PANEL ATTRITION

Cecilia Jona-Lasinio

EIEF – Econometric Reading Group

22 October, 2009
- Sample selection bias (Vella, 1998)
- How to model it? (Heckman, 1976, 1979)
- Accounting for selectivity bias in panel data
  - Attrition (Hausman and Wise, 1979)
- How to test for attrition bias (Nijman and Verbeek, 1992)
- Alternative methods to estimate panel with attrition (Hirano, Imbens, Ridder, Rubin, 2001)
- Open issues
Selection vs Attrition Biases

• In the econometric literature Attrition bias is associated with models of Selection bias.

• Since Heckman (1979), attrition bias has been modeled in the framework of selection bias models.

• However, in the survey sampling literature the recognition of the problem of nonresponse and the bias it can cause dates from much earlier (see Madow et al.. 1983, for a review).
Sample selection bias

**Framework:** analysis of the determinants of wages and labour supply behavior of females (Gronau, 1974; Heckman, 1974)

**Population:** women (working, not working)

**Aim:** look at the determinants of the wages of the working women to draw some conclusions about the determinants of wages for all women.

The *selection bias* is related to the differences between workers and non-workers.
Sample selection bias

Each individual is characterized by her endowment of **observable** and **unobservable** characteristics.

2 possible cases:
- random sub-sample
- non random sub-sample
  (the decision to work is no longer random)

Some of the determinants of the work decision may influence also wages (age, education).
Sample selection bias

Sample selectivity operates through unobservables and their correlation with observables.

The distinction between selection on unobservables and observables is critical to the development of tests for attrition bias and adjustments to eliminate it.
Sample selection model

(1) \( y_i^* = x_i' \beta + \varepsilon_i; \quad i = 1, \ldots, N \)

(2) \( d_i^* = z_i' \gamma + \nu_i; \quad i = 1, \ldots, N \)

(3) \( d_i = 1 \) if \( d_i^* > 0; \quad d_i = 0 \) otherwise

(4) \( y_i = y_i^* \times d_i; \)

Where \( y_i^* \) is a latent endogenous variable with the observed counterpart \( y_i \),
\( d_i^* \) is a latent variable with associated indicator function \( d_i \) reflecting whether
the primary dependent variable is observed or not.

\( x_i \) and \( z_i \) are vectors of exogenous variables, and it is assumed that \( x_i \) is
contained in \( z_i \).

\( \varepsilon_i \) and \( \nu_i \) are zero mean error terms with \( \mathbb{E}[\varepsilon_i \mid \nu_i] \neq 0 \) so that \( \mathbb{E}[x_i \mid \varepsilon_i] \neq 0 \)

\text{OLS provide inconsistent estimates}
1. Heckman (1974) suggested to use a **maximum likelihood estimator** and to assume that $\epsilon_i$ and $v_i$ are i.i.d. N(0,\Sigma) and they are independent of $z_i$.

2. **Two-step estimation** (Heckman, 1976, 1979) where he proposed to introduce a “correction term” accounting for the selection bias.
Selectivity bias in panel data

“In panel data the selection proceeds in **stages**: in every **wave** of the panel some respondents are lost. In order to exploit the unique potential of panel data it is important to assess the biases induced by this progressive selection”, Ridder (1992).

Selectivity bias with respect to time = **Attrition**

Examples: social experimentation, panel surveys
Selectivity bias in panel data

\[ y_{it}^* = x_{it}' \beta + \mu_i + \xi_t + e_{it} \]

\[ d_{it}^* = z_{it}' \gamma + \alpha_i + \psi_t + v_{it} \]

\[ d_{it} = 1 \text{ if } d_{it}^* > 0 \]

\[ y_{it} = y_{it}^* * d_{it} \]
Hausman and Wise (1979) proposed a method to deal with panel attrition that uses a probability model of attrition in conjunction with a traditional random effect model of individual response.

Main findings: the extent of attrition bias depends substantially on the specification of the model used to evaluate the experimental effect.
Attrition model (HW)

\[ y_{it} = X_{it} \beta + \varepsilon_{it}, \quad i=1,..,N; \ t=1,2; \]

\[ \varepsilon_{it} = \mu_i + \eta_t, \quad \varepsilon_{it} \sim N(0, \sigma^2) \]

\[ \mathbb{E}(\varepsilon_{it}) = 0, \quad \mathbb{V}(\varepsilon_{it}) = \sigma^2_{\mu} + \sigma^2_{\eta} = \sigma^2 \]

How much of total variance is accounted for by unobserved individual effect?

\[ \rho_{i2} = \frac{\sigma^2_{\mu}}{\sigma^2_{\mu} + \sigma^2_{\eta}} \]

Suppose that the probability of observing \( y_{i2} \) varies with its value as well as with the values of other variables. Probability of observing \( y_{i2} \) will depend on \( \varepsilon_{i2} \) and OLS will provide biased estimates.
Define \( a_i \) such that \( a_i = 0 \) if attrition occurs in period 2, and \( a_i = 1 \) if attrition does not occur. Suppose that \( y_{i2} \) is observed if \( A_i \geq 0 \), where:

\[
A_i = X_i \xi + W_i \gamma + \varepsilon_{i3},
\]

\[
R_i = [X_{i2}, W_i] \quad \text{and} \quad \delta = [\xi, \gamma]
\]

\[
\Pr (a_i = 1) = \Phi[R_i \delta], \quad \Pr (a_i = 0) = 1 - \Phi[R_i \delta]
\]

where \( \Phi[\ ] \) is the unit normal distribution function.
Attrition model (HW)

\[ y_{i1} = X_{i1}\beta + \varepsilon_{i1}, \]
\[ y_{i2} = X_{i2}\beta + \varepsilon_{i2}, \]
\[ A_i = R_i\delta + \varepsilon_{i3}. \]

HW looked for a way of obtaining asymptotically efficient and consistent estimates of the above parameters and of testing the presence of attrition (\( \rho_{23} = 0 \)).

**ML procedure**: the log likelihood combines the variance components specification of the y with the probit formulation of probabilities. They showed that if \( \rho_{23} = 0 \) then the log likelihood function factors into 2 parts: a normal regression model and a probit equation. In this case GLS is fine.
Empirical results (HW)

From Hausman and Wise, (1979)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Period 1</th>
<th>Period 2</th>
<th>(standard errors)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-experimental average, $\alpha$</td>
<td>6.2638</td>
<td></td>
<td>(0.4517)</td>
</tr>
<tr>
<td>Time effect, $\delta_2$</td>
<td>0.1180</td>
<td></td>
<td>(0.1673)</td>
</tr>
<tr>
<td>Experimental effect, $\xi$</td>
<td>-0.0642</td>
<td></td>
<td>(0.0826)</td>
</tr>
</tbody>
</table>
### Empirical results (HW)

#### Table III

**Parameter Estimates for the Analysis of Variance Specification Combined with the Attrition Model**

<table>
<thead>
<tr>
<th>Analysis of variance</th>
<th>Estimates (standard errors)</th>
<th>Attrition</th>
<th>Estimates (standard errors)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-experimental average, ( \alpha )</td>
<td>6.2636 (.0265)</td>
<td>Constant</td>
<td>-.9210 (.2608)</td>
</tr>
<tr>
<td>Time effect, ( \delta_2 )</td>
<td>.1064 (.0408)</td>
<td>Experimental Effect</td>
<td>.2361 (.1131)</td>
</tr>
<tr>
<td>Experimental effect, ( \xi )</td>
<td>-.1098 (.0453)</td>
<td>Education</td>
<td>.0172 (.0195)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Experience</td>
<td>-.0002 (.0050)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Income</td>
<td>.0934 (.0290)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Union</td>
<td>1.2018 (0.1100)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poor Health</td>
<td>.2715 (.1013)</td>
</tr>
<tr>
<td>Attrition bias parameter ( \rho_{23} )</td>
<td>- .8213 (.0449)</td>
<td>Earnings correlation ( \rho_{12} )</td>
<td>.1697 (.0350)</td>
</tr>
<tr>
<td>Likelihood value</td>
<td>36.24</td>
<td>Earnings variance ( \sigma_n^2 )</td>
<td>.2147 (.0006)</td>
</tr>
</tbody>
</table>

*From Hausman and Wise, (1979)*
Empirical results (HW)

<table>
<thead>
<tr>
<th>Variables</th>
<th>With attrition correction: maximum likelihood estimates (standard errors)</th>
<th>Attrition parameters</th>
<th>Without attrition correction: generalized least squares estimates (standard errors)</th>
<th>Earnings function parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>5.8539 (.0903)</td>
<td>-.6347 (.3351)</td>
<td>5.8911 (.0829)</td>
<td></td>
</tr>
<tr>
<td>Experimental effect</td>
<td>-.0822 (.0402)</td>
<td>.2414 (.1211)</td>
<td>-0.0793 (.0390)</td>
<td></td>
</tr>
<tr>
<td>Time effect</td>
<td>.0940 (.0520)</td>
<td>—</td>
<td>.0841 (.0358)</td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>.0209 (.0052)</td>
<td>-.0204 (.0244)</td>
<td>.0136 (.0050)</td>
<td></td>
</tr>
<tr>
<td>Experience</td>
<td>.0037 (.0013)</td>
<td>-.0038 (.0061)</td>
<td>.0020 (.0013)</td>
<td></td>
</tr>
<tr>
<td>Income</td>
<td>-.0131 (.0050)</td>
<td>.1752 (.0470)</td>
<td>-.0115 (.0044)</td>
<td></td>
</tr>
<tr>
<td>Union</td>
<td>.2159 (.0362)</td>
<td>1.4290 (.1252)</td>
<td>.2853 (.0330)</td>
<td></td>
</tr>
<tr>
<td>Poor health</td>
<td>-.0601 (.0330)</td>
<td>.2480 (.1237)</td>
<td>-.0578 (.0326)</td>
<td></td>
</tr>
</tbody>
</table>

$\hat{\sigma}_n^2 = .1832 \ (0.0057)$  
$\hat{\rho}_{12} = .2596 \ (0.0391)$  
$\rho_{23} = -.1089 \ (1.0429)$  

From Hausman and Wise, (1979)
Empirical evidence (HW)

HW showed that the analysis of variance models which do not account for possible attrition bias lead to parameter estimates that differ substantially from the ML estimates that take into account the attrition bias.

The structural model is less affected by attrition than the analysis of variance model.
Testing attrition bias

• The HW model has been generalized by Nijman and Verbeek (1992). They proposed some simple methods to assess the consistency of the most widely used estimators in panel data with non response.

• But, they did not suggest a unique and efficient method to test for attrition bias. Their main finding is that the more efficient estimates are those on unbalanced panels taking into account attrition.

• The open question is: what is the best method to correct for non response/attrition?
Panel data with attrition can be augmented by replacing the lost units with new units randomly sampled from the original population. This additional samples are called **Refreshment Samples**.

Hirano, Imbens, Ridder and Rubin (2001) show that adding a refreshment sample can reduce the effects of attrition by making the estimate of conventional model more robust.
Most used models for panel data with attrition are:

- Missing At Random (MAR), (Rubin (1976), Little and Rubin (1987));
- Hausman and Wise (HW);
- Additive Non-ignorable (AN), (Irano et al., 2001);
The general model

2 - period panel

\( Z_{it} = \) vector of time-varying variables

\( X_i = \) vector of time-invariant variables

T_1 draw a random sample \( N_P \) [Panel, \( X_i \ Z_{i1} ]

T_2 a subset \( N_{BP} \) [Balanced Panel, \( X_i \ Z_{i1} \ Z_{i2} ]

T_2 refreshment subsample \( N_R [X_i \ Z_{i2} ]\)

\( N_{IP} = N_P - N_{BP} \) [Incomplete Panel]

If \( W_i = 1 \) then \( Z_{i2} \) will be recorded, if \( W_i = 0 \) \( Z_{i2} \) is missing.
Estimating panel data with attrition

\[ f(Z_1, Z_2, X) = \frac{f(Z_1, Z_2, X | W = 1)}{\Pr(W = 1 | Z_1, Z_2, X)} \cdot \Pr(W = 1) \]

Selection on observables (MAR)

\[ \Pr(W = 1 | Z_1, Z_2, X) = \Pr(W = 1 | Z_1, X) \]

Selection on unobservables (HW)

\[ \Pr(W = 1 | Z_1, Z_2 X) = \Pr(W = 1 | Z_2, X) \]
Estimating panel data with attrition

The parametric version of the MAR and HW models allows to specify attrition indicator as:

\[ W_i = 1 \left\{ \pi_0 + \pi_1 \cdot X_i + \pi_2 \cdot Y_{i1} + \eta_i > 0 \right\}, \]

\[ W_i = 1 \left\{ \pi_0 + \pi_1 \cdot X_i + \pi_3 \cdot Y_{i2} + \eta_i > 0 \right\}. \]

The choice between them can be done only on theoretical considerations because they require different exclusion restrictions to be identified. It is difficult to choose one of them a priori.
Estimating panel data with attrition

The refreshment sample allows to make a choice and to estimate a more general model.

\[ W_i = 1 \{ \pi_0 + \pi_1 \cdot X_i + \pi_2 \cdot Y_{i1} + \pi_3 \cdot Y_{i2} + \eta_i > 0 \} \]

Where the attrition decision depends on both \( Y_{i1} \) and \( Y_{i2} \), and the models restrictions become testable. Model parameters can be estimated from the panel thus allowing to produce an estimate of the marginal distribution of \( Z_2 \).

The attrition model can be tested comparing the indirect and direct (refreshment sample) estimates of the marginal distribution.
Estimating panel data with attrition

In presence of refreshment samples MAR and HW models have testable implications and they can be used to build more general models. The probability of response will be

\[ \Pr(W = 1|Z_1 = z_1, Z_2 = z_2) = g(\alpha_0 + \alpha_1 \cdot z_1 + \alpha_2 \cdot z_2 + \alpha_3 \cdot z_1 \cdot z_2), \]

MAR \((\alpha_2=\alpha_3=0)\) and HW \((\alpha_1=\alpha_3=0)\) so that the **Additive Non-ignorable** model will be written as:

\[ \Pr(W = 1|Z_1 = z_1, Z_2 = z_2) = g(\alpha_0 + \alpha_1 \cdot z_1 + \alpha_2 \cdot z_2), \]
Main findings

Adding a refreshment sample allows to reduce the effects of attrition:

• more precise estimates of both MAR and HW
• more general models AN
• provides additional information about the attrition process
Open issues

Figure 1
Attrition Hazards: Sample With No New Entrants
Backup slides