

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 $N \rightarrow \infty$ 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

$$y_{it} = x_{it}\beta + z_i\gamma + \eta_i + v_{it}$$

for $i = 1, \dots, N$ and $t = 1, \dots, T$

$$x_{it} = (x_{1it}, \dots, x_{kit}), \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad z_i = (z_{1i}, \dots, z_{gi}), \quad \gamma = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_g \end{pmatrix}$$

$1 \times k$

$k \times 1$

$1 \times g$

$g \times 1$

y_{it} , η_i , and v_{it} scalars

Observed y_{it} , x_{it} , z_i .

Unobserved η_i , v_{it} .

Stack observations for each individual

$$y_i = X_i\beta + Z_i\gamma + \eta_i + v_i$$

for $i = 1, \dots, N$

$$y_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{pmatrix}, \quad X_i = \begin{pmatrix} x_{1i1} & \dots & x_{ki1} \\ \vdots & \dots & \vdots \\ x_{1iT} & \dots & x_{kiT} \end{pmatrix}, \quad Z_i = \begin{pmatrix} z_{1i} & \dots & z_{gi} \\ \vdots & \dots & \vdots \\ z_{1i} & \dots & z_{gi} \end{pmatrix}, \quad \eta_i = \begin{pmatrix} \eta_i \\ \vdots \\ \eta_i \end{pmatrix}, \quad v_i = \begin{pmatrix} v_{i1} \\ \vdots \\ v_{iT} \end{pmatrix}$$

$T \times 1$ $T \times k$ $T \times g$ $T \times 1$ $T \times 1$

Then stack over individuals

$$y = X\beta + Z\gamma + \eta + v$$

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}, X = \begin{pmatrix} X_1 \\ \vdots \\ X_N \end{pmatrix}, Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_N \end{pmatrix}, \eta = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_N \end{pmatrix}, v = \begin{pmatrix} v_1 \\ \vdots \\ v_N \end{pmatrix}$$

$NT \times 1$ $NT \times k$ $NT \times g$ $NT \times 1$ $NT \times 1$

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

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

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

$$NT \times 1$$

$$NT \times k$$

$$NT \times 1$$

$$NT \times 1$$

$$\begin{aligned}y &= X\beta + \eta + v \\ &= X\beta + u\end{aligned}$$

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

Ordinary least squares

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

Assumption (x_{it} predetermined)

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

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

Assumption (uncorrelated individual effects, or ‘random effects’)

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

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

(or both).

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

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

Panel dimension is then not critical for consistency.

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

Assumption (correlated individual effects, or ‘fixed effects’)

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

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

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

Pooled OLS is subject to the same omitted variable bias as OLS in a single cross-section.

Panel dimension does not change this; but does allow us to transform the model in order to construct consistent estimators.

Classical panel data estimators

Assumption (strict exogeneity)

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

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

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)

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

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).

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

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

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

And

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

Generalised Least Squares

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

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

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

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

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

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

where

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

and

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

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).

Feasible GLS is asymptotically equivalent to true GLS. Hence asymptotically efficient under the classical assumptions.

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

As $T \rightarrow \infty$, $\theta \rightarrow 0$ and $y_{it}^* = y_{it} - \bar{y}_i$. In this case GLS coincides with the simpler Within Groups estimator, and estimation of θ becomes redundant.

Within Groups

Within transformation

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

Key property

$$\bar{\eta}_i = \eta_i \quad \text{so that} \quad \tilde{\eta}_i = \eta_i - \eta_i = 0$$

Example of a transformation that eliminates time-invariant variables.

Transformed model

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

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

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

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

Since the time-invariant individual effects are eliminated by the transformation, the Within estimator remains 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.

But this comes at a price.

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

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

$$y_{it} = x_{it}\beta + z_i\gamma + \eta_i + v_{it}$$

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

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

The Within Groups estimator of β can also be obtained by including a set of N dummy variables, for each individual. Hence also called Least Squares Dummy Variables (LSDV).

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

$$\tilde{x}_{it} = x_{it} - \frac{1}{T}(x_{i1} + \dots + x_{iT}) \quad \tilde{v}_{it} = v_{it} - \frac{1}{T}(v_{i1} + \dots + v_{iT})$$

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

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

Other estimators

Between Groups

OLS on the cross-section equation

$$\bar{y}_i = \bar{x}_i\beta + \bar{\eta}_i + \bar{v}_i \quad i = 1, \dots, N$$

Consistency requires $E[x_{it}\eta_i] = 0$. Not efficient. Used to obtain an estimate of σ_η^2 when implementing feasible GLS.

First-differenced OLS

OLS on the pooled equations

$$\Delta y_{it} = \Delta x_{it}\beta + \Delta v_{it}$$

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

First-differencing is another transformation that eliminates the time-invariant individual effects.

Consistency requires $E[\Delta x_{it}\Delta v_{it}] = 0$. This is implied by (but weaker than) strict exogeneity.

Within Groups is more efficient under classical assumptions (s.t. \tilde{v}_{it} is serially uncorrelated).

First-differenced OLS is more efficient if v_{it} is a random walk (s.t. Δv_{it} is serially uncorrelated).

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.

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

$\hat{\beta}_{GLS}$ (and $\hat{\beta}_{BG}$) is inconsistent if such correlation is present.

Hence these estimators should be similar if there is no correlation with included regressors; but different if there is correlation with included regressors.

Hausman test

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

$$h = \hat{q}' [avar(\hat{q})]^{-1} \hat{q} \stackrel{a}{\sim} \chi^2(k)$$

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

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

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

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.