SDS 383D: Homework 4

Giorgio Paulon

April 4, 2017

Problem 1. Hierarchical models: Math tests

The data set in "mathtest.csv" shows the scores on a standardized math test from a sample of 10th-grade students at 100 different U.S. urban high schools, all having enrollment of at least 400 10th-grade students. (A lot of educational research involves "survey tests" of this sort, with tests administered to all students being the rare exception.)

Let θ_i be the underlying mean test score for school i, and let y_{ij} be the score for the jth student in school i. Starting with the "mathtest.R" script, you'll notice that the extreme school-level averages $\bar{y}_i$ (both high and low) tend to be at schools where fewer students were sampled.

1. *Explain briefly why this would be.*

   In Figure 1, we notice that extreme average values of the scores are obtained for school with few students sampled. This happens because the distribution of $\bar{y}_i$ for each school has a variance of σ^2/n_i. In fact, $\bar{y}_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2/n_i)$. Thus, the smaller n_i, the larger the variability of $\bar{y}_i$ around the grand mean $\frac{1}{I}\sum_{i=1}^{I} \bar{y}_i$.

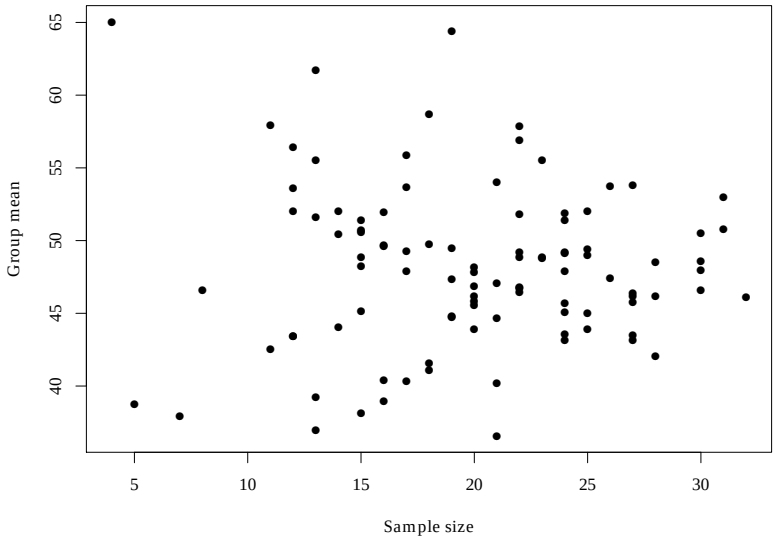

Figure 1: Extreme school-level averages $\bar{y}_i$ for schools with few students sampled.

2. *Fit a normal hierarchical model to these data via Gibbs sampling:*

$$\begin{aligned} y_{ij} &\sim N(\theta_i, \sigma^2) \\ \theta_i &\sim N(\mu, \tau^2\sigma^2) \end{aligned}$$

   Decide upon sensible priors for the unknown model parameters (μ, σ^2, τ^2).

   Let us denote with $i = 1, \dots, I$ the schools and with n_i the number of students for school i. The total number of students is $n = \sum_{i=1}^{I} n_i$.

   The model is the following:

$$y_{ij}|\theta_i, \sigma^2 \overset{\text{iid}}{\sim} N(\theta_i, \sigma^2)$$

along with the priors

$$\begin{aligned}
&\theta_i|\mu,\sigma^2,\tau^2 \sim N(\mu,\tau^2\sigma^2)\\
&\mu \sim \pi(\mu) = \mathcal{I}_{(-\infty,+\infty)}(\mu)\\
&\sigma^2 \sim \pi(\sigma^2) = \frac{1}{\sigma^2}\cdot\mathcal{I}_{[0,+\infty)}(\sigma^2)\\
&\tau^2 \sim \pi(\tau^2) = \mathcal{I}_{[0,+\infty)}(\tau^2).
\end{aligned} \tag{1}$$

In other words, we used a flat prior for μ and for τ^2. The flat prior for μ is motivated by the fact that a proper posterior can be achieved. On the other hand, the flat prior for τ^2 has been chosen according to Gelman (2006). This should work well unless the number of groups I is low (below 5, say). In our case, we dispose of 100 different schools and therefore its use is justified.

The likelihood of the model is

$$\begin{aligned}
L(Y|\boldsymbol{\theta},\sigma^2,\mu,\tau^2) &= \prod_{i=1}^{I}\prod_{j=1}^{n_i}\left\{\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma^2}(y_{ij}-\theta_i)^2\right\}\right\}\\
&\propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\theta_i)^2\right\}.
\end{aligned}$$

Using the prior specification (1) we obtain the following full conditionals:

- the full conditional for each one of the group means θ_i are

$$\begin{aligned}
p(\theta_i|y,\sigma^2,\mu,\tau^2) &\propto L(Y|\theta_i,\sigma^2)p(\theta_i|\mu,\sigma^2,\tau^2)\\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n_i}(y_{ij}-\theta_i)^2\right\}\exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta_i-\mu)^2\right\}\\
&= \exp\left\{-\frac{1}{2\tau^2\sigma^2}\left[\tau^2\sum_{j=1}^{n_i}y_{ij}^2+n_i\tau^2\theta_i^2-2\theta_i\tau^2\sum_{j=1}^{n_i}y_{ij}+\theta_i^2+\mu^2-2\mu\theta_i\right]\right\}\\
&\propto \exp\left\{-\frac{n_i\tau^2+1}{2\tau^2\sigma^2}\left[\theta_i^2-2\frac{\mu+\tau^2n_i\bar{y}_i}{n_i\tau^2+1}\theta_i\right]\right\}\\
&= N\left(\frac{\mu+\tau^2n_i\bar{y}_i}{n_i\tau^2+1};\left(\frac{n_i\tau^2+1}{\tau^2\sigma^2}\right)^{-1}\right)\\
&= N\left(\frac{1}{n_i\tau^2+1}\cdot\mu+\frac{n_i\tau^2}{n_i\tau^2+1}\cdot\bar{y}_i;\left(\frac{n_i}{\sigma^2}+\frac{1}{\tau^2\sigma^2}\right)^{-1}\right),
\end{aligned}$$

  where in the last line we recognize the usual formulation as posterior mean as weighted average of prior mean and data sample mean.

- the full conditional for the grand mean μ is

$$\begin{aligned} p(\mu|y,\boldsymbol{\theta},\sigma^2,\tau^2) &\propto p(\boldsymbol{\theta}|\mu,\tau^2,\sigma^2)p(\mu) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2\tau^2}\sum_{i=1}^{I}(\theta_i-\mu)^2\right\}p(\mu) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2\tau^2}\left(I\mu^2+\sum_{i=1}^{I}\theta_i^2-2\mu\sum_{i=1}^{I}\theta_i\right)\right\} \\ &= N\left(\bar{\boldsymbol{\theta}};\frac{\sigma^2\tau^2}{I}\right). \end{aligned}$$

- the full conditional for the variance σ^2 is

$$\begin{aligned} p(\sigma^2|y,\boldsymbol{\theta},\mu,\tau^2) &\propto p(Y|\boldsymbol{\theta},\sigma^2,\mu,\tau^2)p(\boldsymbol{\theta}|\mu,\tau^2,\sigma^2)p(\sigma^2) \\ &\propto \left(\frac{1}{\sigma^2}\right)^{\frac{n+I}{2}+1}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\theta_i)^2\right\}\exp\left\{-\frac{1}{2\sigma^2\tau^2}\sum_{i=1}^{I}(\theta_i-\mu)^2\right\} \\ &= \left(\frac{1}{\sigma^2}\right)^{\frac{n+I}{2}+1}\exp\left\{-\frac{1}{\sigma^2}\left[\frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\theta_i)^2+\frac{1}{2\tau^2}\sum_{i=1}^{I}(\theta_i-\mu)^2\right]\right\} \\ &= \text{inv-Gamma}\left(\frac{n+I}{2};\frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\theta_i)^2+\frac{1}{2\tau^2}\sum_{i=1}^{I}(\theta_i-\mu)^2\right). \end{aligned}$$

- the full conditional for the variance τ^2 is

$$\begin{aligned} p(\tau^2|y,\boldsymbol{\theta},\mu,\sigma^2) &\propto p(Y|\boldsymbol{\theta},\sigma^2,\mu,\tau^2)p(\boldsymbol{\theta}|\sigma^2,\mu,\tau^2)p(\tau^2) \\ &\propto \left(\frac{1}{\tau^2}\right)^{\frac{I}{2}}\exp\left\{-\frac{1}{2\sigma^2\tau^2}\sum_{i=1}^{I}(\theta_i-\mu)^2\right\} \\ &= \text{inv-Gamma}\left(\frac{I}{2}-1;\frac{\sum_{i=1}^{I}(\theta_i-\mu)^2}{2\sigma^2}\right). \end{aligned}$$

3. *Suppose you use the posterior mean $\hat{\theta}_i$ from the above model to estimate each school-level mean θ_i. Define the shrinkage coefficient κ_i as*

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

which tells you how much the posterior mean shrinks the observed sample mean. Plot this shrinkage coefficient (in absolute value) for each school as a function of that school's sample size, and comment.

In Figure 2, we notice that groups with low sample size get shrunk the most, whereas groups with large sample size hardly get shrunk at all. In fact, the larger the sample size for a group, the more information we have for that group and the less information we need to borrow from the rest of the population.

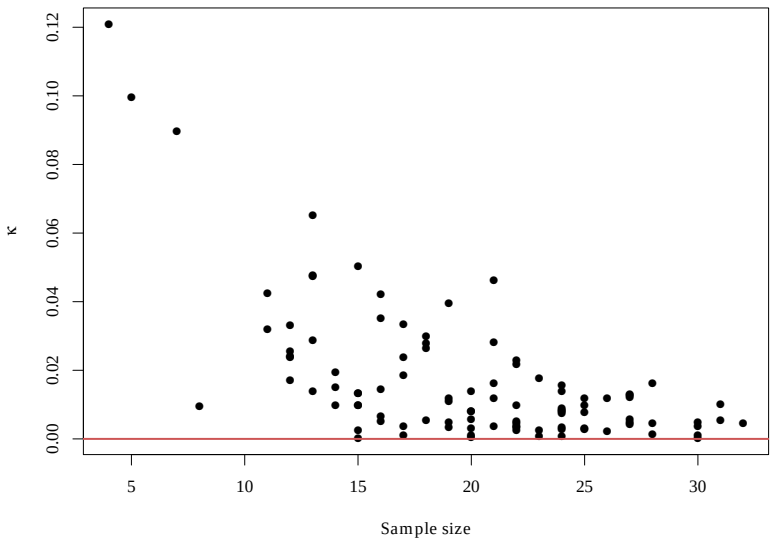

Figure 2: Shrinkage coefficient plotted versus the sample size of the groups.

Problem 2. Hierarchical models: 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,

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

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.

First of all, we analyze this problem with a hierarchical model using the package `lme4`.

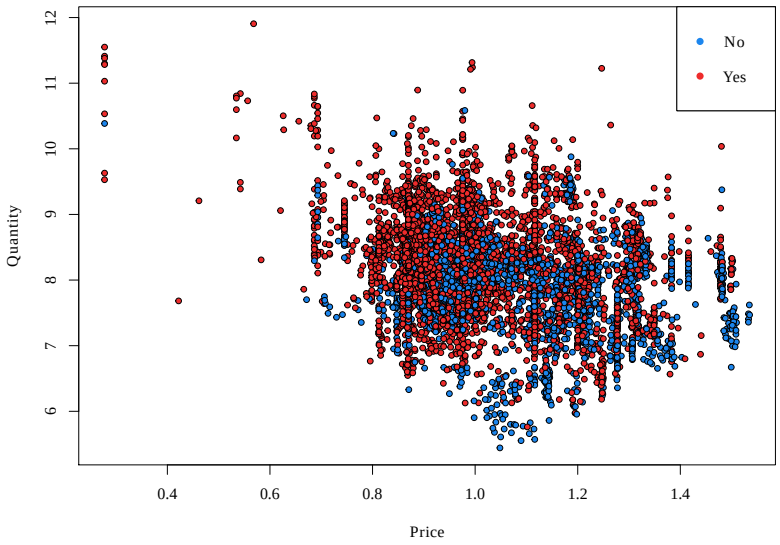

Figure 3: Scatterplot of the data on a log-log scale. The colors correspond to the presence/absence of advertisement.

By a simple descriptive analysis, in Figure 3, we see that there seems to be an effect due to the advertising (red points shifted up). However, we need to be careful because there might be confounding. In particular, red points seem to be shifted on the left, that is, display happens when the cheese is cheaper. Price is a confounder for this model, as it is correlated with both the predictor (display) and the response (quantity).

The necessity of using a hierarchical model is evident in Figure 4. In fact, in some of the groups there are no observations (or only one observation) corresponding to the presence/absence of advertising. Not pooling among the groups can lead to distorted estimates.

Using the command

```
hlm = lmer(logvol ~ (1 + logprice + disp + logprice:disp | store))
```

we fit a hierarchical model to the cheese data, after having log-transformed both price and volume. Remark that the matrix of the covariates is four dimensional: apart from the intercept, we are trying to measure the effect of the log-price, of the advertisement and of their interaction on the log-volume of sales. Thus, the demand curve is allowed to shift and to change shape depending on whether there is a display ad or not. Moreover, the estimates can vary store by store, around a common grand mean of the population. This allows the stores that have different typical volumes to be modelled differently, albeit maintaining a common structure of the population of the stores.

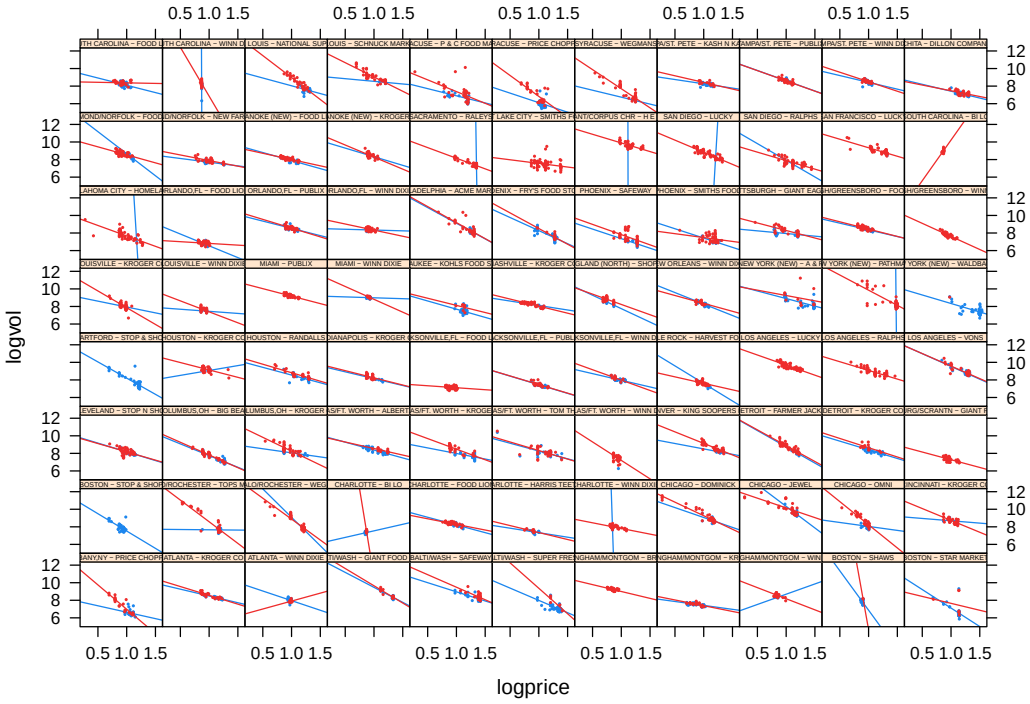

Figure 4: Observation plotted according to the different stores. The colors correspond to the presence/absence (red/blue) of advertisement. The lines represent the OLS estimates with no pooling.

The results are not shown because they are similar to the ones obtained via Gibbs sampling, which is detailed in the following.

We could also fit a fully Bayesian hierarchical model. In fact, let us assume

$$
\begin{aligned}
&\boldsymbol{Y}_i|\boldsymbol{\beta}_i, \boldsymbol{\gamma}_i, \lambda \sim N_{n_i}\left(Z_i\boldsymbol{\gamma}_i, \lambda^{-1}\mathcal{I}_{n_i}\right)\\
&\boldsymbol{\gamma}_i|D \overset{\text{iid}}{\sim} N_q\left(\mathbf{0}, D\right)\\
&D \sim \mathcal{IW}\left(\nu, \Psi\right)\\
&\boldsymbol{\beta} \sim N_p(\mathbf{0}, \lambda_0^{-1}, \mathcal{I}_p)\\
&\lambda \sim \frac{1}{\lambda}\mathcal{I}_{[0,+\infty)}(\lambda).
\end{aligned}
$$

Remark that in the model specified in this way, all of the regression parameters are group dependent (i.e. there are no fixed effects). Z_i denotes the restriction of the covariates to the i^{th} store.

In this case, it is easy to find the full conditional distributions. In fact,

- the full conditional for $\boldsymbol{\beta}$ is

$$
\begin{aligned}
p(\boldsymbol{\beta}|Y, \boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_I, \lambda) &\propto L(Y|\boldsymbol{\beta}, \boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_I, \lambda)p(\boldsymbol{\beta})\\
&\propto \prod_{i=1}^{I}\left\{\exp\left\{-\frac{\lambda}{2}(\boldsymbol{y}_i - X_i\boldsymbol{\beta} - Z_i\boldsymbol{\gamma}_i)^T(\boldsymbol{y}_i - X_i\boldsymbol{\beta} - Z_i\boldsymbol{\gamma}_i)\right\}\right\}\exp\left\{-\frac{\lambda_0}{2}\boldsymbol{\beta}^T\boldsymbol{\beta}\right\}.
\end{aligned}
$$

By defining the shifted quantity $\boldsymbol{w}_i = \boldsymbol{y}_i - Z_i\boldsymbol{\gamma}_i$ we fall in the normal linear model case, that is,

$$p(\boldsymbol{\beta}|Y, \boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_I, \lambda) \propto \exp\left\{-\frac{\lambda}{2}\sum_{i=1}^{I}(\boldsymbol{w}_i - X_i\boldsymbol{\beta})^T(\boldsymbol{w}_i - X_i\boldsymbol{\beta})\right\}\exp\left\{-\frac{\lambda_0}{2}\boldsymbol{\beta}^T\boldsymbol{\beta}\right\}$$
$$\sim N\left(\frac{\lambda}{\lambda+\lambda_0}\left(\sum_{i=1}^{I}X_i^TX_i\right)^{-1}\sum_{i=1}^{I}X_i^T\boldsymbol{w}_i, \frac{1}{\lambda+\lambda_0}\left(\sum_{i=1}^{I}X_i^TX_i\right)^{-1}\right).$$

- the full conditionals for the random effects $\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_I$ are, $\forall i \in \{1, \dots, I\}$,

$$\begin{aligned}
p(\boldsymbol{\gamma}_i|\boldsymbol{y}_i, \boldsymbol{\beta}, \lambda) &\propto p(\boldsymbol{Y}_i|\boldsymbol{\gamma}_i, \boldsymbol{\beta}, \lambda)p(\boldsymbol{\gamma}_i|D) \\
&\propto \exp\left\{-\frac{\lambda}{2}(\boldsymbol{y}_i - X_i\boldsymbol{\beta} - Z_i\boldsymbol{\gamma}_i)^T(\boldsymbol{y}_i - X_i\boldsymbol{\beta} - Z_i\boldsymbol{\gamma}_i)\right\}\exp\left\{-\frac{1}{2}\boldsymbol{\gamma}_i^TD^{-1}\boldsymbol{\gamma}_i\right\} \\
&= \exp\left\{-\frac{\lambda}{2}(\boldsymbol{m}_i - Z_i\boldsymbol{\gamma}_i)^T(\boldsymbol{m}_i - Z_i\boldsymbol{\gamma}_i)\right\}\exp\left\{-\frac{1}{2}\boldsymbol{\gamma}_i^TD^{-1}\boldsymbol{\gamma}_i\right\} \\
&\sim N\left((\lambda Z_i^TZ_i + D^{-1})^{-1}\lambda Z_i^T\boldsymbol{m}_i, (\lambda Z_i^TZ_i + D^{-1})^{-1}\right)
\end{aligned}$$

  where $\forall i \in \{1, \dots, I\}, \boldsymbol{m}_i = \boldsymbol{y}_i - X_i\boldsymbol{\beta}$.

- the full conditional for the precision parameter λ is

$$p(\lambda|Y, \boldsymbol{\beta}, \boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_I) \propto p(Y|\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_I, \boldsymbol{\beta}, \lambda)p(\lambda)$$
$$\sim \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\sum_{i=1}^{I}||\boldsymbol{y}_i - X_i\boldsymbol{\beta} - Z_i\boldsymbol{\gamma}_i||_2^2\right).$$

- the full conditional for the covariance matrix of the random effects is

$$p(D|Y, \boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_I) \propto p(\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_I|D)p(D)$$
$$\sim \mathcal{IW}\left(\nu + I, \Psi + \sum_{i=1}^{I}\boldsymbol{\gamma}_i\boldsymbol{\gamma}_i^T\right).$$

With the model specification just described, we get results that are similar to the ones of the `lme4` package. In particular, in Figure 6 one can see that the partial pooling allows the estimation of the lines also for stores that have no observation with or without advertisement. In Figure 5, instead, the regression lines that are estimated separately for each of the the same stores are shown. In the top left panel, we see a store with a balanced sample size with and without advertising. In this case, the shrinkage effect is less evident. The other three panels show pathological cases in which only a few observations (or none) are present for one of the two groups (usually the non-display group).

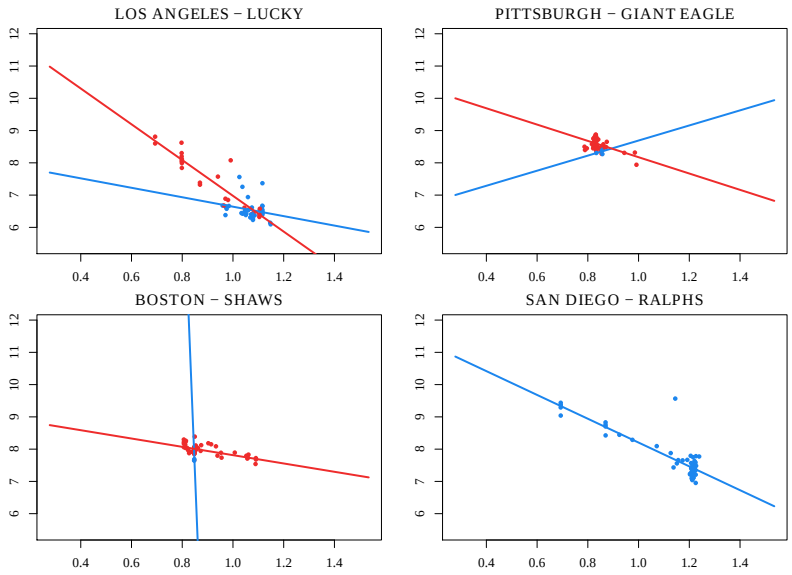

Figure 5: Regression lines fitted for four different stores separately.

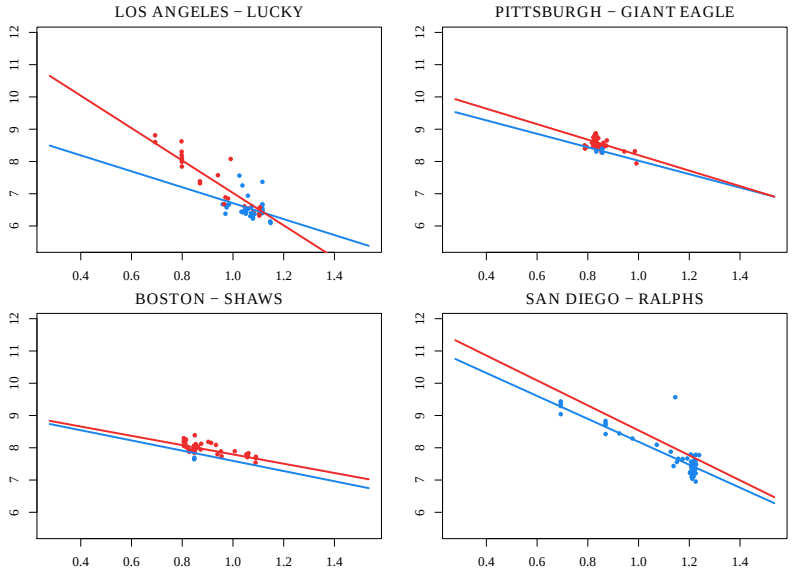

Figure 6: Regression lines fitted for four different stores in the bayesian model (that accounts for partial pooling).

Appendix A

R code

```r
# ====================
# ==== Math Tests ====
# ====================

math <- read.csv(file = 'SDS383D-master/data/mathtest.csv')

y <- math$mathscore
ybar <- aggregate(y, list(math$school), mean)$x
ni <- as.numeric(table(math$school))
n <- sum(ni)
I <- length(unique(math$school))

# Let us see the distribution of the scores for each school
par(mar=c(2,2,1,1))
boxplot(y ~ math$school, col = 'gray', pch = 16, cex = 0.8, lwd = 1.2)
abline(h = mean(y), col = 'indianred3', lwd = 2)

# Let us plot the average scores for each school vs the sample size of that school
par(mar=c(4,4,2,2), family = 'Palatino', cex = 1.1)
plot(ni, ybar, pch = 16, xlab = 'Sample size', ylab = 'Group mean')
# We notice that extreme average values of the scores are obtained for school with few
# students sampled. This happens because the distribution of ybar for each school has a
# variance of sigma^2/ni: the smaller ni, the larger the variability of ybar around the
# grand mean.


# Run the Gibbs Sampler
Niter <- 11000
burnin <- 1000
thin <- 2

# Initialize the chain
thetas.chain <- array(NA, dim = c(Niter, I))
mu.chain <- array(NA, dim = Niter)
sigma2.chain <- array(NA, dim = Niter)
tau2.chain <- array(NA, dim = Niter)
thetas.chain[1,] <- rep(0, I)
mu.chain[1] <- 0
sigma2.chain[1] <- 1
tau2.chain[1] <- 1
```

```
41
42 for (i in 2:Niter){
43   # Update thetas
44   var.post <- tau2.chain[i-1] * sigma2.chain[i-1] / (ni * tau2.chain[i-1] + 1)
45   mean.post <- (mu.chain[i-1] + tau2.chain[i-1] * ni * ybar) / (ni * tau2.chain[i-1] + 1)
46   thetas.chain[i,] <- rnorm(I, mean.post, sqrt(var.post))
47
48   # Update mu
49   theta.bar <- mean(thetas.chain[i,])
50   mu.chain[i] <- rnorm(1, theta.bar, sqrt(sigma2.chain[i-1] * tau2.chain[i-1] / I))
51
52   # Update sigma2
53   S.theta <- sum((thetas.chain[i,] - mu.chain[i])^2)
54   S.y <- sum((y - rep(thetas.chain[i,], times = ni))^2)
55   rate.new <- (1/2) * (S.y + S.theta / tau2.chain[i-1])
56   sigma2.chain[i] <- 1/rgamma(1, (n + I)/2, rate.new)
57
58   # Update tau2
59   rate.new <- S.theta / (2 * sigma2.chain[i])
60   tau2.chain[i] <- 1/rgamma(1, I/2 - 1, rate.new)
61 }
62 # Thin the chains
63 thetas.chain <- thetas.chain[seq(burnin + 1, Niter, by = thin),]
64 mu.chain <- mu.chain[seq(burnin + 1, Niter, by = thin)]
65 sigma2.chain <- sigma2.chain[seq(burnin + 1, Niter, by = thin)]
66 tau2.chain <- tau2.chain[seq(burnin + 1, Niter, by = thin)]
67
68
69 # Let us see how the Bayesian estimates differ from the sample means
70 par(mar=c(4,4,2,2), family = 'Palatino', cex = 1.1)
71 plot(ybar, colMeans(thetas.chain), xlab=bquote(bar(y)), ylab=bquote(hat(theta)), pch = 16)
72 abline(0, 1, col = 'indianred3', lwd = 2)
73 # The slope of this line is smaller than 1, that is, high values of ybar_i correspond to
74 # slightly less high values of the Bayesian estimates of theta_i; low values
75 # of ybar_i correspond to slightly less low values of the Bayesian estimates of
76 # theta_i. This is the shrinkage effect towards the grand mean (partial pooling).
77
78 par(mar=c(4,4,2,2), family = 'Palatino', cex = 1.1)
79 kappa <- (ybar - colMeans(thetas.chain)) / ybar
80 plot(ni, abs(kappa), ylab=bquote(kappa), xlab="Sample size", pch = 16)
81 abline(h=0, lwd = 2, col = 'indianred3')
82 # Groups with low sample size get shrunk the most, whereas groups with large
83 # sample size hardly get shrunk at all. The larger the sample size for a group, the more
84 # information we have for that group and the less information we need to borrow from the
85 # rest of the population.
```

Listing A.1: Math tests analysis.