## Introduction to Panel Data

A panel data set (or just a panel) is a set of data $(y_{it}, x_{it})$ $(i = 1, \ldots, N ; \ \ t = 1, \ldots T)$ with two indices.

Assume that you want to estimate the model

$$y_{it} = X_{it}\beta + \epsilon_{it}$$

where $\text{var}(\epsilon_{it}) = \sigma^2$, $y_{it}$ is a scalar, and $x_{it}$ is $1 \times k$ vector containing $k$ regressors.

There are three well-known (and inexpensive) textbooks on panel data (all quite good and likely available in paperback):

Hsiao: *Analysis of Panel Data,* Cambridge University Press, 1986).

Baltagi: *Econometric Analysis of Panel Data,* Wiley, 1995.

Arellano: *Panel Data Econometrics*, Oxford University Press, 1995.

The Hsiao book is a little dated by now and more focussed on micro-applications, but it is still good to read. Arellano's book is probably the most comprehensive and it is very well written. The only draw-back is that is it covers more ground that you would want if you are new to field. Baltagi's book is also good. This note here is intended to get you started, see also the much-used first-year graduate textbook by Greene, Ch. 14. (By the way: Arellano's book has an appendix on GMM that looks quite accessible.) You should check for new versions and I recommend checking Arellano's web-page (http://www.cemfi.es/ arellano/) for recent work if you need to go deeper.

Examples

1. Assume you have quarterly observations of $y_s, x_s$ for the period (say) $1960.1 \rightarrow 1989.4$. If you want to stress that quarters are different from years, you can write, e.g., the $y_s$ data as

$$y_{it} \text{ where } i \text{ is } year \text{ and } t \text{ is } quarter$$

and you may feel like storing the $y$-data in a matrix

$$Y = [y_{it}] = \begin{bmatrix} y_{1960,1} & y_{1960,2} & \cdots & y_{1960,4} \\ \vdots & & & \vdots \\ y_{1989,1} & \cdots & \cdots & y_{1989,4} \end{bmatrix}$$

Usually one doesn't write the model with quarterly data like this. Example 1 demonstrates that whether you write $y_s$, $s = 1, \ldots, 160 \, (NT)$ or $y_{it}$ $(i = 1, \ldots, 40 \, (N); \; t = 1, \ldots, 4 \, (T))$ is just a convention. Albeit often a practical convention which we use if we want to work with models that display different features for the $t$-index than for the $i$ index.

2. Very typical micro example. Assume that you observe (say) wages $w_{it}$ and experience (and age, education,...) $X_{it}$ for a sample of individuals $i = 1, \ldots, N$ over time $t = 1, \ldots, T$.

This is the type of panel that most people think of when you say "panel data." For this type of data you usually have $N$ large (thousands) and $T$ small (often 2 or 3 and 30–40 at the most).

A much utilized data set in the US is the PSID a panel which have been following a sample of $4000 +$ US families since 1968.

Close to 1000 articles have been published using the PSID.

A lot of theoretical econometric work on panel data is relevant for small $T$, large $N$ and until very recently "panel data" was considered a microeconometrics subfield.

("Small" $T$ and "large" $N$ means that one relies on asymptotic theory derived keeping $T$ fixed and letting $N$ go to infinity.) For this class, however, we are more interested in macro-panels with a rather small $N$ dimension (by the standards of, say, labor-economists). The econometric issues are somewhat different for such panels.

3. A typical macro example is a "consumption function" (simplified a bit)

$$c_{it} = \beta y_{it} + \epsilon_{it}; \quad i = 1, \ldots, 50 \quad t = 1963, \ldots, 1990$$

where $i$ indices U.S. states, $c_{it}$ is state level consumption, $y_{it}$ is state level GDP (called GSP).

This type of panel is often called a "square panel."

If you have the same numbers of observations for each $i$, as implicitly assumed above, we talk about a balanced panel (otherwise unbalanced. In macro, panels are typically balanced).

Estimation

Estimation of the model

$$y_{it} = x_{it}\beta + \epsilon_{it}; \quad var(\epsilon_i) = \sigma^2 \text{ and } \epsilon_{it} \text{ iid}$$

poses no new problems. Since the double index is just a convention you "stack" the data in vector-matrix form

$$y \;=\; X\beta + e,$$

where $y$ and $X$ are of dimension $(NT \times 1)$ and $(NT \times K)$, respectively. Typically the

3

data are stacked as

$$y = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1T} \\ y_{21} \\ \vdots \\ y_{NT} \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1T} \\ x_{21} \\ \vdots \\ x_{NT} \end{pmatrix}$$

or

$$y = \begin{pmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{N1} \\ y_{12} \\ \vdots \\ y_{NT} \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{N1} \\ x_{12} \\ \vdots \\ x_{NT} \end{pmatrix}.$$

(The $Y$ data can be turned into vector-form any which way you want; however, the $y$'s and the $x$'s has to be stacked the same way.) The GAUSS program stacks the data using the RESHAPE command, which follows the convention of first stacking the data from the first row of the matrix, then the second row, etc.

The OLS estimator of $\beta$ is now the standard OLS estimator: $\hat{\beta} = (X'X)^{-1}X'y$.

Typically, we write the data in panel form because we want to model some feature differently for, e.g., states than for time-periods. The simplest example (which is used "all the time" in applied work) is the (one-way) <u>fixed-effect model</u>.

(1) $$y_{it} = \alpha_i + \beta x_{it} + \epsilon_{it} \; ; \quad \epsilon_{it} \; iid(0, \sigma^2).$$

Note that since there is now a constant for each $i$, there can be no constant included in the $x_{it}$ vector.

The model (1) is nothing but a <u>Dummy Variable</u> model. Think of Ex. 1. Usually

the dummy variable model is written as

$$(*) \qquad y_s = \alpha_1 D_{1s} + \alpha_2 D_{2s} + \alpha_3 D_{3s} + \alpha_4 D_{4s} + \beta X_s$$

where

$$D_{is} = \begin{cases} 1 & \text{if period } s \text{ is in the } i'\text{th quarter} \\ 0 & \text{otherwise} \end{cases}$$

The notation in (1) is nothing but a more handy notation for (*). For the model with a dummy variable for each of (say) 2000 individuals in a panel, it becomes very impractical to use the notation (*). Since (1) is a dummy variable model the $X$ matrix for OLS estimation of (1) is

$$X = \begin{pmatrix} 1 & 0 & \ldots & 0 & x_{11} \\ \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & & 0 & x_{1T} \\ 0 & 1 & \ldots & 0 & x_{21} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 1 & & 0 & x_{2T} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & & 1 & x_{N1} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & x_{NT} \end{pmatrix},$$

of dimension NT×(N+K).

The estimator of $(\alpha_1, \ldots, \alpha_N, \beta)$ is $(X'X)^{-1}X'Y$. Simple! <u>Except</u> that your $X'X$ matrix is now $(N + K) \times (N + K)$. In a typical micro panel $(N > 1000)$ you cannot invert $X'X$ by brute force (the computer runs out of memory). So what to do? Make use of the <u>Frisch-Waugh theorem</u>. (This is a result which often comes in handy.)

The **Frisch-Waugh theorem** shows how to estimate the parameter vector $\beta_2$ from the model

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

without estimating the full model in one go, but in such a way that the estimated $\hat{\beta}_2$ is equal to the estimate one would get, had one estimated the full model. Here $X_1$ and $X_2$ are 2 groups of regressors (think of the "dummy-variable part" and the other regressors above as $X_1$ and $X_2$, resp.)

Step 1. Regress $y$ on $X_1$ and calculate the residuals $\tilde{y}$. (The estimated $\tilde{\beta}_1$ is <u>not</u> in general equal to the estimate you would get from the full model.) Regress $X_2$ on $X_1$ (if there are several regressors in $X_2$ one can do it one-by-one) and calculate the residuals $\tilde{X}_2$.

Step 2. Regress $\tilde{y}$ on $\tilde{X}_2$. This gives the OLS estimator $\hat{\beta}_2$ from the full model.

Step 3. (Often not stated as part of the theorem.) You can find $\hat{\beta}_1$ (the OLS estimator from the full model), by regressing $y - X_2\hat{\beta}_2$ on $X_1$.

You can prove the Frisch-Waugh theorem by using partitioned matrices (see Greene's textbook), or more elegantly by using projection theory. The application of Frisch-Waugh's theorem here is to first regress $y$ on the N "dummy-variable columns" of the X-matrix.

Note that the first $N$ columns in $X$ are orthogonal to each other, so it is only the $X_{it}$ column(s) which are not orthogonal to each of the "dummy-variable columns"

Regressing $X_2 = (X_{it})$ ; $i = 1, ..., N$ ; $t = 1, ..., T$ on

$$X_1 = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix},$$

can be done by regressing on each column one-by-one since the columns are <u>orthogonal</u>. (This follows, for example, from the Frisch-Waugh theorem.)

The regression of $X_2$ (for simplicity consider the number of non-dummy regressors to be 1) on $Z_1 = (1, ..., 1, 0, ..., 0, ..., 0, ..., 0)'$ is simply $\frac{Z_1 X_2}{Z_1' Z_1} = \frac{\sum_t X_{1t}}{T} = \frac{1}{T} \sum_t X_{1t}$ often denoted $\overline{X}_1.$ or just $X_1.$ , called the "state-specific mean." (In the case where $i$ is state, otherwise "person-specific mean", etc.) More precisely, it is the "state 1 specific mean", but similar formulas hold for regression on the other columns $i = 2, ..., N$.

The residuals from the regression on $Z_1$ are

$$\begin{aligned} X_{1t} &- X_1. \quad \text{and similarly} \\ X_{it} &- X_i. \quad \text{for general } i \end{aligned}$$

Now you get $\hat{\beta}$ by regressing $y_{it}$ on $(X_{it} - \overline{X}_i.)$ by standard OLS.

Often we do not care about $\hat{\alpha}_i,$ but sometimes we do and we get the estimate from plugging the OLS estimate of $\beta$ into the model

$$y_{it} = \alpha_i + \beta X_{it} + \epsilon_{it}.$$

We have already found $\hat{\beta}$ and we can therefore find $\hat{\alpha}_i$ from the regression model

$$(y_{it} - \hat{\beta} X_{it}) = \alpha_i + u_{it},$$

where you recall that the right-hand side is shorthand for dummy variables. From dummy variable theory we know that the OLS estimator is

$$
\begin{aligned}
\hat{\alpha}_i &= (y_{it} - \hat{\beta}X_{it}) \ \text{ averaged over } \ t \\
&= y_{i\cdot} - \hat{\beta}X_{i\cdot} \ .
\end{aligned}
$$

Note that in the fixed-effect model with dummy variables $\alpha_i$ one cannot estimate the coefficient to a regressor which is only a function of $i$ .

E.g.,

$$
\begin{aligned}
X_{UK,1} &= \alpha_{UK} + X_{UK}\,\beta \\
X_{UK,2} &= \alpha_{UK} + X_{UK}\,\beta \\
X_{UK,3} &= \alpha_{UK} + X_{UK}\,\beta \\
X_{US,1} &= \alpha_{US} + X_{US}\,\beta \\
X_{US,2} &= \alpha_{US} + X_{US}\,\beta \\
X_{US,3} &= \alpha_{US} + X_{US}\,\beta
\end{aligned}
$$

The $X$ matrix is

$$
\begin{pmatrix}
1 & 0 & X_{UK} \\
1 & 0 & X_{UK} \\
1 & 0 & X_{UK} \\
0 & 1 & X_{US} \\
0 & 1 & X_{US} \\
0 & 1 & X_{US}
\end{pmatrix}
$$

and since third column is $X_{UK} \times$ (first column) $+ X_{US} \times$ (second column) the matrix is singular = perfect multicolinearity.

Dummy variables for each time period ("time-fixed effects") are treated the same way and the inclusion of both state- and time-fixed effects is easily handled by removing the time-specific and the state-specific averages sequentially.

In Asdrubali, Sørensen, and Yosha (1996), the inclusion of time-fixed effects is crucial for the interpretation of the regressions as risk sharing estimations. As argued, for example by Cochrane (1991), there are reasons to believe that it is more robust to include a constant in a cross-sectional risk sharing regression than it is to control for aggregate consumption. In this connection, if is important to realize that a panel regression with

time-fixed effects results in an estimate that is a weighted average of cross-sectional regressions run period by period.

Consider the OLS formula:

$$\hat{\beta} \;=\; \Sigma_t \Sigma_i y_{it}(x_{it} - x_{\cdot t})/\Sigma_t \Sigma_i (x_{it} - x_{\cdot t})^2 \;.$$

In the summation over, e.g., the $x$'s we need to sum over all the observations in any order we choose, so we can choose to sum over the $i$-index first. We can rewrite the formula as

$$
\begin{aligned}
\hat{\beta} \;&=\; \frac{1}{\Sigma_t \Sigma_i (x_{it} - x_{\cdot t})^2} \, \Sigma_t \, \Sigma_i \, y_{it}(x_{it} - x_{\cdot t}) \\
&=\; \frac{1}{\Sigma_t \Sigma_i (x_{it} - x_{\cdot t})^2} \, \Sigma_t \, \big[ \, \Sigma_i (x_{it} - x_{\cdot t})^2 \, \Sigma_i y_{it}(x_{it} - x_{\cdot t})/\Sigma_i(x_{it} - x_{\cdot t})^2 \, \big] \\
&=\; \Sigma_t \, w_t \beta_t
\end{aligned}
$$

where the period-by-period weights $w_t$ are given by

$$w_t \;=\; \frac{\Sigma_i (x_{it} - x_{\cdot t})^2}{\Sigma_t \Sigma_i (x_{it} - x_{\cdot t})^2}$$

and the coefficient $\beta_t$ is given by

$$\beta_t \;=\; \Sigma_i \, y_{it}(x_{it} - x_{\cdot t})/\Sigma_i(x_{it} - x_{\cdot t})^2 \;.$$

The $\beta_t$-coefficient is nothing but the coefficient in a cross-sectional OLS regression using period $t$ data. Also note that the $w_t$ weights sum to 1 and that periods with high variance of the regressor ($\Sigma_i(x_{it} - x_{\cdot t})^2$ large) gets more weight.

If you have both time-fixed effects and cross-sectional fixed effects in your model ; i.e.,

$$y_{it} = \mu_i + \nu_t + \beta x_{it} + e_{it} \;,$$

the OLS-estimator can be found by regressing $y_{it}$ on $y_{it} - \overline{y}_{i\cdot} - \overline{y}_{\cdot t} + \overline{y}_{\cdot\cdot}$ on $(X_{it} - \overline{X}_{i\cdot} - \overline{X}_{\cdot t} + \overline{X}_{\cdot\cdot})$ by standard OLS. The last term corrects for the fact that you implicitly have

subtracted the constant twice and if you don't allow for a constant in this OLS regression you have to correct for that.

If you subtract the time and cross-sectional means and then have, say, Stata, run the OLS regression, you need to adjust the standard errors, because you have used $N+T-1$ degrees for freedom before sending the data to the regression. Y

If you have missing data, you need to make sure that all means are calculated on the same data and those are the same used in the regression (so if there one (i,t) pair where, say, y, is missing that has to be dropped before you calculate means for regressors).

## Modeling the variance

In panel data sets it is often reasonable to expect that the variance of the error term is different for each state (person, ...) or time period. For example, considering the value of output of states or countries, there is no doubt that oil-states display higher variance than other states. Or maybe consumption patterns in neighboring states are more similar to each other than they are to those of more distant states. In principle the econometrics is simple. If the $NT \times NT$ matrix of residuals is $\Omega$ then the GLS estimator of $\beta$ is $\hat{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$. The feasible GLS estimator that one usually need to use is a 2-step estimator where one first performs OLS and then estimates $\hat{\Omega}$ and plugs the estimated $\Omega$ into the GLS formula.

Occasionally one may want to estimate $\beta$ by OLS, even if the error covariance matrix is $\Omega$. In this case one still needs to estimate $\Omega$ since the variance of the OLS estimator then is given by $(X'X)^{-1} X' \Omega X (X'X)^{-1}$. In Asdrubali, Sørensen, and Yosha (1996) we estimated a complicated $\Omega$ matrix, but since the estimated $\hat{\Omega}$ was likely to be imprecise, we found it unattractive to invert it and therefore used OLS with "GLS errors." (This is sometimes called a "generalized linear model.") In the article "Consumption and Aggre-

gate Constraints: Evidence from U.S. States and Canadian Provinces" by Ostergaard, Sørensen, and Yosha (JPE 2002) we did invert the $\Omega$ matrix at the prompting of a referee; my software/memory couldn't handle the size of the matrix so I spent a long time figuring out how to do it by splitting it up in smaller sub-matrices—you don't want to know the details but I have the program if you ever need it.

For simpler models, when the dimension of $\Omega$ is too large to keep in memory, one can often transform the data to be independent. (Formally, this means calculate $\tilde{y} = \Omega^{-1/2}y$ and $\tilde{X} = \Omega^{-1/2}X$ and then run OLS on $\tilde{X}$ and $\tilde{y}$.)

As the simplest example, consider the model

$$y_{it} = X_{it}\beta + \epsilon_{it}$$

where $\text{var}(\epsilon_{it}) = \sigma_i^2$. (Each "state" has a different variance.) The model may or may not contain fixed effects, that doesn't affect the following. Such a model is easily estimated by 2-step estimation of Maximum Likelihood (ML). The variance matrix takes the form

$$\Omega = \begin{pmatrix} \sigma_1^2 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_1^2 & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & \sigma_N^2 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & \sigma_N^2 \end{pmatrix}.$$

(Assuming the data are stacked with state 1's data first, then state 2, etc.) The 2-step (feasible GLS) estimator's first step is to estimate the model by OLS, and then estimate the $\Omega$ matrix from the residuals and transform the data to have diagonal covariance matrix. In practice, estimate $\hat{\sigma}_i^2 = \sum_{t=1}^{T} e_{it}^2$, where the $(e_{it})$ matrix is the residuals from an OLS regression; then transform the data to $\tilde{y}_{it} = y_{it}/\hat{\sigma}_i$ and $\tilde{x}_{it} = x_{it}/\hat{\sigma}_i$, and run OLS on the transformed data. (Iterating the process will give the ML estimator.) This is, of

course, a standard correction for heteroskedasticity. In Asdrubali, Sørensen, and Yosha (1996), it was found that correcting for state-specific heteroskedasticity had some effect on the estimates.

One may also want to correct for time-specific heteroskedasticity (similar) or both state- and time-specific heteroskedasticity. In the latter case, the variance matrix becomes more complicated, but again one simply adjust the data by the estimated time- and state-specific standard deviations.

If, say, the $T$ dimension is low, one may want to allow for a totally general pattern of time- (auto-) correlation. Assume now that the data are stacked as above, the variance matrix then takes the form

$$\Omega = \begin{pmatrix} \Gamma & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \Gamma \end{pmatrix}.$$

The $\Omega$ matrix has the form $\Omega = I \otimes \Gamma$, where "$\otimes$" is the so-called Kronecker product, which signifies that each element in the first matrix gets multiplied by the second matrix. (Here $I$ is an $N \times N$ identity matrix and $\Sigma$ is the variance-covariance matrix for the $T$ observations for state $i$, assumed to be the same for each $i$.) The $\Gamma$-matrix contains $T*(T+1)/2$ parameters and cannot be estimated from a single time series (therefore the whole topics of time series analysis is mainly about constructing parameter-parsimonious models for $\Gamma$), but it can be estimated from a panel with $N > T$. The $\Gamma$ matrix is easily estimated after a first step estimation as $\hat{\Gamma}_{st} = \frac{1}{N}\Sigma_{i=1}^{N}e_{is}e_{it}$ ; $s = 1, ..., T$ , $t = 1, ..., T$.

One can fairly easily show that that $(I \otimes \Gamma)^{-1} = I \otimes \Gamma^{-1}$ and one can then use the standard GLS estimator: $\hat{\beta} = (X'\,\Omega^{-1}\,X)^{-1}(X'\,\Omega^{-1}\,y)$. As usual, the problem is how to invert the $NT-$matrix $\Omega$, but if one defines $X_i = x_{it}$ ; $t = 1, ..., T$ and $y_i = y_{it}$ ; $t = 1, ..., T$

then it is easy to realize (using partitioned matrices) that

$$X' \, \Omega^{-1} \, X \; = \; \Sigma_{i=1}^{N} \, X_i' \, \Gamma^{-1} \, X_i \; ,$$

and similarly for $X' \, \Omega^{-1} \, y$. So only matrices of dimension $T \times T$ need to be inverted. For the 2-step estimator the calculation will, of course, involve $\hat{\Gamma}$ rather than $\Gamma$.

In Asdrubali, Sørensen, and Yosha (1996) we postulated a simple AR(1) model for the time series dependence of the residuals. (In our experience with estimating risk sharing panel data models on U.S. states and OECD countries, it has little effect on the estimates to control for auto-correlation.) For cross-correlation between states we estimated a $50 \times 50$ unrestricted variance-covariance matrix (likely to imprecisely estimated). The problem with modeling spatial correlations is that there is not obvious parsimonious model. Some suggestions is to let the variance decline with distance (analogously to the AR(1) model, but is there any really reason to believe that, e.g., Massachusetts is more similar to a nearby rural state like Maine, than it is to another financial center like Chicago? It is hard to know, and hard to test for.

**Random Effect Models**

The most commonly used variance model is the "random effect model." The material is here as a (rigorous) introduction to the topic. The textbooks mentioned contain much more material on random effect models. One reason to cover it here is that it gives a nice example of the way one has to analytically figure out how to invert large covariance matrices in panels.

The (one-way) <u>random-effect</u> model takes the form

$$(*) \qquad\qquad y_{it} = X_{it}\beta + \mu_i + \epsilon_{it} \;\; ; \;\; i = 1, \ldots, N \; , \;\; t = 1, \ldots, T \; ,$$

13

where $X_{it}$ is $1 \times K$, $\mu_i$ is an iid $N(0, \sigma_\mu^2)$ random dummy variable, and $\epsilon_{it}$ is id $N(0, \sigma_\epsilon^2)$. The random effect model has the same <u>interpretation</u> as the fixed effect model, i.e., the random variable $\mu_i$ picks up factors specific for observation $i$, just as the parameters $\alpha_i$ do in the fixed-effect model.

Statistically, (*) is just a suggestive notation for the linear model

$$(**) \qquad\qquad\qquad y_{it} = X_{it}\beta + w_{it} \ ,$$

where

$$Ew_{it}w_{js} = \begin{cases} \sigma_\epsilon^2 + \sigma_\mu^2 & \text{if} \quad i = j \ , \ t = s \\ \sigma_\mu^2 & \text{if} \quad i = j \ , \ t \neq s \\ 0 & \text{otherwise} \end{cases} .$$

Of course, we know how to estimate model (**) — this is (again) just GLS. The random effect model has K+2 parameters; (K $\beta$'s, $\sigma_\epsilon^2$ and $\sigma_\gamma^2$), while the fixed effect model has

$$K + N + 1 \text{ parameters } (\beta, \alpha_i, \sigma_\epsilon^2),$$

which makes the random effect model attractive if the $T$ dimension is short (like 2 or 3).

The random effects model is much more parsimonious than the fixed effects model but it is only consistent if $Ew_{it}X_{it} = 0$. (The standard criterion for exogeneity of the regressors.) When $w_{it} = \mu_i + \epsilon_{it}$ the assumption can also be stated:

$$Ex_{it}\mu_i = 0 \quad \text{and} \quad Ex_{it}\epsilon_{it} = 0.$$

<u>Example</u>

$$w_{it} = \alpha + \beta h_{it} + \mu_i + \epsilon_{it}.$$

$w_{it}$ are hourly wages and $h_{it}$ are hours per week for person $i$ in period $t$. (Do workers who work more hours receive a higher hourly pay?) In such a regression it will typically be

14

hard to defend exogeneity. But now assume that person $i$ (Mr. Smith) for other reasons have a relatively high hourly wage ($\mu_{Smith} > 0$). It is then likely that the high wage induces Mr. Smith to work longer (or shorter) hours. So again the crucial assumption $E\mu_i x_{it} = 0$ is unlikely to hold.

In general, it is often a very strong assumption (meaning that is likely not true) to impose $E\mu_i x_{it} = 0$.

It is usually assumed that $\mu_i$ and $\epsilon_{it}$ are normally distributed. This could also be wrong – e.g., if $\mu_i$ 'picks up' things like wealth, it would be more likely to be log-normal. (If the distribution is known, one may figure out how to solve the problem, but usually one has little to go by, when selecting a model for $\mu_i$.)

The GLS model (**), when the data are "stacked" in the order

$$y = ( y_{11}, \ldots, y_{1T}, y_{21}, \ldots, y_{2T}, \ldots y_{N1}, \ldots, y_{NT} )',$$

has variance matrix $\Omega$ (of dimension $NT \times NT$) defined as

$$\Omega = \begin{pmatrix}
\sigma_\epsilon^2 + \sigma_\mu^2 & \cdots & \sigma_\mu^2 & \cdots & 0 & \cdots & 0 \\
\vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\sigma_\mu^2 & \cdots & \sigma_\epsilon^2 + \sigma_\mu^2 & \cdots & 0 & \cdots & 0 \\
\vdots & & \vdots & \ddots & \vdots & & \vdots \\
0 & \cdots & 0 & \cdots & \sigma_\epsilon^2 + \sigma_\mu^2 & \cdots & \sigma_\mu^2 \\
\vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & \cdots & 0 & \cdots & \sigma_\mu^2 & \cdots & \sigma_\epsilon^2 + \sigma_\mu^2
\end{pmatrix}.$$

Note that $\Omega$ is block diagonal.

Define $H_T$ as a $T \times T$ matrix of ones, i.e.,

$$H_T = \begin{pmatrix} 1 & \ldots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \ldots & 1 \end{pmatrix}.$$

Then $\Omega = \sigma_\mu^2(I_N \otimes H_T) + \sigma_\epsilon^2(I_N \otimes I_T)$.

The $\Omega$-matrix is $NT \times NT$ and typically too large to keep in memory, so in order to do GLS estimation, we need to find a formula for $\Omega^{-1}$ involving matrices of lower

dimensions.

This can be done cleverly in the following fashion: (the notation here follows Baltagi's book.)

Define the $T \times T$ matrix

$$J_T = \begin{pmatrix} \frac{1}{T} & \cdots & \frac{1}{T} \\ \vdots & \ddots & \vdots \\ \frac{1}{T} & \cdots & \frac{1}{T} \end{pmatrix}$$

and

$$E_T = I_T - J_T = \begin{pmatrix} (1 - \frac{1}{T}) & \cdots & -\frac{1}{T} \\ \vdots & & \vdots \\ -\frac{1}{T} & \cdots & (1 - \frac{1}{T}) \end{pmatrix}$$

Now both $J_T$ and $E_T$ are idempotent, e.g.,

$$[J_T J_T]_{ij} = [\textstyle\sum_{t=1}^{T} \frac{1}{T^2}]_{ij} = [T\frac{1}{T^2}]_{ij}$$

$$= [\tfrac{1}{T}]_{ij} = [J_T]_{ij} \ .$$

Also,

$$E_T J_T = (I_T - J_T) J_T = 0$$

so $E_T$ and $J_T$ are orthogonal.

Define $P = I_N \otimes J_T$ and $Q = I_N \otimes E_T$ then $P, Q$ are also idempotent and orthogonal.

Now

$$\Omega = \sigma_\mu^2 (I_N \otimes H_T) + \sigma_\epsilon^2 (I_N \otimes I_T)$$

$$= T\sigma_\mu^2 (I_N \otimes J_T) + \sigma_\epsilon^2 (I_N \otimes E_T) + \sigma_\epsilon^2 (I_N \otimes J_T)$$

$$= (T\sigma_\mu^2 + \sigma_\epsilon^2)(I_N \otimes J_T) + \sigma_\epsilon^2 (I_N \otimes E_T)$$

$$= (T\sigma_\mu^2 + \sigma_\epsilon^2)P + \sigma_\epsilon^2 Q$$

It is now trivial to check that

$$\Omega^{-1} = \frac{1}{T\sigma_\mu^2 + \sigma_\epsilon^2}P + \frac{1}{\sigma_\epsilon^2}Q \ ,$$

by verifying that the $\Omega$ multiplied by the right hand side gives an identity matrix. (This exercise is really very similar to doing a diagonalization of $\Omega$. Try and see this. Hint: The columns of $P$ and $Q$ are the eigenvectors of $\Omega$.)

Given the formula for $\Omega^{-1}$ one can either estimate the model by ML.

## A potential "trap" in dynamic panels

To finish the introduction to panel data, we want to mention a special "problem" in dynamic panel data. Consider the model

$$(*) \qquad\qquad y_{it} = \alpha_i + \gamma y_{it-1} + u_{it} \quad i = 1, \ldots, N \ \ t = 1, \ldots, T$$

$$u_{it} \quad iid(0, \sigma_n^2) \quad |\gamma| < 1$$

Comments: If $N$ fixed, $T \to \infty$ it is essential that $|\gamma| < 1$, in order to use "standard" (non-unit-root type asymptotics). When $T$ is fixed, $N \to \infty$, it actually doesn't matter if $|\gamma|$ is less than 1 (or even explosive). In "square" panels, typical for macro, asymptotic theory of the sort $(N$ fixed, $T \to \infty)$ or $(T$ fixed, $N \to \infty)$ may both provide bad approximations to the actual small sample distribution. Square panels are, therefore, sometimes modelled as $(N \to \infty \ \frac{T}{N} \to K)$, where $K$ is a constant. In general, one always (even if $|\gamma| = 1$ or $|\gamma| > 1$) gets asymptotic normality if $N \to \infty$ (whether $T \to \infty$ or not) since one averages over $N$ independent units. The problem is to find the exact parameters of the asymptotic distribution, and that is at the research frontier for panels with non-stationary data.

Now define (as usual)

$$y_{i\cdot} = \frac{1}{T}\sum_{t=1}^{T} y_{it}$$

17

and define
$$y_{i\cdot-1} = \frac{1}{T}\sum_{t=1}^{T} y_{it-1}.$$

Then

$$\hat{\gamma} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}-y_{i\cdot})(y_{it-1}-y_{i\cdot-1})}{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it-1}-y_{i\cdot-1})^2}$$

or

$$\hat{\gamma}-\gamma = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it-1}-y_{i\cdot-1})u_{it}}{\sum\sum(y_{it-1}-y_{i\cdot-1})^2}.$$

Consider $E(\hat{\gamma}-\gamma)$. By assumption $Ey_{it-1}u_{it}=0$, but what about $Ey_{i\cdot-1}u_{it}$? If $T$ is large $y_{i\cdot-1} \approx Ey_{it}$ by the law of large numbers, and $Ey_{i\cdot-1}u_{it} \approx E(Ey_{it})u_{it}=0$; but if $T$ is small $y_{i-1}$ is <u>not</u> independent of the error term. You can show (do recursive substitution [write $y$'s in terms of $u$'s]) and simplify (here assume $\alpha_i=0$ and $y_{io}=0$ for simplicity).

$$y_{i\cdot-1} = \frac{1}{T}\left[\frac{1-\gamma^{T-1}}{1-\gamma}u_{i1} + \frac{1-\gamma^{T-2}}{1-\gamma}u_{i2} + \ldots + u_{iT-1}\right] \qquad (*)$$

and now

$$plim\ \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{i,t-1}-y_{i\cdot-1})u_{it}$$

$$= -plim_{N\to\infty}\tfrac{1}{N}\sum_{N=1}^{\infty}y_{i\cdot-1}u_{i\cdot}.$$

$$= -\frac{\sigma_\mu^2}{T^2}\frac{(T-1)-T\gamma-\gamma^T}{(1-\gamma)^2}.$$

You get this expression since the correlation of $u_{i\cdot}$ with $y_{i\cdot-1}$, using $(*)$, is

$$\frac{1}{T^2}\sigma_\mu^2\frac{1}{1-\gamma}\left[(1-\gamma^{T-1})+(1-\gamma^{T-2})+\ldots+(1-\gamma)\right]$$

$$\begin{aligned}
&= \tfrac{1}{T^2}\sigma_\mu^2\tfrac{1}{1-\gamma}\left[T-(1+\gamma+\ldots+\gamma^{T-1})\right]\\
&= \tfrac{1}{T^2}\sigma_\mu^2\tfrac{1}{1-\gamma}\left[T-\tfrac{1-\gamma^T}{1-\gamma}\right]\\
&= \tfrac{\sigma_\mu^2}{T^2}\tfrac{1}{(1-\gamma)^2}\left[T(1-\gamma)-(1-\gamma^T)\right]\\
&= \tfrac{\sigma_\mu^2}{T^2}\tfrac{(T-1)-T\gamma+\gamma^T}{(1-\gamma)^2}
\end{aligned}$$

18

Similarly, one can show that

$$plim \frac{1}{NT} \sum \sum (y_{it-1} - y_{i\cdot-1})^2$$

$$= \frac{\sigma_\mu^2}{1-\gamma^2}\{1 - \frac{1}{T} - \frac{2\gamma}{(1-\gamma)^2} \cdot \frac{(T-1) - T\gamma + \gamma^T}{T^2}\}$$

So, in total

$$plim_{N\to\infty}(\hat{\gamma} - \gamma) = -\frac{(1+\gamma)}{T-1}(1 - \frac{1}{T}\frac{1-\gamma^T}{1-\gamma}) \times \{1 - \frac{2\gamma}{(1-\gamma)(T-1)}\left[1 - \frac{1-\gamma^T}{T(1-\gamma)}\right]\}^{-1}.$$

I haven't done the derivation of the last two equations (they are taken from Hsiao's book), you may want to do the exercise of deriving one of them on your own.

If $T = 2$ (for example), this is a <u>huge</u> bias $= -1 + (1 - \frac{1}{2}) = -.5$ (when $\gamma = 0$). Notice, that the bias does not go away if $\gamma = 0$ (the bias remains as long as one attempts to estimate the $\gamma$ parameters). Typical suggested solution:

Difference to get the model

$$\Delta y_{it} = \gamma \Delta y_{it-1} - \Delta u_{it}$$

(which no longer contains fixed effects) AND

(a) estimate $\gamma$ by IV estimation, using $y_{it-2}$ as instrument. OR

(b) estimate $\gamma$ by IV estimation, using $y_{i0}$ as instrument for period 2, $y_{i0}, y_{i1}$ for period 3, $y_{i0}, y_{i1}, y_{i2}$ for period 4, etc.

If the original equation has iid errors, the differenced equation has a unit root in the AR-process of the error-term, which necessitates the use of the twice lagged variables as instruments. One drawback of (a) is that the lagged variable may or may not be a good instrument—if $\gamma$ is small, it is a lousy instrument. Using all lagged variables as in (b) is asymptotically efficient—I don't have practical experience with this estimator (although, if $T$ is large you are using a large amount of instruments and I would tend to

wonder about the quality of the estimated standard errors). If you use this type of panel regressions in serious work you should study the issue more, starting with Arellano's book.

## Panel Unit Roots and Panel Co-Integration (just a few comments on the issues)

These are areas with a lot of current active research. There is a survey article by Banerjee in the Oxford Bulletin of Economics and Statistics (1999) and you may also want to look at Chapter 7 of the Textbook "Applied Time Series Modelling and Forecasting" by Harris and Sollis (Wiley 2003). Peter Pedroni has a lot of articles on these issues and a book in process, check his WEB-page.

Note that if $T$ is small you can use standard panel data methods (I didn't go over it, but see, e.g., "Estimating Vector Autoregressions with Panel Data" by Holtz-Eakin, Newey, and Rosen in Econometrica (1988). If $N$ is large and $T$ is small you will want to use Johansen's co-integration methods).

Levin and Lin (Working Paper UCSD 1992) consider the model:

$$\Delta y_{it} = \alpha_i + \rho y_{it-1} + \Sigma_{k=1}^K \theta_{ik} \Delta y_{it-k} + u_{it} \ .$$

The error terms are supposed to be iid. Levin and Lin show that if $T \to \infty$ and $N \to \infty$ such that $\sqrt{N}/T \to 0$ then

$$T\sqrt{N}\hat{\rho} + 3\sqrt{N} \to N(0, 10.2)$$

if $\rho = 0$ and $\alpha_i = 0$.

Note the correction term $3\sqrt{N}$, which correct for downward bias in $\hat{\rho}$. Also note that the distribution converges at a rate $T$ in the time dimension (as for the Dickey-Fuller unit root test) but only at rate $\sqrt{N}$ across the N iid observations. Intuitively you can think of first $T$ going to infinity for each $i$ which basically gives a Dickey-Fuller unit root distribution and then you take the average across $i$ of those distributions which, since they are iid, always will lead to a normal distribution. However, econometricians deriving distributions has to worry about the rates at which $N$ and $T$ converge.

For an applied economist, the conditions on $N$ and $T$ raise a red flag. In applications you have only a given $N$ and a given $T$ so when will the asymptotic distribution be a reasonable approximation? Other problems: Is is reasonable to have $\rho$ be the same for all states away from the null-hypothesis but still assume the $\theta_i$ parameters are different? And, often the hardest to deal with: what if the shocks to different (states, individuals, ...) are correlated?

A good test for a unit roots that allow for different $\rho_i$ and test if $\rho_i = 0$ for all $i$ is that of Im, Pesaran, and Shin (IPS) (J. of Econometrics 2003). Their test is simply an average of the Dickey-Fuller t-tests for individual $i$'s. Their asymptotic distribution is based on first $T \to \infty$ and then $N \to \infty$; which doesn't have much meaning for an econometrician with a given sample, but Monte Carlo studies show that the test works reasonably well in finite sample. (I have used this test myself.) You need to look up critical value in their paper. (Or you can find a program where it is build in, I myself used a GAUSS program made available by Kao [I had to correct a major error, that is another lesson, don't use free—or even paid—programs without verifying they behave properly]). The IPS test also assume the $i$'s are independent.

A variation that uses P-values for the Dickey-Fuller test is suggested by Maddala and Wu (Oxford Bulletin 1999, this is in the same special issue as the Banerjee survey

article). I don't have experience with this test, but it may be promising.

There will be many papers written that figure out asymptotic distributions under various conditions and you will have to use some judgement (and not just push some button in some computer program). For example, is it reasonable to test if *all i* are stationary at the same time? (Most tests allow for the "nuisance" parameters to be different for different $i$'s.) My view is that the tests are tests for whether "most" $i$'s are stationary, rather than tests for if *all i* are stationary, because it is pretty much impossible to have power against the situation where, e.g., 1 out of 50 states is non-stationary. Of course, if you really believe that each $i$ has a different distribution, then you should test each series one-by-one and forget about the panel. Personally, if I were to seriously use unit root tests in panels, I would use some test such as the ones mentioned but run my own Monte Carlo study in order to examine how they behave in data like the ones I would be analyzing. In general, the fact that there are many different types of panels in terms of unit roots, co-integration, fixed-effects, correlation patterns, etc. makes it hard to rely on one or two simply set of guidelines for how to deal with panel data.