Panel data

Repeated observations on the same cross-section of individual units.

Important advantages relative to pure cross-section data

- possible to control for some unobserved heterogeneity

- possible to model dynamics

Examples

- individual earnings

- household expenditures

- firm investment

- sector productivity

- regional migration

- country income per capita, or growth rates

Dimensions of the panel are important for asymptotic properties of different estimators.

Large N, small T often found in microeconomic data.

Longer T more common with aggregate data.

Semi-asymptotic results let one dimension become large with the other held fixed.

Our emphasis will be on the case where $N \to \infty$ with T fixed, more relevant for microeconometric applications.

Reliance on any asymptotic results is hazardous if neither N nor T is large.

Static linear model

$$y_{it} = x_{it}\beta + w_i\gamma + (\eta_i + v_{it})$$

for $i = 1, ..., N$ and $t = 1, ..., T$

$$x_{it} = (x_{1it}, ..., x_{Kit}),\ \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_K \end{pmatrix},\ w_i = (w_{1i}, ..., w_{Gi}),\ \gamma = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_G \end{pmatrix}$$

$1 \times K$ $\quad$ $K \times 1$ $\quad$ $1 \times G$ $\quad$ $G \times 1$

y_{it}, η_i, and v_{it} scalars

Observed y_{it}, x_{it}, w_i. $\quad$ Unobserved η_i, v_{it}.

Stack T observations for each individual

$$y_i = X_i\beta + W_i\gamma + (\eta_i j_T + v_i)$$

for $i = 1, ..., N$

$$y_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{pmatrix}, \; X_i = \begin{pmatrix} x_{1i1} & \dots & x_{Ki1} \\ \vdots & \ddots & \vdots \\ x_{1iT} & \dots & x_{KiT} \end{pmatrix}, \; W_i = \begin{pmatrix} w_{1i} & \dots & w_{Gi} \\ \vdots & \ddots & \vdots \\ w_{1i} & \dots & w_{Gi} \end{pmatrix}, \; \eta_i j_T = \begin{pmatrix} \eta_i \\ \vdots \\ \eta_i \end{pmatrix}, \; v_i = \begin{pmatrix} v_{i1} \\ \vdots \\ v_{iT} \end{pmatrix}$$

$T \times 1$ $\quad$ $T \times K$ $\quad$ $T \times G$ $\quad$ $T \times 1$ $\quad$ $T \times 1$

j_T is a $T \times 1$ column vector with each element equal to one

Then stack over N individuals

$$y = X\beta + W\gamma + (\eta + v)$$

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}, \; X = \begin{pmatrix} X_1 \\ \vdots \\ X_N \end{pmatrix}, \; W = \begin{pmatrix} W_1 \\ \vdots \\ W_N \end{pmatrix}, \; \eta = \begin{pmatrix} \eta_1 j_T \\ \vdots \\ \eta_N j_T \end{pmatrix}, \; v = \begin{pmatrix} v_1 \\ \vdots \\ v_N \end{pmatrix}$$

$$NT \times 1 \qquad NT \times K \qquad NT \times G \qquad NT \times 1 \qquad NT \times 1$$

Special case: no time-invariant explanatory variables ($G = 0$)

$$y = X\beta + (\eta + v)$$

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}, \; X = \begin{pmatrix} X_1 \\ \vdots \\ X_N \end{pmatrix}, \; \eta = \begin{pmatrix} \eta_1 j_T \\ \vdots \\ \eta_N j_T \end{pmatrix}, \; v = \begin{pmatrix} v_1 \\ \vdots \\ v_N \end{pmatrix}$$

$$NT \times 1 \qquad NT \times K \qquad NT \times 1 \qquad NT \times 1$$

Two important assumptions that we maintain throughout:

$$y_i = X_i\beta + (\eta_i j_T + v_i)$$

Cross-sectional independence: Observations on (y_i, X_i) are independent over $i = 1, ..., N$

Slope parameter homogeneity: The parameters in β are common to all $i = 1, ..., N$

The form of unobserved heterogeneity that we address relates to the individual-specific intercept terms (η_i) in our linear model relating y_{it} to x_{it} (known as 'fixed effects' or 'random effects', depending on whether they are assumed to be correlated or uncorrelated with the explanatory variables in x_{it})

$$y = X\beta + (\eta + v)$$

$$= X\beta + u$$

$$u = \eta + v; \qquad u_{it} = \eta_i + v_{it}$$

Ordinary least squares

$$\widehat{\beta}_{OLS} = (X'X)^{-1}X'y$$

Assumption (x_{it} predetermined)

$$E[x_{it}v_{it}] = 0$$

Properties of $\widehat{\beta}_{OLS}$ then depend on $E[x_{it}\eta_i]$.

Assumption (uncorrelated individual effects, or 'random effects')

$$E[x_{it}\eta_i] = 0$$

Then $\widehat{\beta}_{OLS}$ is a consistent estimator of β as $N \to \infty$ or as $T \to \infty$

(or both).

$E[x_{it}v_{it}] = 0$ and $E[x_{it}\eta_i] = 0 \Longrightarrow E[x_{it}u_{it}] = 0.$

OLS would be consistent under these assumptions for a single cross-section.

Panel dimension is thus not critical here for consistency.

OLS is not efficient in the panel setting, unless $\sigma_\eta^2 = var(\eta_i) = 0$

(and $v_{it} \sim iid(0, \sigma_v^2)$).

Assumption (correlated individual effects, or 'fixed effects')

$$E[x_{it}\eta_i] \neq 0$$

Then $\widehat{\beta}_{OLS}$ is an inconsistent estimator of β as $N \to \infty$ or as $T \to \infty$ (or both).

$E[x_{it}v_{it}] = 0$ and $E[x_{it}\eta_i] \neq 0 \Longrightarrow E[x_{it}u_{it}] \neq 0$.

OLS using the panel (pooled OLS) is subject to the same kind of omitted variable bias as OLS in a single cross-section.

Repeated observations do not change this; but do allow us to transform the model in order to construct consistent estimators.

Panel data is most useful when we suspect that cross-section regression results would be biased, due to (relevant and correlated) omitted variables.

Particularly if it is plausible that important omitted variables are time-invariant (or vary little over the sample period).

And the explanatory variables of interest and the dependent variable vary over time.

Examples:

Do high investment countries tend to have higher per capita income because investment raises income, or because factors like good governance or favorable geography raise both investment and income?

Do high R&D firms tend to have higher TFP because R&D raises TFP, or because good managers both invest in R&D and (independently) achieve high TFP?

Classical panel data estimators

Assumption (strict exogeneity)

$$E[x_{it}v_{is}] = 0 \quad \text{for all } s, t$$

This assumption is crucial for asymptotic properties in the case where $N \to \infty$ with T fixed, although not in the case where $T \to \infty$.

Strict exogeneity rules out 'feedback' from past v_{is} shocks to current x_{it}.

Hence rules out lagged dependent variables.

Assumption (error components)

$$E[\eta_i] = E[v_{it}] = E[\eta_i v_{it}] = 0$$

Assumption (serially uncorrelated shocks)

$$E[v_{it}v_{is}] = 0 \quad \text{for } s \neq t$$

Assumption (homoskedasticity)

$$E[\eta_i^2] = \sigma_\eta^2 \qquad E[v_{it}^2] = \sigma_v^2$$

For the case of uncorrelated individual effects, inefficiency of pooled OLS reflects the serial correlation in $u_{it} = \eta_i + v_{it}$ due to the presence of the time-invariant individual effects (η_i).

$$u_{it} = \eta_i + v_{it} \qquad\qquad u_{i,t-1} = \eta_i + v_{i,t-1}$$

Under the classical assumptions

$$E[u_{it}u_{i,t-1}] = E[\eta_i^2] = \sigma_\eta^2$$

And

$$E[u_{it}^2] = E[\eta_i^2] + E[v_{it}^2] = \sigma_\eta^2 + \sigma_v^2$$

So

$$E[u_i u_i'] = \begin{pmatrix} \sigma_\eta^2 + \sigma_v^2 & \sigma_\eta^2 & \cdots & \sigma_\eta^2 \\ \sigma_\eta^2 & \sigma_\eta^2 + \sigma_v^2 & \dots & \sigma_\eta^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\eta^2 & \sigma_\eta^2 & \dots & \sigma_\eta^2 + \sigma_v^2 \end{pmatrix} = \Omega_i \qquad T \times T$$

And

$$E[uu'] = \begin{pmatrix} \Omega_i & 0 & \cdots & 0 \\ 0 & \Omega_i & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Omega_i \end{pmatrix} = \Omega \qquad NT \times NT$$

Generalised Least Squares

Under the classical assumptions, the GLS (or 'random effects') estimator is consistent and efficient if $E[x_{it}\eta_i] = 0$

$$\widehat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$$

NB. Consistency requires all the explanatory variables to be uncorrelated with the individual effects.

If $E[x_{it}\eta_i] \neq 0$, $\widehat{\beta}_{GLS}$ is inconsistent as $N \to \infty$ with T fixed.

$\widehat{\beta}_{GLS}$ can be obtained using OLS on the transformed model

$$y_{it}^{*} = x_{it}^{*}\beta + u_{it}^{*}$$

where

$$y_{it}^{*} = y_{it} - (1-\theta)\overline{y}_{i}$$

and

$$\theta^2 = \frac{\sigma_v^2}{\sigma_v^2 + T\sigma_\eta^2} \ , \qquad\qquad \overline{y}_i = \frac{1}{T}\sum_{s=1}^{T} y_{is}$$

This transformation is known as 'theta-differencing'.

Feasible GLS uses consistent estimates of σ_η^2 and σ_v^2 to obtain a consistent estimate of θ.

These can be obtained using residuals from the Within Groups and Between Groups estimators (to be discussed below).

Feasible GLS is asymptotically equivalent to true GLS for this model.

Hence feasible GLS is asymptotically efficient under the classical assumptions, when $E[x_{it}\eta_i] = 0$.

$$y_{it}^{*} = y_{it} - (1-\theta)\overline{y}_{i}$$

$$\theta^{2} = \frac{\sigma_{v}^{2}}{\sigma_{v}^{2} + T\sigma_{\eta}^{2}}$$

For $\sigma_{\eta}^{2} = 0$, $\theta = 1$ and $y_{it}^{*} = y_{it}$. Special case where OLS is efficient.

As $T \to \infty$, $\theta \to 0$ and $y_{it}^{*} = y_{it} - \overline{y}_{i}$. In this case GLS coincides with the simpler Within Groups estimator (discussed below), and estimation of θ becomes redundant.

Within Groups

Within transformation

$$\widetilde{y}_{it} = y_{it} - \overline{y}_i \quad , \qquad\qquad \overline{y}_i = \frac{1}{T}\sum_{s=1}^{T} y_{is}$$

Key property

$$\overline{\eta}_i = \eta_i \quad \text{so that} \quad \widetilde{\eta}_i = \eta_i - \eta_i = 0$$

Example of a transformation that eliminates time-invariant variables.

Notice that theta-differencing does not eliminate the time-invariant individuals effects (η_i) from the error term for $\theta \neq 0$ ($\eta_i^* = \eta_i - (1-\theta)\eta_i = \theta\eta_i$)

- hence we require $E[x_{it}\eta_i] = 0$ for GLS to be consistent.

Transformed model

$$\widetilde{y}_{it} = \widetilde{x}_{it}\beta + \widetilde{v}_{it}$$

The Within Groups (or 'fixed effects') estimator is OLS on this transformed model

$$\widehat{\beta}_{WG} = (\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'\widetilde{y}$$

Under classical assumptions, $\widehat{\beta}_{WG}$ is consistent, both for $E[x_{it}\eta_i] = 0$ and for $E[x_{it}\eta_i] \neq 0$

- since the time-invariant individuals effects (η_i) are eliminated from the error term by the within transformation, we do not require $E[x_{it}\eta_i] = 0$ for Within Groups to be consistent.

The Within Groups estimator is thus consistent in the case where some or all of the explanatory variables are correlated with this unobserved heterogeneity.

In some contexts this is a key advantage, relative to cross-section OLS, pooled OLS or GLS.

This illustrates how we can construct consistent estimates using panel data in settings where cross-section OLS would be subject to omitted variables bias.

But note that this advantage comes at a price.

As $N \to \infty$ with T fixed, $\widehat{\beta}_{WG}$ is less efficient than $\widehat{\beta}_{GLS}$ in the case where $E[x_{it}\eta_i] = 0$.

$\widehat{\beta}_{WG}$ is only efficient (under classical assumptions) in the special case where all the explanatory variables are correlated with η_i.

Moreover any observed time-invariant explanatory variables are also eliminated by the transformation, so the Within Groups estimator does not identify the γ parameters in the more general model

$$y_{it} = x_{it}\beta + w_i\gamma + (\eta_i + v_{it})$$

This illustrates that repeated observations (i.e. panel data) are most useful when the variables of interest change over time

- repeated observations are less useful when the variables of interest remain constant over time.

For example, panel data is less successful in controlling for unobserved ability if we want to estimate the effect of schooling on earnings

- years of schooling remain constant for most people once they leave full-time education and join the labour force.

More generally, the Within Groups parameter estimates are likely to be imprecise if there is only limited time-series ('within') variation.

The Within Groups estimator of β can also be obtained by including a set of N dummy variables, for each individual

$$y_{it} = \eta_1 D_{1i} + ... + \eta_N D_{Ni} + x_{it}\beta + v_{it}$$

and using OLS on this model

($D_{1i} = 1$ for the observations on individual 1, and zero otherwise)

Hence Within Groups is also called Least Squares Dummy Variables (LSDV).

Note that, in the case where $N \to \infty$ with T fixed, consistency depends on the strict exogeneity assumption.

$$\widetilde{x}_{it} = x_{it} - \frac{1}{T}(x_{i1} + ... + x_{iT}) \qquad \widetilde{v}_{it} = v_{it} - \frac{1}{T}(v_{i1} + ... + v_{iT})$$

Hence $E[\widetilde{x}_{it}\widetilde{v}_{it}] = 0$ requires $E[x_{it}v_{is}] = 0$ for all s, t unless $T \to \infty$.

This motivated the development of alternative estimators for dynamic panel data models, that are consistent as $N \to \infty$ for fixed T, in the presence of (e.g.) lagged dependent variables.

Other estimators

Between Groups

Between Groups is OLS on the cross-section equation

$$\overline{y}_i = \overline{x}_i\beta + (\overline{\eta}_i + \overline{v}_i) \qquad\qquad i = 1, ..., N$$

Consistency requires $E[x_{it}\eta_i] = 0$.

Between Groups is not efficient

- only used to obtain an estimate of σ_η^2 when implementing feasible GLS.

First-differenced OLS

OLS on the first-differenced equations

$$\Delta y_{it} = \Delta x_{it}\beta + \Delta v_{it} \qquad \text{for } i = 1, ..., N \text{ and } t = 2, ..., T$$

where $\Delta y_{it} = y_{it} - y_{i,t-1}$

First-differencing is another transformation that eliminates the time-invariant individual effects ($\Delta \eta_i = \eta_i - \eta_i = 0$).

Consistency requires $E[\Delta x_{it} \Delta v_{it}] = 0$

- this is implied by (but weaker than) strict exogeneity.

Within Groups is more efficient than first-differenced OLS under classical assumptions, i.e. $v_{it} \sim iid(0, \sigma_v^2)$, serially uncorrelated and homoskedastic.

First-differenced OLS is more efficient if v_{it} is a random walk, i.e. $v_{it} = v_{i,t-1} + \varepsilon_{it}$ with $\varepsilon_{it} \sim iid(0, \sigma_\varepsilon^2)$ s.t. Δv_{it} is serially uncorrelated.

First-differenced OLS (but not Within Groups) would also be consistent as $N \to \infty$ with T fixed in cases where we suspect feedback from second lags or longer lags of v_{is} onto x_{it}, but not from the first lag $(v_{i,t-1})$ onto x_{it} - i.e. where any feedback takes two or more periods to influence x_{it}.

Calculating the feasible GLS estimator

$$y_{it}^* = x_{it}^* \beta + u_{it}^*$$

$$y_{it}^* = y_{it} - (1-\theta)\overline{y}_i$$

$$\theta^2 = \frac{\sigma_v^2}{\sigma_v^2 + T\sigma_\eta^2}$$

σ_v^2 can be estimated consistently using the Within Groups residuals

$$\widehat{\widetilde{v}}_{it} = \widetilde{y}_{it} - \widetilde{x}_{it}\widehat{\beta}_{WG}$$

$$\widehat{\sigma}_v^2 = \frac{\widehat{\widetilde{v}}'\widehat{\widetilde{v}}}{N(T-1) - K}$$

Notice that although we have NT observations and K parameters, we have only $N(T-1)-K$ degrees of freedom for the Within Groups estimator

- we estimate N parameters when we estimate the individual means ($\overline{y}_i$ for $i = 1, ..., N$) used to construct the within transformation

- or, equivalently, when we estimate coefficients on the N individual dummy variables in the Least Squares Dummy Variables representation

- this is also relevant when we estimate the asymptotic variance of the Within Groups estimator using

$$avar(\widehat{\beta}_{WG}) = \sigma_v^2(\widetilde{X}'\widetilde{X})^{-1}$$

σ^2_η can then be estimated consistently using the Between Groups residuals

$$\widehat{\overline{u}}_i = \widehat{(\eta_i + \overline{v}_i)} = \overline{y}_i - \overline{x}_i \widehat{\beta}_{BG}$$

$$\widehat{\sigma}^2_{\overline{u}} = \widehat{(\sigma^2_\eta + \frac{1}{T}\sigma^2_v)} = \frac{\widehat{\overline{u}}' \widehat{\overline{u}}}{N - K}$$

and then

$$\widehat{\sigma}^2_\eta = \widehat{\sigma}^2_{\overline{u}} - \frac{1}{T}\widehat{\sigma}^2_v$$

Testing for correlated individual effects

With fixed T, it is useful to test whether some of the included explanatory variables are correlated with the unobserved individual effects.

$\widehat{\beta}_{WG}$ is consistent whether the individual effects are correlated with the included regressors, or not.

$\widehat{\beta}_{GLS}$ (and $\widehat{\beta}_{BG}$) is consistent only if the individual effects are uncorrelated with all the included regressors; biased and inconsistent otherwise.

Estimates should be similar if η_i is uncorrelated with all the included regressors; but different if η_i is correlated with any of the included regressors.

Hausman test

$$\widehat{q} = \widehat{\beta}_{WG} - \widehat{\beta}_{GLS}$$

$$h = \widehat{q}'[avar(\widehat{q})]^{-1}\widehat{q} \overset{a}{\sim} \chi^2(K)$$

under the null hypothesis that $E[x_{it}\eta_i] = 0$.

$$avar(\widehat{q}) = avar(\widehat{\beta}_{WG}) - avar(\widehat{\beta}_{GLS})$$

An equivalent test can be based on $\widehat{q} = \widehat{\beta}_{WG} - \widehat{\beta}_{BG}$.

These tests require the classical assumptions, under which the FGLS estimator is efficient relative to the Within estimator under the null. Versions robust to heteroskedasticity are now available.