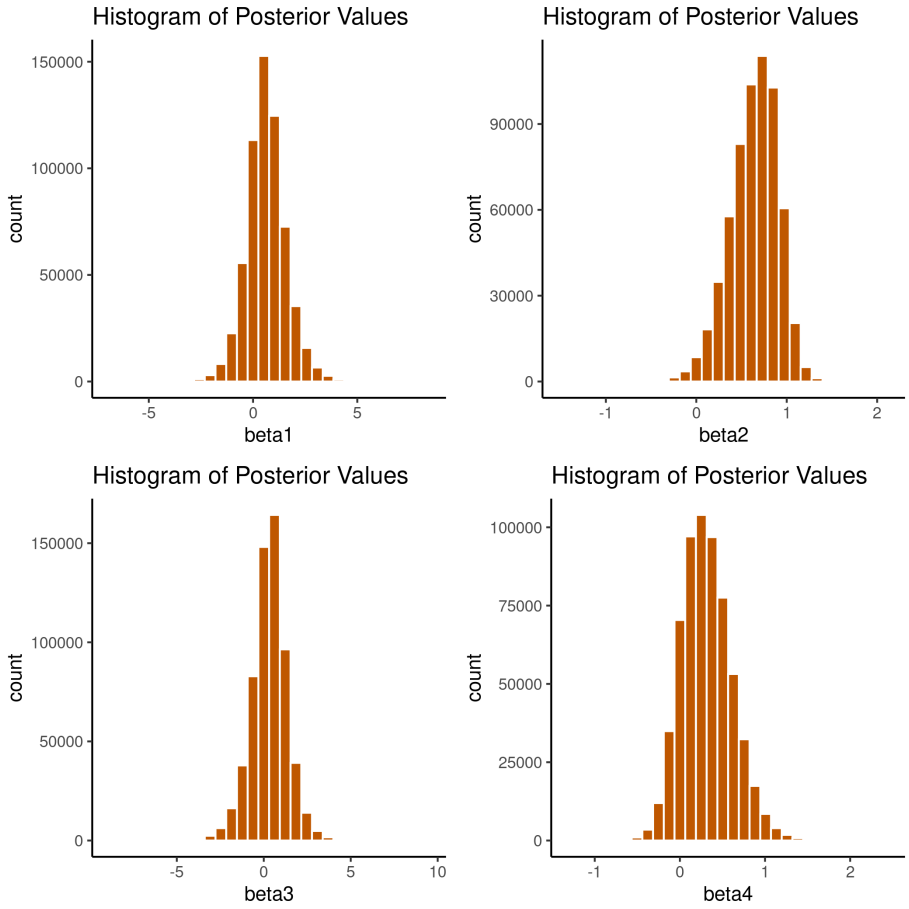

Figure 4: Histogram for β Coefficients.

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.
5. Clearly, we see a case of major model mis-fit in the quite terrible updating of a. 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, as I do not suspect the prior for a to be the main culprit.