

Panel Data

Zhentao Shi

The Chinese University of Hong Kong

Data



- Cross-sectional datasets collected at different time points
- Group-specific intercept (one way to handle endogeneity)
- Real data: <https://www.kaggle.com/code/frankshi0/penn-world-table>

Panel Data Structure

- The same individuals are observed over time $t = 1, \dots, T$
- Independent across $i = 1, \dots, N$

$$Y_{NT \times 1} = \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1T} \\ y_{21} \\ \vdots \\ y_{NT} \end{bmatrix}, \quad X_{NT \times p} = \begin{bmatrix} X'_{11} \\ X'_{12} \\ \vdots \\ X'_{1T} \\ X'_{21} \\ \vdots \\ X'_{NT} \end{bmatrix}$$

$$y_{it} = c + X'_{it}\beta + \alpha_i^* + u_{it}, \quad \varepsilon_{it} = \alpha_i^* + u_{it}$$

Real data: <https://www.kaggle.com/datasets/frankshi0/nber-ces-manufacturing-industry-database/code>

Panel Data Regression

- Temporal observations over $t = 1, \dots, T$ for the same i is viewed as a **group**. Temporal dependence is allowed within the group.

$$y_{it} = c + X'_{it}\beta + \varepsilon_{it}$$

with $E(\varepsilon_{it}) = 0$.

- ε_{it} may be correlated with X_{it} .
- **Composite error**

$$\varepsilon_{it} = \alpha_i^* + u_{it}$$

with $\text{cov}(u_{it}, X_{it}) = 0$

Section 1

Fixed Effect Models

Fixed Effects

- Consistency of OLS counts on $\text{cov}(\varepsilon_{it}, X_{it}) = \text{cov}(\alpha_i^*, X_{it}) + \text{cov}(u_{it}, X_{it}) = 0$.
- Example
 - Production function (Mundlak, 1961)

$$y_{it} = c + X'_{it}\beta + m_i\gamma + u_{it}$$

where m_i is the “management quality” of a firm (usually unobservable).

- $m_i\gamma$, which can be correlated with X_{it} , is captured by α_i^*
- α_i^* can be potentially endogenous (correlated with X_{it})

Static Linear Model

- Model

$$\begin{aligned}y_{it} &= c + X'_{it}\beta + \varepsilon_{it} \\&= (c + \alpha_i^*) + X'_{it}\beta + u_{it} \\&= \alpha_i + X'_{it}\beta + u_{it}\end{aligned}$$

where $\alpha_i = c + \alpha_i^*$ is the new **individual-specific** intercept.
Also called the **fixed effect**.

- β is a p -dimensional **homogenous** slope coefficient
- $(N + p)$ parameters $(\alpha_1, \alpha_2, \dots, \alpha_N; \beta)$

Least Squares Dummy Variable Estimator

- Direct Estimation with N dummy variables

$$y_{it} = \sum_{j=1}^N \alpha_j \mathbf{I}(i = j) + X'_{it} \beta + u_{it}$$

$$Y = D \alpha + X \beta + u, \quad D := I_N \otimes \mathbf{1}_T.$$

The LSDV estimator can be written in closed-form:

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = \left(\begin{bmatrix} D & X \end{bmatrix}' \begin{bmatrix} D & X \end{bmatrix} \right)^{-1} \begin{bmatrix} D & X \end{bmatrix}' Y$$

- The slope coefficient

$$\hat{\beta} = (X' M_D X)^{-1} X' M_D Y,$$

where $M_D := I_{NT} - D (D' D)^{-1} D'$.

- The fixed effects

$$\hat{\alpha} = (D' D)^{-1} D' (Y - X \hat{\beta}).$$

Within-Group Transformation

- Inner-outer optimization: given any β , the OLS estimator

$$\hat{\alpha}_i = T^{-1} \sum_{t=1}^T (y_{it} - X'_{it}\beta) = \bar{y}_i - \bar{X}'_i\beta,$$

where $\bar{y}_i = T^{-1} \sum_{t=1}^T y_{it}$ is the **within group average**, and so is \bar{X}_i .

- Substitute $\hat{\alpha}_i$ into $y_{it} = \alpha_i + X'_{it}\beta + u_{it}$ and rearrange:

$$\tilde{y}_{it} = \tilde{X}'_{it}\beta + \tilde{u}_{it}$$

where $\tilde{y}_{it} = y_{it} - \bar{y}_i$ is the **within-group transformation**, and \tilde{X}_{it} and \tilde{u}_{it} are defined similarly.

- This transformation eliminates the N parameters $(\alpha_i)_{i=1}^N$.

Alternative Interpretation of Within-Group Transformation

- Recall the model

$$y_{it} = \alpha_i + X'_{it}\beta + u_{it}.$$

- For each i , average over $t = 1, \dots, T$:

$$\bar{y}_i = \alpha_i + \bar{X}'_i\beta + \bar{u}_i.$$

- Subtraction:

$$\tilde{y}_{it} = \tilde{X}'_{it}\beta + \tilde{u}_{it}.$$

eliminates the fixed effects.

- No intercept, by construction.

Data in Blocks

- Stack into long vector/matrix of all data

$$\begin{bmatrix} \tilde{y}_1 \\ \tilde{y}_2 \\ \vdots \\ \tilde{y}_N \end{bmatrix}_{NT \times 1} = \begin{bmatrix} \tilde{X}_1 \\ \tilde{X}_2 \\ \vdots \\ \tilde{X}_N \end{bmatrix}_{NT \times p} \underset{p \times 1}{\beta} + \begin{bmatrix} \tilde{u}_1 \\ \tilde{u}_2 \\ \vdots \\ \tilde{u}_N \end{bmatrix}_{NT \times 1}.$$

- Compact expression of data:

$$\tilde{Y} = \tilde{X}\beta + \tilde{u}.$$

Within Estimator

- Within estimator (or equivalently the FE estimator):

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

- The fixed effects are estimated as

$$\hat{\alpha}_i = \bar{y}_i - \bar{X}_i' \hat{\beta}_{FE}$$

- Consistency (T fixed, $N \rightarrow \infty$) is achieved if strict exogeneity holds.

Strict Exogeneity

- A necessary condition for the consistency of $\hat{\beta}_{FE}$:

$$E[(X_{it} - \bar{X}_i)(u_{it} - \bar{u}_i)] = 0.$$

- A sufficient condition is **strict exogeneity**:

$$E[X_{it}u_{is}] = 0 \quad \text{for all } s, t \in \{1, \dots, T\}.$$

- Interpretation: no correlation of the error term u_{is} with the regressor X_{it} in the past, the present, and the future.

Exogeneity

- Notice that

$$\tilde{X}_{it} = X_{it} - \frac{1}{T} \sum_{s=1}^T X_{is} = \left(1 - \frac{1}{T}\right) X_{it} - \frac{1}{T} \sum_{s \neq t} X_{is},$$

is a linear combination of $\{X_{i1}, \dots, X_{iT}\}$.

- **Contemporaneous exogeneity:**

$$E(X_{it}u_{it}) = 0 \quad \text{for all } t \in \{1, \dots, T\}.$$

- **Sequential exogeneity:**

$E(u_{it}X_{is}) = 0$ for all $s \leq t \in \{1, \dots, T\}$. The error term is uncorrelated with the past and present regressor.

- Neither of the above two kinds of exogeneity produces consistent $\hat{\beta}_{FE}$.

Variance Estimation for FE

- Under homoskedasticity:

$$\hat{\sigma}_u^2 = \frac{1}{N(T-1)} \sum_{i=1}^N \sum_{t=1}^T \hat{u}_{it}^2.$$

where $\hat{u}_{it} = \tilde{y}_{it} - \tilde{X}_{it}\hat{\beta}_{FE}$

- Asymptotic normality

$$\left(\hat{\sigma}_u^2 (\tilde{X}'\tilde{X})^{-1} \right)^{-1/2} \left(\hat{\beta}_{FE} - \beta^0 \right) \Rightarrow N(0, I_K).$$

First Difference (FD)

- Alternative way to eliminate FE
- Recall

$$y_{it} = \alpha_i + X'_{it}\beta + u_{it},$$
$$y_{i,t-1} = \alpha_i + X'_{i,t-1}\beta + u_{i,t-1}.$$

- Subtraction:

$$\Delta y_{it} = \Delta X'_{it}\beta + \Delta u_{it},$$

where $\Delta y_{it} = y_{it} - y_{i,t-1}$ is the first-differenced variable, and ΔX_{it} and Δu_{it} are defined similarly.

- Collect the FD variables into ΔY and ΔX :

ΔY is of size $N(T-1) \times 1$,

ΔX is of size $N(T-1) \times p$.

- The FD estimator is:

$$\hat{\beta}_{FD} = (\Delta X' \Delta X)^{-1} \Delta X' \Delta Y.$$

- A necessary condition for consistency is

$$E [\Delta X_{it} \Delta u_{it}] = 0,$$

which is slightly weaker than strict exogeneity.

Section 2

Dynamic Panel Data Models

Dynamic Panel Data

- The simplest dynamic panel model is

$$y_{it} = \alpha_i + \beta y_{i,t-1} + u_{it},$$

where $|\beta| < 1$, $u_{it} \sim \text{iid } (0, \sigma^2)$ over i and t , and $\text{Cov}(u_{it}, y_{i,t-1}) = 0$.

- Replace X_{it} in the static model by the lagged dependent variable $y_{i,t-1}$ to model the dynamic feedback.
- Other regressors can be added on the right-hand side.

FE Estimator

- How about estimate β by the FE estimator?

Let \tilde{Y}_{-1} be the demeaned variable of the lagged dependent variable, and then

$$\hat{\beta}_{FE} = \frac{\tilde{Y}'_{-1} \tilde{Y}}{\tilde{Y}'_{-1} \tilde{Y}_{-1}}$$

- It is easy to see

$$\hat{\beta}_{FE} - \beta_0 = \frac{\tilde{Y}'_{-1} \tilde{U}}{\tilde{Y}'_{-1} \tilde{Y}_{-1}},$$

where strict exogeneity is violated.

Nickell Bias

- Serious consequence: the FE estimator is inconsistent under “small T , large N ” (Nickell, 1981).
- A numerical demonstration of the Nickell bias.
<https://www.kaggle.com/code/jipann/nickell-bias>

Another Angle: First Difference

- Recall

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

- It follows

$$\begin{aligned}\hat{\beta}_{FD} - \beta_0 &= \frac{\sum_{i,t} \Delta y_{i,t-1} \Delta u_{it}}{\sum_{i,t} \Delta y_{i,t-1}^2} \\ &= \frac{\sum_{i,t} (y_{i,t-1} - y_{i,t-2}) (u_{it} - u_{i,t-1})}{\sum_{i,t} (y_{i,t-1} - y_{i,t-2})^2}\end{aligned}$$

Inherent Endogeneity in FD

- The expected value of the numerator is

$$\begin{aligned} & E[(y_{i,t-1} - y_{i,t-2})(u_{it} - u_{i,t-1})] \\ &= E[y_{i,t-1}u_{it}] - E[y_{i,t-1}u_{i,t-1}] - E[y_{i,t-2}u_{it}] + E[y_{i,t-2}u_{i,t-1}] \\ &= 0 - \sigma_u^2 - 0 + 0 = -\sigma_u^2 \neq 0 \end{aligned}$$

- $\hat{\beta}_{FD}$ is inconsistent under finite T .

Remedy: A further Lag as Instrument

- Anderson and Hsiao (1981): $\Delta y_{i,t-2}$ is a valid IV for the regressor $\Delta y_{i,t-1}$
 - $\Delta y_{i,t-2} = y_{i,t-2} - y_{i,t-3}$ is uncorrelated with Δu_{it} .
 - $\Delta y_{i,t-2}$ is correlated with $\Delta y_{i,t-1}$.

- 2SLS:

$$\hat{\beta}_{IV} = \frac{\sum_{i,t} \Delta y_{i,t-2} \Delta y_{i,t}}{\sum_{i,t} \Delta y_{i,t-2} \Delta y_{i,t-1}}.$$

- Consistent and asymptotic normal.

More Lags as IV

- Arellano and Bond (1991):
 $(y_{i,t-2}, y_{i,t-3}, y_{i,t-4}, \dots, y_{i,0})$ are all valid IV.
- Use GMM for estimation.
- The more IVs, the more efficient (in theory).
- Optimal weighting matrix is needed for efficiency.
- The practical issue of “too many instruments.”

Summary

- Panel data
- FE estimator
- RE estimator
- Static panel model
- Dynamic panel model

- Rich information
- Big data