1 Static Panel Data Models

Last Updated: May 7, 2008

1.1 Introduction

This set of lecture notes is derived from *Panel Data Econometrics* by Manuel Arellano.

Panel data has both a cross-sectional and a time series component. The cross-sectional units can be individuals, households, firms or countries. Because panel data allows us to observe the cross-sectional unit in multiple periods, they greatly expand the capabilities of empirical researchers. For example, they enable researchers to study economic dynamics, to decompose variances into systematic and idiosyncratic components and to control for unobserved heterogeneity. Arellano describes these first two tasks as falling under the heading *random effects* and the last task as falling under the heading *fixed effects*.

1.2 Unobserved Heterogeneity

1.2.1 Motivation

We have discussed how often times results from linear regressions cannot be given causal interpretations because, by construction, they assume that the unobservables are not correlated with the observables, whereas, in the structural model of interest this is typically not the case. While, in principle, we can employ instrumental variables techniques to solve these problems, finding valid instruments is a very difficult task. As pointed out by Arellano, one of the major advantages of panel data is that it often enables us identify causal relationships by controlling for time-invariant unobserved heterogeneity. We will also see how we can employ moment restrictions in dynamic models to allow for simultaneous determination of economic outcomes.

We begin by considering a simple linear model

$$y_{i,1} = x_{i,1}\beta + \eta_i + \upsilon_{i,1}.$$

If we assume that $E[\upsilon_{i,1}|\eta_i, x_{i,1}] = 0$ and if η_i is observable then OLS will identify β. However, if η_i is unobserved, then we will require that $E[\upsilon_{i,1}|x_{i,1}] = 0$ and $E[\eta_i|x_{i,1}] = 0$ for OLS to identify β.

However, if we observe the data in two time periods and impose a different assumption, then we can identify β without observing η_i. To see this, note that in the second period we will have that

$$y_{i,2} = x_{i,2}\beta + \eta_i + \upsilon_{i,2}$$

which gives us that

$$\Delta y_{i,2} = \Delta x_{i,2}\beta + \Delta\upsilon_{i,2}.$$

If we assume that $E[\upsilon_{i,t}|x_{i,1}, x_{i,2}] = 0$, for $t = 1, 2$, then OLS of the model in first differences will identify β. Note that this assumption is stronger than $E[\upsilon_{i,t}|x_{i,t}] = 0$ for $t = 1, 2$ because it assumes that $\upsilon_{i,1}$ is mean idenependent of $x_{i,t}$ in both periods, not just the first period.

One important point is that there must be sufficient variation across time in $x_{i,t}$ for first-differencing to work. Otherwise, the data will not be informative of β and panel data techniques will not work. For example, panel data techniques cannot tell us about the returns to schooling.

1.2.2 Fixed Effects Models

We now approach the problem more formally. We assume that we have *i.i.d.* draws of $\{y_{i,1},...,y_{i,T},x_{i,1},...,x_{i,T},\eta_i\}$ and that

$$y_{i,t} = x'_{i,t}\beta + \eta_i + \upsilon_{i,t}.$$

We observe $x_{i,t}$ which is a $k \times 1$ vector of regressors, but we do not observe η_i. We assume that

$$E\left[\upsilon_i|x_i,\eta_i\right] = 0 \tag{A1}$$

where $\upsilon_i = (\upsilon_{i,1},...,\upsilon_{i,T})'$ and $X_i = (x_{i,1},...,x_{i,T})'$. This is a **strict exogeneity** assumption. At times, we will also invoke a homoskedasticity assumption

$$V\left(\upsilon_i|x_i,\eta_i\right) = \sigma^2 I_T. \tag{A2}$$

Let's consider estimation in first diffences. Suppose that $T \geq 2$. Then we will have that

$$\Delta y_{i,t} = \Delta x'_{i,t}\beta + \Delta\upsilon_{i,t} \text{ for } t = 2,...,T$$

which can be written as

$$Dy_i = DX_i\beta + D\upsilon_i$$

where D is the $(T-1) \times T$ difference operator given by

$$D = \begin{pmatrix} -1 & 1 & 0 & ... & 0 & 0 \\ 0 & -1 & 1 & 0 & ... & 0 \\ \vdots & & & \ddots & & \\ 0 & 0 & ... & 0 & -1 & 1 \end{pmatrix},$$

$y_i = (y_{i,1}, ..., y_{i,T})'$ and $X_i = (x_{i,1}, ..., x_{i,T})'$. The OLS estimator of the system in first differences is then given by

$$\widehat{\beta}_{FD} = \left(\sum_{i=1}^{n} (DX_i)' DX_i \right)^{-1} \sum_{i=1}^{n} (DX_i)' Dy_i.$$

This will be an unbiased and consistent estimator because A1 implies that $E[D\upsilon_i | x_i] = 0$. However, it will not be efficient since

$$V(D\upsilon_i | x_i) = \sigma^2 DD'$$

and Aitkens Theorem implies that OLS is not efficient when the covariance matrix of the residuals is not spherical. However, if we pre-multiply the original (not the first-differenced) data by the $(T-1) \times T$ matrix, $A = (DD')^{-1/2} D$, then the error term will become spherical once again and, according to the Gauss-Markov Theorem, OLS based on the transformed data will be efficient. This estimator is given by

$$\widehat{\beta}_{WG} = \left(\sum_{i=1}^{n} X_i' D' (DD')^{-1} DX_i \right)^{-1} \sum_{i=1}^{n} X_i' D' (DD')^{-1} Dy_i.$$

This is the **Generalized Least Squares** (GLS) estimator. Next, note that

$$D'\left(DD'\right)^{-1}D = I_T - ii'/T = Q$$

where i is a $T \times 1$ vector of ones. This matrix takes deviations from means in the sense that

$$\widetilde{y}_i = Qy_i$$

has the elements $\widetilde{y}_{i,t} = y_{i,t} - \overline{y}_i$. Because of this, we can also write this estimator as

$$\widehat{\beta}_{WG} = \left[\sum_{n=1}^{N}\sum_{t=1}^{T}\left(x_{i,t} - \overline{x}_i\right)\left(x_{i,t} - \overline{x}_i\right)'\right]^{-1}\sum_{n=1}^{N}\sum_{t=1}^{T}\left(x_{i,t} - \overline{x}_i\right)\left(y_{i,t} - \overline{y}_i\right).$$

This is the most popular formation of the WG estimator.

Finally, it turns out that estimating the model in levels with a complete set of individual dummy variables is tantamount to GLS estimation of the model in first-differences. Consider a system of T equations in levels

$$y_i = X_i\beta + i\eta_i + \upsilon_i$$

which can be written in stacked form as

$$y = X\beta + C\eta + \upsilon$$

where $y = (y_1', ..., y_N')'$ and $\upsilon = (\upsilon_1', ..., \upsilon_N')'$ are $NT \times 1$ vectors, $X = (X_1', ..., X_N')'$ is an $NT \times k$ matrix, $C = I_N \otimes i$ is an $NT \times N$ matrix and $\eta = (\eta_1, ..., \eta_N)'$ is a $N \times 1$ vector of fixed effects.

According to the Frisch-Waugh Theorem, the OLS estimate of β in this system is

$$\left[X'\left(I_{NT}-C\left(C'C\right)^{-1}C'\right)X\right]^{-1}X'\left(I_{NT}-C\left(C'C\right)^{-1}C'\right)y.$$

The above estimator is the WG estimator because $\left(I_{NT}-C\left(C'C\right)^{-1}C'\right)=I_N\otimes Q$. The estimates of the fixed effects are

$$\widehat{\eta}_i=\frac{1}{T}\sum_{i=1}^{T}\left(y_{i,t}-x'_{i,t}\widehat{\beta}_{WG}\right).$$

Note that, in contrast to $\widehat{\beta}_{WG}$ which is consistent and asymptotically normal under regularity conditions for large N, the estimates of η_i can only be consistent for large T.

1.2.3 Consistent Variance Estimation for Large N

We will now consider how to deal with a violation of A2 due to either heteroskedasticity or serial correlation. Specifically, we will consider a violation of A2 for the transformed residuals, $\upsilon_i^*=A\upsilon_i$. If we define $X_i^*=AX_i$, then we will obtain that

$$\left(\frac{1}{N}\sum_{i=1}^{n}X_i^{*\prime}X_i^*\right)\sqrt{N}\left(\widehat{\beta}_{WG}-\beta\right)=\frac{1}{\sqrt{N}}\sum_{i=1}^{n}X_i^{*\prime}\upsilon_i^*.$$

By the Central Limit Theorem and the Slutsky Theorem, we obtain that

$$\sqrt{N}\left(\widehat{\beta}_{WG}-\beta\right)\xrightarrow{d}N\left(0,E\left[X_i^{*\prime}X_i^*\right]^{-1}E\left[X_i^{*\prime}\upsilon_i^*\upsilon_i^{*\prime}X_i^*\right]E\left[X_i^*X_i^{*\prime}\right]^{-1}\right).$$

We can now based our estimated standard errors off of this asymptotic variance. Note that, within individuals, we can have an arbitrary covariance matrix for the residuals. All that we require is that the residuals, at all points in time, are independent across individuals.

1.2.4 Consistent Variance Estimation for Large T

With large N asymptotics, we can allow for arbitrary time dependence by relying on independence in the cross-section. In contrast, large T asymptotics allows for arbitrary cross-sectional dependence by relying on weak serial correlation. Let $\widehat{\delta}_{WG} = \left(\widehat{\beta}'_{WG}, \widehat{\eta}'\right)$ denote the within group estimator of β and $\eta = (\eta_1, ...\eta_N)'$ and let $w_{i,t} = \left(x'_{i,t}, d'_i\right)$ where d_i is an $N \times 1$ vector with a one in the ith position. Now, we define

$$V = p\lim \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} W'_t \upsilon_t \upsilon_s W_s$$

where

$$W'_t \upsilon_t = \sum_{i=1}^{N} w_{i,t} \upsilon_{i,t}.$$

With T asymptotics, the asymptotic covariance matrix is V. Newey and West suggest estimating V with

$$\widehat{V} = \widehat{\Omega}_0 + \sum_{l=1}^{m} \omega(l, m) \left(\widehat{\Omega}_l + \widehat{\Omega}'_l\right)$$

where $\omega(l, m) = 1 - \left[\frac{l}{m+1}\right]$ and

$$\widehat{\Omega}_l = \frac{1}{T} \sum_{t=l+1}^{T} W'_t \widehat{\upsilon}_t \widehat{\upsilon}'_{t-l} W_{t-l}.$$

The function $\omega(l,m)$ is a weight function that assigns declining weight to the autocovariances as l increases. The bound in the autocovariance term, m, is an increasing function of T. The asymptotic variance of the within group estimator for large T can then be estimated as

$$\widehat{V}\left(\widehat{\delta}_{WG}\right) = T\left(\sum_{t=1}^{T}\sum_{i=1}^{N} w_{i,t}w_{i,t}'\right)^{-1}\widehat{V}\left(\sum_{t=1}^{T}\sum_{i=1}^{N} w_{i,t}w_{i,t}'\right)^{-1}.$$

1.2.5 Maximum Likelihood

We will now consider Maximum Likelihood Estimation of the static linear model with fixed effects. Suppose that the dependent variable is normal conditional on the regressors and fixed effect so that

$$y_i|X_i,\eta_i \sim N\left(X_i\beta+\eta_i i,\sigma^2 I_T\right).$$

The log-likelihood conditional on (X_i,η_i) is then given by

$$\log f\left(y_i|X_i,\eta_i\right) \propto -\frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\upsilon_i'\upsilon_i$$

where $\upsilon_i = y_i - X_i\beta - \eta_i i$. The log-likelihood for the sample is then given by

$$L\left(\beta,\sigma^2,\eta;y,X\right) = \sum_{i=1}^{n}\log f\left(y_i|X_i,\eta_i\right).$$

Clearly, joint maximization of the log-likelihood for the sample of (β,σ^2,η) will yield the WG of β. The MLE of η is given by the formula above. The MLE of the variance of the residuals is

$$\widehat{\sigma}^2 = \frac{1}{NT}\sum_{i=1}^{N}\widehat{\upsilon}_i'\widehat{\upsilon}_i$$

where $\widehat{v}_i = y_i - X_i\widehat{\beta}_{WG} - \widehat{\eta}_i i$.

The MLE estimate of σ^2 is not consistent. To see this, stack the model for all individuals and time periods as

$$y = Z\delta + v$$

where $Z = (X, C)$ and $\delta = (\beta', \eta')'$. Next, write $v = \sigma u$ where $u \sim N(0, I_{NT})$. Then, we can write

$$\widehat{\sigma}^2 = \frac{1}{NT}\widehat{v}'\widehat{v}$$

where $\widehat{v} = y - Z\widehat{\delta}$. If we define $M = I_{NT} - Z(Z'Z)^{-1} Z'$ then we can write

$$\widehat{v}'\widehat{v} = \sigma^2 u'Mu.$$

So, we will obtain that

$$E\left[\widehat{v}'\widehat{v}\right] = E\left[\sum_{i=1}^{N}\widehat{v}_i'\widehat{v}_i\right] = \sigma^2(NT - N - k)$$

since $u'Mu \sim \chi^2_{NT-N-k}$. Next, we will have that

$$p\lim_{N\to\infty}\frac{1}{NT}\sum_{i=1}^{N}\widehat{v}_i'\widehat{v}_i = \frac{T-1}{T}\sigma^2.$$

This is an intriguing results since the MLE of β is consistent but the MLE of σ^2 is not. This is a consequence of the **Incidental Parameters Problem** in which the number of parameters being estimated increases with the sample size.

It turns out that, in the linear model without dynamics, $\overline{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{i,t}$ is a sufficient statistic for η_i. This means that

$$f\left(y_i | X_i, \eta_i, \overline{y}_i\right) = f\left(y_i | X_i, \overline{y}_i\right).$$

To prove this, we start by noting that

$$f\left(y_i | X_i, \eta_i, \overline{y}_i\right) = \frac{f\left(\overline{y}_i | y_i, X_i, \eta_i\right) f\left(y_i | X_i, \eta_i\right)}{f\left(\overline{y}_i | X_i, \eta_i\right)} = \frac{f\left(y_i | X_i, \eta_i\right)}{f\left(\overline{y}_i | X_i, \eta_i\right)}.$$

Next, under our normality assumption, we will have that

$$\overline{y}_i | X_i, \eta_i \sim N\left(\overline{X}_i'\beta + \eta_i, \frac{\sigma^2}{T}\right)$$

and, thus, we obtain

$$f(\overline{y}_i | X_i, \eta_i) \propto -\frac{1}{2}\log\sigma^2 - \frac{T}{2\sigma^2}\overline{\upsilon}_i^2.$$

Taking differences, we obtain that

$$\log f\left(y_i | X_i, \eta_i, \overline{y}_i\right) \propto -\frac{T-1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(\upsilon_{i,t} - \overline{\upsilon}_i\right)^2$$

which does not depend on η_i. The conditional likelihood function is given by

$$L_c\left(\beta, \sigma^2; y, X\right) = \sum_{i=1}^{N}\log f\left(y_i | X_i, \overline{y}_i\right).$$

The maximizer of this function over β is the WG estimator. Interesting, the maximizer of σ^2 is

$$\widetilde{\sigma}^2 = \frac{1}{N(T-1)}\widehat{\upsilon}'\widehat{\upsilon}.$$

Because $\widetilde{\sigma}^2 = \frac{T}{T-1}\widehat{\sigma}^2$, this estimator is consistent for large N.