Econ 2120, Section 6 - Panel Data

Christopher D. Walker, Alice Wu

October 9, 2022

Panel Data and Sampling Process

- Panel data is a data structure where we have multiple measurements of a single cross-sectional unit. Examples:
  - Following individuals/firms over time
  - Collecting data on multiple children within a family
- In our treatment of panel data, the random sampling assumption becomes that $\{(Y_{i1}, \ldots, Y_{iM}, X_{i1}, \ldots, X_{iM})\}_{i=1}^{n}$ is an i.i.d sequence of random vectors:
  - Individuals are iid but dependence across time is allowed.
  - Families are iid but dependence within a family is allowed

Table of Contents

Latent Variable Model

- We assume $\{(Y_{i1}, Y_{i2}, Z_{i1}, Z_{i2}, A_i)\}_{i=1}^n$ are i.i.d.
- A_i is not observed (i.e., it is a *latent variable*). We assume that

$$\mathbb{E}[Y_{it}|Z_{i1}, Z_{i2}, A_i] = g_t(Z_{it}, A_i), \ t = 1, 2$$

- Consequently, an analyst failing to account for A_i may face an omitted variable bias.
- Example:
  - Suppose that Z_{it} is education for twin t, Y_{it} is earnings for twin t, A_i are family-level unobservables.
  - Family level unobservables may correlate with earnings and education, so failing to account for them generates bias.

Assumptions

Assumption 1: Exclusion Restriction

$E[Y_{it}|Z_{i1}, Z_{i2}, A_i] = g_t(Z_{it}, A_i)$ where the effect of Z_{is} on Y_{it} is excluded if $s \neq t$

Assumption 2: Functional Form

$E[Y_{it}|Z_{i1}, Z_{i2}, A_i] = \gamma_{1t} + \gamma_{2t} Z_{it} + \gamma_{3t} A_i$

Assumption 3: Time-invariant Coefficients

$\gamma_{1t} \equiv \gamma_1$, $\gamma_{2t} \equiv \gamma_2$, $\gamma_{3t} \equiv \gamma_3$ for all t

Some comments:

- We imposed the first assumption on the previous slide, however, the other two are new. What do the assumptions rule out?
- Under A1-A3, we have

$$E^*[Y_{it}|1, Z_{i1}, Z_{i2}] = \gamma_1 + \gamma_2 Z_{it} + \gamma_3 E^*[A_i|Z_{i1}, Z_{i2}]$$

Assumptions

Note that we can write

$$E^*[A_i|Z_{i1}, Z_{i2}] = \lambda_0 + \lambda_1 Z_{i1} + \lambda_2 Z_{i2}.$$

Sometimes it might be plausible to impose a symmetry restriction.

Assumption 4: Symmetry

$\lambda_1 = \lambda_2$. In other words, $E^*[A_i|1, Z_{i1}, Z_{i2}] = \lambda_0 + \lambda_1(Z_{i1} + Z_{i2})$

Under A1-A4, we can identify γ_2 using the linear predictors for Y_{it}:

$$E^*[Y_{i1}|1, Z_{it}, Z_{i1} + Z_{i2}] = (\gamma_1 + \gamma_3\lambda_0) + \gamma_2 Z_{i1} + \gamma_3\lambda_1(Z_{i1} + Z_{i2})$$
$$E^*[Y_{i2}|1, Z_{it}, Z_{i1} + Z_{i2}] = (\gamma_1 + \gamma_3\lambda_0) + \gamma_2 Z_{i2} + \gamma_3\lambda_1(Z_{i1} + Z_{i2})$$

We turn to the GLP for inference.

Latent Variable Model vs. GLP

- Define
$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix}, \quad R_i = \begin{pmatrix} 1 & Z_{i1} & (Z_{i1} + Z_{i2}) \\ 1 & Z_{i2} & (Z_{i1} + Z_{i2}) \end{pmatrix}$$
We can find GLP given a weight matrix Φ:
$$E^*_{\Phi}(Y_i \mid R_i) = R_i \beta,$$
where $\beta \in \mathbb{R}^3$.

- Observe the following:
  - If all restrictions in the latent variable model are satisfied by the population distribution, the GLP yields the same $\beta_2 = \gamma_2$ for any Φ.
  - Otherwise, β_2 (GLP) is the best approximation for γ_2 (latent variable model with A1-A4) under a given Φ

Table of Contents

Potential Outcomes

- Suppose economic theory generates the following linear panel data model:

$$Y_{it} = \gamma Z_{it} + A_i + U_{it}, \ \ t = 1, ..., T$$

- We can express this model in terms of **potential outcomes**.
- Let $z \in \mathbb{R}^T$ be treatment assignment. The *potential outcome function* $Y_i(z)$ assigns a $T \times 1$ random vector to each value z:

$$Y_i(z) = \begin{pmatrix} Y_{i1}(z) & \cdots & Y_{iT}(z) \end{pmatrix}'.$$

- The potential outcome model allows us to define causal effects using the notion of a *treatment effect*: $TE_i(z, z') = Y_i(z') - Y_i(z)$ is the treatment effect from z to z'.
- Functionals of the distribution of TE_i are causal estimands. For example, average treatment effect $\mathbb{E}[TE_i(z, z')]$

Connecting Panel Data and Potential Outcomes

- Recall

$$Y_{it} = \gamma Z_{it} + A_i + U_{it}, \ t = 1, ..., T.$$

- The corresponding potential outcome function is

$$Y_{it}(z) = \gamma z_t + A_i + U_{it}, \ t = 1, ..., T, \ z = (z_1, ..., z_T)'.$$

- **A Challenge:** We observe treatment one treatment assignment Z_{it} and the realized outcome $Y_{it} = Y_{it}(Z_{it})$. How can we learn about TE_i?

Strict Exogeneity

- Recall the potential outcomes model

$$Y_{it}(z) = \gamma z_t + A_i + U_{it}, \;\; t = 1, ..., T, \;\; z \in \mathbb{R}^T$$

- We impose the following assumption:

Assumption 1: Strict Exogeneity

Conditional on A_i, the realized treatment Z_i is independent of potential outcomes:

$$\{Y_i(z) : z \in \mathbb{R}^T\} \perp\!\!\!\perp Z_i \,|A_i$$

where $Y_i(z) = (Y_{i1}(z), ..., Y_{iT}(z))$

- We can show that strict exogeneity is equivalent to

$$(U_{i1}, ..., U_{iT}) \perp\!\!\!\perp (Z_{i1}, ..., Z_{iT}) \,|A_i$$

under the model $Y_{it} = \gamma Z_{it} + A_i + U_{it}, \;\; t = 1, ..., T.$

Identification Under Strict Exogeneity

- By strict exogeneity,

$$E[U_{it}|A_i, Z_i] = E[U_{it}|A_i] \ \forall \ t \in \{1, ..., T\}.$$

Assumption 2: Functional Form

$$E[U_{it}|A_i] = \phi_{1t} + \phi_{2t}A_i$$

- Consequently,

$$E[Y_{it}|Z_i, A_i] = \gamma Z_{it} + (A_i + \phi_{1t} + \phi_{2t}A_i)$$

  and γ can be identified if we observe (Z_i, A_i).

- However, we do not observe A_i! So we need to transform the regression function to consistently estimate γ.

Table of Contents

First differences/within: assumptions

- On top of A1 (strict exogeneity) and A2 (linear functional form) in the potential outcome framework,

Assumption 3: Time-Invariance

$$E[U_{it}|A_i] = \phi_1 + \phi_2 A_i$$

- Under A1-A3, we have

$$E[Y_{it}|Z_i, A_i] = \gamma Z_{it} + (\phi_1 + (1 + \phi_2)A_i)$$

First differences/within

$$E[Y_{it}|Z_i, A_i] = \gamma Z_{it} + (\phi_1 + (1+\phi_2)A_i)$$

- **Goal:** find some transformation $f(Y_i)$ such that $E(f(Y_i) \mid Z_i, A_i)$ doesn't depend on A_i
- First differences:
$$E(Y_{it} - Y_{i,t-1} \mid Z_i, A_i) = \gamma(Z_{it} - Z_{i,t-1})$$
- Within:
$$E(Y_{it} - \overline{Y}_i \mid Z_i, A_i) = \gamma(Z_{it} - \overline{Z}_i)$$

Going to data (example: first differences)

1. Stack the model-implied equations

$$E\left(\begin{pmatrix} y_{i2} - y_{i1} \\ \vdots \\ y_{iT} - y_{i,T-1} \end{pmatrix} \mid Z_i\right) = \begin{pmatrix} Z_{i2} - Z_{i1} \\ \vdots \\ z_{iT} - z_{i,T-1} \end{pmatrix} \gamma$$

2. Find Y_i^{new}, R_i, and β such that $E(Y_i^{\text{new}} \mid R_i) = R_i\beta$

$$Y_i^{\text{new}} = \begin{pmatrix} y_{i2} - y_{i1} \\ \vdots \\ y_{iT} - y_{i,T-1} \end{pmatrix}, \quad R_i = \begin{pmatrix} Z_{i2} - Z_{i1} \\ \vdots \\ z_{iT} - z_{i,T-1} \end{pmatrix}, \quad \beta = \gamma$$

3. Use GLS to estimate β

4. From $\hat{\beta}$, get the object of interest

Importance of Strict Exogeneity

- Suppose we only assume that $U_{it} \perp\!\!\!\perp Z_{it} \mid A_i$ for each t
- For example:
$$E(U_{it} \mid Z_i, A_i) = \psi Z_{i,t-1} + \phi_1 + \phi_2 A_i$$
$$E(Y_{it} \mid Z_i, A_i) = \gamma Z_{it} + \psi Z_{i,t-1} + \phi_1 + (1 + \phi_2) A_i$$
- First differences:
$$E(Y_{it} - Y_{i,t-1} \mid Z_i) = \gamma(Z_{it} - Z_{i,t-1}) + \psi(Z_{i,t-1} - Z_{i,t-2})$$
- *In-Class Exercise:* What about within?

Table of Contents

Time-Varying Coefficients

- Suppose we relax A3, under A1-A2. Then we have:

$$E[U_{it}|A_i] = \phi_{1t} + \phi_{2t}A_i$$
$$E[Y_{it}|Z_i, A_i] = \gamma z_t + (\phi_{1t} + (1 + \phi_{2t})A_i)$$

- *In-Class Demonstration:* Verify that first-difference or within estimators won't work when A_i remains unobserved.

- Writing
$$E^*(A_i|1, Z_i) = \lambda_0 + \lambda_1 Z_{i1} + ... + \lambda_T Z_{iT},$$
we obtain the following expression for the linear predictor Y_{it}:
$$E^*(Y_{it}|1, Z_i) = \gamma Z_{it} + \delta_{1t} + \delta_{2t}(\lambda_1 Z_{i1} + ... + \lambda_T Z_{iT})$$

- This leads to the Chamberlain method:
  - Write out the model-implied unrestricted linear predictor and work on the matrix of coefficients that can be manipulated and help you identify some parameters.

Chamberlain method: takeaway

Example from note 6 (T=3):

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \end{pmatrix}, \quad X_i = \begin{pmatrix} 1 \\ Z_{i1} \\ Z_{i2} \\ Z_{i3} \end{pmatrix}, \quad E^*[Y_i'|X_i] = X_i'\Pi$$

where the unrestricted regression coefficients (under the model) are:

$$\Pi = \begin{pmatrix} \delta_{11} & \delta_{12} & \delta_{13} \\ (\gamma + \delta_{21}\lambda_1) & \delta_{22}\lambda_1 & \delta_{23}\lambda_1 \\ \delta_{21}\lambda_2 & (\gamma + \delta_{22}\lambda_2) & \delta_{23}\lambda_2 \\ \delta_{21}\lambda_3 & \delta_{22}\lambda_3 & (\gamma + \delta_{23}\lambda_3) \end{pmatrix}$$

Note: γ is identified as $\pi_{21} - \frac{\pi_{31}}{\pi_{33}}\pi_{22}$