SDS 383D Exercises 5: Hierarchical Linear Models

1 Price elasticity of demand

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.

Answer

My hierarchical model is:

$$
\begin{aligned}
p(\mu) &\propto 1 \\
p(s) &\propto 1/s \\
log\alpha_i \mid \mu, s &\sim \mathcal{N}(\mu, s) \\
log\delta_i \mid s &\sim \mathcal{N}(0, s) \\
m &\sim \mathcal{N}(m_0, v_0) \\
p(v) &\propto 1/v \\
\beta_i \mid m, v &\sim \mathcal{N}(m, v) \\
\lambda_i &\sim \mathcal{N}(0, v) \\
\tau &\sim \text{Inv-Gamma}(a_0, b_0) \\
logQ_{ij} \mid \lambda, k, \alpha_i, \beta_i, P_{ij}, d_{ij} &\sim \mathcal{N}(\log(\alpha_i + \delta d_{ij}) + (\beta_i + \lambda d_{ij}) \log P_{ij}, \tau s)
\end{aligned}
$$

where $m_0 = 0$, $v_0 = 100$, $a_0 = 0.1$ and $b_0 = 1$.

Denote y_{ij} by $\log Q_{ij}$, $\log P_{ij}$ by x_{ij}, $\log \alpha_i$ by a_i and $\log \delta_i$ by θ_i. The posterior distribution can be written as:

$$
\begin{aligned}
p(\mu, s, a, d, m, v, \beta, \lambda, \tau, X) \propto s^{-1} &\times \prod_{i=1}^{k} s^{-\frac{1}{2}} \exp\left(-s^{-1} \frac{(a_i - \mu)^2}{2}\right) \times \prod_{i=1}^{k} s^{-\frac{1}{2}} \exp\left(-s^{-1} \frac{\theta_i^2}{2}\right) \\
&\times \exp\left(\frac{-(m - m_0)^2}{2v_0}\right) \times v^{-1} \times \prod_{i=1}^{k} v^{-\frac{1}{2}} \exp\left(v^{-1} \frac{(\beta_i - m)^2}{2}\right) \\
&\times \prod_{i=1}^{k} v^{-\frac{1}{2}} \exp\left(-v^{-1} \frac{\lambda_i^2}{2}\right) \times \tau^{-a_0 - 1} \exp(-\tau^{-1} b_0) \\
&\times \prod_{i=1}^{k} \prod_{j=i}^{n_i} (\tau s)^{-\frac{1}{2}} \exp\left(-(\tau s)^{-1} \frac{(y_{ij} - a_i - \theta d_{ij} - (\beta_i + \lambda d_{ij}) x_{ij})^2}{2}\right)
\end{aligned}
$$

The complete conditionals are:

$$
\begin{aligned}
\mu \mid \cdots &\sim \mathcal{N}\left(\bar{a}, s\right) \\
s \mid \cdots &\sim \text{Inv-Gamma}\left(\frac{2k+N}{2}, \frac{S_a + \sum_{i=1}^{k} \theta_i^2}{2} + \frac{S_y}{2\tau}\right) \\
a_i \mid \cdots &\sim \mathcal{N}\left(\frac{\tau\mu + n_i \bar{y}_a^{(i)}}{\tau + n_i}, s\left(1 + \frac{n_i}{\tau}\right)^{-1}\right) \\
\theta_i \mid \cdots &\sim \mathcal{N}\left(\frac{N_d^{(i)} \bar{y}_\theta^{(i)}}{\tau + N_d^{(i)}}, s\left(1 + \frac{N_d^{(i)}}{\tau}\right)^{-1}\right) \\
m \mid \cdots &\sim \mathcal{N}\left(\frac{\frac{m_0}{v_0} + \frac{k\bar{\beta}}{v}}{\frac{1}{v_0} + \frac{k}{v}}, \left(\frac{1}{v_0} + \frac{k}{v}\right)^{-1}\right) \\
v \mid \cdots &\sim \text{Inv-Gamma}\left(k, \frac{S_\beta + \sum_{i=1}^{k} \lambda_i^2}{2}\right) \\
\beta_i \mid \cdots &\sim \mathcal{N}\left(\frac{\frac{m}{v} + \frac{S_x^{(i)} \bar{y}_\beta^{(i)}}{\tau s}}{\frac{1}{v} + \frac{S_x^{(i)}}{\tau s}}, \left(\frac{1}{v} + \frac{S_x^{(i)}}{\tau s}\right)^{-1}\right) \\
\lambda_i \mid \cdots &\sim \mathcal{N}\left(\frac{\frac{S_x^{*(i)} \bar{y}_\lambda^{(i)}}{\tau s}}{\frac{1}{v} + \frac{S_x^{*(i)}}{\tau s}}, \left(\frac{1}{v} + \frac{S_x^{*(i)}}{\tau s}\right)^{-1}\right) \\
\tau \mid \cdots &\sim \text{Inv-Gamma}\left(a_0 + \frac{N}{2}, b_0 + \frac{S_y}{2s}\right)
\end{aligned}
$$

where $\bar{a} = \sum_{i=1}^{k} a_i/k$, $S_a = \sum_{i=1}^{k}(a_i - \mu)^2$, $\hat{y}_{ij} = a_i + \theta d_{ij} + (\beta_i + \lambda d_{ij})x_{ij}$, $S_y = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_{ij})^2$, $\bar{y}_a^{(i)} = \sum_{j=1}^{n_i}(y_{ij} - \hat{y}_{ij} + a_i)/n_i$, $\bar{y}_\theta^{(i)} = \sum_{j=1}^{n_i}(y_{ij} - \hat{y}_{ij} + \theta d_{ij})1(d_{ij} = 1)/N_d^{(i)}$, $N_d^{(i)} = \sum_{j=1}^{n_i} 1(z_{ij} = 1)$, $\bar{\beta} = \sum_{i=1}^{k} \beta_i/k$, $S_\beta = \sum_{i=1}^{k}(\beta_i - m)^2$, $\bar{y}_\beta^{(i)} = \sum_{j=1}^{n_i}(y_{ij} - \hat{y}_{ij} + \beta_i x_{ij})x_{ij}/S_x^{(i)}$, $S_x^{(i)} = \sum_{j=1}^{n_i} x_{ij}^2$, $\bar{y}_\lambda^{(i)} = \sum_{j=1}^{n_i}(y_{ij} - \hat{y}_{ij} + \lambda d_{ij} x_{ij})x_{ij} 1(d_{ij} = 1)/S_x^{*(i)}$ and $S_x^{*(i)} = \sum_{j=1}^{n_i} x_{ij}^2 1(d_{ij} = 1)$.

The scatterplot of log sales volume vs log price is shown in Figure 1. The scatterplot grouped by each store with corresponding fitted lines for display or not are shown in Figure 2. The variance of $y = logQ$ is 0.64 while the variance of the residuals is 0.068. It means our fit is able to catch about 89.4% of variation in the response variable, which indicates good fit. As for whether in-store display can affect the the demand is still hard to decide because the price is strongly correlated with the display indicator. Such correlation can be partly explained by noticing that stores tends to put the target goods on sale the advertising period.

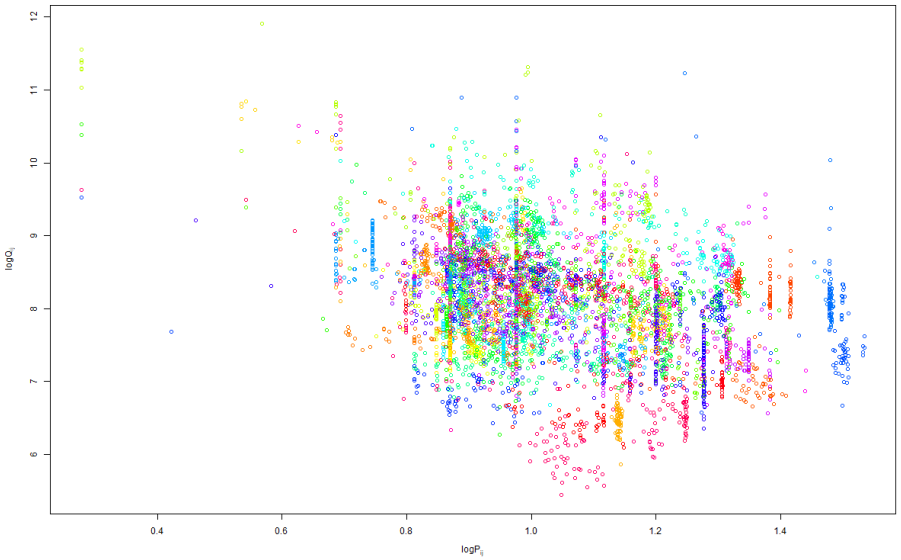

Figure 1: Sales volume vs Price in log scale

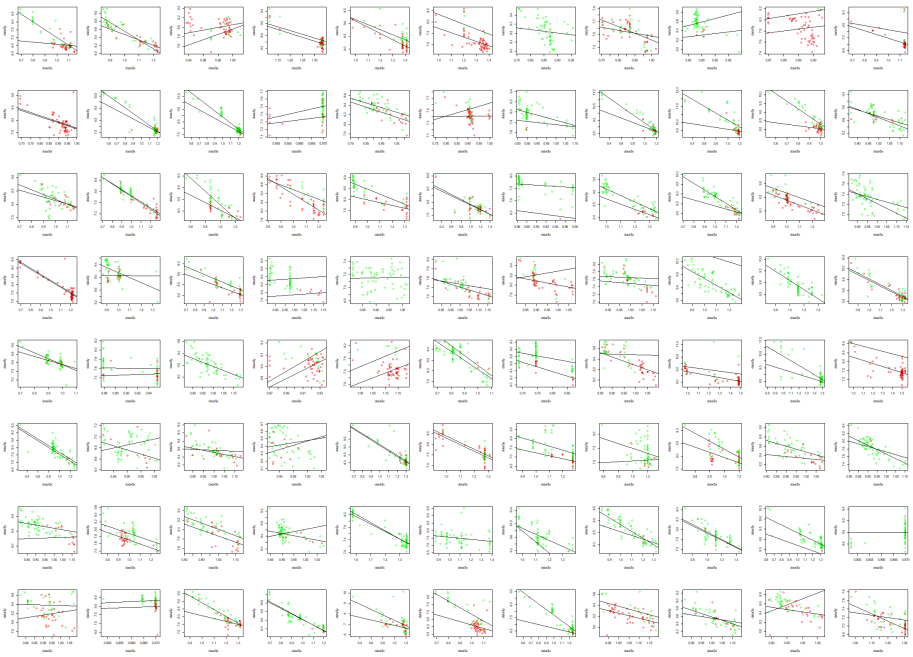

Figure 2: Scatterplot by stores

2 A hierarchical probit model

Read the following paper (or read some distillation of the paper in a book/blog post/slides/etc).

> "Bayesian Analysis of Binary and Polychotomous Response Data." James H. Albert and Siddhartha Chib. *Journal of the American Statistical Association*, Vol. 88, No. 422 (Jun., 1993), pp. 669-679

The paper describes a Bayesian treatment of probit regression (similar to logistic regression) using the trick of *data augmentation*—that is, introducing "latent variables" that turn a hard problem into a much easier one. Briefly summarize your understanding of the key trick proposed by this paper. Then see you if you can apply the trick in the following context, which is more complex than ordinary probit regression.

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:

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

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.

Notes:

1. There are severe imbalances among the states in terms of numbers of survey respondents. Following the last problem, the key is to impose a hierarchical prior on the state-level parameters.

2. The data-augmentation trick from the Albert and Chib paper above is explained in many standard references on Bayesian analysis. If you want to get a quick introduction to the idea, you can consult one of these. A good presentation is in Section 8.1.1 of "Bayesian analysis for the social sciences" by Simon Jackman, available as an ebook through lib.utexas.edu.

3. You are welcome to use the logit model instead of the probit model. If you do this, you'll need to read the following paper, rather than Albert and Chib: Polson, N.G., Scott, J.G. and Windle, J. (2013). Bayesian inference for logistic models using Polya-Gamma latent variables. J. Amer. Statist. Assoc. 108 1339–1349. You can find a routine for simulation Polya-Gamma random variables in the BayesLogit R package and the pypolyagamma python library.

Answer

Data Augmentation

$$
\begin{aligned}
y_{ij} \mid u_{ij} &= I(u_{ij} > 0) \\
u_{ij} \mid z_{ij} &\sim \mathcal{N}(z_{ij}, 1) \\
z_{ij} &= \mu_i + x_{ij}^T \beta_i \\
p(\mu_i) &\propto 1 \\
\beta_i \mid \Sigma &\sim \mathcal{N}(0, \Sigma) \\
p(\Sigma) &\propto 1/\det(\Sigma)
\end{aligned}
$$

where Σ is a diagonal matrix with the form of $\text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$.

The posterior distribution is proportional to

$$
\begin{aligned}
p(\mathbf{u}, \beta \mid \mathbf{y}) \propto \prod_{i=1}^{k} \prod_{j=1}^{n_i} [I(u_{ij} > 0)]^{y_{ij}} [1 - I(u_{ij} \leq 0)]^{1-y_{ij}} \exp\left(-\frac{1}{2}(u_{ij} - z_{ij})^2\right) \\
\times |\Sigma|^{-1} \prod_{i=1}^{k} \times |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\beta_i^T \Sigma^{-1} \beta_i\right)
\end{aligned}
$$

The complete conditionals are

$$
\begin{aligned}
\beta_i \mid \cdots &\sim \mathcal{N}(S_{\beta_i}^{-1} b_{\beta_i}, S_{\beta_i}^{-1}) \\
\mu_i \mid \cdots &\sim \mathcal{N}(b_{\mu_i}/n_i, 1/n_i) \\
u_{ij} \mid y_{ij} = 1, \cdots &\sim \mathcal{N}(z_{ij}, 1) I(u_{ij} > 0) \\
u_{ij} \mid y_{ij} = 0, \cdots &\sim \mathcal{N}(z_{ij}, 1) I(u_{ij} \leq 0) \\
\sigma_l^2 \mid \cdots &\sim \text{Inv-Gamma}\left(\frac{k}{2}, \frac{S_\beta^{(l)}}{2}\right)
\end{aligned}
$$

$S_{\beta_i} = \Sigma^{-1} + \sum_{j=1}^{n_i} x_{ij} x_{ij}^T$, $b_{\beta_i} = \sum_{j=1}^{n_i} (u_{ij} - \mu_i) x_{ij}$, $b_{\mu_i} = \sum_{j=1}^{n_i} u_{ij} - x_{ij}^T \beta_i$ and $S_\beta^{(l)} = \sum_{i=1}^{k} \beta_{il}^2$.

After 1000 iterations (including burn-in stage of 200 iterations), the posterior estimates give an accuracy of 63.7% which is slightly higher than its Frequentist's counterpart, i.e. Logistic regression model of 63.0%.