Using Repeated Cross-sections to explore Movements in and out of poverty

Peter Lanjouw *World Bank*

Jill Luoto *UC Berkeley*

David McKenzie *World Bank*
Goal of this paper

• Explore whether repeated cross-sections which are widely available can be used to provide some reasonable basic descriptives of transitions in and out of poverty.

• Set out an approach which we claim will give upper and lower bounds on mobility.

• Validate these methods by using genuine panel data from Vietnam and Indonesia, generating repeated cross-sections from these panels, and comparing the results of our method to what one would estimate based on the genuine panel.

  • Interested to match overall rates of exit and entry from poverty
  • Also to describe patterns of which groups in society are experiencing more, or less, of such transitions.

  • “profiles” of poverty duration
Our proposed approach

• Combines ideas of poverty-mapping with pseudo-panel ideas.
• Will set out for case of 2 rounds, can be extended easily to multiple rounds.
• Let $x_{i1}$ be characteristics of household $i$ in time period 1, which are observed in both the round 1 and round 2 surveys:
  • All time-invariant characteristics (language, religion, ethnicity)
  • Characteristics of household head if the head doesn’t change across rounds (sex, place of birth, parental education, etc.)
  • Can include time-varying characteristics that can easily be recalled for round 1 in round 2
    • E.g. whether household head was employed in round 1, place of residence in round 1, whether household has a TV in round 1, etc.
Projections

- Project round 1 consumption or income onto $x_{i1}$:

$$y_{i1} = \beta'_1 x_{i1} + \varepsilon_{i1}$$

- Project round 2 consumption or income onto same set of characteristics as they appear at time of second round:

$$y_{i2} = \beta'_2 x_{i2} + \varepsilon_{i2}$$

- Then we are interested in knowing quantities such as:

$$Pr(y_{i1} < p | y_{i2} > p)$$

Don’t observe for the same household
Proposed method

- **Step one:** Use the sample of households observed in round 1, and regress $y_{i1}^1$ on $x_{i1}^1$
  - Obtain the OLS estimator $\hat{\beta}_1$ and the residuals:
    \[ \varepsilon_{i1}^1 = y_{i1}^1 - \hat{\beta}_1' x_{i1}^1 \]
  - Superscript 1 denotes that these are observations for households observed in round 1 only.

- **Step two:** For each household observed in round 2, take a random draw with replacement from the empirical distribution of residuals, then combine with parameter estimate and known $x$ to estimate round 1 income or consumption:
  \[ \hat{y}_{i1}^2 = \hat{\beta}_1' x_{i1}^2 + \varepsilon_{i1}^2 \]
Proposed method

- **Step Three:** calculate movements into and out of poverty using
  \( \hat{y}_{i1}^2 \) in place of the unobserved round 1 variable:

  \[ Pr(\hat{y}_{i1}^2 < p | y_{i2}^2 > p) \]

- **Step Four:** Repeat steps 1-3 \( R \) times, and take average of the
  quantity of interest over the \( R \) replications.
Under what conditions will this be consistent?

• *Condition 1*: the underlying population sampled is the same in round 1 and round 2
  • Requires measure of consumption to be same from round to round,
  • no (non-random) changes in underlying population from births, deaths, migration out of sample…as with pseudo-panels in general, household analysis works best when restricted to households headed by prime age adults
    • If we want to look at *income* mobility, more latitude
Under what conditions will this be consistent?

- **Condition 2**: \( \varepsilon_{i1} \) is independent of \( y_{i2} \). This requires \( \varepsilon_{i1} \) to be independent of \( \varepsilon_{i2} \)

  (otherwise the distribution of \( \varepsilon_{i1|y_{i2}>p} \) is not the same as the unconditional distribution of \( \varepsilon_{i1} \))

  - Won’t hold if:
    - Error term contains individual fixed effect
    - If shocks to consumption or income are non-transitory.

- Expect in many cases this condition to be violated.
  - So long as errors positively correlated (which seems likely in most cases), this will overstate mobility, providing an upper bound on movements into and out of poverty.
  - If errors are negatively correlated then our method wouldn’t provide a bound.
    - We don’t expect a negative correlation on average
    - Will demonstrate this empirically with real panel data.
Lower bound method

• Instead assume the prediction error for household $i$ in round 1 is the same as it is for round 2 (perfect positive autocorrelation).

• *Step One:* for sample of households surveyed in round 2, obtain OLS residuals:

\[ \varepsilon_{i2}^2 = y_{i2}^2 - \hat{\beta}_2'x_{i2}^2 \]

• *Step Two:* then estimate round 1 income or consumption as

\[ \tilde{y}_{i1}^2 = \hat{\beta}_1'x_{i1}^2 + \varepsilon_{i2}^2 \]

• *Step Three:* Use the estimated $y$ from step 2 to calculate poverty dynamic of interest.
Datasets

- Choose two genuine panels from Vietnam and Indonesia:
    - Period over which poverty fell from 58% to 37%, more households exiting poverty than entering
    - Panel of approximately 4800 households
  - Indonesian Family Life Survey 1997 and 2000 (IFLS2 and 3)
    - Static in terms of overall poverty levels, household moving into and out of poverty at similar rates
    - Panel of 7500
Validation of method

- Randomly split each genuine panel into two sub-samples, A and B.
  - Use sub-sample A from round 1 and sub-sample B from round 2 as two repeated cross-sections.
  - Then carry out our method by using sub-sample A to impute round 1 values for sub-sample B, and compare to results we would get using genuine panel for sub-sample B.
Choosing variables

- Consider a hierarchy of models which progressively employ more and more data that is sometimes, but not always, collected retrospectively.
- Since we have the actual panel data to work with, we can force variables to be time-invariant by using round 1 variables.
- Start with a basic “traditional model”, and add more regressors.
Models

1. (Basic Model): gender of head, age of head as of round 1, birthplace of head (rural/urban), whether the head ever attended primary school, education of head’s parents, head’s religion and ethnicity. \((R^2 = 0.20)\)

2. Add locational dummies for where household was living in round 1.

3. Add community variables from round 1 (e.g. village has electricity, village has a stone road, community has a primary school)

4. Head’s sector of work and education in round 1

5. Demographic variables from round 1 (household size, number of children)

6. Household’s assets and housing quality as of round 1 – e.g. did household own TV, radio, what sort of roof and floor did it have? \((R^2 = 0.42)\).
Results: predicting levels of poverty

Table 1: Poverty Headcount:

<table>
<thead>
<tr>
<th>Data Source:</th>
<th>Round 1:</th>
<th>Lower Bound</th>
<th>Truth</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Basic</td>
<td>Full</td>
<td>95% CI</td>
</tr>
<tr>
<td>IFLS</td>
<td>1997 Poverty Rate (P0):</td>
<td>0.147</td>
<td>0.159</td>
<td>0.145</td>
</tr>
<tr>
<td>VLSS</td>
<td>1992 Poverty Rate (P0):</td>
<td>0.611</td>
<td>0.592</td>
<td>0.597</td>
</tr>
</tbody>
</table>

Method seems to be getting levels close
Correlation between residuals

- Recall our claim was that the residuals would likely be positively autocorrelated, making our first method an upper bound, and that this correlation would shrink as we add more variables to the model.
- This is what we see:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indonesia</td>
<td>0.474</td>
<td>0.466</td>
<td>0.464</td>
<td>0.452</td>
<td>0.408</td>
<td>0.348</td>
</tr>
<tr>
<td>Vietnam</td>
<td>0.653</td>
<td>0.575</td>
<td>0.563</td>
<td>0.539</td>
<td>0.523</td>
<td>0.420</td>
</tr>
</tbody>
</table>

Columns 1-6 build increasingly rich models of consumption.
How well does the approach do at estimating overall rates of movements into and out of poverty?

Table 3: Poverty Dynamics from “Pseudo” Panel and Actual Panel Data

<table>
<thead>
<tr>
<th></th>
<th>Lower Bounds</th>
<th>Truth</th>
<th>Upper Bounds</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Indonesia</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1997, 2000 Statuses</td>
<td>Basic Full 95% CI</td>
<td></td>
<td>Basic Full</td>
</tr>
<tr>
<td>Poor, Poor</td>
<td>0.115 0.105 0.047</td>
<td>0.070 0.024 0.037</td>
<td></td>
</tr>
<tr>
<td>Poor, Nonpoor</td>
<td>0.015 0.031 0.065</td>
<td>0.088 0.097 0.090</td>
<td></td>
</tr>
<tr>
<td>Nonpoor, Poor</td>
<td>0.021 0.030 0.065</td>
<td>0.088 0.111 0.099</td>
<td></td>
</tr>
<tr>
<td>Nonpoor, Nonpoor</td>
<td>0.848 0.832 0.759</td>
<td>0.801 0.766 0.774</td>
<td></td>
</tr>
<tr>
<td><strong>Vietnam</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1992, 1998 Statuses</td>
<td>Basic Full 95% CI</td>
<td></td>
<td>Basic Full</td>
</tr>
<tr>
<td>Poor, Poor</td>
<td>0.360 0.322 0.275</td>
<td>0.360 0.227 0.288</td>
<td></td>
</tr>
<tr>
<td>Poor, Nonpoor</td>
<td>0.241 0.274 0.261</td>
<td>0.324 0.331 0.308</td>
<td></td>
</tr>
<tr>
<td>Nonpoor, Poor</td>
<td>0.000 0.039 0.034</td>
<td>0.060 0.138 0.077</td>
<td></td>
</tr>
<tr>
<td>Nonpoor, Nonpoor</td>
<td>0.398 0.366 0.300</td>
<td>0.386 0.305 0.327</td>
<td></td>
</tr>
</tbody>
</table>

For both countries, round 1 year is predicted, round 2 is "truth"
Results seem encouraging

• Bounds not that wide:
  • Full model would lead us to estimate 3-9% of households in Indonesia and 27-31% of households in Vietnam exited poverty over 2 rounds.
  • Genuine panel would say 7-9% in Indonesia and 26-32% in Vietnam

• More detailed model for consumption with higher $R^2$ leads to narrower bounds
  • E.g. bounds of 0.021-0.111 using basic model vs (0.033-0.099) using full model for entry into poverty rate in Indonesia.
What about describing profile of mobility?

- Would like to know which sub-groups of the population are more likely to be entering or exiting poverty.
- E.g. are there regions where more people have entered or exited poverty? How does mobility vary by ethnic group or by education?
- We plot the full model upper bound estimates against the panel estimates to see whether ranking we get seems similar.
Figure 1: Exiting Poverty

Percentage Exiting from Poverty in Vietnam Between Period 1 and 2 by Population Sub Group
Comparing Mobility Based on Panel Data Against Pseudo Panel Data
Pseudo Panel Based on Upper Bound Method and Full Model Specification
Figure 2: Entering Poverty

Percentage Entering into Poverty in Vietnam Between Period 1 and 2 by Population Sub Group
Comparing Mobility Based on Panel Data Against Pseudo Panel Data
Pseudo Panel Based on Upper Bound Method and Full Model Specification
Figure 3: Exiting Poverty: Indonesia

Percentage Exiting from Poverty in Indonesia Between Period 1 and 2 by Population Sub Group
Comparing Mobility Based on Panel Data Against Pseudo Panel Data
Pseudo Panel Based on Upper Bound Method and Full Model Specification
Figure 4: Entering Poverty - Indonesia

Percentage Entering into Poverty in Indonesia Between Period 1 and 2 by Population Sub Group
Comparing Mobility Based on Panel Data Against Pseudo Panel Data
Pseudo Panel Based on Upper Bound Method and Full Model Specification
Do we get the same rankings?

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pearson Correlation</td>
<td>Spearman Rank Correlation</td>
</tr>
<tr>
<td>Lower bound method, basic model</td>
<td>0.399</td>
<td>0.480</td>
</tr>
<tr>
<td>Lower bound method, full model</td>
<td>0.650</td>
<td>0.770</td>
</tr>
<tr>
<td>Upper bound method, basic model</td>
<td>0.271</td>
<td>0.050</td>
</tr>
<tr>
<td>Upper bound method, full model</td>
<td>0.886</td>
<td>0.833</td>
</tr>
</tbody>
</table>
Rankings

- Seems to give relatively similar rankings as genuine panel
- Particularly when upper bound full model is used
- When basic model is used, much less robust conclusions can be drawn
Conclusions/where to from here?

- Genuine panel data is rare, and even the best panels often smaller in scale & frequency than cross-sectional surveys.
- E.g. Indonesia IFLS is one of, if not the, best developing country panel out there
  - But not nationally representative
  - Sample size of around 7000 households
  - Low frequency
  - Vs SUSENAS
    - Annual, nationally representative (and representative at district level), around 200,000 households!

➢ Policymakers and academics do care about movements into and out of poverty- would be nice to be able to say something regularly and in most countries, even if what we can say is relatively basic.
Conclusions/where to?

- We’ve provided a method of using repeated cross-sections to obtain bounds on movements into and out of poverty
  - Validated this against genuine panel data
  - Found the bounds can be narrow enough in practice to be useful
- However, method works best when full range of variables used, some of which are not typically asked retrospectively in surveys
  - But no reason why they can’t be – and much cheaper to add a few of these questions than field a panel

=> Seems worth experimenting with inclusion of some such questions in upcoming surveys.