ECON 6200: Econometrics II

Panel Data

Introduction

Panel data are data that come in a two-dimensional array, most commonly "time" and "unit of observation."

Thus, we have a sample size n and a number of waves T.

With a **balanced** panel, that makes for nT observations.
(We will only consider balanced panels.)

A panel is called "short" if:

- in practice $T \ll n$,
- the asymptoptic framework is to let $n \to \infty$ as T remains fixed.

A "long" panel conversely has n fixed and $T \to \infty$.
It is really a multivariate time series.

Some applications require analyses where both $n \to \infty$ and $T \to \infty$ at the same or different rates.

We will only consider short panels.

Panel Data

All estimators in this chapter are variations on multiple equation common coefficients GMM.

Unlike in the previous chapter, we will develop them from specific to general.

Thus, start with

$$\begin{aligned} Y_{it} &= X_{it}'\beta + \varepsilon_{it} \\ \mathbb{E}(X_{it}\varepsilon_{it}) &= 0 \end{aligned}$$

or equivalently (consolidating equations into vectors and matrices)

$$\begin{aligned} \boldsymbol{Y}_i &= \boldsymbol{X}_i\beta + \boldsymbol{\varepsilon}_i \\ \mathbb{E}(\boldsymbol{X}_i\boldsymbol{\varepsilon}_i) &= \mathbf{0}. \end{aligned}$$

Panel Data

All estimators in this chapter are variations on multiple equation common coefficients GMM.

Unlike in the previous chapter, we will develop them from specific to general.

Thus, start with

$$\begin{array}{rcl} Y_{it} & = & X_{it}'\beta + \varepsilon_{it} \\ \mathbb{E}(X_{it}\varepsilon_{it}) & = & 0 \end{array}$$

or equivalently (consolidating equations into vectors and matrices)

$$\begin{array}{rcl} \boldsymbol{Y}_i & = & \boldsymbol{X}_i\beta + \boldsymbol{\varepsilon}_i \\ \mathbb{E}(\boldsymbol{X}_i\boldsymbol{\varepsilon}_i) & = & \mathbf{0}. \end{array}$$

Then the **Pooled OLS** estimator

$$\hat{\beta}_{pool} \equiv \left(\sum_{i=1}^{n} \boldsymbol{X}_i'\boldsymbol{X}_i\right)^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i'\boldsymbol{Y}_i = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \xrightarrow{p} \beta$$

is unbiased respectively consistent under assumptions mirroring previous results.

(In the above, $\boldsymbol{X}$, $\boldsymbol{Y}$ are pooled data matrices.)

Panel Data

The **Pooled OLS** estimator

$$\hat{\beta}_{pool} \equiv \left(\sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{Y}_i = (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{X}' \boldsymbol{Y}$$

has the same asymptotics as before if we make the same assumptions.

However, the homoskedasticity assumption is extremely strong.

At the very least, we need to use a **cluster-robust variance estimator**

$$\hat{\boldsymbol{V}}_{pool} = (\boldsymbol{X}'\boldsymbol{X})^{-1} \left(\sum_{i=1}^{n} \boldsymbol{X}_i' \hat{\varepsilon}_i \hat{\varepsilon}_i \boldsymbol{X}_i \right) (\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

We had omitted cluster dependent errors, but note the close analogy to FGLS: The central term estimates the $(T \times T)$-matrix $\mathbb{E}(\boldsymbol{X}_i' \hat{\varepsilon}_i \hat{\varepsilon}_i \boldsymbol{X}_i)$.

Panel Data

The **Pooled OLS** estimator

$$\hat{\beta}_{pool} \equiv \left(\sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i' \boldsymbol{Y}_i = (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{X}'\boldsymbol{Y}$$

has the same asymptotics as before if we make the same assumptions.

However, the homoskedasticity assumption is extremely strong.

At the very least, we need to use a **cluster-robust variance estimator**

$$\hat{\boldsymbol{V}}_{pool} = (\boldsymbol{X}'\boldsymbol{X})^{-1} \left(\sum_{i=1}^{n} \boldsymbol{X}_i' \hat{\varepsilon}_i \hat{\varepsilon}_i \boldsymbol{X}_i \right) (\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

We had omitted cluster dependent errors, but note the close analogy to FGLS: The central term estimates the $(T \times T)$-matrix $\mathbb{E}(\boldsymbol{X}_i' \hat{\varepsilon}_i \hat{\varepsilon}_i \boldsymbol{X}_i)$.

But maybe we can do better?

As with GMM, the variance-covariance matrix we try to estimate is "small."

This raises the possibility of using an FGLS-like approach for estimation.

Indeed, we recognize this as doing **efficient** (multiple equation) GMM.

Panel Data

Random Effects

A popular specific model is to assume homoskedasticity but with structure

$$\mathbb{E}(\varepsilon_i \varepsilon_i' | \boldsymbol{X}_i) = \begin{bmatrix} \sigma_u^2 + \sigma_\epsilon^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 + \sigma_\epsilon^2 & \sigma_u^2 & \sigma_u^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \sigma_u^2 + \sigma_\epsilon^2 \end{bmatrix}$$

which has only two degrees of freedom.

Panel Data

Random Effects

A popular specific model is to assume homoskedasticity but with structure

$$\mathbb{E}(\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i' | \boldsymbol{X}_i) = \begin{bmatrix} \sigma_u^2 + \sigma_\epsilon^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 + \sigma_\epsilon^2 & \sigma_u^2 & \sigma_u^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \sigma_u^2 + \sigma_\epsilon^2 \end{bmatrix}$$

which has only two degrees of freedom.

This structure corresponds to assuming that $\varepsilon_{it} = u_i + \epsilon_{it}$, where u_i and ϵ_{it} are uncorrelated.

In other words, the structural equation becomes the random effects model

$$Y_{it} = X_{it}'\beta + u_i + \epsilon_{it}, \text{ where } \mathbb{E}X_{it}u_i = \mathbb{E}X_{it}\epsilon_{it} = 0.$$

Panel Data

The Random Effects estimator is usually understood to be the Feasible GLS estimator for this model (i.e., pre-estimating σ_u and σ_ϵ).

We will get to detailed implementation later because typical first-stage estimates use theory developed later.

Note: Hayashi defines the Random Effects estimator to be two-stage GMM imposing homoskedasticity, i.e. along the lines of Three-Stage Least Squares. This will have more degrees of freedom.

What's important:

- Random Effects has the same identifying assumptions as Pooled OLS.
- It adds an FGLS/TSGMM step. Because within-unit correlation of ε is both salient and (in short panels) easily modelled, this step is usually desired.

Panel Data

Fixed Effects

In

$$Y_{it} = X_{it}'\beta + u_i + \epsilon_{it}, \text{ where } \mathbb{E}X_{it}u_i = \mathbb{E}X_{it}\epsilon_{it} = 0,$$

the condition that $\mathbb{E}X_{it}u_i = 0$ is restrictive.

Think of u_i as unobserved, time-invariant covariates.

Then we say that these cannot be correlated with observed covariates.

Panel Data

Fixed Effects

In

$$Y_{it} = X_{it}'\beta + u_i + \epsilon_{it}, \text{ where } \mathbb{E}X_{it}u_i = \mathbb{E}X_{it}\epsilon_{it} = 0,$$

the condition that $\mathbb{E}X_{it}u_i = 0$ is restrictive.

Think of u_i as unobserved, time-invariant covariates.

Then we say that these cannot be correlated with observed covariates.

Suppose this condition fails.
Is there anything left to identify/estimate?

Panel Data

Fixed Effects

In

$$Y_{it} = X_{it}'\beta + u_i + \epsilon_{it}, \text{ where } \mathbb{E}X_{it}u_i = \mathbb{E}X_{it}\epsilon_{it} = 0,$$

the condition that $\mathbb{E}X_{it}u_i = 0$ is restrictive.

Think of u_i as unobserved, time-invariant covariates.

Then we say that these cannot be correlated with observed covariates.

Suppose this condition fails.
Is there anything left to identify/estimate?

Hint: Consider

$$Y_{it} - Y_{i,t-1} = (X_{it} - X_{i,t-1})'\beta + \epsilon_{it} - \epsilon_{i,t-1}.$$

Panel Data

Consider

$$Y_{it} - Y_{i,t-1} = (X_{it} - X_{i,t-1})'\beta + \epsilon_{it} - \epsilon_{i,t-1}.$$

When can we estimate this by OLS?

- We need that ε_{it} is uncorrelated with past and future t.
- We need $(X_{it} - X_{i,t-1})$ to fulfil a rank condition.
  This will fail with time invariant regressors.

Panel Data

Consider

$$Y_{it} - Y_{i,t-1} = (X_{it} - X_{i,t-1})'\beta + \epsilon_{it} - \epsilon_{i,t-1}.$$

When can we estimate this by OLS?

- We need that ε_{it} is uncorrelated with past and future t.
- We need $(X_{it} - X_{i,t-1})$ to fulfil a rank condition.
  This will fail with time invariant regressors.

This is the basic idea of fixed effects estimation.

Fixed effects can be "differenced away" in numerous ways:

- First differencing ("between estimator"),
- demeaning ("within estimator"),
- adding an indicator of each unit ("dummy variables regression").

Are these numerically the same?

Panel Data

Demeaning and dummy variable regression are numerically the same.

They also correspond to the "classic" FE estimator.

All methods are the same from a pure identification point of view:
They impose the same assumptions, e.g. the same restrictions on $(X_{it} - X_{i,t-1})$.

The methods differ by what weighting matrix they imply.

They are asymptotically equivalent if this matrix is pre-estimated.

Panel Data

Some more remarks on the classic Fixed Effects estimator:

- The implied weighting matrix $\boldsymbol{M} = \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'$ is efficient if the idiosyncratic error ϵ_{it} is homoskedastic and uncorrelated.
  (In contrast, the between transformation is efficient if ϵ_{it} follows a random walk in direction t. However, this is a less salient assumption.)
- Indeed, can show that FE equals TSLS (or really SUR), thinking of the cross-equation restrictions in

  $$\mathbb{E}X_{is}\epsilon_{it} = 0, \forall s, t$$

  as overidentifying restrictions/instruments.
- The estimator necessarily has higher (asymptotic) variance than pooled OLS. This is because OLS algebra applies, but demeaning reduces the sum of squared deviations of any r.v.
- For variance estimation, a d.f. adjustment of $T(T-1)$ is not negligiblble for realistic T and is therefore recommended including under asymptotic justification.
- In the error model that motivates Random Effects, the FE estimator can be used to estimate σ^2_ϵ. Doing this first and then backing out an estimate of σ^2_u is the standard approach.

Panel Data

Formal Statement

The following result is technically extremely close to earlier ones.

Assume:

1. $Y_{it} = X_{it}' + u_i + \epsilon_{it}, t = 1, \ldots, T, T \geq 2$.
2. $(\boldsymbol{X}_i, \boldsymbol{\epsilon}_i)$ are i.i.d.
   We do not need this for u_i. Why?
3. $\mathbb{E}(X_{is}\epsilon_{it}) = 0, \forall s, t = 1, \ldots, T$.
   This is stronger than for consistency of OLS. Why?
4. $\boldsymbol{Q} \equiv \mathbb{E}(\bar{\boldsymbol{X}}_i'\bar{\boldsymbol{X}}_i)$ is p.d., where $\bar{\boldsymbol{X}}_i \equiv (X_{i2} - \overline{X}_i, \ldots, X_{i2} - \overline{X}_i)'$.
   This excludes time-invariant regressors. Why?
5. $\mathbb{E}\epsilon_{it}^4 < \infty$, $\mathbb{E}\|X_{it}\|^4 < \infty$.

Panel Data

Formal Statement

The following result is technically extremely close to earlier ones.

Assume:

1. $Y_{it} = X_{it}' + u_i + \epsilon_{it}, t = 1, \ldots, T, T \geq 2$.
2. $(\boldsymbol{X}_i, \epsilon_i)$ are i.i.d.
   We do not need this for u_i. Why?
3. $\mathbb{E}(X_{is}\epsilon_{it}) = 0, \forall s, t = 1, \ldots, T$.
   This is stronger than for consistency of OLS. Why?
4. $\boldsymbol{Q} \equiv \mathbb{E}(\bar{\boldsymbol{X}}_i' \bar{\boldsymbol{X}}_i)$ is p.d., where $\bar{\boldsymbol{X}}_i \equiv (X_{i2} - \overline{X}_i, \ldots, X_{i2} - \overline{X}_i)'$.
   This excludes time-invariant regressors. Why?
5. $\mathbb{E}\epsilon_{it}^4 < \infty$, $\mathbb{E}\|X_{it}\|^4 < \infty$.

Theorem

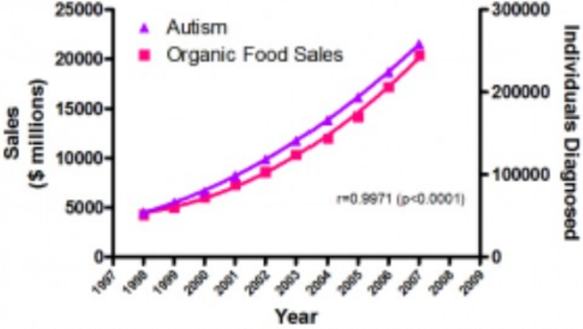
Under the above assumptions,

$$\begin{aligned} \sqrt{n}(\hat{\beta}_{\text{FE}} - \beta) &\overset{d}{\to} N(0, \boldsymbol{Q}^{-1}\Omega\boldsymbol{Q}^{-1}) \\ \Omega &\equiv \mathbb{E}(\bar{\boldsymbol{X}}_i'\epsilon_i\epsilon_i'\bar{\boldsymbol{X}}_i) \end{aligned}$$

Panel Data

Two-Way Fixed Effects

While we do not do time series in this lecture, you may have heard that undetrended time series can exhibit extreme (and extremely confounded) correlations.



Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Special Education Programs, Data Analysis System (DANS), OMB# 1820-0043: "Children with Disabilities Receiving Special Education Under Part B of the Individuals
with Disabilities Education Act

Panel Data

Two-Way Fixed Effects

While we do not do time series in this lecture, you may have heard that undetrended time series can exhibit extreme (and extremely confounded) correlations.

This motivates to also detrend our panel and estimate

$$Y_{it} = X_{it}\beta + u_i + \nu_t + \epsilon_{it}.$$

- This is called the Two-Way Fixed Effects (TWFE) model.
- With $T = 2$ (only), it is equivalent to "Difference-in-Difference" estimation.
- The best known estimator is pooled OLS after a double-within transformation.
- This model is the subject of an active literature. Do consult with an econometrician before using it.

Panel Data

Dynamic Panels

Next, consider

$$Y_{it} = \alpha Y_{i,t-1} + X_{it}\beta + u_i + \epsilon_{it}.$$

- Such dynamic panels are of great empirical importance.
- That we consider only one lag of the outcome is strictly for exposition.
- In practice, if data have a trend, it would be essential to separately model that to avoid spurious correlation.

Panel Data

Dynamic Panels

Next, consider

$$Y_{it} = \alpha Y_{i,t-1} + X_{it}\beta + u_i + \epsilon_{it}.$$

- Such dynamic panels are of great empirical importance.
- That we consider only one lag of the outcome is strictly for exposition.
- In practice, if data have a trend, it would be essential to separately model that to avoid spurious correlation.

The qualitatively new problem is that, even under the preceding assumptions, we have that Y_{it} is an endogenous regressor in the transformed (demeaned, first differenced, or otherwise) data.

This is most easily seen with first differencing:

$$\mathbb{E}(\Delta Y_{i2} \Delta\epsilon_{i3}) = \mathbb{E}((Y_{i2} - Y_{i1})(\epsilon_{i3} - \epsilon_{i2}))$$
$$= \mathbb{E}(Y_{i2}\epsilon_{i3}) - \mathbb{E}(Y_{i1}\epsilon_{i3}) - \mathbb{E}(Y_{i2}\epsilon_{i2}) + \mathbb{E}(Y_{i1}\epsilon i2) = 0 - 0 - \sigma_\epsilon^2 + 0 = -\sigma_\epsilon^2.$$

Panel Data

From finding that (for first-differenced data)

$$\mathbb{E}(\Delta Y_{i,t-1}\Delta\epsilon_{i,t}) = -\sigma_\epsilon^2,$$

we expect FE estimation to be biased and inconsistent. In fact, assuming stationarity ($|\alpha| < 1$), the asymptotic bias of $\hat{\alpha}$ can be computed as

$$\text{plim}\,\hat{\alpha}_{\text{FE}} = \alpha - \frac{1+\alpha}{2\alpha/(1-\alpha) + (T-1)/(1-\alpha^{T-1})},$$

and a bias of the same order (i.e., $O(1/T)$) is inherited by $\hat{\beta}_{\text{FE}}$.

Note that the bias is negative (this is expected from the first display) and not particularly small even for moderate T.

We conclude that FE estimation is theoretically problematic in this setting.

Note:

- The above analysis is due to Nickell (1981) and occasionally referred to as "Nickell critique/bias."
- Some practitioners deliberately ignore the problem because the fixes have their own issues.
  Not a discussion I will get into here.

Panel Data

What can we do about that? Instrumental Variables! Write

$$\begin{aligned} \mathrm{cov}(Y_{i,t-2}, \Delta Y_{i,t-1}) &\neq 0 \\ \mathbb{E}(Y_{i,t-1}\Delta\epsilon_t) &= \mathbb{E}(Y_{i,t-2\epsilon_t}) - \mathbb{E}(Y_{i,t-2\epsilon_{i,t-1}}) = 0. \end{aligned}$$

This means that if $T \geq 3$, we have an instrument.

(More generally, T must exceed the number of lags of Y used by at least 2.)

This gives rise to the (Anderson-)Hsiao estimator.

An important limitation is that the estimator is sensitive to misspecification, i.e. ϵ_{it} must really be uncorrelated and the correct number of lags specified.

Panel Data

What can we do about that? Instrumental Variables! Write

$$\begin{aligned} \mathrm{cov}(Y_{i,t-2}, \Delta Y_{i,t-1}) &\neq 0 \\ \mathbb{E}(Y_{i,t-1}\Delta\epsilon_t) &= \mathbb{E}(Y_{i,t-2\epsilon_t}) - \mathbb{E}(Y_{i,t-2\epsilon_{i,t-1}}) = 0. \end{aligned}$$

This means that if $T \geq 3$, we have an instrument.

(More generally, T must exceed the number of lags of Y used by at least 2.)

This gives rise to the (Anderson-)Hsiao estimator.

An important limitation is that the estimator is sensitive to misspecification, i.e. ϵ_{it} must really be uncorrelated and the correct number of lags specified.

Finally, the above algebra also implies that $Y_{i,t-3}, Y_{i,t-4}, \ldots$ are valid and relevant (also getting weak with distance) instruments.

So we can do (overidentified) Multiple Equation GMM.

This is the Arellano-Bond estimator.

It exists as one-step (as in TSLS) and two-step (as in efficient GMM) estimator.