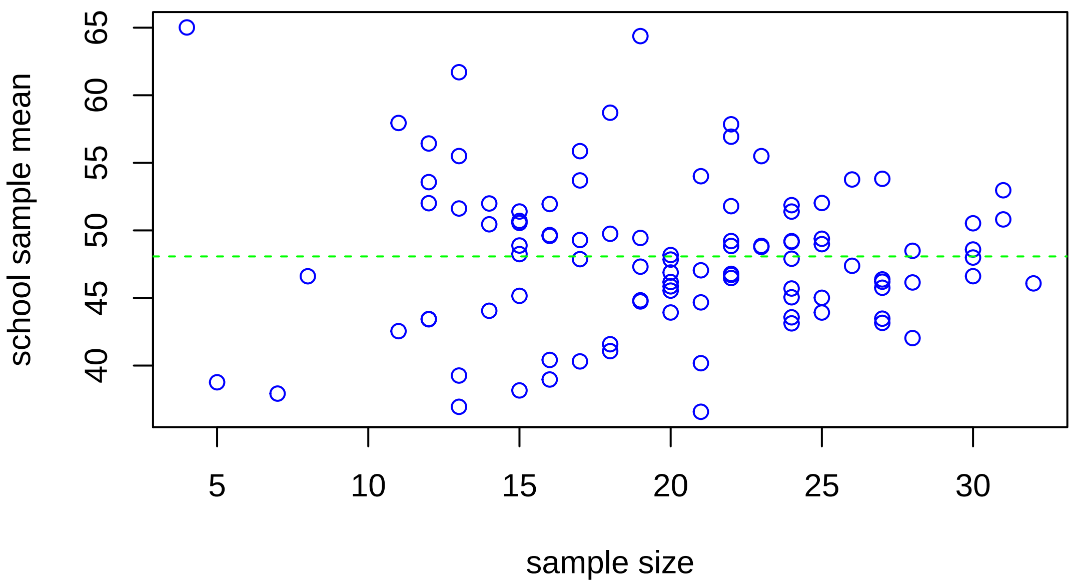
Math Tests

1. With smaller sample sizes, the standard deviation of the sampling distribution is larger, so we expect the sample means with smaller sample sizes to be farther from their respective means than those associated with larger sample sizes. We should not be surprised, therefore, that the smaller sample sizes also tend to lead to sample means farther from the grand mean, which is shown by the green dashed line. The following plot shows this:



2. Let $i \in \{1 \dots n\}$ denote the school, $j \in \{1, \dots, m_i\}$ the observations from school i, and M the total number of observations. The model is as follows:

$$\begin{aligned} y_{ij} &\sim N(\theta_i, \sigma^2) \\ \theta_i &\sim N(\mu, \tau^2\sigma^2) \\ \left.\begin{aligned} \mu &\sim 1 \\ \sigma^2 &\sim 1/\sigma^2 \\ \tau^2 &\sim 1 \end{aligned}\right\} &\text{ improper} \end{aligned}$$

The prior for σ^2 is Jeffrey's prior, as we saw in class. The prior for μ is flat, which we also saw in class. I chose a flat prior for τ^2, as well, after skimming the Gelman article. I decided to use that instead of an inverse gamma since, as we shall see, the flat prior updates to an inverse gamma anyway, and I chose a simpler non-informative prior to avoid choosing hyper-hyperparameters. With this setup, the joint posterior distribution, with $\theta := \langle \theta_1, \ldots, \theta_n \rangle$ is

$$
\begin{aligned}
p(\theta, \mu, \sigma^2, \tau^2 | y) &\propto p(y|\theta, \mu, \sigma^2, \tau^2) p(\theta, \mu, \sigma^2, \tau^2) \\
&= p(y|\theta, \mu, \sigma^2, \tau^2) p(\theta|\mu, \sigma^2, \tau^2) p(\mu) p(\sigma^2) p(\tau^2) \\
&\propto p(y|\theta, \mu, \sigma^2, \tau^2) p(\theta|\mu, \sigma^2, \tau^2) p(\sigma^2) \\
&= \left(\prod_{i=1}^{n} \prod_{j=1}^{m_i} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y_{ij} - \theta_i)^2}{2\sigma^2} \right) \right) \cdot \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma\tau} \exp\left(\frac{-(\theta_i - \mu)^2}{2\sigma^2\tau^2} \right) \right) \cdot \frac{1}{\sigma^2} \\
&\propto \left(\frac{1}{\sigma^2} \right)^{(M+n)/2+1} \left(\frac{1}{\tau^2} \right)^{n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \sum_{j=1}^{m_i} (y_{ij} - \theta_i)^2 \right) \exp\left(-\frac{1}{2\sigma^2\tau^2} \sum_{i=1}^{n} (\theta_i - \mu)^2 \right)
\end{aligned}
$$

We can now write the conditional distributions. As we saw in class,

$$\theta_i | y, \mu, \sigma^2, \tau^2 \sim N\Big(\Big)$$

We can get the others with the usual methods.

$$f(\sigma^2 | \ldots) \propto \left(\frac{1}{\sigma^2} \right)^{(M+n)/2+1} \exp\left(-\frac{1}{2\sigma} \left(\sum_{i=1}^{n} \sum_{j=1}^{m_i} (y_{ij} - \theta_i)^2 + \sum_{i=1}^{n} \frac{(\theta_i - \mu)^2}{\tau^2} \right) \right)$$

whence

$$\sigma^2 | \ldots \sim \mathrm{IG}\left(\frac{M+n}{2}, \ \ \frac{1}{2} \left(\sum_{i=1}^{n} \sum_{j=1}^{m_i} (y_{ij} - \theta_i)^2 + \sum_{i=1}^{n} \frac{(\theta_i - \mu)^2}{\tau^2} \right) \right)$$

To get the conditional distribution for μ, let us first note that

$$
\begin{aligned}
\sum_{i=1}^{n} (\theta_i - \mu)^2 &= \sum_{i=1}^{n} (\mu^2 - 2\mu\theta_i + \theta_i^2) \\
&= n\mu^2 - 2n\mu\bar{\theta} + K
\end{aligned}
$$

for some K not involving μ. Thus,

$$
\begin{aligned}
f(\mu | \ldots) &\propto \exp\left(-\frac{1}{2\sigma^2\tau^2} \sum_{i=1}^{n} (\mu - \theta_i)^2 \right) \\
&\propto \exp\left(-\frac{n}{\sigma^2\tau^2} (\mu^2 - 2\mu\bar{\theta}) \right)
\end{aligned}
$$

whence

$$\mu | \ldots \sim N\left(\bar{\theta}, \frac{\sigma^2\tau^2}{n} \right)$$

Finally,

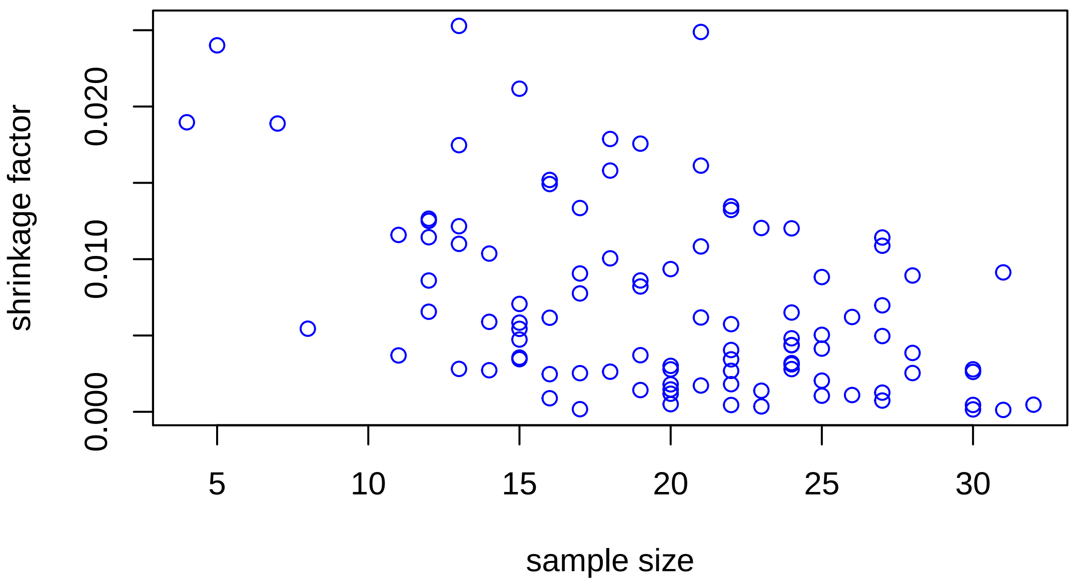
$$f(\tau^2 | \ldots) \propto \left(\frac{1}{\tau^2} \right)^{n/2} \exp\left(-\frac{1}{2\sigma^2\tau^2} \sum_{i=1}^{n} (\theta_i - \mu)^2 \right)$$

whence

$$\tau^2 | \ldots \sim \mathrm{IG}\left(\frac{n}{2}, \sum_{i=1}^{n} \frac{(\theta_i - \mu)^2}{2\sigma^2} \right)$$

With these distributions, we can set up a Gibbs sampler to sample from the posterior distribution. (See the file named Math.r.)

3. After running the Gibbs sampler, we get a sample for the posterior estimation for each θ_i. From this we can define the shrinkage factors $(y_{ij} - \hat{\theta}_i)/y_{ij}$, whose absolute values are plotted below.



Price elasticity of demand

Let $i \in \{1, \ldots, m := 88\}$ index the stores, and let n_i denote the number of observations for store i. Let n denote the total number of observations; that is, $n := \sum_{i=1}^{m} n_i$. For each i, let X_i be the $n_i \times 4$ matrix with an intercept, the log-price, display marker, and price/display interaction term for each observation of store i. We shall create a model with unmodeled and modeled effects as follows.

$$
\begin{aligned}
y_i &\sim N\big(X_i\beta + X_i\gamma_i, \sigma^2 I\big), \text{ where } y_i \text{ is the log-volume vector for store } i \\
\sigma^2 &\sim 1/\sigma^2, \text{ where we use } \lambda := 1/\sigma^2 \text{ when convenient} \\
\beta &\sim N(\mu, \Sigma), \text{ where } \beta \text{ is the coefficient vector for overall effects (not store specific)} \\
\gamma_i &\sim N(0, T), \text{ where } \gamma_i \text{ is the coefficient vector for store } i\text{'s offset from the average effects} \\
T &\sim IW(\nu, V)
\end{aligned}
$$

The hyperparameters for β and T are unmodeled. To run a Gibbs sampler, we need the full conditionals for the modeled (hyper)parameters.

1. For λ, we need the full likelihood (using all stores's data), which is proportional to

$$
\lambda^{n/2} \exp\left(-\frac{\lambda}{2}\sum_{i=1}^{m}\Big(y_i - X_i\beta - X_i\gamma_i\Big)^{\mathsf{T}}\Big(y_i - X_i\beta - X_i\gamma_i\Big)\right)
$$

Multiplying this by the prior gives us

$$
f(\lambda|\ldots) \propto \lambda^{n/2-1} \exp\left(-\frac{\lambda}{2}\sum_{i=1}^{m}\Big(y_i - X_i\beta - X_i\gamma_i\Big)^{\mathsf{T}}\Big(y_i - X_i\beta - X_i\gamma_i\Big)\right)
$$

which is a gamma density, and thus we see that

$$
\sigma^2|\ldots \sim IG\left(\frac{n}{2}, \frac{1}{2}\sum_{i=1}^{m}\Big(y_i - X_i\beta - X_i\gamma_i\Big)^{\mathsf{T}}\Big(y_i - X_i\beta - X_i\gamma_i\Big)\right)
$$

2. For each γ_i, we need the store-specific likelihood function, which is proportional to

$$
\exp\left(-\frac{\lambda}{2}\Big(y_i - X_i\beta - X_i\gamma_i\Big)^{\mathsf{T}}\Big(y_i - X_i\beta - X_i\gamma_i\Big)\right)
$$

Thus, the conditional posterior density for γ_i is

$$
\begin{aligned}
f(\gamma_i|\ldots) &\propto \exp\left(-\frac{1}{2}\gamma_i^{\mathsf{T}}T^{-1}\gamma_i\right)\exp\left(-\frac{\lambda}{2}\Big(y_i - X_i\beta - X_i\gamma_i\Big)^{\mathsf{T}}\Big(y_i - X_i\beta - X_i\gamma_i\Big)\right) \\
&\propto \exp\left(-\frac{1}{2}\gamma_i^{\mathsf{T}}T^{-1}\gamma_i - \frac{\lambda}{2}\Big(\gamma_i^{\mathsf{T}}X_i^{\mathsf{T}}X_i\gamma_i - 2\gamma_i^{\mathsf{T}}\big(X_i^{\mathsf{T}}y_i - X_i^{\mathsf{T}}X_i\beta\big)\Big)\right) \\
&= \exp\Big(\gamma_i^{\mathsf{T}}\Big(T^{-1} + \lambda X_i^{\mathsf{T}}X_i\Big)\gamma_i - 2\gamma_i^{\mathsf{T}}\lambda\Big(X_i^{\mathsf{T}}y_i - X_i^{\mathsf{T}}X_i\beta\Big)\Big).
\end{aligned}
$$

This is proportional to a normal density; in particular,

$$
\gamma_i|\ldots \sim N(\kappa^*, T^*),
$$

where

$$
\begin{aligned}
T^* &:= \big(T^{-1} + \lambda X_i^{\mathsf{T}}X_i\big)^{-1} \\
\kappa^* &:= \lambda T^*\big(X_i^{\mathsf{T}}y_i - X_i^{\mathsf{T}}X_i\beta\big)
\end{aligned}
$$

3. To get the conditional for β, we follow a pattern similar to the simplification seen above for γ_i, but with the roles of β and γ_i switched and a summation over all stores involved. We therefore see that the likelihood is, up to proportionality, equal to

$$\exp\left(-\frac{\lambda}{2}\sum_{i=1}^{m}\left(\beta^\intercal X_i^\intercal X_i\beta - 2\beta^\intercal\big(X_i^\intercal y_i - X_i^\intercal X_i\gamma_i\big)\right)\right)$$

The prior for β reduces by proportionality to

$$\exp\left(-\frac{1}{2}\left(\beta^\intercal\Sigma^{-1}\beta - 2\beta^\intercal\Sigma^{-1}\mu\right)\right)$$

We thus get

$$\begin{aligned} f(\beta|\ldots) &\propto \exp\left(-\frac{1}{2}\left(\beta^\intercal\Sigma^{-1}\beta - 2\beta^\intercal\Sigma^{-1}\mu\right)\right)\exp\left(-\frac{\lambda}{2}\sum_{i=1}^{m}\left(\beta^\intercal X_i^\intercal X_i\beta - 2\beta^\intercal\big(X_i^\intercal y_i - X_i^\intercal X_i\gamma_i\big)\right)\right) \\ &= \exp\left[-\frac{1}{2}\left(\beta^\intercal\left(\Sigma^{-1} + \lambda\sum_{i=1}^{m}X_i^\intercal X_i\right)\beta - 2\beta^\intercal\left(\Sigma^{-1}\mu + \lambda\sum_{i=1}^{m}\big(X_i^\intercal y_i - X_i^\intercal X_i\gamma_i\big)\right)\right)\right] \end{aligned}$$

This is the kernel of a multivariate normal density, so

$$\beta|\ldots \sim N(\mu^*, \Sigma^*),$$

where

$$\begin{aligned} \Sigma^* &:= \left(\Sigma^{-1} + \lambda\sum_{i=1}^{m}X_i^\intercal X_i\right)^{-1} \\ \mu^* &:= \Sigma^*\left(\Sigma^{-1}\mu + \lambda\sum_{i=1}^{m}\big(X_i^\intercal y_i - X_i^\intercal X_i\gamma_i\big)\right) \end{aligned}$$

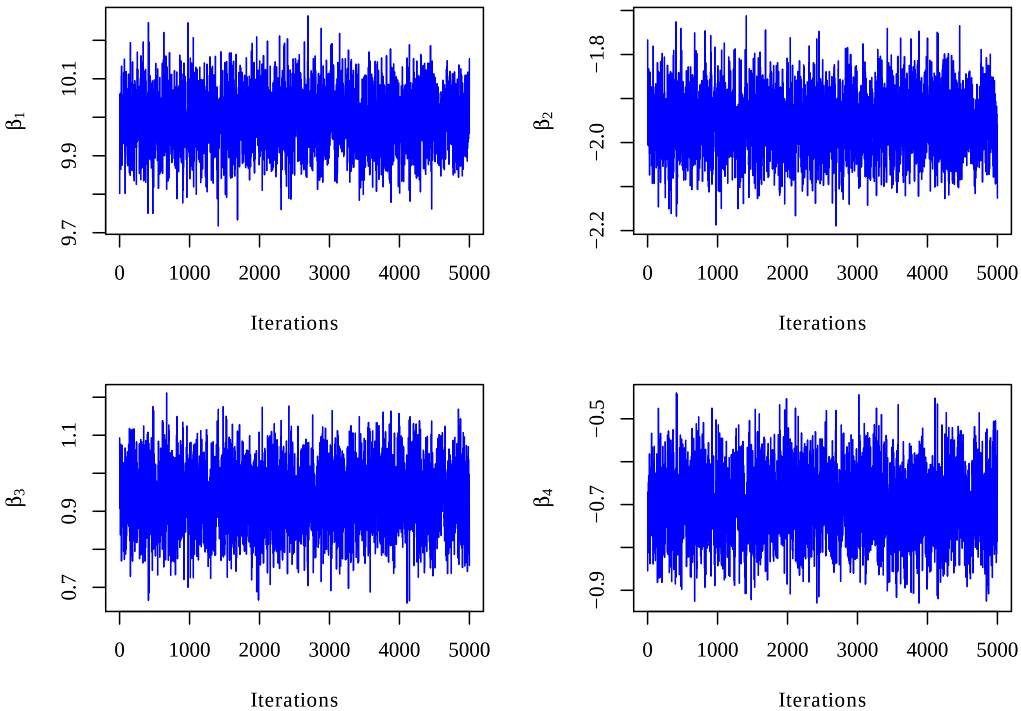
4. For T, first recall the identity $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ whenever A and B are conformable for both products. Recall also that $x^\intercal Ax = \mathrm{tr}(x^\intercal Ax)$ for any quadratic form $x^\intercal Ax$. We need only the part of the full likelihood that involves T, and we combine that with the prior to get

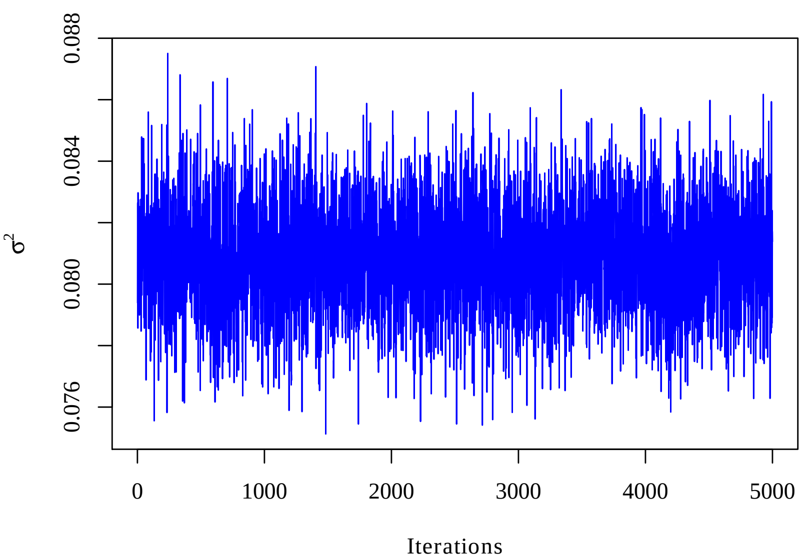
$$\begin{aligned} f(T|\ldots) &\propto |T|^{-(\nu+3)/2}\exp\left(-\frac{1}{2}\mathrm{tr}\big(VT^{-1}\big)\right)|T|^{-m/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{m}\gamma_i^\intercal T^{-1}\gamma_i\right) \\ &= |T|^{-(\nu+m+3)/2}\exp\left(-\frac{1}{2}\mathrm{tr}\big(VT^{-1}\big) - \frac{1}{2}\sum_{i=1}^{m}\gamma_i^\intercal T^{-1}\gamma_i\right) \\ &= |T|^{-(\nu+m+3)/2}\exp\left(-\frac{1}{2}\mathrm{tr}\big(VT^{-1}\big) - \frac{1}{2}\mathrm{tr}\left(\sum_{i=1}^{m}\gamma_i\gamma_i^\intercal T^{-1}\right)\right) \\ &= |T|^{-(\nu+m+3)/2}\exp\left[-\frac{1}{2}\mathrm{tr}\left(\left(V + \sum_{i=1}^{m}\gamma_i\gamma_i^\intercal\right)T^{-1}\right)\right] \end{aligned}$$

and therefore

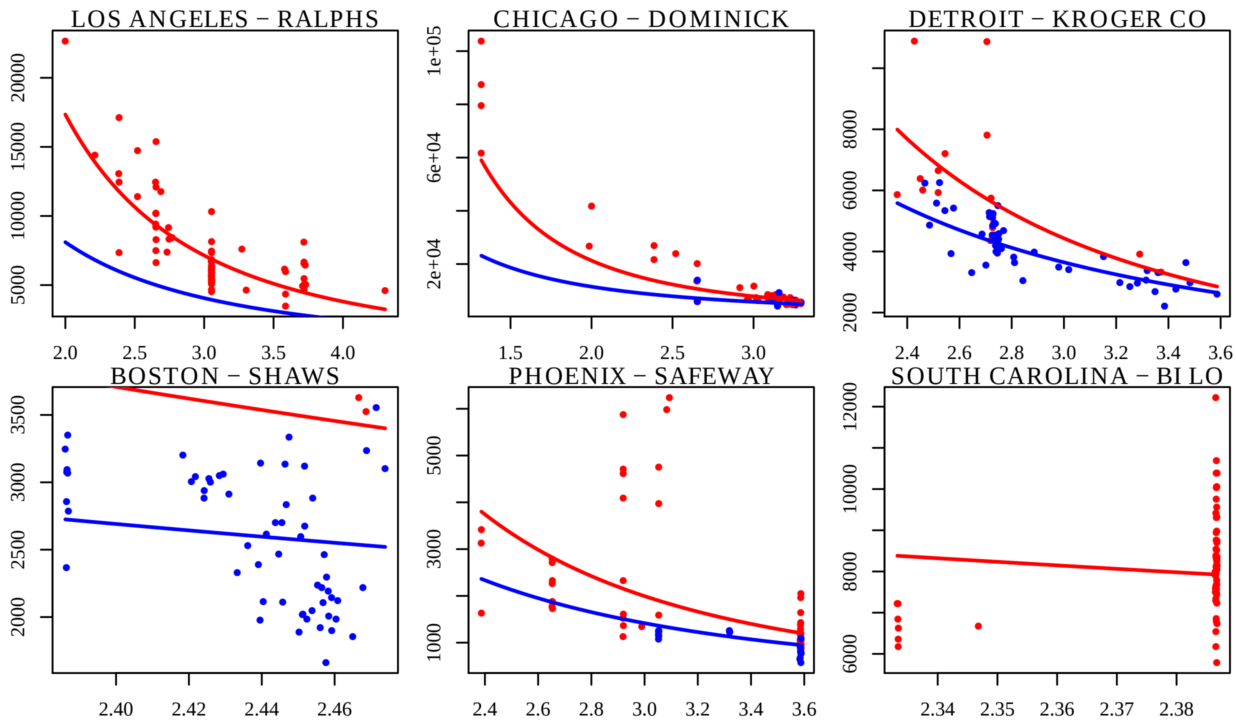
$$T|\ldots \sim IW\left(\nu + m,\ V + \sum_{i=1}^{m}\gamma_i\gamma_i^\intercal\right).$$

After running the Gibbs sampler, we can graph the samples for the parameters. Shown below are the plots for β and σ^2.



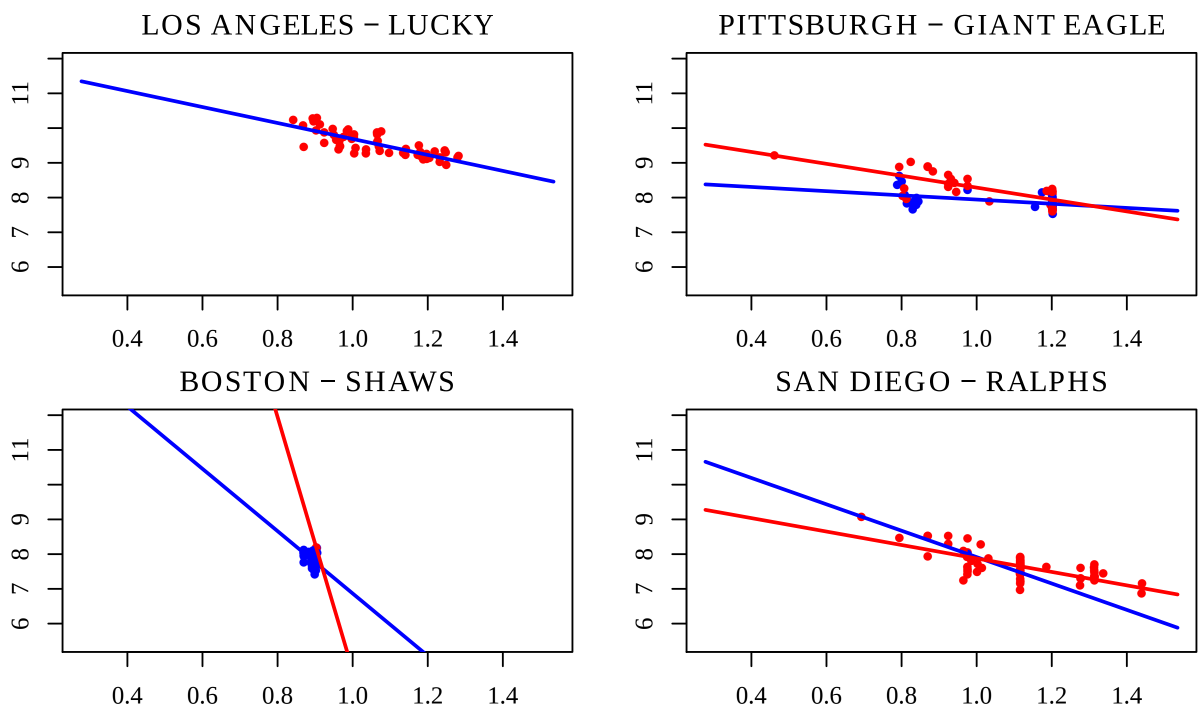


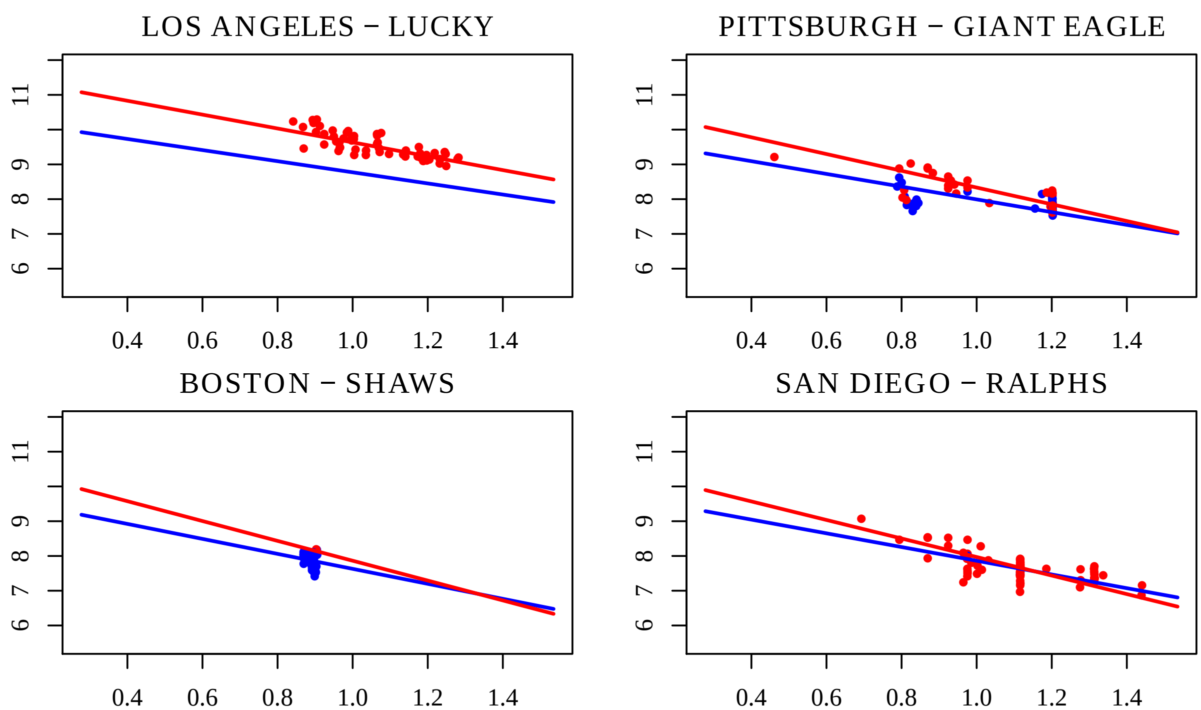
Putting the pieces together, we can graph the modeled demand curves. Below are six such curves.



There is shrinkage toward a common mean, but this is not completely evident here.

To see this, let us instead plot the log demand curves for four stores. Shown first are the curves as estimated by ordinary least squares, and second those estimated by our model.





Polling Data

As suggested in the paper, let us introduce normally-distributed latent variables z_{ij} such that

$$y_{ij} = \begin{cases} 0 & \text{if } z_{ij} < 0, \text{ and} \\ 1 & \text{if } z_{ij} \geq 0, \end{cases}$$

where $i \in \{1, \ldots, N\}$ indexes the states, and $j \in \{1, \ldots, n_i\}$ indexes the voters in state i. The model I use is as follows:

$$\begin{aligned} z_i := \langle z_{i1}, \ldots, z_{in_i} \rangle &\sim N_{n_i}\big(X_i \beta + S_i \mu_i, I\big) \\ \mu_i &\sim N(0, \sigma^2) \\ \beta &\sim 1 \\ \sigma^2 &\sim 1 \end{aligned}$$

where X_i is the $n_i \times p$ matrix of observed data for state i, including an intercept column, and $S_i := \langle 1, \ldots, 1 \rangle^\intercal$ has length n_i, so that μ_i encodes the state-specific offset from the overall mean already included in β.

To run a Gibbs sampler, we need the conditionals for the parameters z_{ij}, μ_i, σ^2, and β. Note that the only effect of y_{ij} on any of these is through z_{ij}, except for z_{ij} itself, so we need not include these in the conditionals (and are therefore not part of the elided parameters in the conditionals below).

1. For z_{ij}, we know that the result will be from a normal distribution, but we know which side of 0 it must be if we are given y_{ij}, and therefore we draw z_{ij} from a truncated normal distribution:

$$z_{ij} | y_{ij}, \ldots \sim \mathbb{1}(y_{ij} = 1) N^+\big(\mu_i + X_{ij}\beta, 1\big) + \mathbb{1}(y_{ij} = 0) N^-\big(\mu_i + X_{ij}\beta, 1\big)$$

where $\mathbb{1}$ denotes the usual indicator function and $N^\pm$ denotes the truncated positive/negative normal distributions.

2. Let us now consider μ_i:

$$\begin{aligned} f(\mu_i | \ldots) &\propto \exp\left(-\frac{\mu_i^2}{2\sigma^2}\right) \exp\left(-\frac{1}{2}\big(z_i - X_i\beta - S_i\mu_i\big)^\intercal \big(z_i - X_i\beta - S_i\mu_i\big)\right) \\ &\propto \exp\left(-\frac{\mu_i^2}{2\sigma^2}\right) \exp\left(-\frac{1}{2}\big(-2S_i^\intercal z_i \mu_i + 2S_i^\intercal X_i \beta\mu_i + S_i^\intercal S_i \mu_i^2\big)\right) \\ &= \exp\left(-\frac{\mu_i^2}{2\sigma^2}\right) \exp\left(-\frac{1}{2}\big(-2S_i^\intercal (z_i - X_i\beta\mu_i)\mu_i + n_i\mu_i^2\big)\right) \\ &= \exp\left(-\frac{1}{2}\big(-2S_i^\intercal (z_i - X_i\beta\mu_i)\mu_i + \big(1/\sigma^2 + n_i\big)\mu_i^2\big)\right) \end{aligned}$$

and thus

$$\mu_i \sim N\big(m, v\big)$$

where

$$\begin{aligned} v &:= \frac{1}{\frac{1}{\sigma^2} + n_i} \\ m &:= v \cdot S_i^\intercal\Big(z_i - X_i\beta\Big) \end{aligned}$$

3. Since σ^2 has a flat prior, we update using the likelihood:

$$f(\sigma^2|\ldots) \propto \left(\frac{1}{\sigma^2}\right)^{N/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\mu_i^2\right)$$

and thus

$$\sigma^2 \sim IG\left(\frac{N}{2},\ \frac{1}{2}\sum_{i=1}^{N}\mu_i^2\right).$$

4. Finally, let us find the conditional for β:

$$\begin{aligned}
f(\beta|\ldots) &\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{N}\big(z_i - X_i\beta - S_i\mu_i\big)^{\mathsf{T}}\big(z_i - X_i\beta - S_i\mu_i\big)\right)\\
&\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{N}\Big(\beta^{\mathsf{T}}X_i^{\mathsf{T}}X_i\beta - 2\beta^{\mathsf{T}}X_i^{\mathsf{T}}(z_i - S_i\mu_i)\Big)\right)\\
&= \exp\left[-\frac{1}{2}\left(\beta^{\mathsf{T}}\left(\sum_{i=1}^{N}X_i^{\mathsf{T}}X_i\right)\beta - 2\beta^{\mathsf{T}}\left(\sum_{i=1}^{N}X_i^{\mathsf{T}}(z_i - S_i\mu_i)\right)\right)\right]
\end{aligned}$$

and thus

$$\beta|\ldots \sim N_p\big(\nu^*, T^*\big)$$

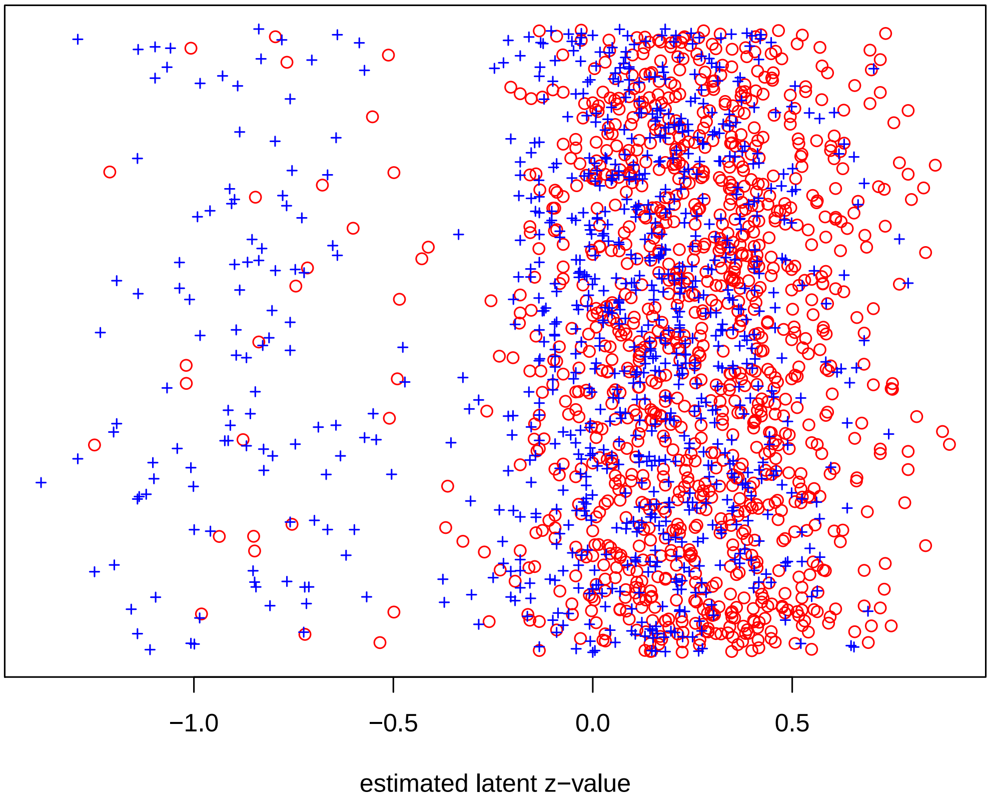
where

$$\begin{aligned}
T^* &= \left(\sum_{i=1}^{N}X_i^{\mathsf{T}}X_i\right)^{-1}\\
\nu^* &= T^*\left(\sum_{i=1}^{N}X_i^{\mathsf{T}}(z_i - S_i\mu_i)\right)
\end{aligned}$$

I ran a Gibbs sampler and found the following estimations of β and the state-specific offsets μ_i. Also included are the states' percentage vote for Bush (obtained from http://uselectionatlas.org) and the survey sample sizes for each state. The state abbreviations are colored according to whether Bush (red) or Dukakis (blue) won that state's popular vote.

β entries		state-specifc offsets			
covariate	coefficient	state	offset	% vote for Bush	sample size
(Intercept)	0.19418972	AL	0.360683013	59.17	31
High School	0.14364084	AR	−0.087962626	59.95	12
Some College	0.31703079	AZ	0.136267915	56.37	33
Bachelor's	0.19152460	CA	−0.087579088	51.13	190
30–44	−0.18185872	CO	−0.060548254	53.06	25
45–64	−0.04245047	CT	0.048397242	51.98	26
65+	−0.14226771	DC	−0.067773062	14.30	3
female	−0.05983321	DE	−0.1660797	55.88	7
black	−1.06729879	FL	0.073958582	60.87	126
		GA	0.143667878	59.75	49
		IA	−0.290966274	44.50	20
		ID	−0.119581042	62.08	4
		IL	−0.186759186	50.69	85
		IN	0.111364807	59.84	34
		KS	0.222963504	55.79	27
		KY	0.115600941	55.52	31
		LA	0.085911795	54.27	32
		MA	−0.298473475	45.38	50
		MD	−0.367909905	51.11	31
		ME	−0.122347547	55.34	13
		MI	−0.124830356	53.57	83
		MN	−0.301704627	45.90	42
		MO	−0.058555656	51.83	37
		MS	0.270982167	59.89	28
		MT	−0.08473819	52.07	5
		NC	0.253209685	57.97	44
		ND	−0.020347265	56.03	7
		NE	−0.002584641	60.15	17
		NH	0.026341716	62.49	3
		NJ	0.02246534	56.24	67
		NM	−0.137791838	51.86	23
		NV	0.04755717	58.86	4
		NY	−0.278085596	47.52	163
		OH	0.169239756	55.00	101
		OK	0.065538495	57.93	18
		OR	−0.25194075	46.61	18
		PA	−0.112536358	50.70	95
		RI	−0.152126119	43.93	12
		SC	0.230830879	61.50	30
		SD	−0.040964541	52.85	6
		TN	0.425473533	57.89	45
		TX	0.072999311	55.95	129
		UT	0.347507491	66.22	11
		VA	0.279423687	59.74	58
		VT	0.090704516	51.10	2
		WA	−0.053957025	48.46	54
		WI	−0.1963418	47.80	54
		WV	0.00779423	47.46	28
		WY	0.077987609	60.53	2

We can see the model's prediction for individuals in the following graph. Dukakis voters are represented by $+$, and $\circ$ represents Bush voters.



The vertical axis shows jitter only. Note that the the densest area is near $z = 0.25$, and most Dukakis voters are incorrectly predicted to vote for Bush, although their z-values are not far from 0. Several error rates in the model fit are given in the following table.

population	n	% error rate
all	2015	36.6
Bush Voters	1124	13.8
Dukakis Voters	891	65.4
$\|z\| > 0.25$	994	29.2
$\|z\| > 0.5$	350	18.9