7 Hierarchical Models for Combining Data

Hierarchical refers to the situation where there may be a natural structure to the data, e.g.: Individuals within regions within countries. In exploiting this we can manage instances where the parameter space is large and we may wish to reduce the dimension through distributional assumptions. For example, when we considered random effects, α_i was potentially high dimensional, but the distributional assumption that $\alpha_i \sim N(0, \sigma_\alpha^2)$ allowed us to reduce the dimension of the parameter space.

7.1 Multilevel Data

We now have new interpretations of within and between variability:

- **Within variability** Variation of individual-level data around individual time means in a panel data model. Variation of individual-level data around village means in a two-level model.
- **Between variability** Variation of individual time means around the overall mean in a panel data model. Variation of village means around the overall mean in a two-level model.

Considering a population with J groups and n_j individuals per group:

$$\begin{aligned} \{y_1, \ldots, y_{n_j}\} &\overset{\text{iid}}{\sim} p(y_j|\theta_j) \quad \text{within group} \\ p(\theta_1, \ldots, \theta_J) &= \prod_{j=1}^{J} p(\theta_j|\phi) \quad \text{between groups} \\ \phi &\sim p(\phi) \quad \text{prior} \end{aligned}$$

Within group parameters $\theta_j = (\mu_j, \sigma^2)$, between group parameters $\phi = (\psi, \tau^2)$. $p(\theta_j|\phi)$ describes heterogeneity between group means, while $p(y_j|\theta_j)$ describes within group variability. We assume the within group sampling variability is constant across groups. We can use Gibbs Sampling[1] to approximate the posterior distribution

$$p(\mu_1, \ldots, \mu_J, \sigma^2, \psi, \tau^2|y_1, \ldots, y_J)$$

7.1.1 Posterior Distributions

We can write this posterior distribution using Bayes rule:

$$\begin{aligned} p(\mu_1, \ldots, \mu_J, \sigma^2, \psi, \tau^2|y_1, \ldots, y_J) \propto & \underbrace{p(y_1, \ldots, y_J|\mu_1, \ldots, \mu_J, \sigma^2, \psi, \tau^2)}_{\text{likelihood}} \\ & \times \underbrace{p(\mu_1, \ldots, \mu_J|\psi, \tau^2)}_{\text{posterior for } \mu} \times \underbrace{p(\psi)p(\tau^2)p(\sigma^2)}_{\text{priors}} \end{aligned} \tag{7.1}$$

Note that there would normally be a large set of parameters to estimate here, by imposing the hierarchical structure we only need to estimate within gropu means (governed by the priors).

$$= \underbrace{\left[\prod_{j=1}^{J}\prod_{i=1}^{n_j} p(y_{ij}|\mu_j, \sigma^2)\right]}_{A} \underbrace{\left[\prod_{j=1}^{J} p(\mu_j|\psi, \tau^2)\right]}_{B} \underbrace{p(\psi)p(\tau^2)p(\sigma^2)}_{C}$$

[1] Train has a concise intro to Gibbs and other MCMC

- A is the conditional likelihood, since we have independence across both groups and individuals it can be expressed as 2 products.
- B is the prior for the parameters, by independence across groups we write this as a product.
- C is the prior for the hyperparameters

Explanation. ***How does the form of B relate to the distinction between FE and RE in the classical panel data model?***
RE can be represented by the full hierarchical model, we set the hyperprior mean to $\bar{y}$ (the empirical Bayes approach) and get estimates equal to classical RE estimates. The RE model involved a hyperprior that gives a distribution with a common mean for the μ_j's.
The FE model does not have a hyperprior, ie C is not included. For example each fixed effect μ_j has a prior distribution with mean 0 and variance ∞. □

If the sample size is small, the estimated variance of the group means $\bar{y}_j$ - the estimator of μ_j - will be large. It might be that the data supports some degree of pooling or shrinkage across groups to get a better estimate of μ_j[2].

7.1.2 Empirical Bayes

The form of the posterior in (7.1) can cause computational problems given the number of parameters and potentially small sample size. An Empirical Bayes approach represents an approximation to full Bayesian model.
Consider the following hierarchical model for group means:

$$\begin{aligned} \bar{y}_j|\mu_j &\overset{\text{iid}}{\sim} N(\mu_j, \sigma^2) \quad j = 1, \dots, J \quad (\sigma^2 \text{ known}) \\ \mu_j &\overset{\text{iid}}{\sim} N(\psi, \tau^2) \quad j = 1, \dots, J \quad (\tau^2 \text{ and } \psi \text{ unknown}) \end{aligned}$$

We can write the posterior distribution for μ_j as:

$$p(\mu_j|\bar{y}_j, \psi, \tau^2) = \frac{f(\bar{y}_j|\mu_j)p(\mu_j|\psi, \tau^2)}{\int f(\bar{y}_j|\mu_j)p(\mu_j|\psi, \tau^2)d\mu_j}$$

Here we have assumed σ^2 is known, so no prior is required, however there is no prior for ψ and τ^2. We can construct the posterior using estimates of these hyperparameters from the data:

$$p(\mu_j|\bar{y}_j, \hat{\psi}, \hat{\tau}^2) \sim N(\hat{S}\bar{y}_j + (1 - \hat{S})\hat{y}_j, (1 - \hat{S})\hat{\sigma}^2)$$

The EV estimator of the mean of the posterior distribution is a weighted average of the sample mean and the prior mean, where the weight is the shrinkage factor $\hat{S}$:

7.1.3 Panel Data

Consider the following unobserved linear panel data model:

$$y_{it} = \mu + \delta_i + \omega_{it}$$

The individual unobserved effects are $\mu_i = \mu + \delta_i$. Let us assume:

$$\begin{aligned} var(y_{it}) &= var(\delta_i) + var(\omega_{it}) \\ &= \tau^2 + \sigma^2 \end{aligned}$$

[2] See Stein's paradox which defines situations in which there are estimators better than the arithmetic average.

Here we're making the same assumption as in RE, that $\delta_i t$ and ω_{it} are independent. This gives us a Bayesian representation for the RE estimator:

$$y_{it}|\mu_i \sim N(\mu_i, \sigma^2) \tag{7.2}$$
$$\mu_i \sim N(\mu, \tau^2) \tag{7.3}$$
$$\mu_i|y \sim N(\hat{\mu}_i, \sigma^2 + \tau^2) \tag{7.4}$$

Note that unlike in RE we need to make assumptions on the full distributions from the outset, we can's just use moments as before.
(7.2) is the likelihood, we assume normal data here. (7.3) is the normal prior for individual effects, and (7.4) is the posterior distribution for μ_i. $\hat{\mu}_i$ is the RE estimator and the mean of the posterior distribution for μ_i. It is also an EB estimator.

Claim 7.1.1. The product of two univariate Gaussian PDFs is proportional to a Gaussian PDF.

Proof. Let $f(x)$ and $g(x)$ be two Gaussian PDFs with means μ_f and μ_g and variances σ_f^2 and σ_g^2 respectively:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma_f^2}} \exp\left(-\frac{(x-\mu_f)^2}{2\sigma_f^2}\right) \quad \text{and} \quad g(x) = \frac{1}{\sqrt{2\pi\sigma_g^2}} \exp\left(-\frac{(x-\mu_g)^2}{2\sigma_g^2}\right)$$

Their product is:

$$f(x)g(x) = \frac{1}{2\pi\sigma_f\sigma_g} \exp - \left(\frac{(x-\mu_f)^2}{2\sigma_f^2} + \frac{(x-\mu_g)^2}{2\sigma_g^2}\right)$$

Examine the term in the exponent:

$$\begin{aligned}
\frac{(x-\mu_f)^2}{2\sigma_f^2} + \frac{(x-\mu_g)^2}{2\sigma_g^2} &= \frac{(\sigma_g^2(x-\mu_f)^2 + \sigma_f^2(x-\mu_g)^2)}{2\sigma_f^2\sigma_g^2} \\
&= \frac{(\sigma_g^2+\sigma_f^2)x^2 - 2(\sigma_g^2\mu_f + \sigma_f^2\mu_g)x + \sigma_f^2\mu_g^2 + \sigma_g^2\mu_f^2}{2\sigma_f^2\sigma_g^2} \\
&= \frac{x^2 - 2\frac{\sigma_g^2\mu_f+\sigma_f^2\mu_g}{\sigma_g^2+\sigma_f^2}x + \frac{\sigma_f^2\mu_g^2+\sigma_g^2\mu_f^2}{\sigma_g^2+\sigma_f^2}}{2\frac{\sigma_f^2\sigma_g^2}{\sigma_g^2+\sigma_f^2}}
\end{aligned}$$

This is a quadratic in x, and is thus also a Gaussian function. To pin down the new parameters we just need to complete the square and compare with the standard form of the Gaussian PDF. Since a term ε can be added that is independent of x to complete the square, we can write the exponent wlog as:

$$\frac{\left(x - \frac{\sigma_g^2\mu_f+\sigma_f^2\mu_g}{\sigma_g^2+\sigma_f^2}\right)^2}{2\frac{\sigma_f^2\sigma_g^2}{\sigma_g^2+\sigma_f^2}} + \varepsilon$$

Thus we have:

$$\mu_{fg} = \frac{\sigma_g^2\mu_f + \sigma_f^2\mu_g}{\sigma_g^2+\sigma_f^2} = \frac{\mu_f/\sigma_f^2 + \mu_g/\sigma_g^2}{1/\sigma_f^2 + 1/\sigma_g^2} \quad \text{and} \quad \sigma_{fg}^2 = \frac{\sigma_f^2\sigma_g^2}{\sigma_g^2+\sigma_f^2}$$

□

We can now derive (with some abuse of notation) the posterior distribution for μ_i:

$$\begin{aligned} p(\mu_i|y) &\propto p(y|\mu_i)p(\mu_i) \\ &= N(\mu_i, \sigma^2)N(\mu, \tau^2) \end{aligned}$$

Using 7.1.1 we can write the mean of the posterior distribution as:

$$\mathbb{E}[\mu_i|y] = \frac{\frac{\mu}{\tau^2} + \frac{\mu_i}{\sigma^2}}{\frac{1}{\tau^2} + \frac{1}{\sigma^2}}$$

We can write our finite sample estimate $\hat{\mu}_i$ as:

$$\begin{aligned} \hat{\mu}_i &= \frac{\frac{\bar{y}_i}{\hat{\sigma}^2} + \frac{\bar{y}}{\hat{\tau}^2}}{\frac{1}{\hat{\sigma}^2} + \frac{1}{\hat{\tau}^2}} \\ &= \frac{\hat{\sigma}^2\bar{y}_i + \hat{\tau}^2\bar{y}}{\hat{\sigma}^2 + \hat{\tau}^2} \\ &= (1 - \hat{S})\bar{y}_i + \hat{S}\bar{y} \quad \text{where} \quad \hat{S} = \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \hat{\tau}^2} \end{aligned}$$

For $0 < \hat{S} < 1$, $\hat{\mu}_i$ is a compromise between the pooled estimator $\hat{S} = 1$ and the (individual means) FE estimator $\hat{S} = 0$.

Explanation. ***What happens when*** $\tau^2 \to 0$***?***
As $\tau^2 \to 0$, there is no longer any variation between individuals, so it is optimal to use the fully pooled model. Clearly we can see:

$$\lim_{\tau^2 \to 0} \hat{S} = 1 \quad \Rightarrow \quad \lim_{\tau^2 \to 0} \hat{\mu}_i = \bar{y}$$

that our estimator of μ_i becomes the population mean.
What happens when $\tau^2 \to \infty$***?***
As $\tau^2 \to \infty$, the individual means are too noisy to be informative, so it is optimal to use the individual means estimator (FE). This is analogous to RE → FE when $\sigma_\alpha^2 \to \infty$¿

$$\lim_{\tau^2 \to \infty} \hat{S} = 0 \quad \Rightarrow \quad \lim_{\tau^2 \to \infty} \hat{\mu}_i = \bar{y}_i$$

□

Explanation. ***How does*** (7.1) ***differ from the RE distribution?***
The distribution we considered in RE was $\alpha_i \sim i.i, d, (0, \sigma_\alpha^2)$, however the mean is unimportant since any mean would be absorbed by the constant terms.
We are making the same *i.d.d.* assumption here, with similar assumptions on the variance. The difference is that now we require the full distribution, not just the moments. □

How do Bayesians conceptualise fixed versus random effects estimators when all effects are random? The classical fixed vs RE dichotomy is not relevant. Here the distinction is:

- RE: Hierarchical prior $\mu_i \sim N(\mu, \tau^2)$
- FE: Non-hierarchical independence prior for each μ_i

7.2 Model Averaging

The problem: Estimation and inference on the determinants of y, where the set of regressors is large. Let θ denote the parameters of the regressors included in a model, we can estimate the posterior density as $p(\theta|y, M_j)$, where M_j is the jth model. Suppose there are K potential regressors, model M_j is described by a $K \times 1$ binary vector γ_j where $\gamma_{jk} = 1$ if regressor k is included in model M_j, and $\gamma_{jk} = 0$ otherwise.
The model space M is thus the set of all 2^K possible models. *How do we account for model uncertainty in making unconditional (on the space of models) inference on any given element of θ?*

7.2.1 Marginal Likelihood, Prior and Posterior Odds and the Bayes Factor

Marginal likelihood (integrated over the parameter space):

$$\ell(y|M_j) = \int_\theta p(y|\theta, M_j)\ell(\theta|M_j)d\theta$$

For two models M_i and M_j the posterior odds is given by:

$$\frac{p(M_i|y)}{p(M_j|y)} = \frac{p(M_i)}{p(M_j)}\frac{\ell(y|M_i)}{\ell(y|M_j)}$$

Using Bayes theorem we can write the posterior model probabilities as:

$$p(M_j|y) = \frac{p(M_j)\ell(y|M_j)}{\ell(y)} \propto p(M_j)\ell(y|M_j)$$

If the set of models is exhaustive, we can write:

$$p(M_j|y) = \frac{p(M_j)\ell(y|M_j)}{\sum_{i=1}^{2^K} p(M_i)\ell(y|M_i)}$$

We can compute the posterior distribution of θ given model M_j as:

$$p(\theta|y, M_j) = \frac{\ell(y|M_j, \theta)p(\theta|M_j)}{p(y|M_j)}$$

And the unconditional posterior distribution of θ is:

$$p(\theta|y) = \sum_{j=1}^{2^K} p(\theta|y, M_j)p(M_j|y)$$

We can thus think of the posterior model probabilities as the weights $p(M_j|y)$ that we should attach to the posterior distributions $p(\theta|y, M_j)$ when averaging over the model space.

7.2.2 Prior

How do we get priors for:

- The model space $p(M_j)$
- Elements of θ that are model=specific (EG regression parameters)
- Elements of θ that are common to all models (EG variance parameters)

Here we will focus on $p(M_j)$. Recall that a linear model is described by a set of binary variables responsible for including/excluding regressors. An independent Bernoulli prior for a given model $p(M_j|\pi)$ can be written as:

$$p(M_j|\pi) = p(\gamma|\pi) = \prod_{k=1}^{K} \pi_k^{\gamma_{jk}} (1-\pi_k)^{1-\gamma_{jk}}$$

where π_k is the independent prior probability that regressor k is included in the model. For example a uniform prior across all elements, $\pi = 0.5$ implies the number of covariates should be large. Note the limitation of this approach, we would expect the presence of some regressors to be correlated with the presence of others.

Definition 7.2.1: Prior Inclusion Probability

$$p(\gamma_{jk} = 1|y) = \sum_{j=1}^{2^K} \mathbb{1}(\gamma_{jk} = 1|y, M_j) p(M_j|y)$$

Note that this is essentially a counting rule, we are summing the probabilities of a model being selected which includes $\gamma_{jk} = 1$. This can be thought of as a measure of how important regressor k is in the model space.