Amber Day Spring 22

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 misfit?

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

Proposed Model: $y_{ij} = x_{ij}^T \beta_i z_i + \varepsilon_{ij}$

y_{ij}: log of volume (Q_{ij})

$z_i = \begin{cases} 1 & \text{data is from store } i \\ 0 & \text{otherwise} \end{cases}$

x_{ij}^T: [I P D PD] covariates of model for observation ij

I: intercept; P: ln(price); D: display; PD: interaction of P and D

β_i: Store dependent coefficients

ε: Error

$i = 1, \ldots, 88$ store index

$j = 1, \ldots, n_i$ observation index

number of observations for store i, $\sum_{i=1}^{88} n_i = N = 5555$

Model Specification:

$y_{ij} \sim N(x_{ij}^T \beta_i, \sigma_i^2)$

$\beta_i \sim N(\mu, \Sigma)$

* In each regression the features compete to explain the y_{ij}. So the parameters describing the features are assumed to be correlated. That's why it makes sense for us to put a MVN prior on the β_i's. At the end of the day we expect these β_i's to be correlated.

Priors:

$\sigma_i^2 \propto \frac{1}{\sigma_i^2}$

* Alternative suggestion $\sigma_i^2 \sim IG(\frac{\nu}{2}, \frac{\nu}{2})$ where $\frac{\nu}{2}, \frac{\nu}{2}$ have priors

* Alternative suggestion $\mu \sim N(\mu_0, \text{diag}(\hat{\sigma}_1, \ldots, \hat{\sigma}_4))$ where $\hat{\sigma}_j \sim IG(a_0, b_0)$ set to $\frac{1}{2}$ or have half-cauchy

$\mu \sim N(\mu_0, I)$

coefficients from a normal lin reg on all of the data $y_{ij} = x_{ij}^T \beta + \varepsilon$

$\Sigma \sim \text{InvWish}(4, I)$

* Prior of convenience, conjugate to variance matrix (Σ).

Conditionals:

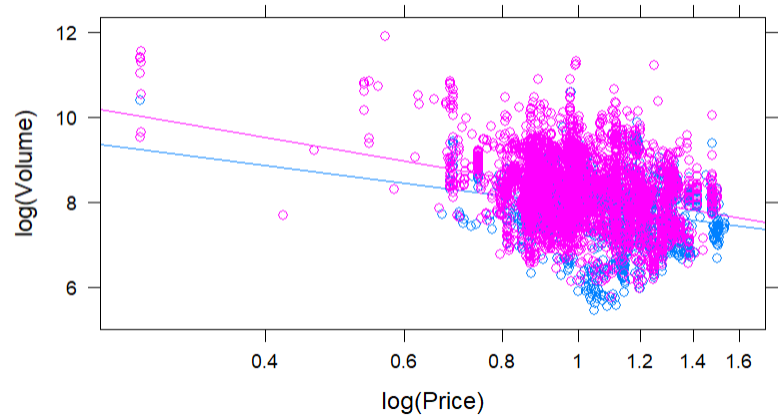
$\sigma^2 | \ldots \sim IG\left(\frac{N}{2}, \frac{1}{2} \sum_{i=1}^{88} \sum_{j=1}^{n_i} (y_{ij} - x_{ij}' \beta_i)\right)$

$\theta | \ldots \sim N\left((I + 88\Sigma^{-1})^{-1} (\theta_0 + \Sigma^{-1} \sum_{i=1}^{88} \beta_i), (I + 88\Sigma^{-1})^{-1}\right)$

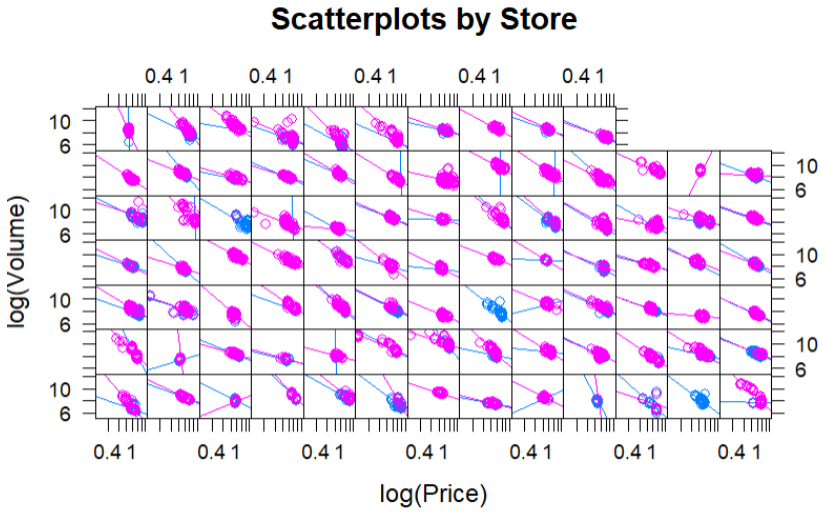
$\Sigma | \ldots \sim \text{InvWish}\left(I + \sum_{i=1}^{88} (\beta_i - \theta)(\beta_i - \theta)', 92\right)$

$\beta_i | \ldots \sim N\left((\Sigma^{-1} + \sigma^2 \sum_{j=1}^{n_i} x_{ij} x_{ij}')^{-1} (\Sigma^{-1}\theta + \sigma^2 \sum_{j=1}^{n_i} y_{ij} x_{ij}), (\Sigma^{-1} + \sigma^2 \sum_{j=1}^{n_i} x_{ij} x_{ij}')^{-1}\right)$

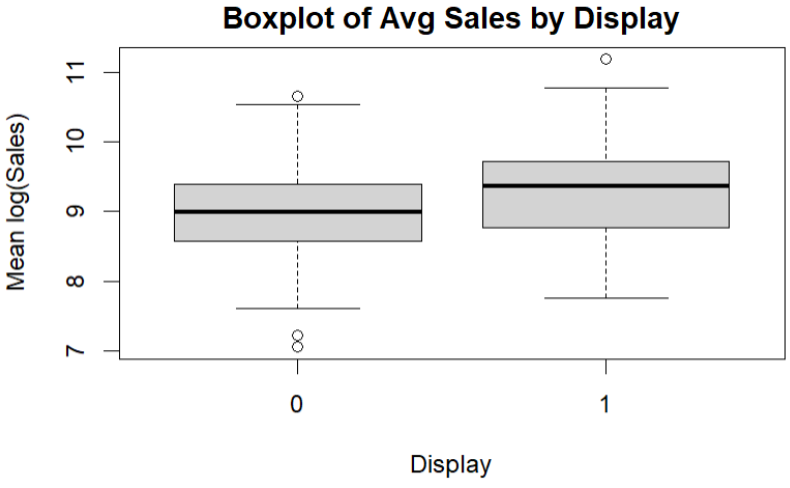
Exploring the data:



$$color = \begin{cases} pink\ has\ display \\ blue\ has\ no\ display \end{cases}$$

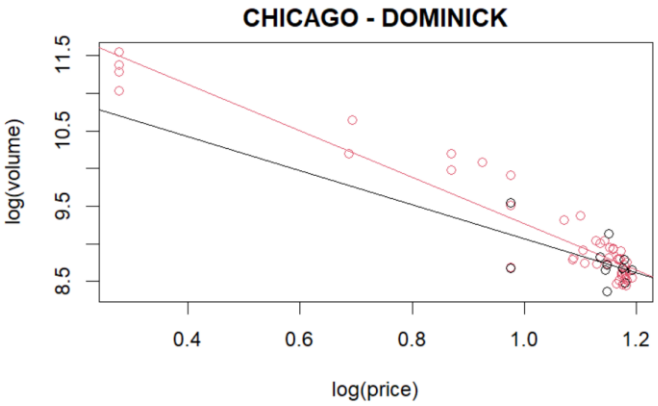


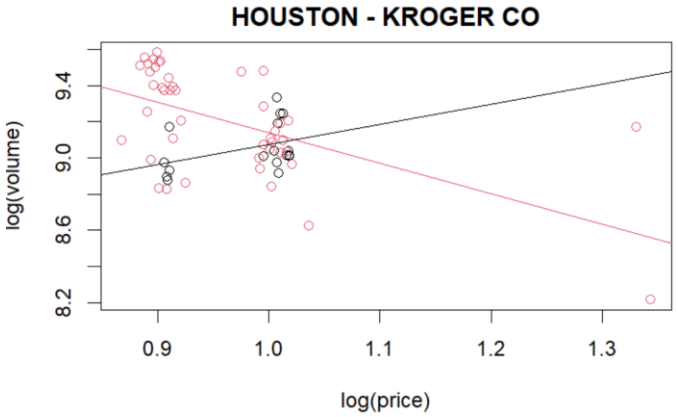
$$color = \begin{cases} pink\ has\ display \\ blue\ has\ no\ display \end{cases}$$

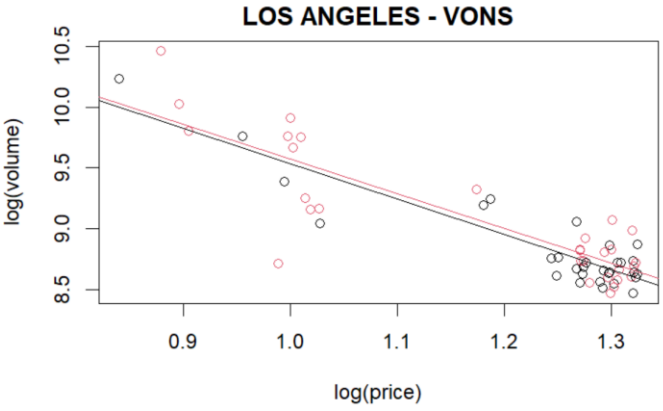


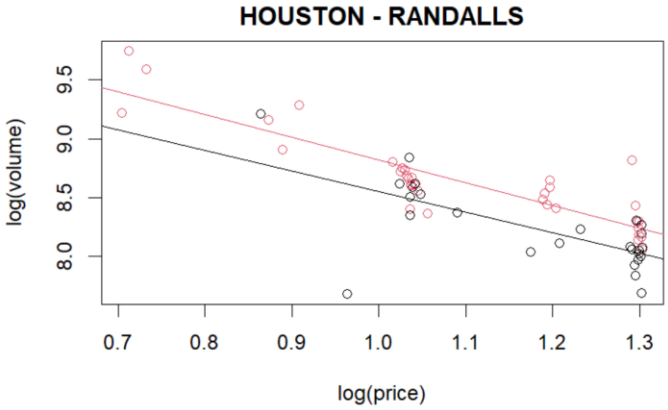
Boxplot of average ln $(sales)$ by display

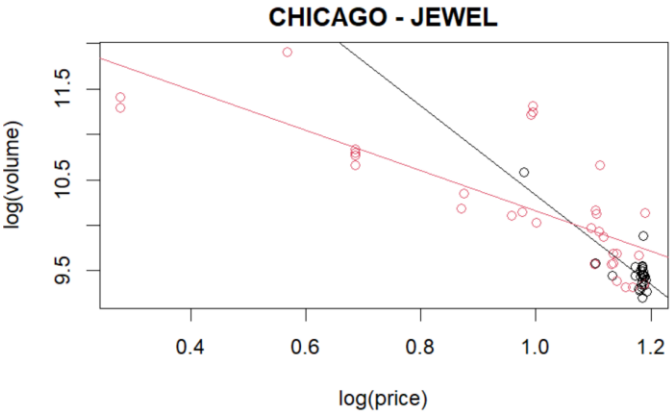
Plot of fit for stores. Red is for data with a display and black is for data without a display

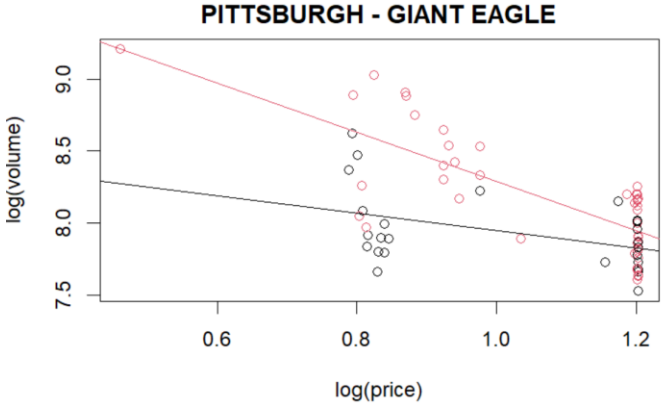


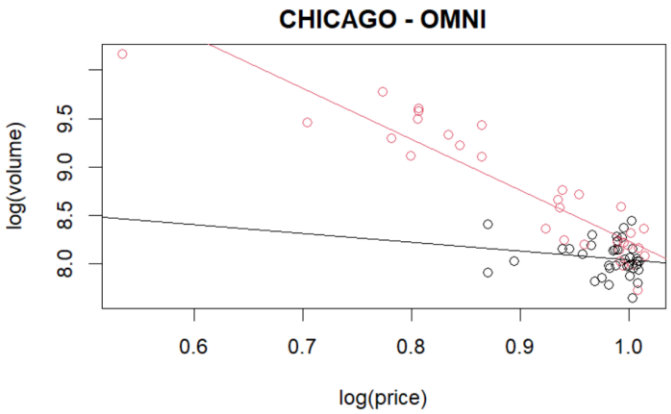


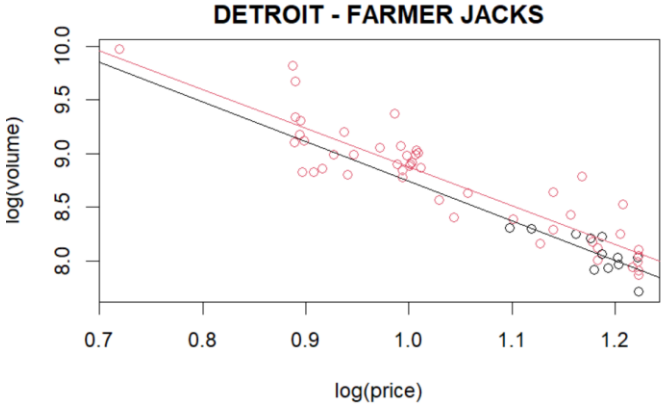


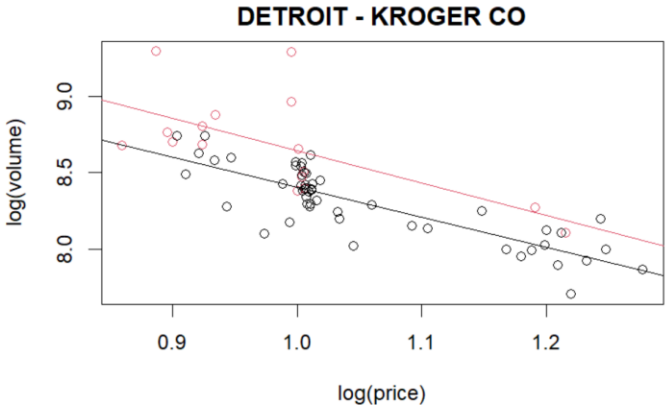


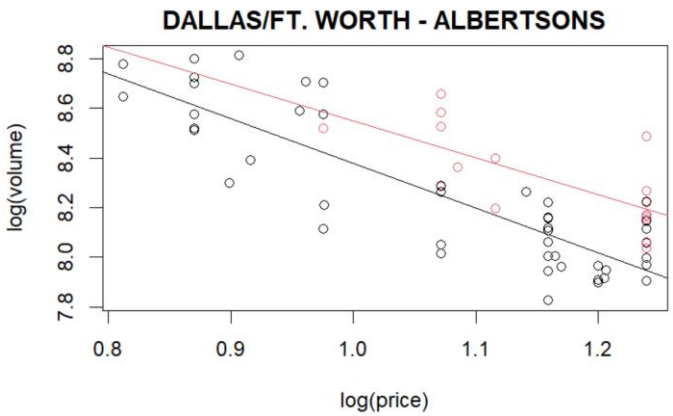


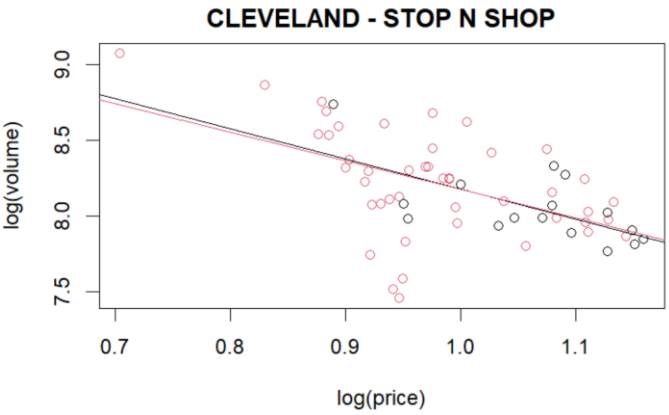


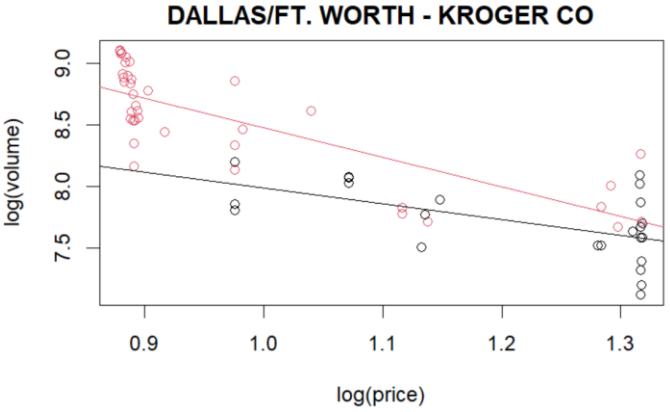


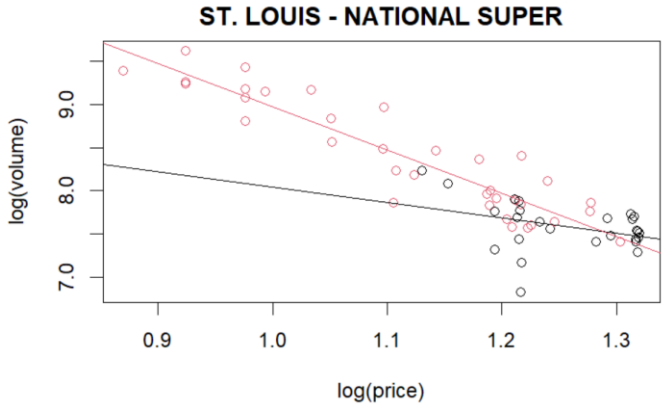


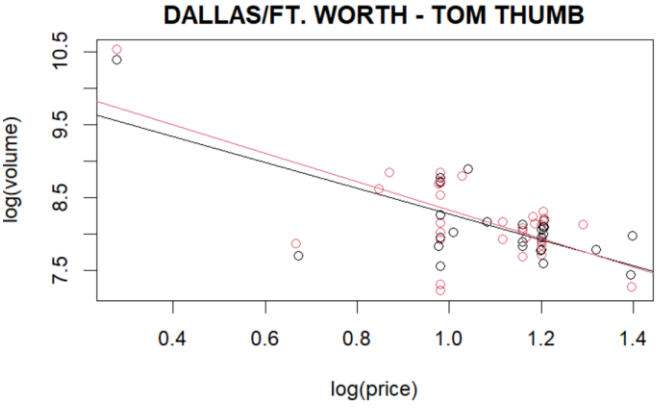

DETROIT - KROGER CO
log(volume)
log(price)

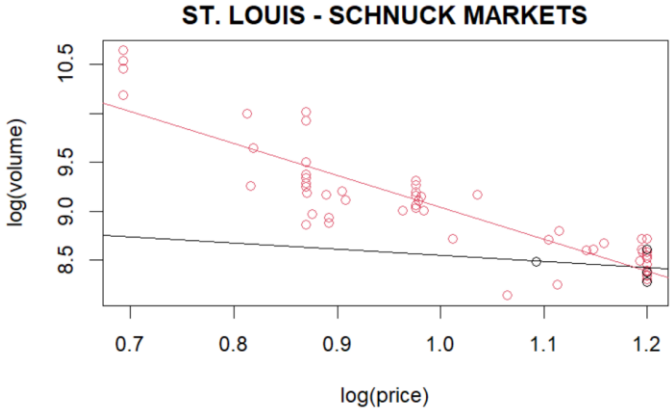

DALLAS/FT. WORTH - ALBERTSONS
log(volume)
log(price)

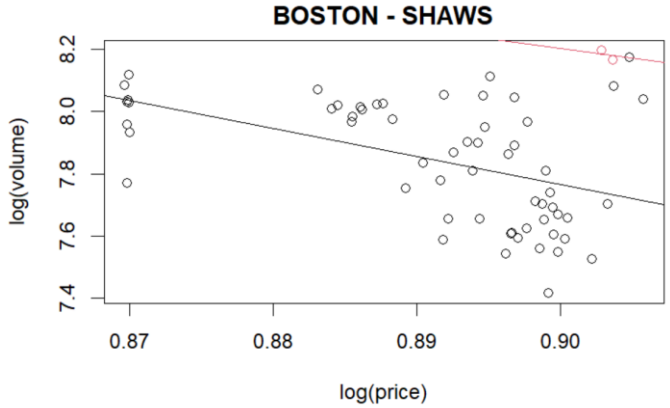

CLEVELAND - STOP N SHOP
log(volume)
log(price)

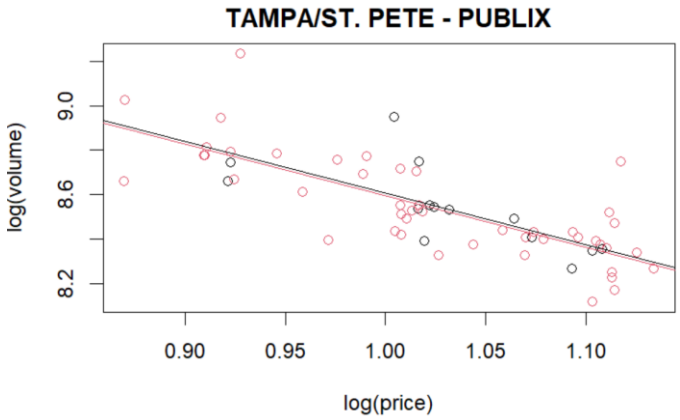

DALLAS/FT. WORTH - KROGER CO
log(volume)
log(price)

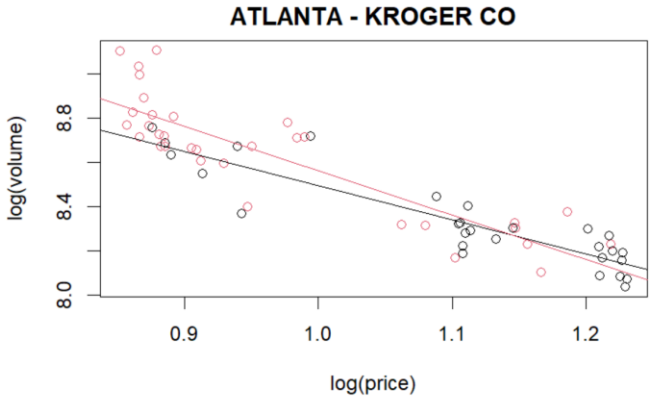

ST. LOUIS - NATIONAL SUPER
log(volume)
log(price)

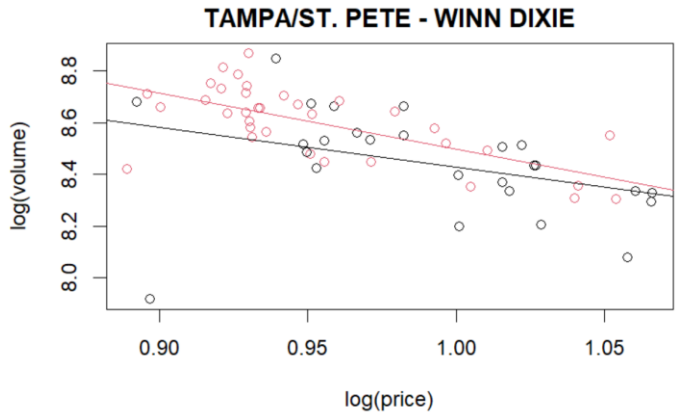

DALLAS/FT. WORTH - TOM THUMB
log(volume)
log(price)

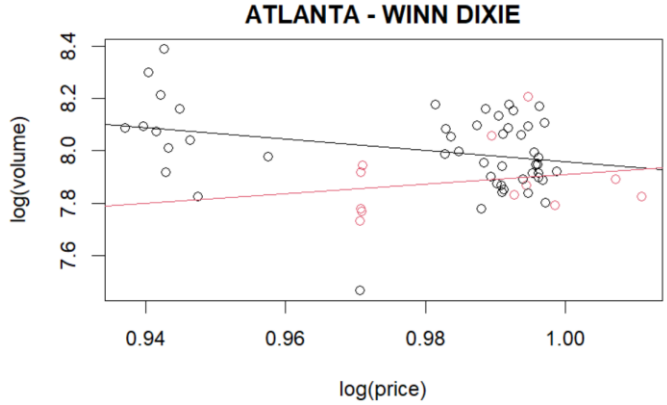

ST. LOUIS - SCHNUCK MARKETS
log(volume)
log(price)

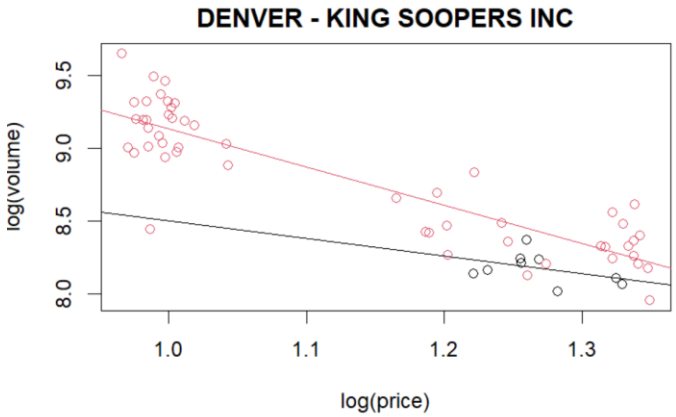

BOSTON - SHAWS
log(volume)
log(price)

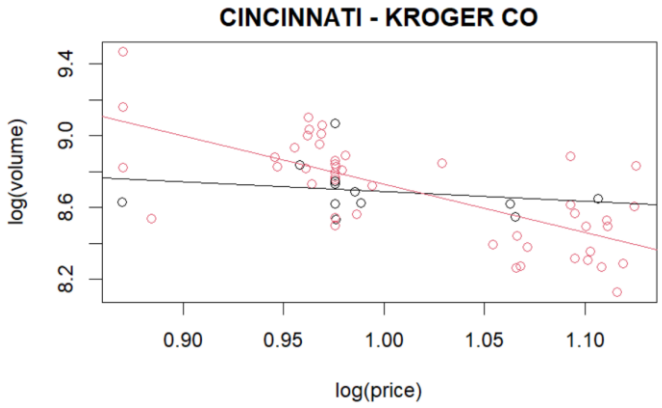

TAMPA/ST. PETE - PUBLIX
log(volume)
log(price)

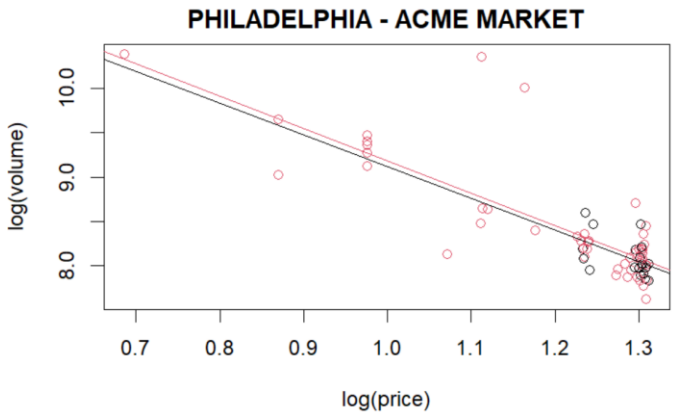

ATLANTA - KROGER CO
log(volume)
log(price)

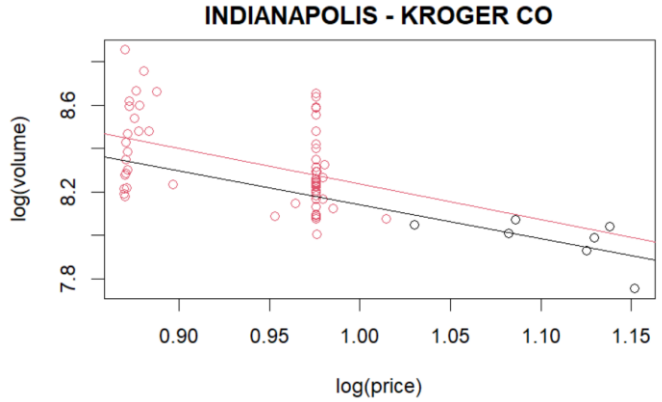

TAMPA/ST. PETE - WINN DIXIE
log(volume)
log(price)


ATLANTA - WINN DIXIE
log(volume)
log(price)


DENVER - KING SOOPERS INC
log(volume)
log(price)


CINCINNATI - KROGER CO
log(volume)
log(price)


PHILADELPHIA - ACME MARKET
log(volume)
log(price)


INDIANAPOLIS - KROGER CO
log(volume)
log(price)

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:

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

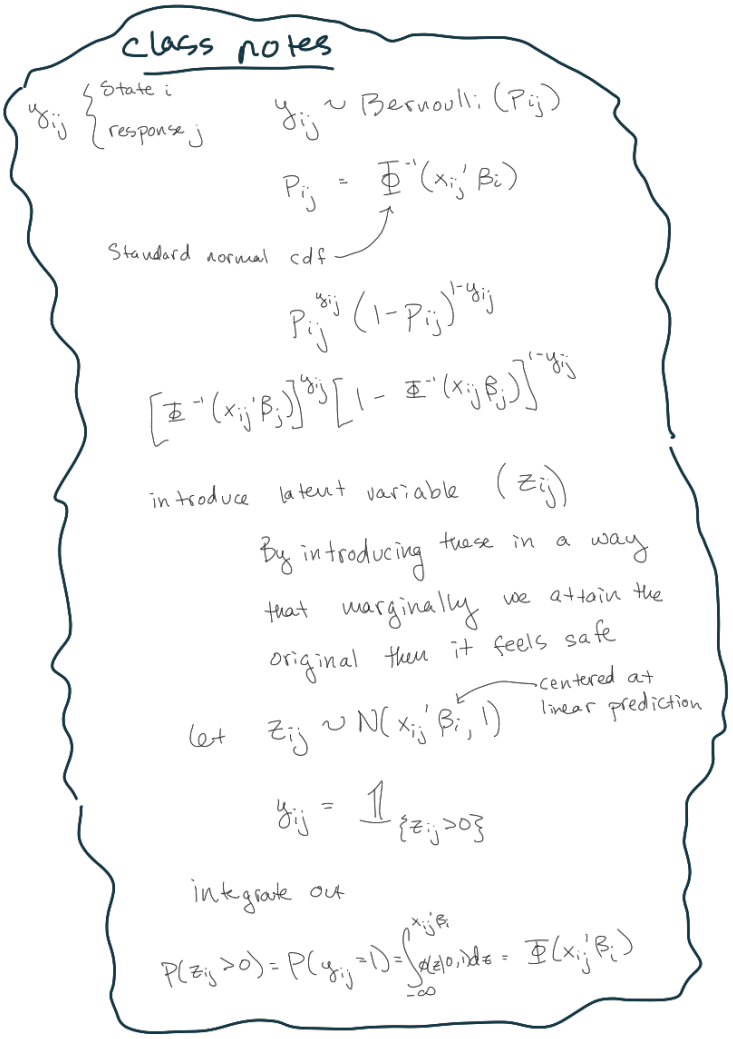
Here y_{ij} is the response (Bush=1, other=0) for respondent j in state i; $\Phi(\cdot)$ is the probit link function, i.e. the CDF of the standard normal distribution; μ_i is a state-level intercept term; x_{ij} is a vector of respondent-level demographic predictors; and β_i is a vector of regression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



Using age, education, race and sex we can convert the categorical data of age and education into 3 indicator variables each.
This helps us to see the marginal effect of education and age on voting outcome

Model:

$$P(y_{ij} = 1) = \Phi(x_{ij}' \alpha_i)$$

$$\alpha_i \sim N(\beta_i^*, B^*)$$

$$\beta_i^* \sim N(0, 10000 \cdot I_{p+1})$$

$$B^* \sim \text{Inv-Wish}(p+2, I_{p+1})$$

for states $i = 1, \dots, n$
and individuals $j = 1, \dots, N_i$ in each state
with p covariates and an intercept in the model.

Posteriors:

$$\alpha_i | \dots \sim N(\tilde{\beta}, \tilde{B})$$

$$\tilde{B} = [(B^*)^{-1} + x_i' x_i]^{-1}$$

$$\tilde{\beta} = \tilde{B}[(B^*)^{-1}\beta_i^* + x_i' z_i]$$

$$z_{ij} | \dots \sim \begin{cases} N(x_{ij}' \alpha_i, 1)|_{[0,\infty)}, & y_{ij} = 1 \\ N(x_{ij}' \alpha_i, 1)|_{(-\infty,0]}, & y_{ij} = 0 \end{cases}$$

$$\beta_i^* | \dots \sim N(A^{-1}b, A^{-1})$$

$$A = (B^*)^{-1} + (10\,000 \cdot I_{p+1})^{-1}$$

$$b' = \alpha_i' (B^*)^{-1}$$

$$B^* | \dots \sim \text{Inv-Wish}\left(n + p + 2, I_{p+1} + \sum_{i=1}^{n} (\alpha_i - \beta_i^*)(\alpha_i - \beta_i^*)'\right)$$

α_i = mean of non-black, 65+ year old, no high school, male propensity to vote for Bush in state i

β_2: Measures the marginal change in voting propensity for Bush as a woman in state i

β_3: Measures the marginal change in voting propensity for Bush as a black individual in state i

β_4: Measures the marginal change in voting propensity for Bush for someone with a bachelor's degree in state i

β_5: Measures the marginal change in voting propensity for Bush for someone with a high school diploma in state i

β_6: Measures the marginal change in voting propensity for Bush for someone with some college completed in state i

β_7: Measures the marginal change in voting propensity for Bush for someone age 18-29 in state i

β_8: Measures the marginal change in voting propensity for Bush for someone age 30-44 in state i

β_9: Measures the marginal change in voting propensity for Bush for someone age 45-64 in state i

Posterior mean estimates for coefficients