

Lecture Note 8: Hierarchical and Panel Models (revised 03/9/08)

A Simple Hierarchical Model

Suppose we have data on multiple individuals, and for each individual we observe some quantity multiple times. We suppose that the quantity has a different mean for each individual. A simple model that we could start with is:

$$\begin{aligned}\alpha_i | \mu_\alpha, \sigma_\alpha, \sigma &\stackrel{\text{iid}}{\sim} N(\mu_\alpha, \sigma_\alpha^2), & i = 1, \dots, n; \\ y_{it} | \alpha_i, \sigma, \mu_\alpha, \sigma_\alpha &\stackrel{\text{iid}}{\sim} N(\alpha_i, \sigma^2), & i = 1, \dots, n, \quad t = 1, \dots, T.\end{aligned}$$

Here i indexes individuals, and t indexes time periods. We only directly observe the $\{y_{it}\}$.

A crude estimate of each α_i could be obtained by averaging within individual:

$$\hat{\alpha}_i = \frac{1}{T} \sum_{t=1}^T y_{it},$$

but if T was small, then these estimates would be very imprecise. Since the α_i are linked through their common distribution $N(\mu_\alpha, \sigma_\alpha^2)$, a full analysis that took account of this structure could use the data more effectively.

This type of model is often called a hierarchical model, because there are multiple “levels,” one determining the α_i , and another determining the y_{it} . We sometimes distinguish between “parameters” α_i, σ and “hyperparameters” μ_α and σ_α .

In economics we often refer to models with this type of grouping structure as panel data models. If we want to derive asymptotic approximation theory for estimators in this model, we would need to specify whether $n \rightarrow \infty$, or $T \rightarrow \infty$, or both. The asymptotic approximations could be quite different depending on whether we regards n as “large” and T fixed, or T as “large” with n fixed.

To do a Bayesian analysis, we need to complete the model by specifying prior distributions on $\mu_\alpha, \sigma_\alpha$, and σ . It turns out that in some hierarchical models, we need to be careful about the choice of priors. One issue is that using improper priors for both variance parameters may lead to an improper posterior. To avoid this, we will specify $\mu_\alpha, \sigma_\alpha, \sigma$ as independent, with diffuse but proper priors for σ_α and σ . In particular, we assume

$$p(\mu_\alpha, \sigma_\alpha, \sigma) = p(\mu_\alpha)p(\sigma_\alpha)p(\sigma),$$

with $p(\mu_\alpha) \propto 1$ and

$$\frac{1}{\sigma_\alpha^2} \sim \frac{\chi_1^2}{.01},$$

$$\frac{1}{\sigma^2} \sim \frac{\chi_1^2}{.01}.$$

This is equivalent to using $\mathcal{G}(1/2, .01/2)$ priors for the two precisions.

We now need to figure out how to simulate the posterior distribution of $\mu_\alpha, \sigma_\alpha, \sigma, \alpha_1, \dots, \alpha_n$. We will use a Gibbs sampling algorithm. Our blocking will be:

1. $(\alpha_1, \dots, \alpha_n)$
2. σ
3. $\mu_\alpha, \sigma_\alpha$.

So, for each of the three blocks, we need to determine the form of the full conditional density. It is useful to start with the complete joint density of all the random variables. For simplicity, let $\alpha = (\alpha_1, \dots, \alpha_n)$ denote all the α_i , and let $y = (y_{11}, \dots, y_{1T}, \dots, y_{nT})$ denote all the y_{it} . Then we can write

$$p(y, \alpha, \sigma, \mu_\alpha, \sigma_\alpha) = p(\mu_\alpha, \sigma_\alpha, \sigma) p(\alpha | \mu_\alpha, \sigma_\alpha, \sigma) p(y | \alpha, \sigma, \mu_\alpha, \sigma_\alpha).$$

Notice that this simplifies substantially: since $\mu_\alpha, \sigma_\alpha, \sigma$ are independent in the prior,

$$p(\mu_\alpha, \sigma_\alpha, \sigma) = p(\mu_\alpha) p(\sigma_\alpha) p(\sigma).$$

Also, from the initial statement of the model,

$$p(\alpha | \mu_\alpha, \sigma_\alpha, \sigma) = p(\alpha | \mu_\alpha, \sigma_\alpha),$$

(because the α_i are independent of σ) and

$$p(y | \alpha, \sigma, \mu_\alpha, \sigma_\alpha) = p(y | \alpha, \sigma).$$

So we can write

$$p(y, \alpha, \sigma, \mu_\alpha, \sigma_\alpha) = p(\mu_\alpha) p(\sigma_\alpha) p(\sigma) p(\alpha | \mu_\alpha, \sigma_\alpha) p(y | \alpha, \sigma).$$

This type of simplification due to conditional independence is typical in hierarchical models.

Now, let's figure out the full conditional densities for the Gibbs sampler.

1. α :

We need to derive

$$p(\alpha | y, \sigma, \mu_\alpha, \sigma_\alpha).$$

By the usual definition of conditional densities,

$$p(\alpha|y, \sigma, \mu_\alpha, \sigma_\alpha) \propto p(y, \alpha, \sigma, \mu_\alpha, \sigma_\alpha).$$

Substituting for the joint density, and then dropping terms that do not involve α ,

$$\begin{aligned} p(\alpha|y, \sigma, \mu_\alpha, \sigma_\alpha) &\propto p(\mu_\alpha)p(\sigma_\alpha)p(\sigma)p(\alpha|\mu_\alpha, \sigma_\alpha)p(y|\alpha, \sigma) \\ &\propto p(\alpha|\mu_\alpha, \sigma_\alpha)p(y|\alpha, \sigma) \end{aligned}$$

Note that

$$p(\alpha|\mu_\alpha, \sigma_\alpha) = \prod_{i=1}^n \phi(\alpha_i|\mu_\alpha, \sigma_\alpha^2),$$

(where ϕ denotes a normal density), and

$$p(y|\alpha, \sigma) = \prod_{i=1}^n \prod_{t=1}^T \phi(y_{it}|\alpha_i, \sigma^2).$$

Both terms factor in α_i . So the α_i are conditionally independent of each other, and we can draw for them individually. Looking at the density for an individual α_i ,

$$p(\alpha_i|y, \sigma, \mu_\alpha, \sigma_\alpha) \propto \frac{1}{\sqrt{2\pi\sigma_\alpha^2}} \exp\left(-\frac{1}{2\sigma_\alpha^2}(\alpha_i - \mu_\alpha)^2\right) \cdot \prod_{t=1}^T \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_{it} - \alpha_i)^2\right).$$

This has the form of the posterior in a normal model for y_{i1}, \dots, y_{iT} , when the variance is known, and the mean α_i has a $N(\mu_\alpha, \sigma_\alpha^2)$ prior distribution. So our results from LN4 apply, and the density for α_i will be normal.

2. σ :

We want to derive

$$p(\sigma|\mu_\alpha, \sigma_\alpha, \alpha, y).$$

We can simplify this as:

$$\begin{aligned} p(\sigma|\mu_\alpha, \sigma_\alpha, \alpha, y) &\propto p(y, \alpha, \sigma, \mu_\alpha, \sigma_\alpha) \\ &= p(\mu_\alpha)p(\sigma_\alpha)p(\sigma)p(\alpha|\mu_\alpha, \sigma_\alpha)p(y|\alpha, \sigma) \\ &\propto p(\sigma)p(y|\alpha, \sigma). \end{aligned}$$

The first term corresponds to the Gamma prior for σ^{-2} . The second term can be written as

$$p(y|\alpha, \sigma) = \prod_{i=1}^n \prod_{t=1}^T \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_{it} - \alpha_i)^2\right).$$

This has the form of a $N(0, \sigma^2)$ likelihood for $(y_{it} - \alpha_i)$. That is, it corresponds to the model:

$$(y_{it} - \alpha_i) \stackrel{\text{iid}}{\sim} N(0, \sigma^2).$$

Since we can use the current value for the α_i within the Gibbs sampler to form the $(y_{it} - \alpha_i)$ terms, we can use the results of LN4 to obtain the scaled chi-square posterior for σ^{-2} .

3. $\mu_\alpha, \sigma_\alpha$:

Write

$$\begin{aligned} p(\mu_\alpha, \sigma_\alpha | y, \alpha, \sigma) &\propto p(y, \alpha, \sigma, \mu_\alpha, \sigma_\alpha) \\ &= p(\mu_\alpha) p(\sigma_\alpha) p(\sigma) p(\alpha | \mu_\alpha, \sigma_\alpha) p(y | \alpha, \sigma) \\ &\propto 1 \cdot p(\sigma_\alpha) p(\alpha | \mu_\alpha, \sigma_\alpha) \end{aligned}$$

Notice that

$$p(\alpha | \mu_\alpha, \sigma_\alpha) = \prod_{i=1}^n \phi(\alpha_i | \mu_\alpha, \sigma_\alpha^2).$$

This has the form of the likelihood for the model:

$$\alpha_i | \mu_\alpha, \sigma_\alpha \stackrel{\text{iid}}{\sim} N(\mu_\alpha, \sigma_\alpha^2).$$

So we have yet another normal likelihood, and since we have chosen the prior for $(\mu_\alpha, \sigma_\alpha)$ in a conjugate form, we can use the results of LN4 to obtain the density.

Linear Panel Data Model

A commonly used model in econometrics has

$$y_{it} = \alpha_i + x'_{it}\beta + \epsilon_{it},$$

where x_{it} does *not* contain a constant term.

This is not a complete model yet, because we have not specified the distribution of the ϵ_{it} and the α_i . We could make this similar to the simple hierarchical model of the previous section, by specifying that

$$\alpha_i | X, \beta, \mu_\alpha, \sigma_\alpha, \sigma \stackrel{\text{iid}}{\sim} N(\mu_\alpha, \sigma_\alpha^2),$$

and

$$\epsilon_{it} | X, \beta, \mu_\alpha, \sigma_\alpha, \sigma, \alpha \stackrel{\text{iid}}{\sim} N(0, \sigma^2).$$

Here, we use α to denote the entire vector $(\alpha_1, \dots, \alpha_n)$ as before, and use X to denote all the covariates $\{x_{it}\}$. If we further specify the prior distribution for $(\mu_\alpha, \sigma_\alpha, \beta, \sigma)$, then we could devise a Gibbs sampler in a similar way to the previous example.

There are a couple of different blockings that would work here. One possibility is to rewrite the model as

$$y_{it} = \mu_\alpha + x'_{it}\beta + v_i + \epsilon_{it},$$

$$v_i | X, \beta, \mu_\alpha, \sigma_\alpha, \sigma \stackrel{\text{iid}}{\sim} N(0, \sigma_\alpha^2).$$

Then, defining $\tilde{x}'_{it} = (1, x'_{it})$, and $\tilde{\beta} = (1, \beta)'$, we could write

$$y_{it} = \tilde{x}'_{it}\tilde{\beta} + v_i + \epsilon_{it}.$$

So (provided the priors had a convenient form) we could devise a Gibbs sampling with the blocking

1. μ_α, β
2. $v = (v_1, \dots, v_n)$
3. σ
4. σ_α .

Alternatively, we could keep the original form of the model, and block as: (α) , (β) , $(\mu_\alpha, \sigma_\alpha)$, and (σ) .

However, we should be aware that the assumption

$$\alpha_i | X, \beta, \mu_\alpha, \sigma_\alpha, \sigma \stackrel{\text{iid}}{\sim} N(\mu_\alpha, \sigma_\alpha^2),$$

is not innocuous. In particular, it defines the α_i to be independent of the X conditional on the parameters.

To see why this might be a problem, let us consider a concrete example.

Cobb-Douglas Production Function

Consider a firm which produces output using a technology described by a Cobb-Douglas production function:

$$Y(K, L) = AK^\gamma L^{\gamma^2}. \tag{1}$$

Here K is capital input, L is labor input, and Y is output. If $\gamma_1 + \gamma_2 < 1$ (so there are nonincreasing returns to scale), we might suppose that the firm chooses capital and labor to maximize profits, taking output price p_y and input prices p_k, p_l as given. In this case, factor demands for K and L will be increasing in A , since firms will hire inputs until their marginal revenue product equals their price.

We are interested in connecting this economic model to empirical data on firms. We will be working with data based on many firms, so we need to specify the production functions for *each* firm. Suppose we can write, for $i = 1, \dots, n$,

$$Y_i(K, L) = A_i K^{\gamma_1} L^{\gamma_2}.$$

This expresses the notion that each firm has a Cobb-Douglas production function with common coefficients γ_1, γ_2 . However, there are differences in how efficient the firms are, which arises from variation in A_i across firms.

Notice that K, L are not “data” but are simply arguments in the function $Y_i(K, L)$. We will use K_i and L_i to denote the amounts of capital and labor actually chosen by the firm, and Y_i to denote the actual output of the firm. Then the Cobb-Douglas model implies that

$$Y_i = A_i K_i^{\gamma_1} L_i^{\gamma_2},$$

or taking logs:

$$\log Y_i = \log A_i + \gamma_1 \log K_i + \gamma_2 \log L_i.$$

To simplify notation, let us define $y_i \equiv \log Y_i$, and similarly for A_i, K_i , and L_i . Then we can write

$$y_i = \gamma_0 + \gamma_1 k_i + \gamma_2 l_i + u_i,$$

where $\gamma_0 \equiv E(a_i)$, and $u_i = a_i - \gamma_0$. This looks like a classical regression model for y_i given k_i and l_i . However, in the classical regression model, the error term is assumed to satisfy

$$u_i | k, l, \gamma, \sigma \stackrel{\text{iid}}{\sim} N(0, \sigma^2),$$

where k and l refer to all the observations on the inputs, and γ is the vector $\gamma_0, \gamma_1, \gamma_2$.

So we would be assuming that u_i is independent of k_i and l_i . But recall that under price-taking and profit-maximization, we would expect that k_i and l_i are related quite strongly to efficiency a_i and hence to u_i . So, if we define u_i as (demeaned) efficiency, then we cannot assume it is independent of the regressors. If we just view this as a statistical model, with disturbances u_i independent of regressors, then γ gives the conditional mean of y_i given a constant, k_i and

l_i , but cannot be interpreted as parameters of the production function. Moreover, if we don't assume u_i is independent of the inputs, it's not clear how to estimate the production function parameters unless we make quite strong assumptions about factor demand.

One possible solution emerges if the firms are observed in multiple time periods. Suppose for each firm, we observe output and measured inputs in each of T years. We will denote these observations by (y_{it}, k_{it}, l_{it}) , for $i = 1, \dots, n$, $t = 1, \dots, T$. Suppose that our previous model continues to hold, so that

$$y_{it} = a_{it} + \gamma_1 k_{it} + \gamma_2 l_{it}.$$

Here a_{it} is interpreted as a measure of firm i 's efficiency at time t . If we write $a_{it} \equiv \alpha_i + u_{it}$, then we can write our model as

$$y_{it} = \alpha_i + \gamma_1 k_{it} + \gamma_2 l_{it} + u_{it}.$$

We can simplify the notation slightly by writing

$$y_{it} = \alpha_i + x'_{it}\beta + u_{it},$$

where

$$\beta = (\gamma_1, \gamma_2)', \quad x_{it} = (k_{it}, l_{it})'.$$

We could interpret α_i as capturing firm-specific inputs, such as management quality, which do not change over time. We might then assume that

$$u_{it} | X \alpha_1, \dots, \alpha_n, \beta, \sigma \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2).$$

(Formally, this should also be conditional on any parameters involved in the distribution of α_i as well, but we haven't defined those yet, so we will be a little loose here.) This is a strong assumption, but in some cases we might argue that u_{it} reflects idiosyncratic shocks to production that cannot be forecasted by the firm when it makes its production decisions.

Even after all this, though, we would definitely not want to assume that the α_i are independent of the regressors. So our earlier assumption, that

$$\alpha_i | X, \gamma, \sigma, \mu_\alpha, \sigma_\alpha \stackrel{\text{iid}}{\sim} N(\mu_\alpha, \sigma_\alpha^2)$$

would not be appropriate. Since the factor inputs k_{it}, l_{it} depend on α_i , the distribution of α_i conditional on the inputs should depend on them.

One possibility is to model the α_i as:

$$\alpha_i | X, \xi, \sigma_\alpha, \gamma, \sigma \stackrel{\text{iid}}{\sim} N(\bar{x}'_i \xi, \sigma_\alpha^2),$$

where

$$\bar{x}_i = \frac{1}{T} \sum_{t=1}^T x_{it}.$$

This is a type of “correlated random effects” model. The α_i are the random effects, and they are allowed here to be correlated with the regressors. As an exercise, you should try to work out a Gibbs sampling algorithm for this model.