Panel Data Regression

090507
Panel Data

Up to now we have analyzed either

- **Cross-sectional data** (data collected on several individuals/units at one point in time)

  or

- **Time series data** (data collected on one individual/unit over several time periods)

What if we have a combination of these two types of data?
Panel data are repeated cross-sections over time, in essence there will be space as well as time dimensions.

Other names are *pooled data, micropanel data, longitudinal data, event history analysis* and *cohort analysis*.
Panel Data Examples

The individuals/units can for example be workers, firms, states or countries

- Annual unemployment rates of each state over several years
- Quarterly sales of individual stores over several quarters
- Wages for the same worker, working at several different jobs
Panel Data Examples

Some american surveys:

- The National Longitudinal Survey of Youth (NLSY) tracks labor market outcomes for thousands of individuals, beginning in their teenage years.

- The Panel Survey of Income Dynamics (PSID) since 1968 collects data on 5000 families about various socioeconomic and demographic variables.

- The Survey of Income and Program Participation (SIPP), conducts interviews four times a year about the respondents economic conditions.
Panel Data

Potential gains

- take heterogeneity into account, get individual-specific estimates
- especially suitable to study dynamics of change
- study more sophisticated behavioral models
- minimize bias due to aggregation

However, panel data also increases the complexity of the analysis.
Panel Data

- Balanced/unbalanced
- Short panel/long panel
Two kinds of models:

**FIXED EFFECTS MODELS**

**RANDOM EFFECTS MODELS**

The two types of analyses make conceptually contrasting assumptions about effects as either random or fixed.
Example with 2 explanatory variables:

\[ Y_{it} = \beta_1 + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it} \]

Notice the subscript index \textit{it}

- \textit{i} stands for the \textit{i} \text{th} cross-sectional unit, \( i = 1, \ldots, N \)
- \textit{t} stands for the \textit{t} \text{th} time period, \( i = 1, \ldots, T \)
Pooled OLS Regression

Treats all observation as equivalent and OLS as usual

\[ Y_{it} = \beta_1 + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it} \]

In this case the error term captures "everything"

Naive, ignores time and space
Several kinds of fixed effects models, differs in the assumptions about

- The intercept
- The slope coefficients
Fixed Effects Models with Dummy Variables

<table>
<thead>
<tr>
<th></th>
<th>Varies over individuals</th>
<th>Varies over time</th>
</tr>
</thead>
<tbody>
<tr>
<td>The intercept</td>
<td>√</td>
<td>—</td>
</tr>
<tr>
<td>The slope coefficients</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Different intercepts for different individuals $\beta_{1i}$

$$Y_{it} = \beta_{1i} + \beta_{2}X_{2it} + \beta_{3}X_{3it} + u_{it}$$

but each individuals intercept does not vary over time

If the number of individuals is $N = 4$

$$Y_{it} = \alpha_{1} + \alpha_{2}D_{2i} + \alpha_{3}D_{3i} + \alpha_{4}D_{4i} + \beta_{2}X_{2it} + \beta_{3}X_{3it} + u_{it}$$
Fixed Effects Models with Dummy Variables

<table>
<thead>
<tr>
<th></th>
<th>Varies over individuals</th>
<th>Varies over time</th>
</tr>
</thead>
<tbody>
<tr>
<td>The intercept</td>
<td>—</td>
<td>✓</td>
</tr>
<tr>
<td>The slope coefficients</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Different intercepts for different time periods instead $\beta_{1t}$

$$Y_{it} = \beta_{1t} + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it}$$

If the number of time periods is $T = 20$

$$Y_{it} = \lambda_1 + \lambda_2 D_{2t} + \lambda_3 D_{3t} + \ldots + \lambda_{20} D_{20t} + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it}$$
Fixed Effects Models with Dummy Variables

<table>
<thead>
<tr>
<th></th>
<th>Varies over individuals</th>
<th>Varies over time</th>
</tr>
</thead>
<tbody>
<tr>
<td>The intercept</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>The slope coefficients</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Different intercepts for different individuals AND time periods $\beta_{1it}$

$$Y_{it} = \beta_{1it} + \beta_{2}X_{2it} + \beta_{3}X_{3it} + u_{it}$$

For $N = 4$ and $T = 20$

$$Y_{it} = \alpha_{1} + \alpha_{2}D_{2i} + \alpha_{3}D_{3i} + \alpha_{4}D_{4i} + \lambda_{1} + \lambda_{2}D_{2t} + \lambda_{3}D_{3t} + \ldots + \lambda_{20}D_{20t} + \beta_{2}X_{2it} + \beta_{3}X_{3it} + u_{it}$$
Fixed Effects Models with Dummy Variables

<table>
<thead>
<tr>
<th></th>
<th>Varies over individuals</th>
<th>Varies over time</th>
</tr>
</thead>
<tbody>
<tr>
<td>The intercept</td>
<td>√</td>
<td>—</td>
</tr>
<tr>
<td>The slope coefficients</td>
<td>√</td>
<td>—</td>
</tr>
</tbody>
</table>

Both intercepts and slopes varies over individuals, introduces a lot of dummy variables

\[
Y_{it} = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \beta_2 X_{2it} + \beta_3 X_{3it} \\
+ \gamma_1 D_{2i} X_{2it} + \gamma_2 D_{2i} X_{3it} + \gamma_3 D_{3i} X_{2it} + \gamma_4 D_{3i} X_{3it} \\
+ \gamma_5 D_{4i} X_{2it} + \gamma_6 D_{4i} X_{3it} + u_{it}
\]

the number of interaction terms is number of dummy variables × number of explanatory variables
Fixed Effects Models with Dummy Variables

Both intercepts and slopes varies over individuals and time

requires even more variables
Fixed Effects Models

Cautions

- a lot of dummy variables
  ⇒ less df
  ⇒ increased risk of multicollinearity

- have to reflect on the assumptions about the error term $u_{it}$
  - heteroscedasticity?
  - autocorrelation?

  easily gets complicated when both time and space dimensions
Random Effects Models

Now, in the Random Effects Model

\[ Y_{it} = \beta_{1i} + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it} \]

the intercepts/effects \( \beta_{1i} \) are assumed to be random variables with mean value

\[ E(\beta_{1i}) = \beta_1 \]

and the intercept value for individual \( i \) can be expressed as

\[ \beta_{1i} = \beta_1 + \varepsilon_i \quad i = 1, \ldots, N \]

where \( E(\varepsilon_i) = 0 \) and \( Var(\varepsilon_i) = \sigma^2_\varepsilon \)
Random Effects Models

each individual in the sample is considered to be a drawing from an infinite (or "close to") population of individuals which share the common mean value $\beta_1$

$$Y_{it} = \beta_1 + \beta_2 X_{2it} + \beta_3 X_{3it} + \varepsilon_i + u_{it}$$

$$Y_{it} = \beta_1 + \beta_2 X_{2it} + \beta_3 X_{3it} + w_{it}$$

The error term $w_{it}$ consists of two components, random effects models are sometimes called error components models
Random Effects Models

Assumptions about the error components

\[ \varepsilon_i \sim N \left(0, \sigma^2_{\varepsilon} \right) \]

\[ E(\varepsilon_i \varepsilon_j) = 0 \quad \text{for } i \neq j \]

\[ u_{it} \sim N \left(0, \sigma^2_u \right) \]

\[ E(u_{it} u_{is}) = E(u_{it} u_{it}) = E(u_{it} u_{js}) = 0 \quad \text{for } i \neq j \quad t \neq s \]

\[ E(\varepsilon_i u_{it}) = 0 \]
Random Effects Models

\[
\begin{align*}
E(w_{it}) &= 0 \\
\text{Var}(w_{it}) &= \sigma^2_\varepsilon + \sigma^2_u \\
\text{Corr}(w_{it}, w_{is}) &= \frac{\sigma^2_\varepsilon}{\sigma^2_\varepsilon + \sigma^2_u}
\end{align*}
\]
Random Effects vs Fixed Effects

depends on

- whether or not the individuals can be viewed as a random sample from a large population

\[ E(\varepsilon_iX_i) = 0? \]

If yes: random effects, if no: fixed effects

- the relation between \( T \) and \( N \)
  - for large \( T \) and small \( N \) not a big difference
  - for small \( T \) and large \( N \) random effects estimators are more efficient than fixed effects (if the assumptions hold)