

SDS 383D: Exercises 4 – Hierarchical Models

April 10, 2017

Professor Scott

Spencer Woody

Problem 1

Math Tests

We have a model where y_{ij} is the test score of the j th student in school i , with indices $i = 1, 2, \dots, I$ and $j = 1, 2, \dots, N_i$, so N_i is the sample size for school i and there are $N = \sum_{i=1}^I$ total test scores. Let $\lambda = 1/\sigma^2$ and $\gamma = 1/\tau^2$ be the precision parameters. Further, let $y_i = [y_{i1}, y_{i2}, \dots, y_{iN_i}]^T$ and $y = [y_1^T, y_2^T, \dots, y_I^T]^T$ and $\theta = [\theta_1, \theta_2, \dots, \theta_I]^T$. As we can see in Figure 1, schools with smaller sample sizes tend to have more extreme average test scores.

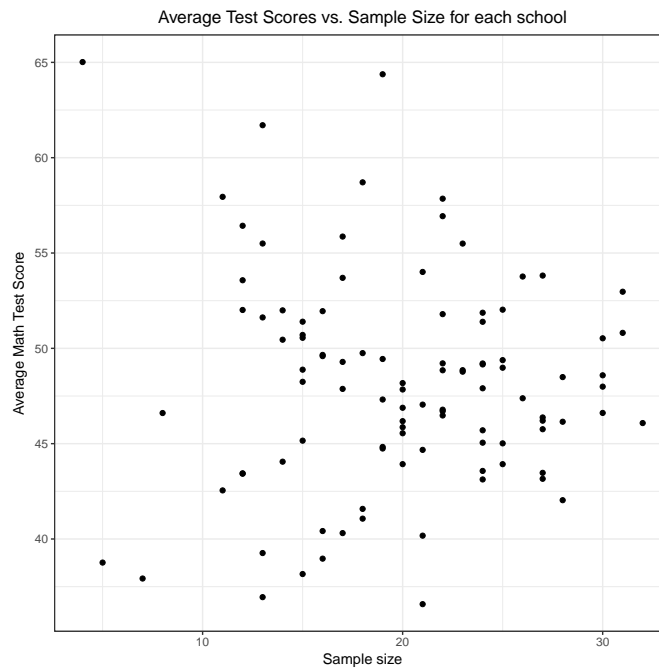


Figure 1: Scatter plot of sample size and average test scores

The hierarchical model for these data is

$$(y_{ij}|\theta_i, \lambda) \sim \mathcal{N}(\theta_i, \lambda^{-1})$$

$$(\theta_i|\mu, \lambda, \gamma) \sim \mathcal{N}(\mu, (\lambda\gamma)^{-1}).$$

We set the priors

$$\pi(\mu) \propto 1, \quad -\infty < \mu < \infty$$

$$\pi(\lambda) \propto \lambda^{-1}, \quad \lambda > 0$$

$$\pi(\gamma) \propto 1, \quad \gamma > 0,$$

that is to say, In order to implement the Gibbs sampler, we need the posterior full conditionals for each θ_i , μ , λ , and γ .

- For each θ_i ,

$$f(\theta_i|y_i, \mu, \lambda, \gamma) \propto f(y_i|\theta_i, \lambda) \cdot f(\theta_i|\mu, \lambda, \gamma)$$

$$\sim \mathcal{N}\left((N_i\lambda + \lambda\gamma)^{-1} \cdot (N_i\lambda\bar{y}_i + \lambda\gamma\mu), (N_i\lambda + \lambda\gamma)^{-1}\right),$$

which we know from the normal-normal conjugacy derived in Exercises 1.

- For μ ,

$$\begin{aligned}
 \pi(\mu|y, \theta, \lambda, \gamma) &\propto f(\theta|\lambda, \gamma, \mu) \cdot \pi(\mu) \\
 &\propto \left(\prod_{i=1}^I \exp \left[-\frac{1}{2} \lambda \gamma (\theta_i - \mu)^2 \right] \right) \cdot 1 \\
 &= \exp \left[-\frac{1}{2} \lambda \gamma \sum_{i=1}^I (\theta_i - \mu)^2 \right] \\
 &= \exp \left[-\frac{1}{2} \lambda \gamma \sum_{i=1}^I (\theta_i^2 - 2\theta_i \mu + \mu^2) \right] \\
 &\propto \exp \left[-\frac{1}{2} \lambda \gamma (I\mu^2 - 2I\bar{\theta}\mu) \right] \\
 &\sim \mathcal{N}(\bar{\theta}, (I\lambda\gamma)^{-1}).
 \end{aligned}$$

- For λ ,

$$\begin{aligned}
 \pi(\lambda|y, \mu, \gamma, \theta) &\propto f(y|\lambda, \theta) \cdot f(\theta|\lambda, \gamma, \mu) \cdot \pi(\lambda) \\
 &\propto \left(\prod_{i=1}^I \prod_{j=1}^{N_i} \lambda^{1/2} \exp \left[-\frac{1}{2} (y_{ij} - \theta_i)^2 \right] \right) \cdot \left(\prod_{i=1}^I \lambda^{1/2} \exp \left[-\frac{1}{2} \lambda \gamma (\theta_i - \mu)^2 \right] \right) \cdot \lambda^{-1} \\
 &= \lambda^{(N+I)/2-1} \exp \left[-\frac{1}{2} \left(\sum_{i=1}^I \sum_{j=1}^{N_i} (y_{ij} - \theta_i)^2 + \gamma \sum_{i=1}^I (\theta_i - \mu)^2 \right) \lambda \right] \\
 &\sim \text{Gamma} \left(\frac{N+I}{2}, \frac{1}{2} \left[\sum_{i=1}^I \sum_{j=1}^{N_i} (y_{ij} - \theta_i)^2 + \gamma \sum_{i=1}^I (\theta_i - \mu)^2 \right] \right).
 \end{aligned}$$

- For γ ,

$$\begin{aligned}
 \pi(\gamma|y, \mu, \lambda, \theta) &\propto f(\theta|\lambda, \gamma, \mu) \cdot \pi(\gamma) \\
 &\propto \left(\prod_{i=1}^I \gamma^{1/2} \exp \left[-\frac{1}{2} \lambda \gamma (\theta_i - \mu)^2 \right] \right) \cdot 1 \\
 &= \gamma^{I/2} \exp \left[-\frac{1}{2} \lambda \sum_{i=1}^I (\theta_i - \mu)^2 \cdot \gamma \right] \\
 &\sim \text{Gamma} \left(\frac{I}{2} + 1, \frac{1}{2} \lambda \sum_{i=1}^I (\theta_i - \mu)^2 \right).
 \end{aligned}$$

Table 1: 95% posterior credible intervals

	2.5%	50%	97.5%
μ	47.03	48.10	49.18
λ	0.0111	0.0118	0.0126
γ	2.43	3.49	5.03

Given the posterior mean $\hat{\theta}_i$ as an estimate of θ_i , define the shrinkage coefficient

$$\kappa_i = \frac{\bar{y}_i - \hat{\theta}_i}{\bar{y}_i},$$

which is a measure incomplete pooling. Figure 2 shows the absolute shrinkage coefficient for each school as a function of sample size. As sample size increases, the shrinkage decreases because we are gaining precision in estimating the school-level mean θ_i .

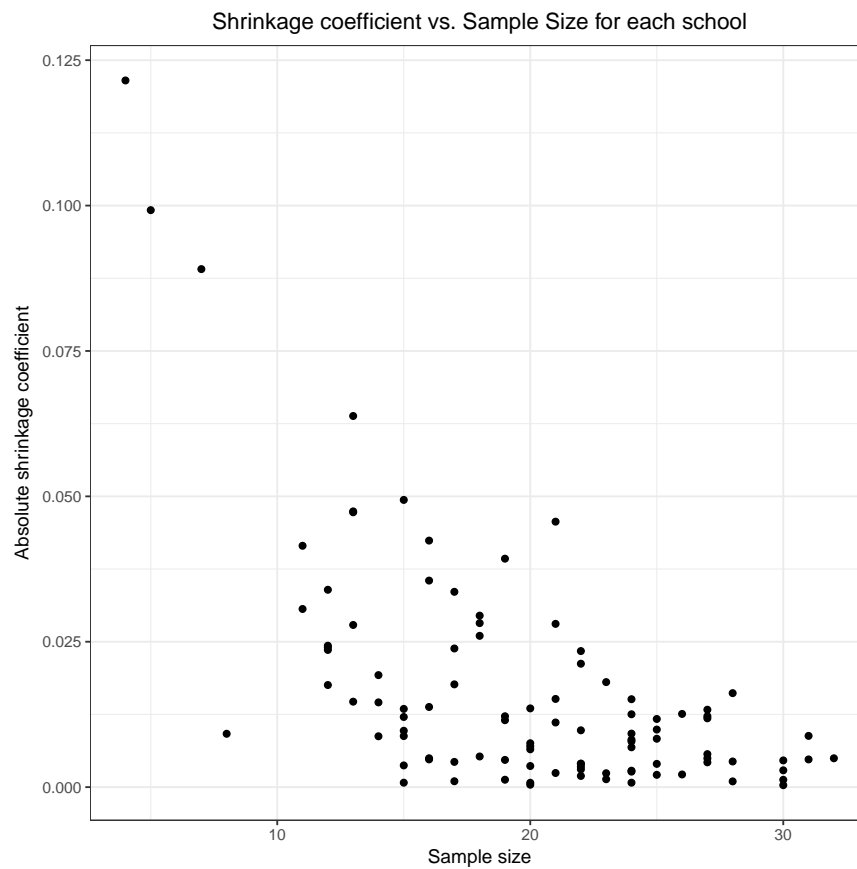


Figure 2: Absolute shrinkage coefficient as a function of sample size

Problem 2

Price elasticity of demand

Here we model the demand curve for cheese, which is given by

$$Q = \alpha P^\beta,$$

where Q is the quantity of cheese demanded, P is price, β is a parameter for the *price elasticity of demand* and α is a (rather unremarkable) scaling parameter. Note that if we take a logarithmic transform of the equation in our demand model, we obtain the linear relationship

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

Figure 3 shows all the data with a fitted OLS line, and Figure 4 shows the data on a store-by-store level with the same OLS line from all data on each panel. The fact that the OLS line performs poorly on any given individual store's data suggests that a hierarchical approach would be beneficial. The hierarchical linear model for the quantity of cheese sold for the t th observation at store i is

$$y_{it} = \alpha_i + \beta_i x_{it} + \gamma_i z_{it} + \theta_i z_{it} x_{it} + \epsilon_{it},$$

where x_{it} is the log-price of cheese and z_{it} is an indicator variable taking on a value of 1 when the display is shown, and 0 otherwise.

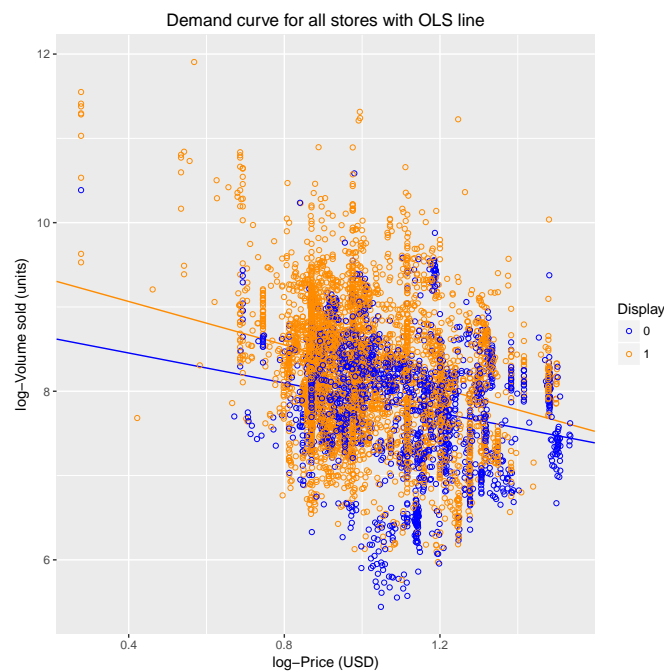


Figure 3: Scatterplot for data from all stores with OLS line

Using frequentist REML to build this model we obtain these results,

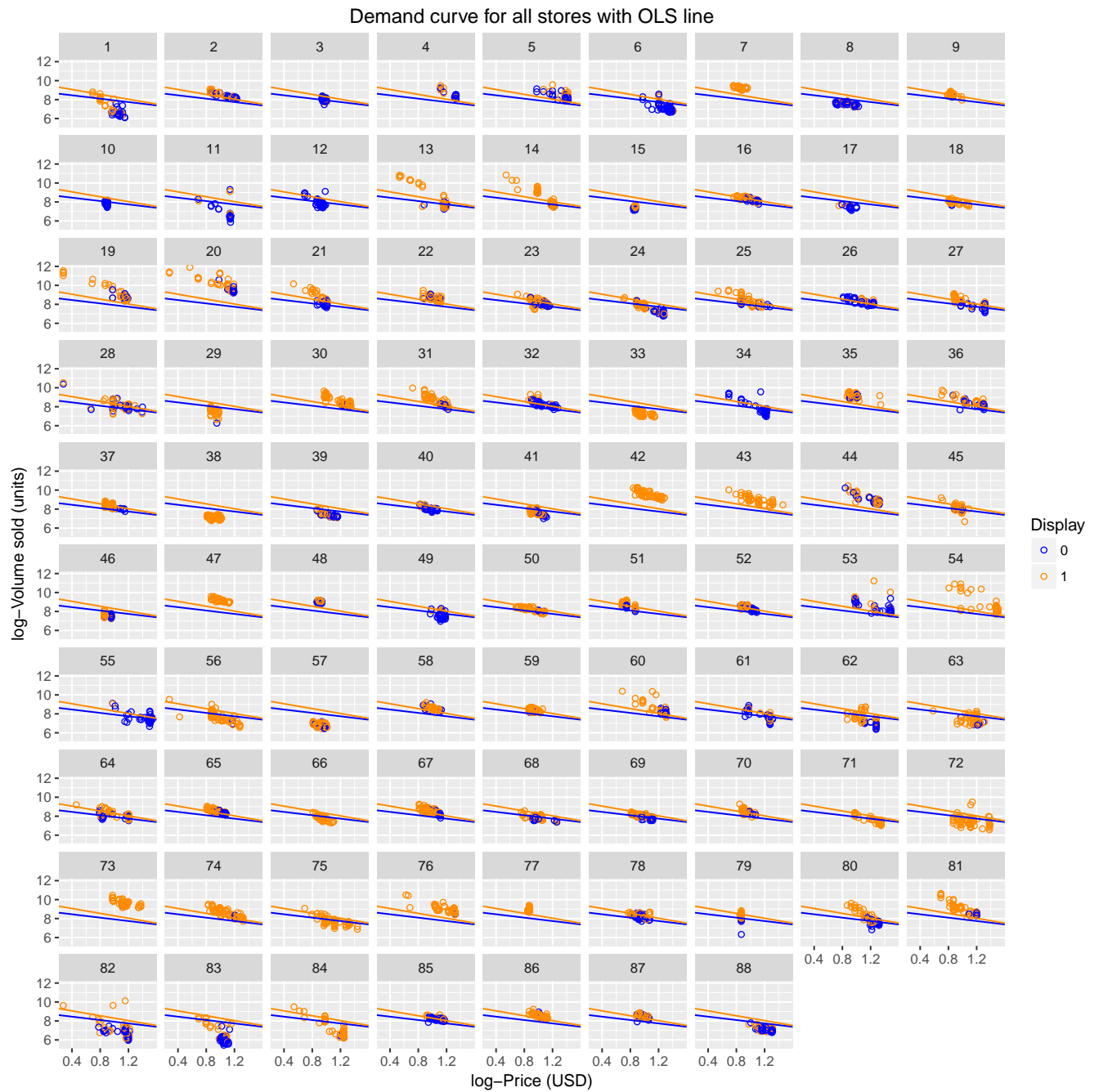


Figure 4: Scatterplot for data from all stores with OLS line

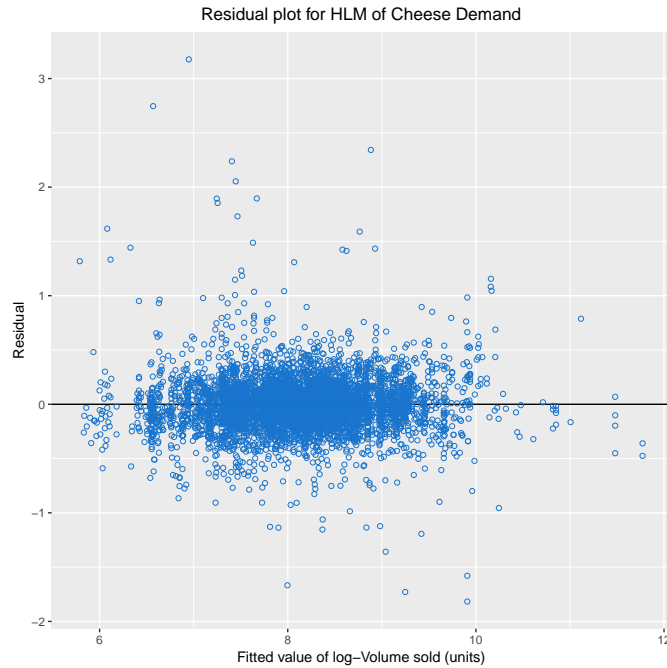


Figure 5: Residual plot using HLM and REML method

*Full Bayesian***Model specification**

Here we specify a general Bayesian hierarchical linear model. Let y_i be a n_i -length vector representing the responses of group i . There are $N = \sum_i^I n_i$ total responses. X_i is the $n_i \times p$ design matrix for the observations in group i , and Z_i is a $n_i \times q$, $q \leq p$ matrix whose columns are a subset of the columns of X_i , and this represents the subject-level effects, sometimes called “random effects.”. Then the responses y_i are distributed as:

$$y_i | \beta, b_i, \lambda \sim \mathcal{N}_{n_i}(X_i \beta + Z_i b_i, \lambda^{-1} \mathcal{I}_{n_i})$$

$$b_i | D \stackrel{\text{iid}}{\sim} \mathcal{N}_q(0, D)$$

Note that the responses y_{it} for subject i are therefore assumed to iid, and also note two results of this model,

$$E(y_i | b_i) = X_i \beta + Z_i b_i$$

$$E(y_i) = E(E(y_i | b_i)) = X_i \beta,$$

or in other words, The priors are

$$\pi(\lambda) \propto \lambda^{-1}$$

$$\pi(\beta) \propto 1$$

$$\pi(D) \sim \text{IW}(\nu, \Psi).$$

To implement a Gibbs sampler, we need the full conditional posterior distributions for b_i , λ , β , and D .

- For each b_i , first define $v_i := y_i - X_i\beta$,

$$\begin{aligned}
 p(b_i|y_i, \lambda, \beta, D) &\propto p(y_i|\beta, b_i, \lambda)p(b_i|D) \\
 &\propto \exp\left[-\frac{1}{2}\lambda(y_i - X_i\beta - Z_i b_i)^T(y_i - X_i\beta - Z_i b_i)\right] \cdot \exp\left[-\frac{1}{2}b_i^T D^{-1} b_i\right] \\
 &= \exp\left[-\frac{1}{2}\lambda(Z_i b_i - v_i)^T(Z_i b_i - v_i)\right] \cdot \exp\left[-\frac{1}{2}b_i^T D^{-1} b_i\right] \\
 &\propto \exp\left[-\frac{1}{2}b_i^T (\lambda Z_i^T Z_i + D^{-1}) b_i - 2b_i^T \lambda Z_i^T v_i\right] \\
 &\propto \exp\left[-\frac{1}{2}\left(b_i - [\lambda Z_i^T Z_i + D^{-1}]^{-1} \lambda Z_i^T v_i\right)^T (\lambda Z_i^T Z_i + D^{-1}) \left(b_i - [\lambda Z_i^T Z_i + D^{-1}]^{-1} \lambda Z_i^T v_i\right)\right] \\
 &\sim \mathcal{N}\left([\lambda Z_i^T Z_i + D^{-1}]^{-1} \lambda Z_i^T v_i, [\lambda Z_i^T Z_i + D^{-1}]^{-1}\right) \\
 &\sim \mathcal{N}\left([\lambda Z_i^T Z_i + D^{-1}]^{-1} \lambda Z_i^T (y_i - X_i\beta), [\lambda Z_i^T Z_i + D^{-1}]^{-1}\right).
 \end{aligned}$$

- For λ ,

$$\begin{aligned}
 \pi(\lambda|y, \beta, b) &\propto p(y|\lambda, \beta) \cdot \pi(\lambda) \\
 &= \left(\prod_{i=1}^I \lambda^{n_i/2} \exp\left[-\frac{1}{2}\lambda(y_i - X_i\beta - Z_i b_i)^T(y_i - X_i\beta - Z_i b_i)\right]\right) \cdot \lambda^{-1} \\
 &\sim \text{Gamma}\left(\frac{N}{2}, \frac{1}{2} \sum_{i=1}^I \|y_i - X_i\beta - Z_i b_i\|_2^2\right)
 \end{aligned}$$

- For β , define $w_i := y_i - Z_i b_i$.

$$\begin{aligned}
 \pi(\beta|y, \lambda, b) &\propto p(y|\lambda, \beta) \cdot \pi(\beta) \\
 &\propto \left(\prod_{i=1}^I \exp\left[-\frac{1}{2}\lambda(y_i - X_i\beta - Z_i b_i)^T(y_i - X_i\beta - Z_i b_i)\right]\right) \cdot 1 \\
 &= \prod_{i=1}^I \exp\left[-\frac{1}{2}\lambda(X_i\beta - w_i)^T(X_i\beta - w_i)\right] \\
 &\propto \prod_{i=1}^I \exp\left[-\frac{1}{2}\lambda\left(\beta^T X_i^T X_i \beta - 2\beta^T X_i^T w_i\right)\right] \\
 &= \exp\left(-\frac{1}{2}\lambda\left[\beta^T \left(\sum_{i=1}^I X_i^T X_i\right) \beta - 2\beta^T \sum_{i=1}^I X_i^T w_i\right]\right) \\
 &= \exp\left(-\frac{1}{2}\lambda\left[\beta^T \left(\sum_{i=1}^I X_i^T X_i\right) \beta - 2\beta^T \sum_{i=1}^I X_i^T (y_i - Z_i b_i)\right]\right) \\
 &\sim \mathcal{N}\left(\left[\sum_{i=1}^I X_i^T X_i\right]^{-1} \sum_{i=1}^I X_i^T (y_i - Z_i b_i), \left[\lambda \sum_{i=1}^I X_i^T X_i\right]^{-1}\right).
 \end{aligned}$$

- For D ,

$$\begin{aligned}\pi(D|b) &\propto p(b|D) \cdot \pi(D) \\ &\propto \left(\prod_{i=1}^I [\det(D)]^{-1/2} \exp \left[-\frac{1}{2} b_i^T D^{-1} b_i \right] \right) \cdot [\det(D)]^{-\frac{\nu+q+1}{2}} \exp \left[-\frac{1}{2} \text{tr}(\Psi D^{-1}) \right] \\ &\sim \text{IW} \left(I + \nu, \Psi + \sum_{i=1}^I b_i b_i^T \right)\end{aligned}$$

The most computationally intensive part of this Gibbs sampler scheme is sampling each b_i , and I chose to do this by exploiting a block-diagonal matrix of each Z_i and drawing each b_i simultaneously as a long vector called b . For this application specifically, the X_i and Z_i are identical, with a column of 1's for the intercept, a column of log-prices, a column of indicator variables for display, and a column of interaction terms for log-price and display. We run 6000 iterations of the Gibbs sampler with the first 1000 draws discarded as burn-in. The `mix` folder within the `img` folder shows traceplots of λ , each component in β , and four randomly selected columns of posterior draws of b , which all show a good degree of mixing. Histograms for λ and each component of β are shown below. Figure 8 shows a grid of plots, each of which has 95% credible intervals of all the subject-level effects on a given covariate terms, arranged in increasing order by posterior median. Note that on the x -axis is different for each plot in order to have each one ordered by posterior median.

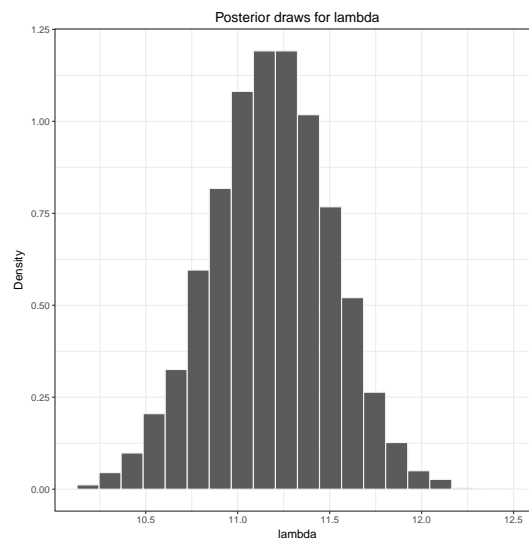


Figure 6: Histogram of posterior draws of λ

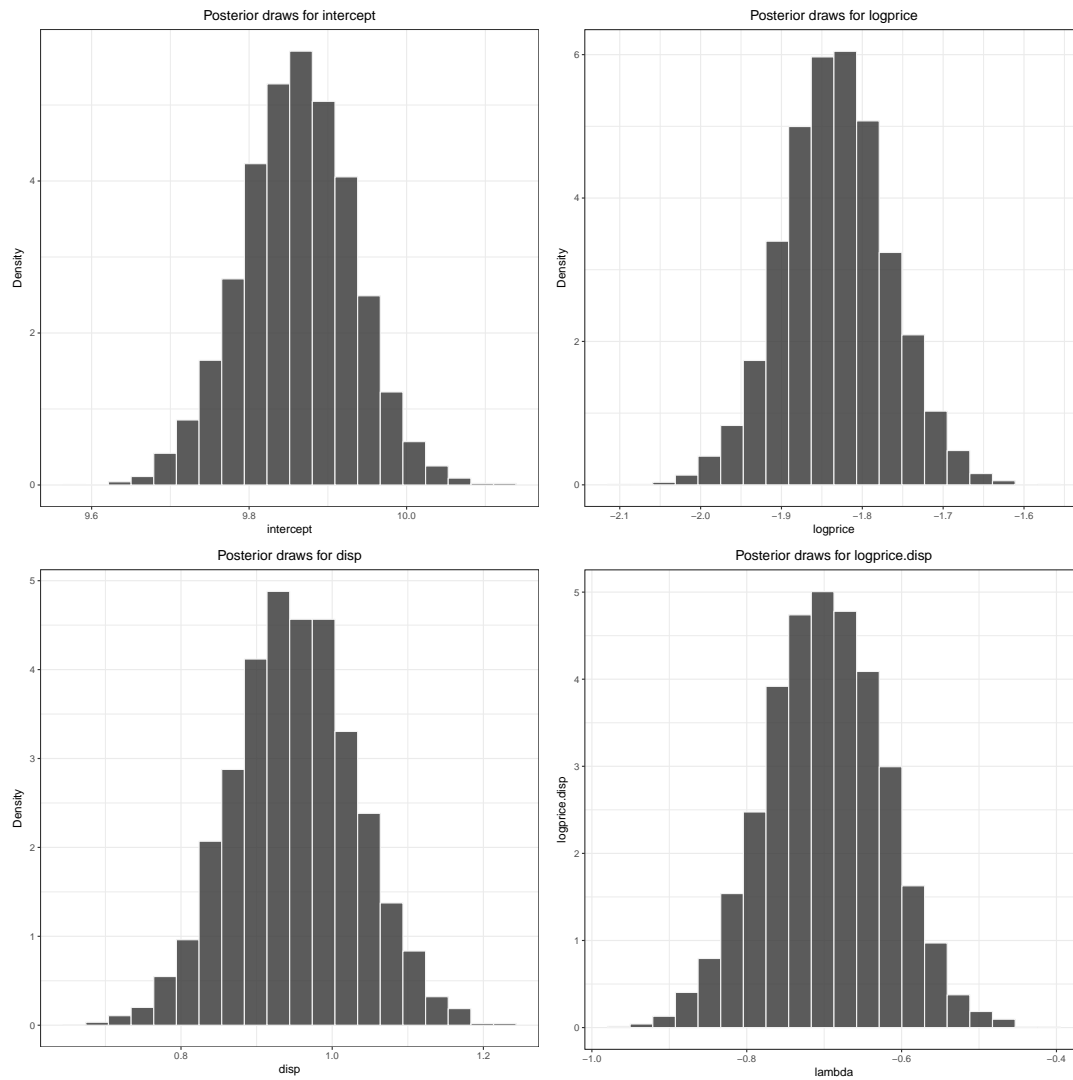


Figure 7: Histogram of posterior draws of each term in β

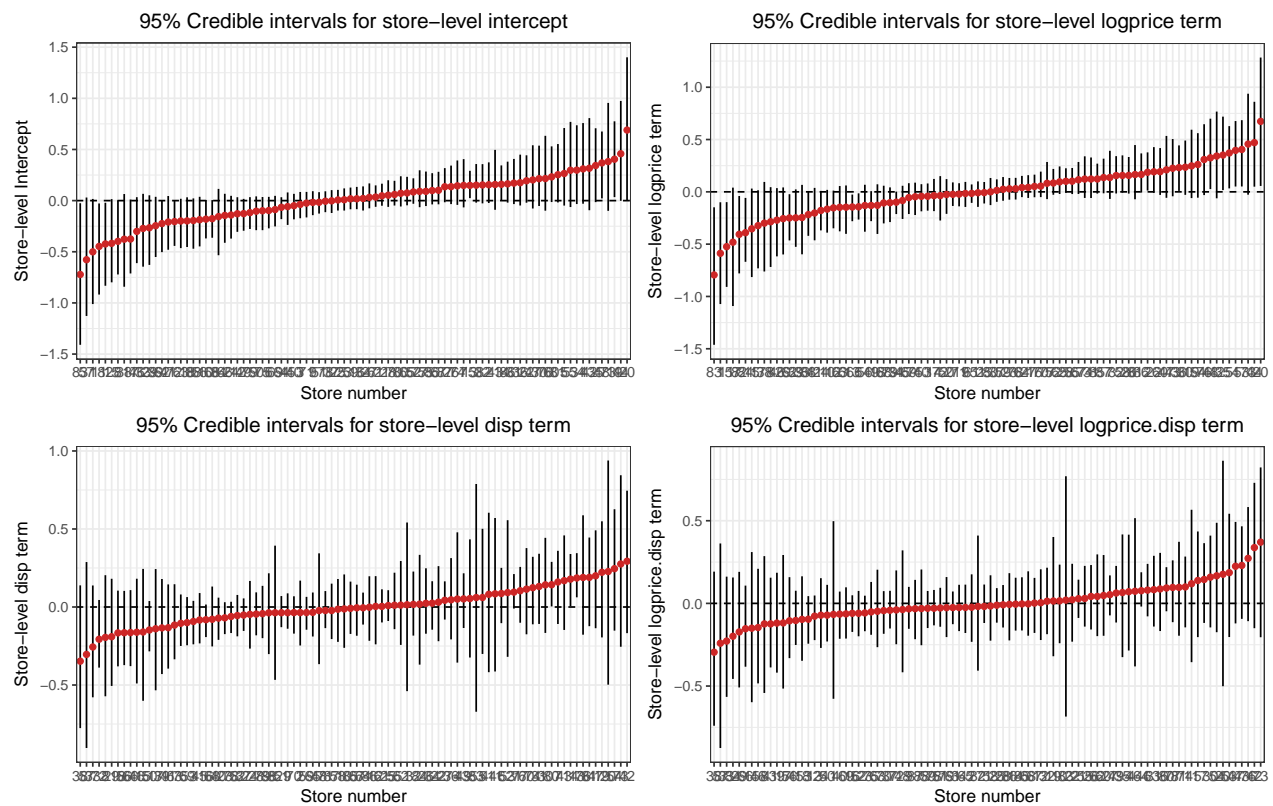


Figure 8: Ordered 95% credible intervals of store-level each store

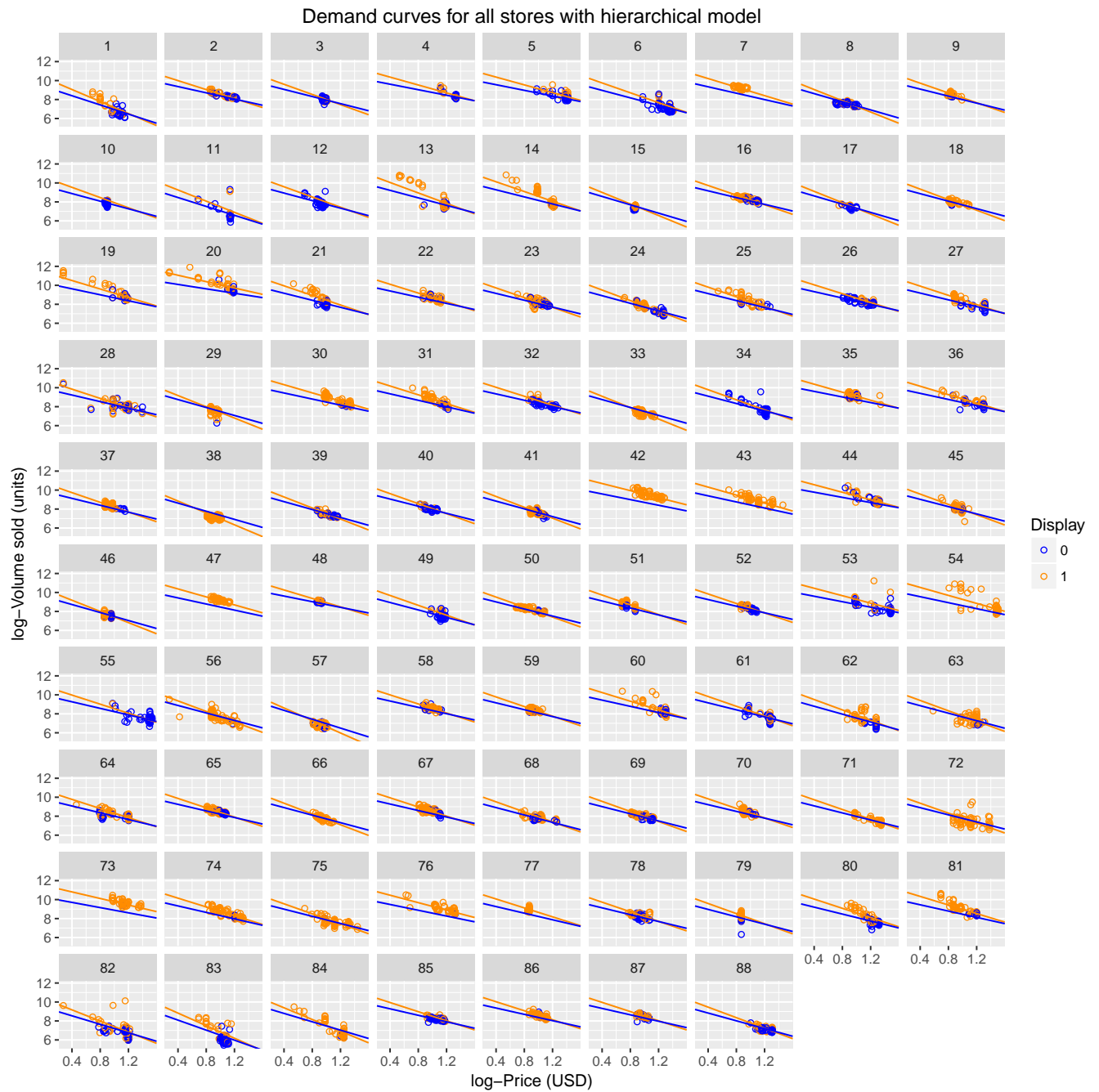


Figure 9: Each store's demand curves with fitted line from Bayesian hierarchical model

Problem 3

A hierarchical probit model via data augmentation

For this model we model y_{ij} , the j th binary 0-1 response, $j \in \{1, 2, \dots, n_i\}$, within group $i \in \{1, 2, \dots, I\}$ through the utilization of data augmentation whereby we introduce a latent variable z_{ij} ,

$$(z_{ij}|\beta, \gamma_i) \sim N(x_{ij}^T\beta + w_{ij}^T\gamma_i, 1)$$

$$y_{ij} = \mathbf{1}(z_{ij} > 0) = \begin{cases} 1 & \text{if } z_{ij} > 0 \\ 0 & \text{if } z_{ij} \leq 0 \end{cases},$$

where x_{ij} is a vector of covariate features and w_{ij} is a subset of these features whose effects vary at the subject level, captured through γ_i . We can see that this implies a probit link function so that

$$p_{ij} = P(y_{ij} = 1) = \Phi(x_{ij}^T\beta + w_{ij}^T\gamma_i),$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution. Let z_i be the n_i -length vector of responses from subject i , and similarly, X_i is a $n_i \times p$ design matrix of subject i and W_i is a $n_i \times q$ design matrix with $q \leq p$ whose columns are a subset of the columns of X_i , so then we see that

$$(z_i|\beta, \gamma_i) \sim \mathcal{N}_{n_i}(X_i\beta + W_i\gamma_i, \mathcal{I}_{n_i}), \quad (1)$$

and furthermore we model the subject-level responses as coming from a multivariate normal distribution

$$\gamma_i \stackrel{\text{iid}}{\sim} \mathcal{N}_q(0, D)$$

where D is some $q \times q$ covariance matrix.