Linear Panel Models

Johannes Gräb

July 17, 2006
What is Panel Data?

- Combination of cross section and time series data
- Repeated observations on the same cross section, observed for several time periods
- Terminology:
  - Longitudinal data
  - Repeated measures
  - Pooled cross section and time series data
Motivation for Using Panel Data

- Increased precision in estimation
  - More information
  - More degrees of freedom
- Solving the omitted variables problem
- Allowing for unobserved heterogeneity
- Examination of issues that cannot be studied in either cross-sectional or time-series settings alone
  - e.g. female labor force participation
The Problem of Omitted Variables

- Modelling causal relationship often suffers from endogeneity caused by omitted variables.
- OLS estimates are biased if the omitted/unobserved variable is correlated with the regressor and a determinant of the dependent variable.
- Instrumental variable methods allow for correcting of omitted variable bias, but it is often difficult to obtain valid instruments.
- Having a panel it is in some cases possible to overcome the omitted variable problem.
Basic Framework of the Panel Model

- Panel data analysis uses a wide range of models and estimators
- The framework for discussion is a regression model of the form:
  \[ y_{it} = x_{it}' \beta + c_i + \varepsilon_{it} \quad i = 1, ..., N, \quad t = 1, ..., T \]
  - \( c_i \) is an of unobserved individual or group specific variable
  - The unobserved individual variable is assumed to be constant over time
Overview of Models and Estimators

<table>
<thead>
<tr>
<th>Assumed Model</th>
<th>Pooled</th>
<th>Random Effects</th>
<th>Fixed Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Estimator of $\beta$</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pooled OLS</td>
<td>Consistent</td>
<td>Consistent</td>
<td>Inconsistent</td>
</tr>
<tr>
<td>Between</td>
<td>Consistent</td>
<td>Consistent</td>
<td>Inconsistent</td>
</tr>
<tr>
<td>Within/LSDV</td>
<td>Consistent</td>
<td>Consistent</td>
<td>Consistent</td>
</tr>
<tr>
<td>First Differences</td>
<td>Consistent</td>
<td>Consistent</td>
<td>Consistent</td>
</tr>
<tr>
<td>Random Effects</td>
<td>Consistent</td>
<td>Consistent</td>
<td>Inconsistent</td>
</tr>
</tbody>
</table>
Pooled Model

- The most restrictive model is the pooled model.
- There is assumed to be no unobserved individual heterogeneity, so the model reduces to:

\[ y_{it} = x_{it}'\beta + c + \varepsilon_{it} \]

- The model underlies the usual assumptions for cross section analysis.
Pooled OLS Estimator (POLS)

- Consistent, if:
  - the pooled model is appropriate
  - \( E(x_t \epsilon_t) = 0 \) for all \( t \) ⇒ contemporaneous exogeneity
  - no perfect collinearity

- Correct statistical inference, if:
  - \( E(\epsilon_t^2 | x_t) = \sigma^2 \) for all \( t \) ⇒ Homoskedasticity
  - \( E(\epsilon_t \epsilon_s | x_t, x_s) = 0 \) for all \( t \neq s \) ⇒ no serial correlation

- The standard errors can be adjusted for heteroskedasticity and serial correlation to assure correct inference
Fixed Effects and Random Effects Models

- The fixed and random effects model assume the existence of unobserved individual heterogeneity:

\[ y_{it} = x_{it}'\beta + c_i + \varepsilon_{it} \]

- Main question: Is the unobserved heterogeneity correlated with the explanatory variables?
- The fixed effects model (FE) treats \( c_i \) as a variable that is partially correlated with the observed regressors
- The random effects model (RE) treats \( c_i \) as independently distributed of the regressors
Least Squares Dummy Variable Estimator (LSDV)

- Traditional approach to estimate “fixed” effects models
- LSDV approach: $c_i$ is to be estimated along with $\beta$ coefficients
- Introduction of $N$ dummy variables, for each specific unit:

$$y_{it} = x_{it}'\beta + \sum_{i=1}^{N} c_i + \varepsilon_{it}$$

- Introduction of $T-1$ time effects
Within Estimator

- Transforms the data to solve out the unobserved effect
- Original model:
  \[ y_{it} = x_{it}'\beta + c_i + \varepsilon_{it} \]
- Take averages over time dimension:
  \[ \bar{y}_i = \bar{x}_i'\beta + c_i + \bar{\varepsilon}_i. \]
- Substraction yields:
  \[ y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)'\beta + (\varepsilon_{it} - \bar{\varepsilon}_i) \]
Within Estimator

- Measures the variation in data only over time
- Time invariant explanatory variables drop out
- FE estimators are consistent and efficient, if
  - the regressors are strictly exogenous:
    \[ E[\varepsilon_{it}|x_{i1}, \ldots, x_{iT}] = 0 \quad t = 1, 2, \ldots, T \]
  - the idiosyncratic error terms \( \varepsilon_{it} \) are homoskedastic and serially uncorrelated
Conclusions About Fixed Effects Estimators

- **Advantages**
  - Consistent estimates, if unobserved heterogeneity is correlated with regressors
  - FE estimators solve omitted variable bias

- **Drawbacks**
  - Time invariant regressors fall out
  - Tremendous loss of information $\Rightarrow$ less degrees of freedom

- FE estimators should not be used unless it is necessary
Between Estimator

- Exploits only variation over cross sections
- Reduces panel to a cross section of averages over time
- Take averages over time dimension:
  \[
  \bar{y}_i = \bar{x}_i \beta + c_i + \bar{\varepsilon}_i.
  \]
- Estimator not very useful:
  - Inconsistent in a FE model
  - Inefficient in a RE model
Random Effects Estimator

- The RE estimator models the individual effects as purely random across cross sectional units

\[ y_{it} = x_{it}'\beta + c_i + \varepsilon_{it} \]

- The composite error is: \( \nu_{it} = (c_i + \varepsilon_{it}) \)

- RE estimator takes into account information about the structure of the error term
**Random Effects Estimator**

- RE estimator is a pooled GLS estimator:

\[
y_{it} - \hat{\lambda} \bar{y}_i = (x_{it} - \hat{\lambda} \bar{x}_i)' \beta + (v_{it} - \hat{\lambda} \bar{v}_i)
\]

\[
\hat{\lambda} = 1 - \frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + T \sigma_c^2}}
\]

- RE estimator measures variation over time and over cross sections

- RE estimator is a matrix weighted average of the FE- and the BE- estimators
Pooled OLS versus Random Effects versus Fixed Effects

- If there is no individual heterogeneity, pooled OLS will be consistent and efficient.
- Wald test for $H_0 : \sigma_c^2 = 0$ in the RE model.
- If there is heterogeneity, there exists a trade off between bias and precision.
  - If $c_i$ is uncorrelated with $x_i$, RE will be more efficient.
  - If $c_i$ is correlated with $x_i$, RE will be biased, but FE consistent.
- In some cases it is advisable to use RE, even if it will be biased.
Hausman Test

- The Hausman test on fixed effects, is a test on the difference of the estimators.
- Idea: If $c_i$ is uncorrelated with $x_i$, both estimates are consistent and the difference is expected to be relatively small.
- Test statistic of a single coefficient:

$$T = \frac{\hat{\delta}_{FE} - \hat{\delta}_{RE}}{\sqrt{\hat{\text{Var}}(\hat{\delta}_{FE}) - \hat{\text{Var}}(\hat{\delta}_{RE})}}$$
Procedure to Estimate Panel Data

1. Run a FE model and a RE model
2. Run a Hausman test
3. If the difference between the $\beta'$s is significantly different, FE will be appropriate
4. If not, test for $\sigma_c^2 = 0$
5. If $\sigma_c^2 = 0$ is rejected, RE will be appropriate
6. If not, POLS will be fine
Example: Hours and Wages

- How large is the responsiveness of labor supply to wages?

\[ \ln \text{hrs}_{it} = c_i + \beta \ln \text{wg}_{it} + \varepsilon_{it} \]

- Cross section analysis finds relatively small positive response
- Cross section estimates might be biased because of unobserved heterogeneity
- An unobserved individual effect that might be positively correlated with wage is the desire to work
# Hours and Wages

<table>
<thead>
<tr>
<th></th>
<th>POLS</th>
<th>Between</th>
<th>Within</th>
<th>RE-GLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c$</td>
<td>7.442</td>
<td>7.483</td>
<td>7.220</td>
<td>7.346</td>
</tr>
<tr>
<td>$\beta$</td>
<td>.083</td>
<td>.067</td>
<td>.168</td>
<td>.119</td>
</tr>
<tr>
<td>Robust se</td>
<td>.030</td>
<td>.024</td>
<td>.085</td>
<td>.051</td>
</tr>
<tr>
<td>Default se</td>
<td>.009</td>
<td>.020</td>
<td>.019</td>
<td>.014</td>
</tr>
<tr>
<td>$R^2$</td>
<td>.015</td>
<td>.021</td>
<td>.016</td>
<td>.014</td>
</tr>
<tr>
<td>$\sigma_c$</td>
<td>.000</td>
<td>–</td>
<td>.181</td>
<td>.161</td>
</tr>
<tr>
<td>$\sigma_\epsilon$</td>
<td>.283</td>
<td>–</td>
<td>.232</td>
<td>.233</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>0</td>
<td>–</td>
<td>1</td>
<td>.585</td>
</tr>
<tr>
<td>$N$</td>
<td>5320</td>
<td>532</td>
<td>5320</td>
<td>5320</td>
</tr>
</tbody>
</table>
Instrumental Variable Regression in a Panel Model

- Consider you want to estimate the effect of a time invariant regressor, when the appropriate model is a FE model.
- RE estimation is biased, while in FE estimation time invariant regressors fall out.
- Hausman and Taylor (HT) considered a model that preserves the advantages of both estimators:
  - It allows for correlation between individual effects and regressors (FE).
  - It identifies the effects of time invariant regressors (RE).
- HT suggest an instrumental variable approach.
Hausman Taylor Model

They consider a model of the form:

\[ y_{it} = x'_{1it} \beta_1 + x'_{2it} \beta_2 + w'_{1i} \gamma_1 + w'_{2i} \gamma_2 + c_i + \varepsilon_{it} \]

- \( x_{1it} \) are time-varying variables, uncorrelated with \( c_i \)
- \( x_{2it} \) are time-varying variables, correlated with \( c_i \)
- \( w_{1i} \) are time-invariant variables, uncorrelated with \( c_i \)
- \( w_{2i} \) are time-invariant variables, correlated with \( c_i \)
Hausman and Taylor considered the following steps:

1. Within estimation produces consistent estimators of $\beta_1$ and $\beta_2$

2. Regress $w_{1i}, w_{2i}$ using $w_{1i}, x_{1it}$ as instruments on the group means of the within residuals

3. Use the residual variances in step 1 and step 2 to obtain the FGLS weight $\hat{\lambda}$ and perform GLS transformation for all variables

4. Use the weighted instrumental variable estimators to obtain the coefficients of interest by instrumental variables regression
Dynamic Panel Models

- Hausman and Taylors instrumental variable approach as a basis for dynamic panel models
  \[ y_{it} = \gamma y_{i,t-1} + x_{it}' \beta + c_i + \varepsilon_{it} \]
- Introduced estimators are all biased in this dynamic model
- Arellano and Bond proposed an IV approach
- First-Differences model:
  \[ y_{it} - y_{i,t-1} = \gamma (y_{i,t-1} - y_{i,t-2}) + (x_{it} - x_{i,t-1})' \beta + (\varepsilon_{it} - \varepsilon_{i,t-1}) \]
- For example, \( y_{i,t-2} \) as an instrument for \( (y_{i,t-1} - y_{i,t-2}) \)
