ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Dang, Hai-Anh; Lanjouw, Peter F.

Working Paper

Measuring Poverty Dynamics with Synthetic Panels Based on Repeated Cross-Sections

IZA Discussion Papers, No. 15827

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Dang, Hai-Anh; Lanjouw, Peter F. (2022) : Measuring Poverty Dynamics with Synthetic Panels Based on Repeated Cross-Sections, IZA Discussion Papers, No. 15827, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/272454

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15827

Measuring Poverty Dynamics with Synthetic Panels Based on Repeated Cross-Sections

Hai-Anh H. Dang
Peter F. Lanjouw

DECEMBER 2022

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15827

Measuring Poverty Dynamics with Synthetic Panels Based on Repeated Cross-Sections

Hai-Anh H. Dang
World Bank, Indiana University, IZA and Vietnam National University

Peter F. Lanjouw
VU University Amsterdam and Tinbergen Institute

DECEMBER 2022

Any opinions expressed in this paper are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but IZA takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

The IZA Institute of Labor Economics is an independent economic research institute that conducts research in labor economics and offers evidence-based policy advice on labor market issues. Supported by the Deutsche Post Foundation, IZA runs the world's largest network of economists, whose research aims to provide answers to the global labor market challenges of our time. Our key objective is to build bridges between academic research, policymakers and society.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Measuring Poverty Dynamics with Synthetic Panels Based on Repeated Cross-Sections*

Panel data are rarely available for developing countries. Departing from traditional pseudo-panel methods that require multiple rounds of cross-sectional data to study poverty mobility at the cohort level, we develop a procedure that works with as few as two survey rounds and produces point estimates of transitions along the welfare distribution at the more disaggregated household level. Validation using Monte Carlo simulations and real cross-sectional and actual panel survey data— from several countries, spanning different income levels and geographical regions—perform well under various deviations from model assumptions. The method could also inform investigation of other welfare outcome dynamics.

JEL Classification: C53, D31, I32, O15

Keywords: transitory and chronic poverty, income mobility, consumption, cross sections, synthetic panels, household surveys

Corresponding author:
Hai-Anh H. Dang
Data Production & Methods Unit Development Data Group World Bank
1818 H St. N.W.
Washington, D.C. 20433
USA
E-mail: hdang@worldbank.org

* We would like to thank the editor Climent Quintana-Domeque, two anonymous reviewers, Francois Bourguignon, Fiona Burlig, Alan Dorfman, Chris Elbers, Francisco Ferreira, Gary Fields, Paul Glewwe, Bill Greene, Bo Honore, Stephen Jenkins, Dean Jolliffe, Aart Kraay, Christoph Lakner, Yue Man Lee, Michael Lokshin, Andy McKay, David McKenzie, David Newhouse, Reema Nayar, Franco Peracchi, Tuoc Van Phan, Menno Pradhan, Sergiy Radyakin, Carolina Sanchez-Paramo, Erik Thorbecke, Renos Vakis, Roy van der Weide, Nobuo Yoshida, and participants at meetings of the Econometric Society in Asia (Singapore) and Latin America (Medellin), International Association for Applied Econometrics (London), International Conference on Panel Data (London), North East Universities Development Consortium (MIT), and conferences and seminars at Cornell, IFPRI, Oxford, University of New South Wales, and World Bank for helpful discussions on earlier versions of this paper. We further thank Renos Vakis and Leonardo Lucchetti for their help with the Peruvian data. We also thank the UK Foreign Commonwealth and Development Office (FCDO) for funding assistance through its various programs with the World Bank, including the Strategic Research Program (SRP), Knowledge for Change (KCP) program, and the Data and Evidence for Tackling Extreme Poverty (DEEP) Research Program.

I. Introduction

Following the steady progress of the past few decades in global poverty reduction, policy makers in both richer and poorer countries are devoting more attention to the nuanced dynamics underlying poverty and income mobility (e.g., Stiglitz, 2013; Piketty, 2014; World Bank, 2017). Measuring and tracking economic mobility, especially for the lower income groups, are increasingly regarded as essential for improving shared prosperity.[1] Indeed, a better understanding of the factors that help households escape poverty, or induce them to remain in or fall into poverty, would lead to a more effective and efficient fight against poverty. Panel data are traditionally employed to answer these questions. Collecting such data, however, can be very costly and can pose a number of logistical and capacity-related challenges. The scarcity of panel data has thus rendered the analysis of welfare dynamics difficult, if not impossible, in many developing country settings.

To overcome the non-availability of (actual) panel data, there have been a variety of efforts to develop pseudo-panels (or synthetic panels) out of multiple rounds of cross-sectional data (see, e.g., Deaton (1985), Pencavel (2007), Inoue (2008), and Juodis (2018)). Notably, since cross-section samples are typically refreshed each time that the surveys are fielded, these synthetic panels are possibly less exposed to the concerns surrounding attrition and measurement error that are often leveled at panel data.[2] Yet, because of their emphasis on cohorts rather than the household or individual, synthetic panel methods have not been widely applied to the study of poverty

[1] For example, Reeves (2020) calls for using mobility metrics as the "measure of the nation" for the U.S. Poverty mobility also stands out in a December 2013 address by US President Obama to the Center for American Progress (https://www.whitehouse.gov/the-press-office/2013/12/04/remarks-president-economic-mobility). See also Baulch (2011) for a collection of studies on poverty dynamics for developing countries.

[2] See, for example, Glewwe and Jacoby (2000) and Kalton (2009) respectively for overviews of the advantages and disadvantages of cross sections and panel data in developing and richer country contexts. See also Lee, Ridder and Strauss (2017) for a recent study that investigates the impacts of measurement errors on poverty mobility using several rounds of panel data from South Korea.

dynamics. Two notable exceptions are Bourguignon, Goh and Kim (2004) and Güell and Hu (2006) who construct synthetic panels at the household level. However, these two approaches require certain assumptions that may not always be easily satisfied in available cross sections: the former requires at least three rounds of cross section data and assumes a first-order auto-regression (AR(1)) process through which past household or individual incomes (earnings) can affect present outcomes; the latter is exclusively restricted to duration analysis.

Building on a poverty imputation technique described in Elbers *et al.* (2003), a recent paper by Dang, Lanjouw, Luoto, and McKenzie (2014) constructs synthetic panels from as few as two rounds of household-level cross sectional data that can provide lower-bound and upper-bound estimates of poverty transitions. Drawing on validation data from Vietnam and Indonesia, this paper finds that the "true" estimates of poverty mobility (as revealed by the actual panel data) are generally sandwiched between the upper bounds and lower bounds derived from the synthetic panels. However, this method's practical appeal is limited since it often yields rather wide bounds on estimated mobility, and these can be narrowed only if certain key statistical parameters can be imported from externally available panel data.[3]

We propose a significant refinement to the method introduced by Dang *et al.* (2014) to analyze mobility using *only* commonly available cross-sectional survey data. Our new method is predicated on some additional but fairly standard assumptions that allow us to move beyond *bound* estimates to actual *point* estimates of poverty mobility. This offers greater accuracy, easier interpretation, and potentially much wider application. In particular, we can easily investigate multiple measures

[3] This method focuses on constructing the synthetic panels from two or more rounds of cross sections, each of which has consumption data. See also Gibson (2001) for a somewhat related study on how panel data on a subset of individuals can be used to infer chronic poverty for a larger sample. More broadly, this method is related to the literature on identifying the bounds on the joint distribution for outcomes in different samples (see, e.g., Cross and Manski, 2002) and the statistical literature on imputing missing data (see, e.g., Little and Rubin, 2020). See also Ridder and Moffitt (2007) and Dang, Jolliffe, and Carletto (2019) respectively for reviews on the econometrics of data combination and poverty imputation.

of poverty dynamics, such as the population shares in different poverty categories in both survey periods considered together (i.e., unconditional or joint probabilities) or the population shares in different poverty status categories in one period given their welfare status in the other period (i.e., conditional probabilities). We further provide new formulae for the standard errors on point estimates.

We also make additional contributions on both the time dimension and deeper treatment of income mobility. In particular, we extend the existing method to settings where more than two rounds of data are available to investigate richer inter-temporal profiles of movement into and out of poverty. Our framework also permits more general analysis of mobility among different income groups, rather than just the *2x2* poverty transition matrix. This expands analysis of mobility from merely focusing on the lower part of the income distribution to its entire range and offer relevant inputs for policy advice. For example, as living standards are rising globally and the global poverty rate has been decreasing, more attention is being focused on the vulnerable population groups that are currently not poor but have a high risk of falling into poverty (e.g., World Bank (2017)). As another example, it is common practice to present a 5x5 transition matrix to examine income mobility where this is permitted by available panel data (e.g., Fields (2001)).

On the empirical front, we first validate our estimates with Monte Carlo simulations for various data situations, including settings where variables are only partially observed to one where they are fully observed, as well as different sample sizes. We further implement a number of "stress tests" of the estimators under deviations from the model assumptions. We subsequently validate our proposed methods with multiple rounds of cross sectional and panel survey data from several countries including Bosnia-Herzegovina, Lao PDR, Peru, the United States, and Vietnam. These countries represent diverse settings ranging from developing to high-income countries in different

geographical locations, covering both household income (the US) and household consumption data (the remaining countries). We find that our synthetic panel estimates are close to those derived from panel data—often lying within the 95 percent confidence intervals (CIs) or even one standard error of the latter in many cases.

Recent validations and applications of (earlier versions of) our synthetic panel methods by various researchers for different country contexts ranging from India to Africa, Latin America, and Europe have been yielding encouraging results (Ferreira *et al.*, 2012; Beegle *et al.*, 2016; UNDP, 2016; OECD, 2018; Dang *et al.*, 2019; Salvuci and Tarp, 2021). Even in those cases where our synthetic panel estimates fall outside the CIs surrounding the true panel estimates, the observed qualitative patterns of poverty mobility are generally quite similar between the panel and synthetic panel estimates. Herault and Jenkins (2019) and Garces-Urzainqui (2017) similarly document examples where strict statistical criteria are not satisfied, but the qualitative conclusions needed for policy design remain fairly robust.[4]

This paper consists of six sections. We discuss the basic framework and theoretical results in the next section, and the Monte Carlo simulation exercise in Section III. Our data are described in Section IV and we report on the empirical validations using actual panel data in Section V. Section VI offers concluding remarks. We leave most of the technical details to Appendix 1, describe in

[4] Herault and Jenkins (2019) also suggest that their poverty mobility estimates based on household survey data from Australia and Great Britain are less accurate than those using data from lower-income countries in other studies. Yet, two notable features stand out from their validation study that may contribute (to some extent) to the smaller accuracy in their study. One, their estimated R^2's using household survey data from Australia and Great Britain hover around 0.1-0.2 for regressions with more than 30 independent variables (regressors), which are generally lower than those shown in previous studies using much fewer regressors. For example, our estimated R^2's are predominantly between 0.2 and 0.5 for regressions using seven regressors only (Appendix 3, Table 3.1). Second, between two-thirds and three-fourths of the estimated coefficients on these regressors in Herault and Jenkins (2019) are statistically insignificant, which stand in contrast to the generally strongly statistically significant estimated coefficient in other validation studies. Indeed, adding more regressors in a misspecified model could result in less accurate estimates for both the correlation coefficients and the income model as a whole (Snijders and Bosker, 1994; Nakagawa and Schielzeth, 2013; De Luca *et al.*, 2018). A deeper concern raised by Herault and Jenkins (2019), and also echoed in Garcés Urzainqui (2017) and Colgan (2022), relates to the potential sensitivity of results to cohort definition.

more detail the Monte Carlo simulation in Appendix 2, offer more data description, robustness checks, and additional estimation results in Appendix 3, and summarize the estimation procedures in Appendix 4.

II. Analytical Framework for Point Estimates on Poverty Mobility

II.1. Basic Framework

Let y_{ij} represent household consumption or income in survey round j for household i, where $i = 1, \ldots, N$, and $j = 1$ or 2. Let x_{ij} be a vector of time-invariant household characteristics that are observed in both survey rounds. Subject to data availability, these characteristics can include such variables as sex, ethnicity, religion, language, place of birth, and parental education as well as variables that can be converted into time-invariant versions based, for example, on information about household heads' age and education. The vector x_{ij} can also include time-varying household characteristics if retrospective questions about the round-1 values of such characteristics are asked in the second round survey.

Consider the following projection of household consumption (or income) on household characteristics for survey round j

$$y_{ij} = \beta_j' x_{ij} + \varepsilon_{ij} \quad (1)$$

We are interested in knowing such quantities of poverty dynamics as

$$P(y_{i1} \sim z_1 \ \ and \ y_{i2} \sim z_2) \quad (2)$$

or

$$P(y_{i1} \sim z_1 \ \ | y_{i2} \sim z_2) \quad (3)$$

where the vector x_{ij} includes a vector of ones, z_j is the poverty line in period j, and the relation sign ($\sim$) indicates either the larger sign ($>$) or smaller or equal sign ($\leq$). For example, $P(y_{i1} \leq z_1 \ and \ y_{i2} > z_2)$ represents the probability that household i is poor in the first period but nonpoor

in the second period (considered together for two periods), and $P(y_{i2} > z_2 | y_{i1} \leq z_1)$ represents the probability that household *i*, who are poor in the first period, escape poverty in the second period. These probabilities can also be interpreted as population quantities; for example, $P(y_{i2} > z_2 | y_{i1} \leq z_1)$ correspond to the percentage of poor households in the first period that escape poverty in the second period. We also refer to those who are either poor or non-poor in both periods as the immobile, and those who escape or fall into poverty over time respectively as the upward and downward mobile. For convenience, we also refer to quantities (2) and (3) respectively as unconditional mobility and conditional mobility.[5]

If panel data are available, we can easily estimate these quantities; otherwise, we can use synthetic panels for this purpose. To further operationalize the framework, we make the following two assumptions.

Assumption 1: *The underlying population sampled is the same in survey round 1 and survey round 2.*

Assumption 1 ensures that the distributions of the time-invariant household characteristics in the two survey rounds would be the same. As such, these time-invariant household characteristics can be employed as the connectors of household consumption between the two periods (i.e., $x_{i1} \equiv x_{i2}$). Coupled with Equation (1), this assumption implies that households in period 2 with identical characteristics to those of households in period 1 would have achieved the same consumption levels in period 1 and vice versa (given the same error term). Assumption 1 will be violated if the underlying population changes due to major events as births, deaths, or migration; these events can be caused by natural disasters or economic crises or simply because the two survey rounds are

[5] We restrict our discussion in this paper to a money-metric measure of poverty; for a multidimensional measure see Alkire and Foster (2011). Also see Calvo and Dercon (2009) and Foster (2009) for discussion on other definitions of chronic poverty.

too far apart. We can thus test this assumption by examining whether the observable time-invariant characteristics of the population of interest change significantly from one survey round to the next.

Assumption 2: $\boldsymbol{\varepsilon_{i1}}$ *and* $\boldsymbol{\varepsilon_{i2}}$ *have a bivariate normal distribution with zero mean, (the partial) correlation coefficient ρ, and standard deviations* $\boldsymbol{\sigma_{\varepsilon_1}}$ *and* $\boldsymbol{\sigma_{\varepsilon_2}}$ *respectively.*

Since we often convert household consumption y_{ij} to the logarithmic scale for better analysis, the normality assumption for log consumption is equivalent to the log-normality assumption of consumption. Unlike Assumption 1, the assumption of joint normality is widely used in practice but cannot be tested without panel data. We come back to relaxing this assumption in the empirical analysis. The partial (conditional) correlation coefficient ρ is usually non-negative in most household surveys, for several reasons. First, since household poverty status tends to be strongly related over time, the joint probability that a household is poor in both survey rounds considered *together* is expected to be higher than the product of the probability that this household is poor in one round and poor in round two, respectively. Second, if shocks to consumption or income (for example, finding or losing a job) have some persistence, and consumption reacts to these income shocks, then consumption errors will also exhibit positive autocorrelation. And finally, although some households may experience negatively correlated incomes over time (e.g., reducing expenditure in one period in order to prepare for a wedding in the next), factors leading to such a correlation are unlikely to apply to the majority of households at the same time.[6]

Assumption 2 is also simpler and less data-demanding than the assumptions typically employed in other pseudo panel models that analyze multiple rounds of repeated cross sections. Put differently, we assume that no cohort (or time) specific effects exist; neither do we explicitly

[6] Assumption 2 is basic and fairly standard for most analysis of household consumption (income) data. We return to discussing other aspects such as potential heterogeneity and heteroscedasticity of the error terms with Monte Carlo simulation in Section III and Appendix 2.

assume individual level heterogeneity (as in, for example, Inoue (2008)). While we acknowledge that this rather simplistic departure from the literature could result in more restrictive analysis, it is motivated by the dearth of (even cross-sectional) survey data we typically face with in practice, particularly for poorer countries.[7] Notably, in situations where only two rounds of repeated cross sections exist, Assumption 2 is crucial for implementing our proposed model. But we return to relax this assumption and allow for the cohort fixed effects in the error terms in an alternative approach in Proposition 2. We further examine heterogeneity analysis in Section V.3.

If ρ is known, we can estimate quantity (2) by

$$P(y_{i1} \sim z_1 \; and \; y_{i2} \sim z_2) = \Phi_2\left(d_1 \frac{z_1 - \beta_1' x_{ij}}{\sigma_{\varepsilon_1}}, d_2 \frac{z_2 - \beta_2' x_{ij}}{\sigma_{\varepsilon_2}}, \rho_d\right) \quad (4)$$

where $\Phi_2(.)$ stands for the standard bivariate normal cumulative distribution function (cdf), d_j is an indicator function that equals 1 if the household is poor and equals -1 if the household is non-poor in period j, and $\rho_d = d_1 d_2 \rho$.

We discuss next our point estimates method which addresses the limitations of, and significantly extends, the bounds method introduced in Dang *et al.* (2014).

II.2. Theoretical Estimates for ρ

We offer the following proposition to obtain ρ, which helps provide the point estimate for poverty mobility.

Proposition 1- Point estimate of ρ
Given Equation (1) and Assumptions 1 and 2, and assuming that the simple (unconditional) correlation coefficient between household consumption in two survey rounds $\rho_{y_{i1}y_{i2}}$ is known, the partial correlation coefficient ρ is given by

$$\rho = \frac{\rho_{y_{i1}y_{i2}} \sqrt{var(y_{i1})\, var(y_{i2})} - \beta_1' var(x_i) \beta_2}{\sigma_{\varepsilon_1} \sigma_{\varepsilon_2}} \quad (5)$$

[7] Serajuddin *et al.* (2015) find that, over the period 2002-11, more than one-third (57) of the 155 countries for which the World Bank monitors poverty data have only one poverty data point or no data at all. Even where countries collect data on poverty, these data may not be comparable over time. Indeed, Beegle *et al.* (2016) point out that around half of 48 Sub-Saharan African countries did not have two comparable household surveys for the period 1990-2012.

Central to the estimation of ρ in Proposition 1 is the value of $\rho_{y_{i1}y_{i2}}$. We propose next a simple way to approximate this parameter based on cohort-level averages from the survey data.

Lemma 1- Approximation of $\boldsymbol{\rho_{y_{i1}y_{i2}}}$
Assume the following simple linear projection of household consumption between period 1 and period 2

$$y_{i2} = \delta y_{i1} + \eta_{i2} \quad (6)$$

where δ is a scalar, η_{i2} is the random error term. Further assume there are no other control (x_{ij}) variables in Equation (6) and x_{ij} have no cohort-specific first moment. Also assume that the sample size of each household survey round is large enough (or $N \to \infty$) and the number of cohorts (C) constructed from the survey data is fixed. The simple correlation coefficient $\rho_{y_{i1}y_{i2}}$ can then be approximated with the synthetic panel cohort-level simple correlation coefficient $\rho_{y_{c1}y_{c2}}$, where c indexes the cohorts constructed from the household survey data.

See Appendix 1 for further discussion.

We can rely on the existing literature on pseudo-panel data to construct cohorts. For example, cohorts can be based on age (Deaton, 1985; Pencavel, 2007) or some combination of age and other characteristics such as education (e.g., Blundell *et al.*, 1998) or region (e.g., Propper *et al.*, 2001). In the same spirit, other time-invariant characteristics such as gender or ethnicity may also qualify as candidates for cohort construction. The implicit assumption underlying traditional pseudo-panel analysis is that cohort dummy variables have a strong relationship with household consumption.[8] The assumption stated in Lemma 1 on a fixed number of cohorts is standard in the traditional pseudo-panel literature (Moffitt, 1993; Verbeek and Vella, 2005; Joudis, 2018) and helps preclude

[8] In addition, we can obtain good estimates of correlation at the cohort-level aggregated data if the individual data within a cohort show very similar values (or the intraclass correlation is close to 1 (Snijders and Bosker, 2011)). Furthermore, if these cohort dummy variables do not capture any variation in household consumption, the synthetic panel cohort-level simple correlation coefficient $\rho_{y_{c1}y_{c2}}$ would simply be 0. In the extreme case, consumption (or poverty) mobility can happen entirely within cohorts, but this case would be easily detected since it results in $\rho_{y_{c1}y_{c2}}$ being equal to 1 (i.e., since cohort means remain unchanged across the two survey rounds). We return to more discussion in the next section on Monte Carlo simulation.

measurement errors with cohort means. It is also defined as the Type 1 asymptotics of pseudo panel data (Verbeek, 2008).

Notably, given the small number of cohorts in practice (where we may have only two rounds of repeated cross sections), we do not include other control variables in Equation (6). Similar to Assumption 2 that is discussed earlier, Equation (6) represents a simplification of the typical linear dynamic model employed in the pseudo-panel literature due to data constraints (see, e.g., Moffitt (1993)). The assumption that x_{ij} have no cohort-specific first moment helps ensure that δ is consistently estimable when it is linked to Equation (1) (Inoue, 2008). Lemma 1 can be straightforwardly extended to multiple waves to obtain $\rho_{y_{cj}y_{ck}}$ for any pair of survey rounds *j* and *k*, but a longer time interval between survey rounds tends to decrease $\rho_{y_{cj}y_{ck}}$. For example, Kopczuk, Saez, and Song (2010) find that the (rank) correlation of earnings decreases over longer time intervals for panel data from the US Social Security Administration between 1937 and 2004.[9]

Alternatively, we can instead assume that the number of cohorts is large enough (instead of being fixed). This is the Type 2 asymptotics of pseudo panel data, which was proposed by Deaton (1985) and subsequently used in various studies including Verbeek and Nijman (1993) and Collado (1997). Using this different assumption allows us to employ a richer assumption for the error terms that includes the cohort fixed effects in the error term, which we refer to as Assumption 3 below.

Assumption 3*: Further assume that the error terms* $\boldsymbol{\varepsilon_{i1}}$ *and* $\boldsymbol{\varepsilon_{i2}}$ *include a cohort fixed effects.*

This offers another way of estimating ρ.

Proposition 2- Alternative estimate of $\boldsymbol{\rho}$
Given Equation (1) and Assumptions 1, 2, and 3, and assume that the sample size of each household survey round is large enough (or $N \to \infty$*) and the number of cohorts (C) constructed*

[9] This result also holds for actual panel data from various other countries such as China, India, Peru, Vietnam, and the U.K. (Chaudhuri and Ravallion, 1994; Khor and Pencavel, 2006; Jenkins, 2011; our estimates).

from the survey data is large enough (or $C \to \infty$). The partial correlation coefficient ρ can then be estimated from a modified version of Equation (1) where all the variables are aggregated to the cohort level

$$y_{cj} = \beta_j' x_{cj} + \varepsilon_{cj} \quad (7)$$

where the error term ε_{cj} includes a cohort fixed effect τ_c and the error ν_{cj}.

Proof
See Appendix 1.

Notably, the different assumptions over whether the number of cohorts is fixed (Lemma 1) or goes to infinity (Proposition 2) result in two different ways to construct cohorts. Lemma 1 suggests that we can construct cohorts based on age (and perhaps interacted with another variable), but Proposition 2 suggests that we can construct cohorts based on a combination of all the different values of the time-invariant variables in x_{ij}. The latter approach provides many more cohorts than the former, if there are enough time-invariant variables. For instance, using the US's PSID data in 2007-2009 with a sample size of around 3,400 observations, the number of constructed cohorts is 31 with Lemma 1 (using age as the cohort variable, with a restriction of heads between age 25 and 55), but the corresponding figure using Proposition 2 (for a combination of age, gender, years of schooling, ethnicity, and urban residence) is 1,120. Given a typical sample size of 5,000 for most current household surveys, the number of cohorts (and cohort cell sizes) can be slightly larger.

While it appears reasonable to assume that *N* tending to infinity with most current household surveys, there is no consensus in the literature on how large cohort sizes should be. Monte Carlo simulations by Verbeek and Nijman (1992) suggests that cohort sizes of 100 to 200 are sufficient, while Devereux (2007) argues for larger cohort sizes in the thousands. Khan (2021) offers a new metric to calculate cell sizes; yet, it is a complex function that is sensitive to variations within and across cohorts, over time for cohorts, as well as autocorrelation and covariance of the control variables. Indeed, our validation results, shown in Section V, suggest that we can obtain reasonably

good estimates for total sample sizes ranging from slightly more than 1,300 observations (Bosnia-Herzegovina) to 9,100 observations (Peru) and the results do not appear to strongly depend on the sample sizes.

We note the caveat that Lemma 1 provides approximates of $\rho_{y_{i1}y_{i2}}$ and ρ and Proposition 2 is based on asymptotic theory (using a large number of cross sections), and how well these estimates turn out to be in practice (using only two rounds of cross sections) is an empirical issue. A simple (but partial) diagnostic test for Proposition 1 to work is that the cohort-level simple correlation coefficient $\rho_{y_{c1}y_{c2}}$ is statistically different from 0; the corresponding test for Proposition 2 is that β_j in Equation (7) are jointly statistically different from 0. We offer a sample Stata command in Appendix 4 to estimate Equation (7). Our preferred method for the empirical illustrations in this paper is Proposition 1 and Lemma 1, since this approach lays out more clearly the relationship between $\rho_{y_{i1}y_{i2}}$ and ρ. But we also use Proposition 2 for alternative estimates.[10]

II.3. Mobility for Three (or More) Periods or Consumption Groups

We next provide Proposition 3, which shows the asymptotics of the point estimates in Equation (4).

Proposition 3- Asymptotic results for point estimates for 2 periods
Assume that Equation (1) and Assumptions 1 and 2 hold, and assume further that all the standard regularity conditions are satisfied for Equations (1), (i.e., $X'\varepsilon/N \xrightarrow{p} 0$ *and* $X'X/N \xrightarrow{p} M$ *finite and positive definite).*[11] *Let P be the population parameter of interest (e.g.,* $P = P(y_{i1} < z_1 \text{ and } y_{i2} > z_2)$ *for household i, i=1,…, N),* d_j *an indicator function that equals 1 if the household is poor and equals -1 if the household is non-poor in period j, j= 1, 2,* $\rho_d = d_1 d_{\ 2}\rho$, and $\rho_{y_{i1}y_{i2},d} = d_1 d_{\ 2}\rho_{y_{i1}y_{i2}}$, *and the (^) sign represent the estimate. Our point estimates are distributed as*

$$\sqrt{n}\left[P - \Phi_2\left(d_1\frac{z_1-\widehat{\beta}_1'x_{ij}}{\widehat{\sigma}_{\varepsilon_1}}, d_2\frac{z_2-\widehat{\beta}_2'x_{ij}}{\widehat{\sigma}_{\varepsilon_2}}, \hat{\rho}_d\right)\right] \sim N\,(0,\ V) \qquad (8)$$

[10] We further discuss theoretical bounds on ρ and another way to approximate it in Appendix 1.
[11] As is the usual practice, vectors of time-invariant characteristics x_i's (kx1) are transposed into row vectors and stacked on top of each other to form the matrix X (nxk), and the vectors of error terms ε (nx1) are formed similarly from the scalars ε_i's.

where $\widehat{\Phi}_2(.) = \Phi_2\left(d_1 \frac{z_1 - \widehat{\beta}_1' x_{ij}}{\widehat{\sigma}_{\varepsilon_1}}, d_2 \frac{z_2 - \widehat{\beta}_2' x_{ij}}{\widehat{\sigma}_{\varepsilon_2}}, \hat{\rho}_d\right)$ *is the estimated quantities of poverty dynamics for household i.*

The covariance-variance matrix V can be decomposed into two components, one due to sampling errors and the other due to model errors assuming these two errors are uncorrelated such that $V = \Sigma_s + \Sigma_m$.

Proof

See Appendix 1.

Several remarks are in order for this proposition. First, given a better fit for our regressions in Equation (1), the model-based variances (i.e., synthetic panel estimates in our case) are usually smaller than the design-based variances (i.e., weighted estimates based on panel data) (Matloff, 1981; Binder and Roberts, 2009). Furthermore, a larger sample size would reduce the sampling variance; thus, this points to the advantages of cross sections over panel data when the former have larger sample sizes than the latter (see Appendix 3 for more discussion). While the reduction of variance can vary depending on the specific model or datasets under consideration (Binder and Roberts, 2009), our estimation results (Table 3) show that the model-based variances for the synthetic panels can hover around 10-50 percent of those for the design-based variances for different countries.[12]

Second, we can use data either from the first or the second survey round as the base year for Proposition 3, given the following identity

$$P(y_{i1} \leq z_1 \; and \; y_{i2} > z_2) \equiv P(y_{i2} > z_2 \; and \; y_{i1} \leq z_1) \qquad (9)$$

We provide next Proposition 4 that further extends Proposition 3 to settings with more than two consumption groups.

[12] Our results are consistent with the findings in Binder and Roberts (2009), where the largest reduction in variances appear to depend on other factors and not just sample size differences. In particular, the reduction in variances for the synthetic panels are rather similar for Lao PDR and Peru, despite the ratio of the sample size for the cross sections over that of the actual panel is four times and 1.6 times for Peru and Lao PDR respectively.

Proposition 4- Asymptotic results for point estimates for mobility between different groups for two periods

Given the same assumptions in Proposition 3, let P^{lm} represent household i's (i=1,..., N) probability of moving from consumption group l in period 1 to consumption group m in period 2, that is $P^{lm} = P(z_1^{l-1} < y_{i1} \leq z_1^{l} \text{ and } z_2^{m-1} < y_{i2} \leq z_2^{m})$, where l, m= 1,...,k, and the z_j are the thresholds that separate the different consumption groups, with $z_j^0 = -\infty$ and $z_j^k = \infty$, for period j, j= 1, 2. D*efining $F^{l,m}$ as $\Phi_2\left(\frac{z_1^l - \beta_1' x_{ij}}{\sigma_{\varepsilon_1}}, \frac{z_2^m - \beta_2' x_{ij}}{\sigma_{\varepsilon_2}}, \rho\right)$, and the (^) sign represent the estimate, our point estimates are distributed as*

$$\sqrt{n}\left[P^{lm} - (\hat{F}^{l,m} - \hat{F}^{l,(m-1)} - \hat{F}^{(l-1),m} + \hat{F}^{(l-1),(m-1)})\right] \sim N\,(0,\ V) \qquad (10)$$

Proof

See Appendix 1.

We provide in Appendix 1 several additional theoretical results. These include Corollary 3.1 (which provides the asymptotic results for conditional probabilities) and Proposition 5 (which extends Proposition 3 to the general setting where there are three or more survey rounds, i.e., $j \geq 3$).

III. Monte Carlo Simulation

We first start in this section with assuming that both Assumptions 1 and 2 are satisfied before examining situations where these assumptions can be relaxed. Assume that household *i*'s consumption can be generated for both periods using the following model

$$y_{i1} = \alpha_1 + \beta_{11}x_{i1} + \beta_{12}x_{i2} + \beta_{13}x_{i3} + \beta_{14}x_{i4} + \beta_{15}x_{i5} + \beta_{16}x_{i6} + \beta_{17}x_{i7} + \beta_{18}x_{i8} + v_{i1} \quad (11)$$
$$y_{i2} = \alpha_2 + \beta_{21}x_{i1} + \beta_{22}x_{i2} + \beta_{23}x_{i3} + \beta_{24}x_{i4} + \beta_{25}x_{i5} + \beta_{26}x_{i6} + \beta_{27}x_{i7} + \beta_{28}x_{i8} + v_{i2} \quad (12)$$

where the x_i's are household head's time-invariant characteristics, and v_i's the random error terms. We choose eight regressors for Equations (11) and (12) to better mimic situations where we can employ up to seven time-invariant regressors when working with real household survey data (Appendix 3, Table 3.1).

Also assume the following parameter values

$\alpha_1 = 1\,;\ \ \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = \beta_{15} = \beta_{16} = \beta_{17} = \beta_{18} = 1$
$\alpha_2 = 1.5\,;\ \ \beta_{21} = 1.2, \beta_{22} = 1.1, \beta_{23} = 1.05, \beta_{24} = 1.3, \beta_{25} = 0.9, \beta_{26} = 1.15, \beta_{27} = 1.4, \beta_{28} = 0.6$
and

$x_{i1} \sim N\,(0,\ 2.5), x_{i2} \sim N\,(0,\ 5), x_{i3} \sim N\,(0,\ 6), x_{i4} \sim N\,(0,\ 4), x_{i5} \sim N\,(0,\ 1), x_{i6} \sim N\,(0,\ 3),$
$x_{i7} \sim N\,(0,\ 2), x_{i8} \sim N\,(0,\ 1)$

$$\begin{pmatrix} v_{i1} \\ v_{i2} \end{pmatrix} \sim BVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 6.5 & 1 \\ 1 & 6.5 \end{pmatrix} \right)$$

where N(0, c) stands for the normal distribution with mean 0 and variance *c*; *BVN(., .)* similarly represents the bivariate normal distribution with the vector of mean 0 and the given variance-covariance matrix. Without loss of generality, we assume a certain degree of correlation over time for the error terms *v*'s (ρ=0.15), which may be caused by time-varying factors such as unexpected shocks. Given these parameter values, we can calculate that $var(y_{i1})$= 27, $var(y_{i2})$= 38.6, $\rho_{y_{i1}y_{i2}}$= 0.89, as well as a range of values for R^2 with each equation.

The values for β's are motivated by the estimates for Equation (1) using real household survey (Appendix 3, Table 3.1). For example, the ratios of the estimated coefficients between the two repeated cross sections range from 0.60 to 1.11 for Vietnam during 2006-2008 and 0.73 to 1.39 for the US during 2007-2009.

We choose the value of 0.15 as a lower value of ρ (say, rather than 0) for two main reasons. First, in theory, it tends to be positive as earlier discussed with Assumption 2. Indeed, a zero correlation coefficient implies perfect income mobility between two periods (i.e., an average household's income in the second period has no relationship with its income in the first period), which rarely occurs perhaps except under extremely special circumstances such as overnight regime change. Second, empirical evidence using actual panel data from various countries suggest that ρ is often (much) larger than this value. For example, Khor and Pencavel (2006) estimate ρ to be 0.54 for China, and range from 0.62 to 0.78 for various richer countries such as Denmark, France, Germany, Italy, Sweden, the UK, and the US in the late 1980s and early 1990s. Estimates by Dang *et al.* (2014) put ρ at 0.39 for Nepal (1995/96- 2003/04) and 0.50 for Indonesia (1997-

2000). Our estimation results (Table 1) suggest ρ ranges from 0.43 to 0.70 for countries with different income levels, such as Bosnia-Herzegovina, Lao PDR, the United States, Peru, and Vietnam during the 2000s. We discuss in more detail the Monte Carlo simulation procedures in Appendix 2.

We examine three main different data situations. These range from the most data-scarce situation where we only observe x_1 (i.e., ρ= 0.88 and is almost identical to $\rho_{y_{i1}y_{i2}}$) to a typical setting with a few such variables (i.e., ρ=0.58), and to an unusual setting where we fully observe all the x's (i.e., ρ=0.15). These data situations correspond to Models 1, 5, and 8 in Table 2.1, which also provides the values for ρ for additional data situations. We provide simulation results for these models at three different sample sizes N= 1,000 (small), 4,000 (medium), and 10,000 (large), with 1,000 simulations for each model run. We fix the poverty line in period 2 at the 30th percentile, and then graph in Figure 2 the true percentage of households that are poor in both periods (solid line), its 95 percent CIs (shaded bands), and the estimated percentage using simulated data (dashed line) against the whole spectrum of poverty rates in the first period.

Figure 1 shows that estimated poverty rates closely track the true rates and fall within their 95 percent CIs. Unsurprisingly, more time-invariant variables result in better predictions. Indeed, the dashed lines are almost indistinguishable from the solid line for the graphs where ρ = 0.15 or even where ρ = 0.58. When very limited information exists on time-invariant variables (ρ = 0.88, or both the R^2 equal 0.09), estimates (partially) fall outside the 95 percent confidence interval for the middle part of the distribution for mid-sized or unusually large sample sizes (N= 4,000 or 10,000), but still compares favorably well to true poverty rates for small sample sizes (N= 1,000). Varying the model parameters or the poverty lines gives us similar results (not shown).

The results remain robust where we relax Assumptions 1 and 2 in various ways. These include situations where the time-invariant household characteristics x's have different distributions or are correlated with each other, or where ρ may vary for different population groups, or the errors are heteroskedastic errors (Appendix 2).

IV. Data

To validate our method with real survey data, we analyze household panel survey data from Bosnia-Herzegovina (Bosnia-Herzegovina Living Standards Measurement Survey, BLSMS) in 2001-2004, Lao PDR (Expenditure and Consumption Survey, LECS) in 2002/03-2007/08, the United States (Panel Study of Income Dynamics, PSID) in 2005, 2007, and 2009, Peru (Peruvian National Household Survey, ENAHO) in 2004, 2005, and 2006, and Vietnam (Vietnam Household Living Standards Survey, VHLSS) in 2004, 2006, and 2008. The number of households comprises 2,376 households for Bosnia-Herzegovina, 6,500 households for the LECS, 9,189 households for each round of the VHLSSs, more than 5,000 households for the PSIDs, and almost 20,000 households for the ENAHOs. These data are of high quality and are typically employed to produce estimates of poverty and income for these countries. We discuss further details of these datasets in Appendix 3.

Consistent with the literature on pseudo-panel data, we restrict the household heads' age range to 25-55 for the first survey round and adjust this appropriately for later survey rounds to ensure stable household formation (e.g., looking at the age cohort 27-57 if the next survey round is two years later). While this age range can be extended to include older people, it may be ill-advised to include those who are younger, at least since most household heads tend to be older than 25 in all the countries we look at. The time-invariant variables that we use include the household head's age, years of schooling, ethnicity (i.e., whether belonging to ethnic majority groups), and whether

the household resides in urban areas.[13] We provide the estimated parameters for Equation (1) in Appendix 3, Table 3.1.

V. Empirical Validation for Poverty Dynamics

Our Monte Carlo simulation suggests that the proposed method works reasonably well even when our assumptions are not fully satisfied (Section III and Appendix 2). We examine in this section how well the method performs with real household survey data.

V.1. Testing Assumptions and Estimates for ρ

Regarding testing for Assumptions 1 and 2 using the real household survey data, since Assumption 1 is automatically satisfied for true panel data, we only test it for the cross-sectional components with Lao PDR, Peru, and Vietnam. The t-tests for the null hypothesis that the distributions of the time-invariant variables are the same across survey rounds are not rejected at the 5 percent level for the latest period for Peru and Vietnam, but not for Lao PDR. While this suggests that Lao PDR may not offer the best data for validation purposes, we still show validation results for this country since these differences may not be practically very large (e.g., half a year of schooling between the two rounds).[14] Assumption 2 is not testable for the cross sections, but can be tested using the actual panels. Formal multivariate normality tests, including the Doornik-Hansen (2008) test, reject the hypothesis of univariate or bivariate normality distribution for all the countries. Nevertheless, plotting the estimated error terms (ε_{ij}) for both the cross sections and the panel data against the normal distribution (see Appendix 3, Figure 3.1 for Lao PDR, Peru, and

[13] In contexts where there is (much) migration, the urban residence dummy variable may not satisfy Assumption 1. We return to testing this assumption with real household survey data in Section V.1.

[14] Assumption 1 is also satisfied for Vietnam in 2004-06, and mostly satisfied for Peru in 2004-05 except for heads' years of schooling and urban residence. However, similar to Lao PDR, these differences appear not very large (e.g., a difference of 0.2 years of schooling between two rounds). Our earlier Monte Carlo simulation results suggest that this assumption can be violated to some extent.

Vietnam only to save space), suggests that these approximate the latter fairly closely in practice for each year.[15]

We next discuss the estimates for ρ. After obtaining an estimate for $\rho_{y_{i1}y_{i2}}$ from the synthetic panels based on age cohorts (using Lemma 1)—which are all highly statistically significant with p-values less than 0.01—we provide the synthetic panels estimates for ρ (using Proposition 1) in Table 1, column Method 1. Estimates using the synthetic panels deviate from those using the actual panels from 0.02 (the US during 2005-2009) to 0.16 (Vietnam during 2004-2008) in absolute terms, corresponding to a range of 4 to 28 percent in relative terms. This is within the range of $\pm$30%, where our Monte Carlo simulation (Appendix 2) indicates that estimates remain robust. Furthermore, estimates for ρ are smaller than those for $\rho_{y_{i1}y_{i2}}$, which is consistent with our earlier theoretical discussion.[16]

Alternatively, we also estimate ρ using Proposition 2. Assumption 3 for the cohort fixed effects is satisfied for all the countries, except for Lao PDR so Proposition 2 does not apply for this country. The estimates for ρ (column Method 2 for the synthetic panels) are somewhat better than the estimates using Method 1 for three countries: Peru, Vietnam, and the US, but are worse for Bosnia-Herzegovina. Note, however, that estimates for ρ are just an intermediate input in the estimation of poverty mobility, which is the focus of our analysis.

V.2. Overall Poverty Mobility

It can be useful to briefly examine the performance of the bound estimates first. To save space, we show in Table 2 the bound estimates for unconditional poverty dynamics using the latest two

[15] Still, as an alternative to making a parametric bivariate normal distribution as in Assumption 2, we also experiment with relaxing this assumption and employ a copula approach. Estimation results are rather similar and are further discussed in Appendix 3.

[16] Estimation results using an alternative method (Corollary 1.1 in Appendix 1) are very similar to those using Proposition 1, with the differences being at most 0.01.

survey rounds available for each country. While all the true poverty rates are reassuringly encompassed within the estimated bounds, the bound estimates are generally quite wide and can be hard to interpret. For example, the true upward and downward mobility rates for Vietnam are respectively 5.9 percent and 4.9 percent. Yet, the bound estimates for both upward mobility and downward mobility for this country are almost identical at [0.5, 9.8] and [0.6, 9.9].[17] This points further to the value of seeking improvements on the bound estimates.

We show the point estimates for the same countries and periods in Table 3, using data in the second survey round (x_{i2}) as the base year for predictions. To evaluate the goodness-of-fit for estimation results, we show comparison with the 95% CIs and one standard error around estimates based on the panels. We also consider the efficiency of the synthetic panel estimates by looking at the proportion of the overlap between the 95 percent CIs of the synthetic panel estimates and the true estimates over the 95 percent CI of the synthetic panel estimates. The larger this overlap, the more efficient the synthetic panel estimates are; for instance, an overlap of 100 percent indicates that the 95 percent CI of the synthetic panel estimates falls well within that of the true estimates. We show both the averaged proportions of the overlap (or mean coverage) for all the dynamics calculations and the number of times that the overlap reaches 100 percent.

Results appear very encouraging with the synthetic panel point estimates being close to the true point estimates and lying within the 95 percent CIs around the true estimates for all the cases (i.e., 20 out of 20). Furthermore, more than half of the synthetic panel point estimates fall within one standard error of the actual panel estimates (i.e., 11 out of 20). For the efficiency tests, the mean coverage ranges from 83 percent to 100 percent and there is 100 percent overlap for more than four fifths (i.e., 17 out of 20) of the cases.

[17] We provide the bound estimates for the conditional mobility rates in Appendix 3, Table 3.2. These bounds form even wider intervals than those shown in Table 2.

In addition, for the US and Bosnia-Herzegovina where the sample size is the same for both actual panel and synthetic panel estimates, the standard errors for the latter are smaller than those for the former, which is consistent with our earlier discussion. We discuss in Appendix 3 various robustness checks including using data in the first survey round as the base year, or data in earlier survey rounds (x_{i1}), or bootstrap standard errors, or using a copula approach as an alternative to assuming a bivariate normal distribution.[18]

V.3. Further Extensions

We further extend the proposed method to provide estimates for population sub-groups; it is important to do so for at least two reasons. First, policy makers are usually interested in focusing on smaller population groups rather than the whole population in designing social safety net programs; and second, synthetic panels usually have larger sample sizes than panel data, which can help improve estimate accuracy. We plot the estimated rates with their 95 percent CIs for the absolute measures of poverty dynamics against the true rates for the population categorized by ethnicity (i.e., ethnic minority groups), gender of household heads (i.e., female-headed households), education achievement (i.e., primary education or higher, lower secondary education or higher), and residence areas (i.e., urban households or regions the household live in) for Peru in Figure 2. Not surprisingly, the 95 percent CIs for synthetic panels estimates are much smaller than those for the true rates with the gaps between the standard errors amplified roughly twice (i.e., multiplied by 1.96). Our estimates appear to be reasonably good, and fall within the 95 percent CIs for the true rates around half of the times for the immobile; the corresponding figure is three-fourths or more for the mobile.

[18] We provide the point estimates for the conditional mobility rates in Appendix 3, Table 3.3. Estimation results are, unsurprisingly, slightly less accurate than those in Table 3 since both the numerators and denominators in the ratios in Corollary 3.1 are estimated.

We next show in Table 4 the estimated consumption quintile transition matrix using data from Vietnam in 2006-2008, where the actual and synthetic panel estimates are shown in panel A and panel B respectively. Estimates are off with some of the row and column totals (which sum up to 20 percent by definition), but we focus on the inner transitions since the former do not offer as much insight into mobility as the latter.[19] Estimation results are, again, rather encouraging with the majority (i.e., four-fifths) of the inner transitions falling within the 95 percent CIs of the true estimates, which are presented in bold. These estimates also pass the 100 percent mark of the coverage test. Other useful statistics that can be calculated from Table 4, panel B, include the percentages of the population that have seen either an improvement or a decline or remained in the same quintile over time, which are respectively 24.7 percent, 27.3 percent, and 48 percent. These estimates are within the 95 CIs around those based on the actual panels. Furthermore, some of the remaining estimates that fall just outside these 95 percent CIs around the true estimates appear practically close to the latter (e.g., the transition from quintile 3 to quintile 4 or from the richest quintile to quintile 2). We further discuss the poverty mobility estimates for three periods in Appendix 3.

VI. Discussion and Conclusion

Panel data currently are still unavailable in a large majority of developing countries, and this situation may exist for quite some time. In the absence of panel data, our proposed method offers a means to construct synthetic panels that allow study of poverty and welfare dynamics. While our estimates are not perfect, Monte Carlo simulations and analysis using real household survey data

[19] The row or column totals should sum up to 20 percent by definition and serve mostly as an indicator of prediction accuracy for these totals only. In addition, it may be useful to highlight the fact that our validation is predicated on the assumption that the true panel data for Vietnam have good quality. If the mobility in the true panel data is partly caused by spurious changes due to measurement errors (or attrition bias) in household consumption, our estimates based on the synthetic panel data would be more accurate since cross sections are free of such data issues.

indicates that they perform reasonably well under various deviations from the model assumptions. Moreover, synthetic panels are constructed from cross sections, which are not affected by issues specific to actual panels such as attrition and measurement errors.

Our proposed method need not be restricted only to the analysis of poverty transition, and may be further applied to other dynamics analysis, such as labor transitions or health consumption. In fact, there have recently been promising extensions of our method to other topics such as intergenerational mobility (see, e.g., Foster and Rothbaum, 2015), shared prosperity (Dang and Lanjouw, 2016), or more extensive analysis of welfare dynamics along the whole income distribution (see, e.g., Bourguignon *et al.*, 2019).

However, one should attempt to check the underlying assumptions before constructing synthetic panels. In particular, the explanatory power of the income model and sensitivity to cohort definition, including the two proposed methods to estimate ρ, should be investigated. Since our proposed estimates for the correlation coefficients are based on practical approximation (as well as asymptotic theory), they may be biased in surveys with small sample sizes. Extra care should be taken to validate estimation results wherever possible (say, by using older panel data for the same country) before producing new estimates.

References

Alkire, Sabina and James E. Foster. (2011). "Counting and Multidimensional Poverty Measurement." *Journal of Public Economics*, 95: 476-487.

Beegle, Kathleen, Luc Christiaensen, Andrew Dabalen, and Isis Gaddis. (2016). *Poverty in a Rising Africa*. Washington, DC: The World Bank.

Binder, David A. and Georgia Roberts. (2009). "Design- and Model-Based Inference for Model Parameters". In D. Pfeffermann and C.R. Rao. *Handbook of Statistics, Vol. 29B- Sample Surveys: Inference and Analysis*. North-Holland: Elsevier.

Blundell, Richard, Alan Duncan, and Costas Meghir. (1988). "Estimating Labor Supply Responses Using Tax Reforms". *Econometrica*, 66(4): 827- 861.

Bourguignon, Francois, Chor-Ching Goh, and Dae Il Kim. (2004). "Estimating Individual Vulnerability to Poverty with Pseudo-Panel Data", *World Bank Policy Research Working Paper No. 3375*. Washington DC: The World Bank.

Bourguignon, Francois, Hector Moreno, and Hai-Anh Dang. (2019). "On the Construction of Synthetic Panels". Working paper. Paris School of Economics.

Calvo, César and Stefan Dercon. (2009). "Chronic Poverty and All That: The Measurement of Poverty Over Time". In Tony Addison, David Hulme, and Ravi Kanbur. (Eds.) *Poverty Dynamics: Interdisciplinary Perspectives*. Oxford University Press: New York.

Chaudhuri, S. and M. Ravallion, (1994). "How Well Do Static Indicators Identify the Chronically Poor?" *Journal of Public Economics*, 53, 367-394.

Colgan, Brian. (2022). "EU-SILC and the potential for synthetic panel estimates." *Empirical Economics*. Doi: https://doi.org/10.1007/s00181-022-02277-7

Collado, M.D. (1997). "Estimating Dynamic Models from Time Series of Independent Cross-Sections". *Journal of Econometrics*, 82, 37–62.

Cross, Philip J. and Charles F. Manski. (2002). "Regressions, Short and Long". *Econometrica*, 70(1): 357-368.

Dang, Hai-Anh and Peter Lanjouw. (2013). "Measuring Poverty Dynamics with Synthetic Panels Based on Cross-Sections". World Bank Policy Research Working Paper # 6504. Washington DC: The World Bank.

---. (2016). "Toward a New Definition of Shared Prosperity: A Dynamic Perspective from Three Countries". In Kaushik Basu and Joseph Stiglitz. (Eds). *Inequality and Growth: Patterns and Policy*. Palgrave MacMillan Press.

Dang, Hai-Anh, Dean Jolliffe, and Calogero Carletto. (2019). "Data Gaps, Data Incomparability, and Data Imputation: A Review of Poverty Measurement Methods for Data-Scarce Environments". *Journal of Economic Surveys*, 33(3): 757-797.

Dang, Hai-Anh, Peter Lanjouw, Jill Luoto, and David McKenzie. (2014). "Using Repeated Cross-Sections to Explore Movements in and out of Poverty". *Journal of Development Economics,* 107: 112-128.

De Luca, Giuseppe, Jan R. Magnus, and Franco Peracchi. (2018). "Balanced variable addition in linear models." *Journal of Economic Surveys*, 32(4): 1183-1200.

Deaton, Angus. (1985). "Panel Data from Time Series of Cross-Sections". *Journal of Econometrics*, 30: 109- 126.

Devereux, Paul J. (2007). "Small-Sample Bias in Synthetic Cohort Models of Labor Supply". *Journal of Applied Econometrics*, 22: 839-848.

Khan, Rumman. (2021). "Assessing Sampling Error in Pseudo-Panel Models". *Oxford Bulletin of Economics and Statistics*, 83(3): 742-769.

Doornik, Jurgen A. and Henrik Hansen. (2008). "An Omnibus Test for Univariate and Multivariate Normality". *Oxford Bulletin of Economics and Statistics*, 70, 927–939.

Elbers, Chris, Jean O. Lanjouw, and Peter Lanjouw. (2003). "Micro-level Estimation of Poverty and Inequality". *Econometrica*, 71(1): 355-364.

Ferreira, Francisco H. G., Julian Messina, Jamele Rigolini, Luis-Felipe López-Calva, Luis Felipe López-Calva, and Renos Vakis. (2012). *Economic Mobility and the Rise of the Latin American Middle Class.* Washington DC: World Bank.

Fields, Gary S. (2001). *Distribution and development: a new look at the developing world*. Canbridge: MIT Press.

Foster, James E. (2009). "A Class of Chronic Poverty Measures". In Tony Addison, David Hulme, and Ravi Kanbur. (Eds.) *Poverty Dynamics: Interdisciplinary Perspectives*. Oxford University Press: New York.

Foster, James E. and Jonathan Rothbaum. (2015). "Using Synthetic Panels to Estimate Intergenerational Mobility". *Working paper No. 013/2015*. Espinosa Yglesias Research Centre.

Garces-Urzainqui, David. (2017). Poverty Transitions Without Panel Data? An Appraisal of Synthetic Panel Methods. *Paper presented at the 7th Meeting of the Society for the Study of Economic Inequality*, New York city.

Gibson, John (2001) "Measuring Chronic Poverty without a Panel". *Journal of Development Economics,* 65(2): 243-266.

Glewwe, Paul and Hanan Jacoby. (2000). "Recommendations for Collecting Panel Data". In Margaret Grosh and Paul Glewwe. (Eds). *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Measurement Study*. Washington DC: The World Bank.

Güell, Maia and Luojia Hu. (2006). "Estimating the Probability of Leaving Unemployment Using Uncompleted Spells from Repeated Cross-Section Data". *Journal of Econometrics*, 133: 307–341.

Herault, Nicolas and Stephen Jenkins. (2019). "How Valid are Synthetic Panel Estimates of Poverty Dynamics?" *Journal of Economic Inequality,* 17(1): 51-76.

Inoue, Atsushi. (2008). "Efficient Estimation and Inference in Linear Pseudo-Panel Data Models". *Journal of Econometrics*, 142: 449- 466.

Jenkins, Stephen. P. (2011). *Changing Fortunes: Income Mobility and Poverty Dynamics in Britain.* Oxford: Oxford University Press.

Juodis, Artūras. (2018). "Pseudo Panel Data Models with Cohort Interactive Effects". *Journal of Business and Economic Statistics*, 36(1): 47-61.

Kalton, Graham. (2009). "Designs for Surveys over Time". In D. Pfeffermann and C.R. Rao. *Handbook of Statistics, Vol. 29A- Sample Surveys: Design, Methods and Applications.* North-Holland: Elsevier.

Khor, Niny and John Pencavel. (2006). "Income Mobility of Individuals in China and the United States." *Economics of Transition*, 14(3): 417-458.

Kopczuk, Wojciech, Emmanuel Saez, and Jae Song. (2010). "Earnings inequality and mobility in the United States: evidence from social security data since 1937." *Quarterly Journal of Economics*, 125(1): 91-128.

Lee, Nayoung, Geert Ridder, and John Strauss. (2017). "Estimation of Poverty Transition Matrices with Noisy Data". *Journal of Applied Econometrics,* 32: 37–55.

Little, Roderick J. A. and Donald B. Rubin. (2020). *Statistical Analysis with Missing Data*. 3rd Edition. New Jersey: Wiley.

Matloff, Norman S. (1981). "Use of Regression Functions for Improved Estimation of Means". *Biometrika*, 68(3): 685-689.

Moffitt, Robert. (1993). "Identification and Estimation of Dynamic Models with a Time Series of Repeated Cross- Sections". *Journal of Econometrics*, 59: 99-123.

Nakagawa, Shinichi, Paul CD Johnson, and Holger Schielzeth. (2017). "The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded." *Journal of the Royal Society Interface*, 14(134): 20170213.

OECD. (2018). A Broken Social Elevator? How to Promote Social Mobility. Paris: OECD Publishing. https://doi.org/10.1787/9789264301085-en

Pencavel, John. (2007). "A Life Cycle Perspective on Changes in Earnings Inequality among Married Men and Women". *Review of Economics and Statistics*, 88(2): 232-242.

Piketty, Thomas. (2014). *Capital in the Twenty-First Century*. USA: Belknap Press.

Propper, Carol, Hedley Rees, and Katherine Green. (2001). "The Demand for Private Medical Insurance in the UK: A Cohort Analysis". *Economic Journal*, 111: C180-C200.

Reeves, Richard. (2020). "Biden should restore the Office of Economic Opportunity abolished by Reagan". https://www.brookings.edu/opinions/biden-should-restore-the-office-of-economic-opportunity-abolished-by-reagan/

Ridder, Geert and Robert Moffitt. (2007). "The Econometrics of Data Combination". In Heckman and Leamer. (Eds). *Handbook of Econometrics*, Volume 6B. Elservier: the Netherlands.

Salvucci, Vincenzo, and Finn Tarp. (2021). "Poverty and vulnerability in Mozambique: An analysis of dynamics and correlates in light of the Covid-19 crisis using synthetic panels." *Review of development economics*, 25(4): 1895-1918.

Serajuddin, Umar, Hiroki Uematsu, Christina Wieser, Nobuo Yoshida, and Andrew Dabalen. (2015). "Data deprivation: another deprivation to end." *World Bank Policy Research Paper no. 7252*, World Bank, Washington, DC.

Snijders, Tom AB, and Roel J. Bosker. (1994). "Modeled variance in two-level models." *Sociological methods & research*, 22(3): 342-363.

---. (2011). *Multilevel analysis: an introduction to basic and advanced multilevel modelling*. Sage Publications, London.

Stiglitz, Joseph E. (2013). *The Price of Inequality- How Today's Divied Society Endangers Our Future*. New York: W. W. Norton & Company.

United Nations Development Programme (UNDP). (2016). *Multidimensional Progress: Well-being beyond Income*. New York: United Nations Development Programme.

Verbeek, Marno. (2008). "Synthetic Panels and Repeated Cross-sections", pp.369-383 in L. Matyas and P. Sevestre (eds.) *The Econometrics of Panel Data*. Berlin: Springer-Verlag.

Verbeek, Marno and T. Nijman. (1992). “Can Cohort Data Be Treated as Genuine Panel Data?” *Empirical Economics*, 17: 9- 23.

Verbeek, Marno and T. Nijman. (1993). “Minimum MSE Estimation of a Regression Model with Fixed Effects from a Series of Cross-Sections”. *Journal of Econometrics*, 59, 125–13.

Verbeek, Marno and Francis Vella. (2005). “Estimating Dynamic Models from Repeated Cross-sections”. *Journal of Econometrics*, 127, 83-102.

World Bank. (2017). *Monitoring Global Poverty: Report of the Commission on Global Poverty*. Washington, DC: The World Bank.

Table 1: Estimated ρ from Actual Panels and Synthetic Panels for Different Countries

Country	Survey Year	Actual panels		Synthetic panels		
				Method 1		Method 2
		$\rho_{y_{i1}y_{i2}}$	ρ	$\rho_{y_{i1}y_{i2}}$	ρ	ρ
Bosnia-Herzegovina	2001 2004	0.48	0.45	0.43	0.40	0.61
Lao PDR	2002-03 2007-08	0.51	0.43	0.56	0.46	N/A
Peru	2004 2005	0.82	0.64	0.82	0.69	0.67
	2005 2006	0.82	0.66	0.80	0.63	0.68
	2004 2006	0.79	0.63	0.73	0.51	0.68
Vietnam	2004 2006	0.81	0.66	0.85	0.73	0.61
	2006 2008	0.78	0.63	0.85	0.76	0.62
	2004 2008	0.75	0.58	0.84	0.74	0.47
United States	2005 2007	0.76	0.66	0.89	0.84	0.72
	2007 2009	0.82	0.70	0.86	0.79	0.74
	2005 2009	0.72	0.57	0.71	0.59	0.56

Note: The synthetic panel estimates are based on cross sectional data except for Bosnia-Herzegovina and the US, where these estimates are based on two rounds of actual panel data. $\rho_{y_{i1}y_{i2}}$ is the simple correlation across two survey rounds for household consumption for all countries except for the US, where it is the correlation for household income. ρ is the partial correlation, conditional on household head's gender, years of schooling, ethnicity, and residence areas. All estimates for $\rho_{y_{i1}y_{i2}}$ and ρ are significant at the 0.01 level. Household heads' ages are restricted to between 25 and 55 in the first survey round and adjusted accordingly for the second survey round.

Table 2: Estimated Bounds on Poverty Dynamics Based on Synthetic Data for Two Periods, Joint Probabilities (Percentage)

Poverty Status	Bosnia- Herzegovina		Lao PDR		Peru		United States		Vietnam	
First Period & Second Period	2001- 2004		2002/03- 2007/08		2005-06		2007-09		2006-08	
	Actual panel	Synthetic panel	Actual panel	Synthetic panel	Actual Panel	Synthetic Panel	Actual Panel	Synthetic Panel	Actual Panel	Synthetic Panel
Poor, Poor	10.3	[4.7, 18.0]	13.8	[8.5, 24.1]	29.9	[23.2, 40.7]	6.0	[2.5, 8.8]	9.9	[4.7, 14.0]
	(1.7)		(1.2)		(1.3)		(0.4)		(0.8)	
Poor, Nonpoor	12.6	[2.8, 16.1]	14.3	[2.3, 17.9]	11.6	[2.5, 20.0]	3.8	[0.6, 6.9]	5.9	[0.5, 9.8]
	(1.2)		(1.1)		(0.9)		(0.3)		(0.5)	
Nonpoor, Poor	10.5	[2.2, 15.6]	10.9	[0.5, 16.1]	8.9	[0.2, 17.7]	4.6	[1.4, 7.7]	4.9	[0.6, 9.9]
	(1.4)		(1.0)		(0.8)		(0.4)		(0.5)	
Nonpoor, Nonpoor	66.5	[63.7, 77.0]	61.0	[57.5, 73.1]	49.7	[39.1, 56.6]	85.7	[82.9, 89.2]	79.3	[75.5, 84.8]
	(2.2)		(1.6)		(1.6)		(0.6)		(1.0)	
N	1342	1342	1989	3215	2250	9084	3368	3368	2723	3701

Note: Synthetic panels are constructed from cross sections for Lao PDR, Peru, and Vietnam and from panel halves for Bosnia-Herzegovina and the US. Predictions are obtained using the estimated parameters from the first and second survey rounds on data in the second survey round. The estimated bounds are shown in brackets under the "Synthetic Panel" for each country. All numbers are weighted using household weights for Peru, and population weights for other countries. Poverty rates are in percent. Household heads' ages are restricted to between 25 and 55 for the first survey round and adjusted accordingly with the year difference for the second survey round.

Table 3: Poverty Dynamics Based on Synthetic Panel Data for Two Periods, Joint Probabilities (Percentage)

Poverty Status	Bosnia- Herzegovina		Lao PDR		Peru		United States		Vietnam	
First Period & Second Period	2001- 2004		2002/03- 2007/08		2005-06		2007-09		2006-08	
	Actual panel	Synthetic panel	Actual panel	Synthetic panel	Actual Panel	Synthetic Panel	Actual Panel	Synthetic Panel	Actual Panel	Synthetic Panel
Poor, Poor	10.3	8.2	13.8	13.2	29.9	30.9	6.0	6.2	9.9	9.6
	(1.7)	(0.2)	(1.2)	(0.4)	(1.3)	(0.4)	(0.4)	(0.2)	(0.8)	(0.3)
Poor, Nonpoor	12.6	12.6	14.3	13.2	11.6	12.3	3.8	3.2	5.9	4.9
	(1.2)	(0.3)	(1.1)	(0.1)	(0.9)	(0.1)	(0.3)	(0.1)	(0.5)	(0.1)
Nonpoor, Poor	10.5	12.1	10.9	11.4	8.9	10.0	4.6	4.0	4.9	5.0
	(1.4)	(0.2)	(1.0)	(0.2)	(0.8)	(0.1)	(0.4)	(0.1)	(0.5)	(0.1)
Nonpoor, Nonpoor	66.5	67.2	61.0	62.2	49.7	46.8	85.7	86.6	79.3	80.4
	(2.2)	(0.6)	(1.6)	(0.6)	(1.6)	(0.4)	(0.6)	(0.3)	(1.0)	(0.4)
Goodness-of-fit Tests										
Within 95% CI	4/4		4/4		4/4		4/4		4/4	
Within 1 standard error	2/4		4/4		2/4		1/4		2/4	
Mean coverage (percent)	100		100		91.6		83.0		100	
Coverage of 100%	4/4		4/4		3/4		2/4		4/4	
N	1342	1342	1989	3215	2250	9084	3368	3368	2722	3701

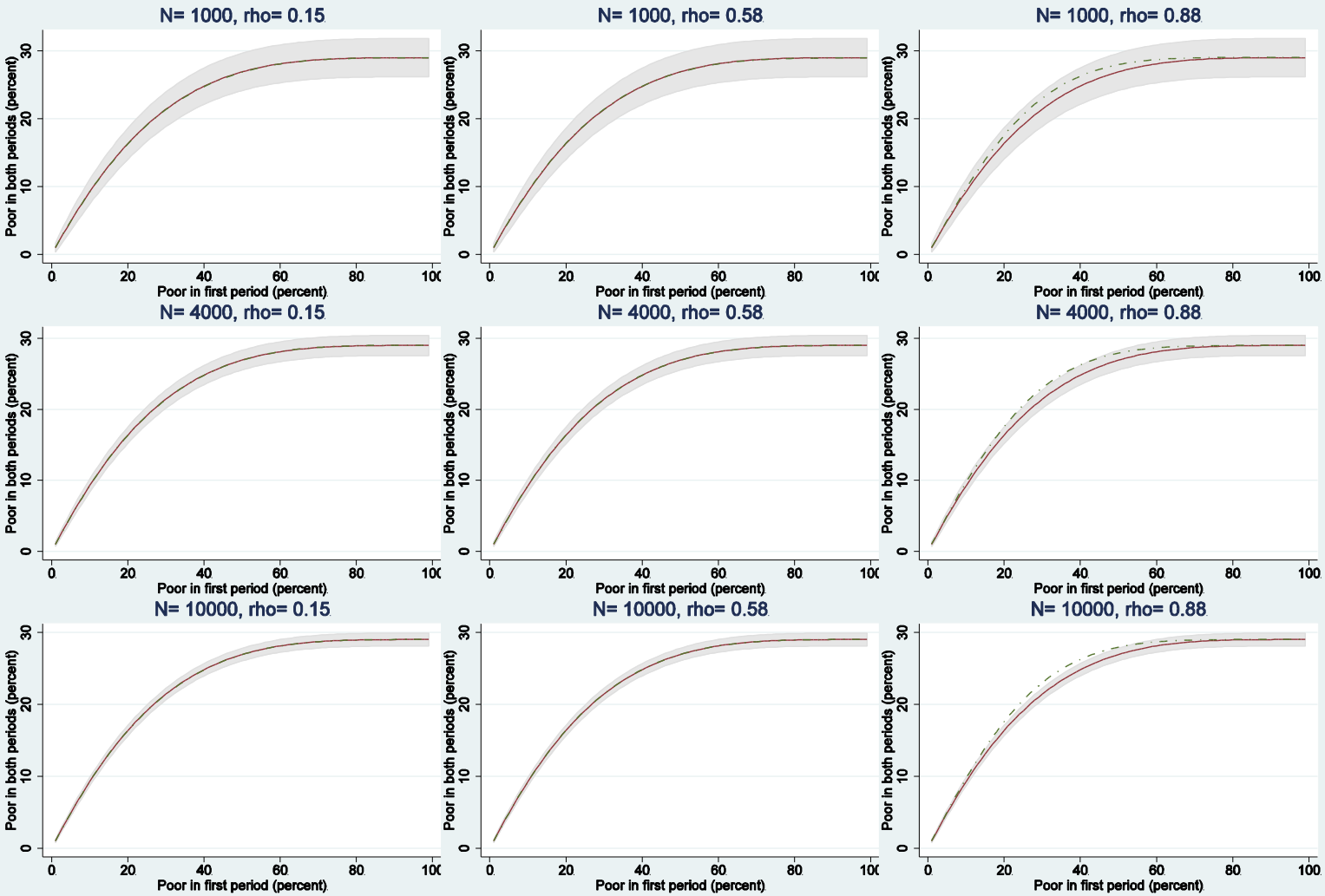
Note: Synthetic panels are constructed from cross sections for Lao PDR, Peru, and Vietnam and from panel halves for Bosnia-Herzegovina and the US. Predictions are obtained using the estimated parameters from the first and second survey rounds on data in the second survey round. Standard errors are obtained adjusting for complex survey design for all countries, except for the US PSID. All numbers are weighted using household weights for Peru, and population weights for other countries. Poverty rates are in percent. Household heads' ages are restricted to between 25 and 55 for the first survey round and adjusted accordingly with the year difference for the second survey round. The "Within 95% CI" row shows the number of times that the estimates based on the synthetic panels fall within the 95% confidence interval (CI) of the estimates based on the actual panels; the "Within 1 standard error" row shows a similar figure but using one standard error around the estimates based on the actual panels. The "Mean coverage (percent)" row shows the mean proportion of the 95% CI around the synthetic panel estimates that overlap with those based on the actual panels; the "Coverage of 100%" row shows a similar figure for the number of times that the former fall completely inside the latter.

Table 4: Consumption Dynamics for Two Periods, Vietnam 2006-2008 (Percentage)

			2008					
			Poorest	**Quintile 2**	**Quintile 3**	**Quintile 4**	**Richest**	**Total**
Panel A: True Panels	**2006**	**Poorest**	12.7	4.7	1.7	0.6	0.2	19.7
			(0.8)	(0.4)	(0.3)	(0.2)	(0.1)	(0.9)
		Quintile 2	4.8	7.5	4.6	2.0	0.6	19.6
			(0.4)	(0.6)	(0.5)	(0.3)	(0.1)	(0.9)
		Quintile 3	1.8	5.2	6.9	4.6	1.5	20.0
			(0.3)	(0.5)	(0.5)	(0.5)	(0.2)	(0.9)
		Quintile 4	0.6	2.0	5.0	7.8	4.8	20.2
			(0.2)	(0.3)	(0.5)	(0.6)	(0.5)	(0.9)
		Richest	0.1	0.6	1.8	4.9	12.9	20.5
			(0.1)	(0.2)	(0.3)	(0.5)	(0.7)	(0.8)
		Total	20.0	20.0	20.0	20.0	20.0	100
			(1.0)	(0.9)	(0.9)	(0.9)	(0.9)	
			2008					
			Poorest	**Quintile 2**	**Quintile 3**	**Quintile 4**	**Richest**	**Total**
Panel B: Synthetic Panels	**2006**	**Poorest**	**13.7**	3.6	**1.6**	**0.4**	**0.0**	19.2
			(0.3)	(0.0)	(0.0)	(0.0)	(0.0)	(0.3)
		Quintile 2	**5.6**	5.4	**4.5**	**2.2**	0.3	17.8
			(0.1)	(0.0)	(0.0)	(0.0)	(0.0)	(0.1)
		Quintile 3	**2.3**	**4.5**	**6.4**	5.6	**1.5**	20.4
			(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.1)
		Quintile 4	**0.6**	**2.1**	**5.1**	**8.5**	**5.2**	21.4
			(0.0)	(0.0)	(0.0)	(0.1)	(0.0)	(0.1)
		Richest	**0.0**	0.3	**1.4**	**5.4**	**14.0**	21.1
			(0.0)	(0.0)	(0.0)	(0.0)	(0.2)	(0.2)
		Total	22.2	15.8	18.9	22.2	20.9	100
			(0.3)	(0.1)	(0.1)	(0.1)	(0.2)	

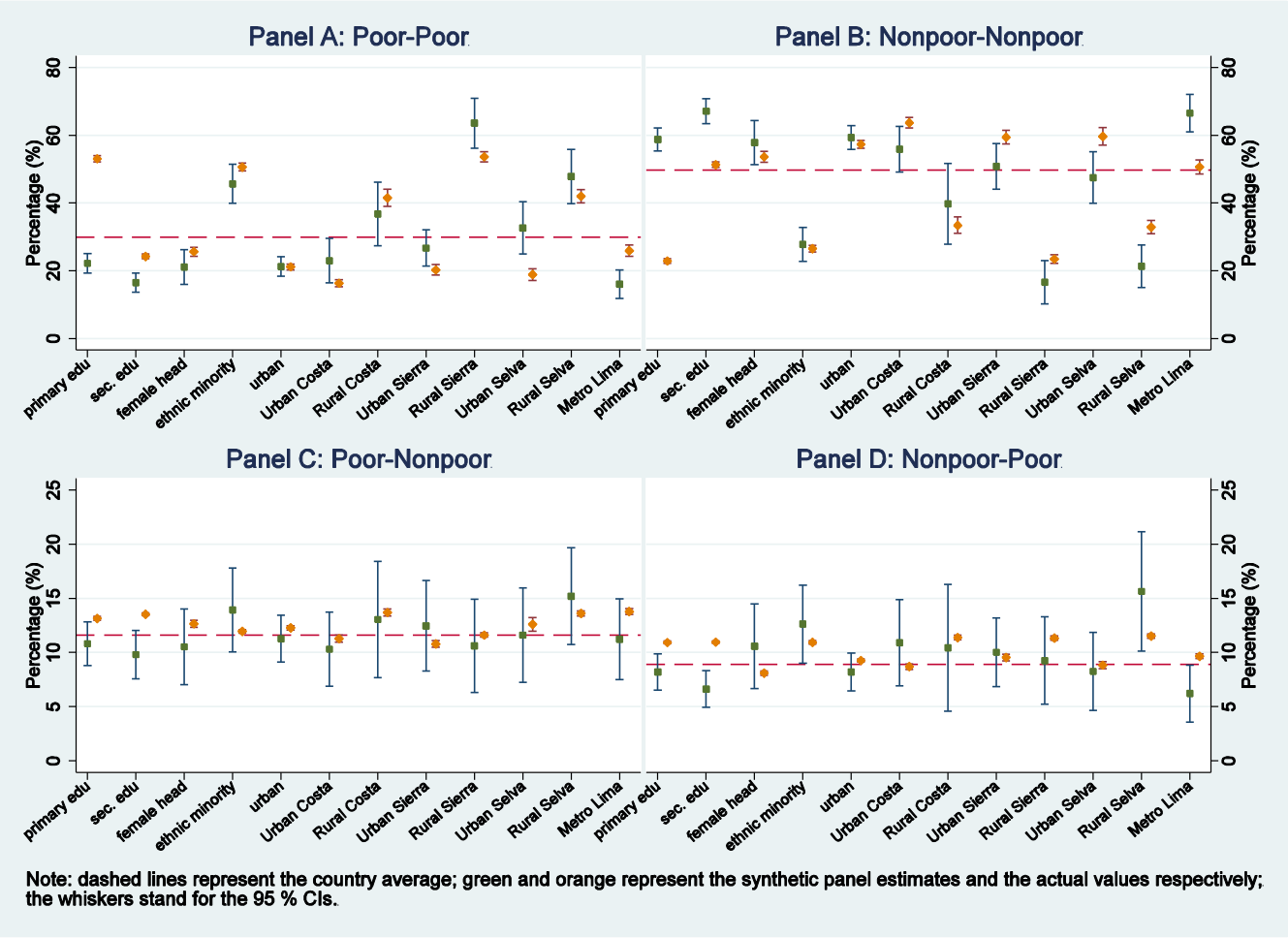
Note: Synthetic panels are constructed from cross sections for Vietnam. Predictions are obtained using the estimated parameters from the first and second survey rounds on data in the second survey round. Standard errors are obtained adjusting for complex survey design. All numbers are weighted using population weights. Poverty rates are in percent. Household heads' ages are restricted to between 25 and 55 for the first survey round and adjusted accordingly with the year difference for the second survey round. Joint probabilities are shown. Estimates based on the synthetic panels that fall within the 95% CI of those based on the actual panels are shown in bold.

Figure 1: Predicted Poverty Rates vs. True Poverty Rates for Two Periods Based on Simulated Data



Note: the true percentage of households that are poor in both periods and its 95 percent CIs are respectively represented by the solid line and shaded bands; the estimated percentage using simulated data is represented by the dashed line.

Figure 2: Profiles of Poverty Mobility, Peru 2005- 2006



Note: dashed lines represent the country average; green and orange represent the synthetic panel estimates and the actual values respectively; the whiskers stand for the 95 % CIs.

Appendixes for online publication

Appendix 1: Additional Theoretical Results and Proofs

Proof of Proposition 1- Point estimate of ρ

If panel data were available, the simple correlation coefficient for household consumption between the two survey rounds would be

$$\rho_{y_{i1}y_{i2}} = \frac{cov(y_{i1},y_{i2})}{\sqrt{var(y_{i1})\,var(y_{i2})}} = \frac{cov(\beta_1{}'x_{i1}+\varepsilon_{i1},\beta_2{}'x_{i2}+\varepsilon_{i2})}{\sqrt{var(y_{i1})\,var(y_{i2})}} = \frac{\beta_1{}'var(x_i)\beta_2+\rho\sqrt{\sigma_{\varepsilon_1}^2\sigma_{\varepsilon_2}^2}}{\sqrt{var(y_{i1})\,var(y_{i2})}} \quad (1.1)$$

where the second equality follows from Equation (1). The third equality follows from replacing time-invariant household characteristics x_{i1} and x_{i2} with x_i, given Assumption 1 that the underlying population being sampled in survey rounds 1 and 2 are the same, and rewriting the covariance of the error terms using Assumption 2. Solving for ρ from the above equality, we have

$$\rho = \frac{\rho_{y_{i1}y_{i2}}\sqrt{var(y_{i1})\,var(y_{i2})}-\beta_1{}'var(x_i)\beta_2}{\sigma_{\varepsilon_1}\sigma_{\varepsilon_2}} \quad (1.2)$$

Corollary 1.1- A convenient approximation of ρ

Let R_j^2, for j= 1, 2, represent the coefficients of determination obtained from estimating Equations (1). If β_1 and β_2 are close in values such that they approximate one another (i.e., $\beta_1 \approx \beta_2$), the partial correlation coefficientρcan also be approximated by

$$\rho \approx \frac{\rho_{y_{i1}y_{i2}}-\sqrt{R_1^2R_2^2}}{\sqrt{1-R_1^2}\sqrt{1-R_2^2}}$$

Proof

See Appendix 1.

Several remarks are in order. While the proposed method of estimating ρ given in (5) directly follows from our framework provided by Equation (1) and provides more accurate results, Corollary 1.1 provides a shorthand calculation for ρ, as well as some intuition into the process. It suggests that, if the estimated parameters in Equation (1) are close to each other, the partial correlation coefficient for household consumption can be interpreted as the simple correlation coefficient purged of (the geometric mean of) its multiple correlation with household (time-invariant) characteristics in the two survey rounds, and then reweighted by (the geometric mean of) the shares of the unexplained predicted errors. Our validation exercise, to be discussed below, shows that in practice these two formulae yield very similar estimates for ρ.[20] Furthermore, it is

[20] Another way, still, to estimate ρ is using the recursion formula for partial correlation coefficients provided by Anderson (2003, p.41); however, this formula requires many more calculations than the formulae provided above, so we do not discuss it further here. We abuse the notations $var(x_i)$ and $var(y_i)$ to refer to both the population true quantities and their sample estimates to keep the expressions simpler. Similarly, we subsequently use N to refer to both the total population and the sample survey. Since the variance-covariance matrix of the time-invariant household characteristics $var(x_i)$ in Equation (1.2) is the same for each round of true panel data, but can vary for the cross sectional surveys, it may be useful to separately try the variance from each survey round to see if there is any difference in poverty estimates. In our empirical estimates, discussed below, these variance-covariance matrices are very similar between survey rounds and result in almost identical estimation results.

straightforward to show from Proposition 1 that the simple (unconditional) correlation coefficient $\rho_{y_{i1}y_{i2}}$ can provide an upper value for ρ.[21]

Proof of Corollary 1.1- A convenient approximation of ρ

Rewriting equality (5), we have

$$\rho = \frac{\rho_{y_{i1}y_{i2}}\sqrt{var(y_{i1})\,var(y_{i2})} - \beta_1{}'var(x_i)\beta_2}{\sigma_{\varepsilon_1}\sigma_{\varepsilon_2}} = \frac{\rho_{y_{i1}y_{i2}} - \sqrt{\frac{\beta_1{}'var(x_i)\beta_2\beta_1{}'var(x_i)\beta_2}{var(y_{i1})\,var(y_{i2})}}}{\sqrt{1-R_1^2}\sqrt{1-R_2^2}}$$
$$\approx \frac{\rho_{y_{i1}y_{i2}} - \sqrt{\frac{\beta_1{}'var(x_i)\beta_1\beta_2{}'var(x_i)\beta_2}{var(y_{i1})\,var(y_{i2})}}}{\sqrt{1-R_1^2}\sqrt{1-R_2^2}} = \frac{\rho_{y_{i1}y_{i2}} - \sqrt{R_1^2R_2^2}}{\sqrt{1-R_1^2}\sqrt{1-R_2^2}} \quad (1.3)$$

where the second equality follows from dividing both the numerator and the denominator by $\sqrt{var(y_{i1})\,var(y_{i2})}$ and using the definition of R^2 as the proportion of the total variation in y that is accounted for by the variation in the regressors (or equivalently, the proportion of the total variation in y that is not accounted for by the variation in the regressors is obtained by subtracting R^2 from 1). The third equality follows from replacing $\beta_2\beta_1'$ with $\beta_1\beta_2'$, given that $\beta_1 \approx \beta_2$. The last equality follows again from the definition for R^2.[22]

Strictly speaking, we can make the less restrictive assumption that $\beta_1\beta_2' \approx \beta_2\beta_1'$ instead of assuming $\beta_1 \approx \beta_2$ for Corollary 1.1, but we use the latter for convenience. Note that for the three-variable case, the two formulae in (1.2) and (1.3) are identical.

Lemma 1- Approximation of $\rho_{y_{i1}y_{i2}}$

Our discussion for the validity of Lemma 1 here simply serves as a straightforward application of the established results in the literature on pseudo-panel data to a context with only two rounds of cross section. Consider the following simple linear projection for household consumption in period 2 on household consumption in period 1 (that have no control variables)

$$y_{i2} = \delta y_{i1} + \eta_{i2} \quad (1.4)$$

where y_{it} is household i's consumption in period t, t= 1, 2, and η_{i2} is a random error term. Similar to Assumption 2, Equation (1.4) represents a simplification of the typical linear dynamic model employed in the pseudo-panel literature where there are no additional control variables (x_{ij}) in Equation (1.4) (see, e.g., Moffitt (1993, pp. 100)). Note that combined with Equation (1), we can also rewrite Equation (1.4) as

$$y_{i2} = \delta\beta_1'x_{i1} + \delta\varepsilon_{i1} + \eta_{i2} \quad (1.5)$$

The assumption that x_{ij} have no cohort-specific first moment helps ensure that δ is consistently estimable when it is linked to Equation (1) (Inoue, 2008, Theorem 1).

[21] Another lower bound for ρ can be obtained by just implementing the same procedures in Proposition 1, where the predicted error terms are obtained from estimating Equation (1) (including the age variable). But this lower bound provides a downward biased estimates and tends to be (much closer to) 0. See Dang and Lanjouw (2013) for more discussion on these bounds.

[22] An alternative proof that employs the familiar expression that links the simple and partial correlation coefficients for bivariate normal variables is provided in Dang and Lanjouw (2013).

We assume the Type 1 asymptotics of pseudo panel data (as defined by Verbeek (2008, pp. 373)), that is the sample size of each household survey round is large enough (or $N \to \infty$) and the number of cohorts (C) constructed from the survey data is fixed.

In the absence of panel data we do not observe y_{i1} for the same household, and we only have two repeated cross sections. Our objective of obtaining the simple correlation coefficient $\rho_{y_{i1}y_{i2}}$ in this case is closely related to getting a consistent estimate for δ, since by definition $\rho_{y_{i1}y_{i2}} = \frac{cov(y_{i1},y_{i2})}{\sqrt{var(y_{i1})\,var(y_{i2})}} = \sqrt{\frac{var(y_{i1})}{var(y_{i2})}}\,\delta$. A consistent estimate for δ can be obtained by instrumenting for y_{i1} with the cohort dummy variables interacted with the time dummy variables, as long as these instrumental variables are relevant and exogenous. Thus estimation of δ in Equation (1.4) is identical to estimation of δ in Equation (1.6) below, where we apply OLS to the same model where all variables are aggregated to the cohort level (see, e.g., Moffitt (1993, pp.108) or Verbeek (2008, pp. 373))

$$y_{c2} = \delta' y_{c1} + \eta_{c2} \tag{1.6}$$

Consequently, we can consistently estimate $\rho_{y_{i1}y_{i2}}$ in a similar way as

$$\rho_{y_{i1}y_{i2}} = \rho_{y_{c1}y_{c2}} = \frac{cov(y_{c1},y_{c2})}{\sqrt{var(y_{c1})\,var(y_{c2})}} \tag{1.7}$$

More generally, for any two survey periods *j* and *k*, we can obtain the simple correlation coefficient $\rho_{y_{ij}y_{ik}}$ as

$$\rho_{y_{ij}y_{ik}} = \rho_{y_{cj}y_{ck}} = \frac{cov(y_{cj},y_{ck})}{\sqrt{var(y_{cj})\,var(y_{ck})}} \tag{1.8}$$

But note that $\rho_{y_{ij}y_{ik}}$ tends to decrease for longer time intervals between the two survey rounds. This can be seen more clearly when we write out Equation (1.4) for periods *j* and *k*

$$y_{ik} = \delta^{k-j}\beta_j' x_{ij} + \sum_{\tau=0}^{k-j-1}\delta^{\tau}\eta_{i,k-\tau} \tag{1.9}$$

Since δ is often less than 1 under most normal circumstances, δ^{k-j} becomes smaller for larger *k-j* (i.e., a longer time interval) (see Dang and Lanjouw (2017, Proposition 5) for related discussion). Using panel data from the U.S. Social Security Administration between 1937 and 2004, Kopczuk, Saez, and Song (2010) find the (rank) correlation of earnings decrease over longer time intervals. This result also holds for actual panel data from various other countries such as China, India, Peru, Vietnam, and the U.K. (Chaudhuri and Ravallion, 1994; Khor and Pencavel, 2006; Jenkins, 2011; our estimates).

Proof of Proposition 2- Alternative estimate of ρ

The proof of Proposition 2 also builds on the established results in the literature on pseudo-panel data to a context with only two rounds of cross section, and aggregate variables to the cohort level in a similar spirit to (1.5). But the key difference is that we rely on Type 2 asymptotics for Proposition 2 (instead of using Type 1 asymptotics as with Lemma 1).

Assume that the sample size of each household survey round is large enough (or $N \to \infty$) and the number of cohorts (C) constructed from the survey data is large enough (or $C \to \infty$), we aggregate all the variables in Equation (1) (instead of Equation (1.4) in Lemma 1). Writing out the error term ε_{cj} we have

$$y_{cj} = \beta_j' x_{cj} + \tau_c + \nu_{cj} \tag{2.1}$$

It follows that we can consistently estimate ρ as

$$\rho = \frac{Cov(\varepsilon_{c1},\,\varepsilon_{c2})}{\sqrt{var(\varepsilon_{c1})\,var(\varepsilon_{c2})}} = \frac{\sigma_\tau^2}{\sigma_\tau^2+\sigma_\nu^2} \tag{2.2}$$

In practice, it is rather straightforward to estimate ρ in Equation (2.2) using standard statistical softwares as earlier discussed.

Proposition 3- Asymptotic results for point estimates for 2 periods

Assume that Equation (1) and Assumptions 1 and 2 hold, and assume further that all the standard regularity conditions are satisfied for Equations (1), (i.e., $X'\varepsilon/N \xrightarrow{p} 0$ *and* $X'X/N \xrightarrow{p} M$ *finite and positive definite).*[23] *Let P be the population parameter of interest (e.g.,* $P = P(y_{i1} < z_1 \text{ and } y_{i2} > z_2)$ *for household i, i=1,..., N),* d_j *an indicator function that equals 1 if the household is poor and equals -1 if the household is non-poor in period j, j= 1, 2,* $\rho_d = d_1 d_{\ 2}\rho$, and $\rho_{y_{i1}y_{i2},d} = d_1 d_{\ 2}\rho_{y_{i1}y_{i2}}$, *and the (^) sign represent the estimate. Our point estimates are distributed as*

$$\sqrt{n}\left[P - \Phi_2\left(d_1\frac{z_1-\widehat{\beta}_1{}'x_{ij}}{\widehat{\sigma}_{\varepsilon_1}}, d_2\frac{z_2-\widehat{\beta}_2{}'x_{ij}}{\widehat{\sigma}_{\varepsilon_2}}, \hat{\rho}_d\right)\right] \sim N\,(0,\ V) \tag{3.1}$$

where $\widehat{\Phi}_2(.) = \Phi_2\left(d_1\frac{z_1-\widehat{\beta}_1{}'x_{ij}}{\widehat{\sigma}_{\varepsilon_1}}, d_2\frac{z_2-\widehat{\beta}_2{}'x_{ij}}{\widehat{\sigma}_{\varepsilon_2}}, \hat{\rho}_d\right)$ *is the estimated quantities of poverty dynamics for household i.*

The covariance-variance matrix V can be decomposed into two components, one due to sampling errors and the other due to model errors assuming these two errors are uncorrelated such that $V = \Sigma_s + \Sigma_m$.

Proof

See the proof for the general case with Proposition 5.

Proposition 4- Asymptotic results for point estimates for mobility between different groups for two periods

Given the same assumptions in Proposition 3, let P^{lm} *represent household i's (i=1,..., N) probability of moving from consumption group l in period 1 to consumption group m in period 2, that is* $P^{lm} = P(z_1^{l-1} < y_{i1} \le z_1^{l} \text{ and } z_2^{m-1} < y_{i2} \le z_2^{m})$, *where l, m= 1,...,k, and the* z_j *are the thresholds that separate the different consumption groups, with* $z_j^0 = -\infty$ *and* $z_j^k = \infty$, *for period j, j= 1, 2.* D*efining* $F^{l,m}$ *as* $\Phi_2\left(\frac{z_1^l-\beta_1{}'x_{ij}}{\sigma_{\varepsilon_1}}, \frac{z_2^m-\beta_2{}'x_{ij}}{\sigma_{\varepsilon_2}}, \rho\right)$, *and the (^) sign represent the estimate, our point estimates are distributed as*

$$\sqrt{n}\left[P^{lm} - (\hat{F}^{l,m} - \hat{F}^{l,(m-1)} - \hat{F}^{(l-1),m} + \hat{F}^{(l-1),(m-1)})\right] \sim N\,(0,\ V) \tag{4.1}$$

Proof

Given *g* consumption groups in each period, there are *gxg* transitions in total. The formulae for the standard errors for the general case can be far more complicated than those for mobility for three periods or more. Thus we suggest estimation of the standard errors by the bootstrap method.[24]

---

[23] As is the usual practice, vectors of time-invariant characteristics x_i's (kx1) are transposed into row vectors and stacked on top of each other to form the matrix *X* (nxk), and the vectors of error terms ε (nx1) are formed similarly from the scalars ε_i's.

[24] Also note that our empirical estimates, discussed later, point to little, if any, difference between the standard errors estimated using the analytical formulae offered in Proposition 3 and those using the bootstrap approach.

The consistency part of this proof is similar as that for Proposition 5 below. The second term in the square bracket in Equation (4.1) directly follows from the definition of the bivariate normal probability function. Note that $F^{l,m}= 0$, for *l, m= 1,..., k,* where either *l* or *m* equals 0, and reduces to a univariate normal probability when either *l* or *m* equals *k* (e.g., $F^{1k} = \Phi\left(\frac{z_1^1-\beta_1{'}x_{ij}}{\sigma_{\varepsilon_1}}\right)$). We provide a few examples of these special cases as a result of Proposition 4.

$$P^{15} = P(y_{i1} < z_1^1 \ and \ y_{i2} > z_2^4) = \Phi_2\left(\frac{z_1^1-\beta_1{'}x_{i2}}{\sigma_{\varepsilon_1}}, -\frac{z_2^4-\beta_2{'}x_{i2}}{\sigma_{\varepsilon_2}}, -\rho\right) \quad (4.2)$$

$$P^{35} = P(z_1^2 < y_{i1} < z_1^3 \ and \ y_{i2} > z_2^4) = \Phi_2\left(\frac{z_1^3-\beta_1{'}x_{i2}}{\sigma_{\varepsilon_1}}, -\frac{z_2^4-\beta_2{'}x_{i2}}{\sigma_{\varepsilon_2}}, -\rho\right) - \Phi_2\left(\frac{z_1^2-\beta_1{'}x_{i2}}{\sigma_{\varepsilon_1}}, -\frac{z_2^4-\beta_2{'}x_{i2}}{\sigma_{\varepsilon_2}}, -\rho\right) \quad (4.3)$$

$$P^{55} = P(y_{i1} > z_1^4 \ and \ y_{i2} > z_2^4) = \Phi_2\left(-\frac{z_1^4-\beta_1{'}x_{i2}}{\sigma_{\varepsilon_1}}, -\frac{z_2^4-\beta_2{'}x_{i2}}{\sigma_{\varepsilon_2}}, \rho\right) \quad (4.4)$$

To save space, we state and prove below Proposition 5 for the general case of the k survey rounds. Proposition 3 with the two survey rounds follows as a special case.

Proposition 5- Asymptotic results for point estimates for k periods

Assume that the same assumptions in Proposition 3 hold and extend to all k periods. Let P represent household i's (i=1,..., N) quantity of poverty dynamics (e.g., $P(y_{i1} \sim z_1 \ , y_{i2} \sim z_2, y_{i3} \sim z_3, \ldots, y_{ik} \sim z_k)$*), d_j an indicator function that equals 1 if the household is poor and equals -1 if the household is non-poor in period j, and* $d_{jl} = d_j d_l$*, our point estimates are distributed as*

$$\sqrt{n}\left[P - \Phi_k\left(d_1\frac{z_1-\hat{\beta}_1{'}x_{ij}}{\hat{\sigma}_{\varepsilon_1}}, \ldots, d_k\frac{z_k-\hat{\beta}_k{'}x_{ij}}{\hat{\sigma}_{\varepsilon_k}}, \hat{\Sigma}_\rho\right)\right] \sim N\,(0,\ V) \quad (5.1)$$

The covariance-variance matrix V can be decomposed into two components, one due to sampling errors and the other due to model errors assuming these two errors are uncorrelated such that $V = \Sigma_s + \Sigma_m$*. The first component* Σ_s *is due to the sampling errors and can be estimated using the bootstrap method, and the second component* Σ_m *is due to the model errors.*

Proof

To save space, we show the proof of the general case of the *k* survey rounds. Proposition 3 with the two survey rounds follows as a special case.

Given that household consumption can be explained by household characteristics in Equations (1) and the standard regularity conditions are satisfied, our estimator $\widehat{\Phi}_2(.)$ is a continuous and differentiable function of $\hat{\beta}_m, \hat{\sigma}_{\varepsilon_m}, \hat{\rho}_{y_{im}y_{in},d}$, for *m= 1,..., k-1*, *n=m+1,..., k*, and $j \neq m,\ n$, which are consistent estimators of the parameters. Thus $\widehat{\Phi}_2(.)$ is a consistent estimator of $\Phi_2(.)$.

We can then decompose the variance for $P - \widehat{\Phi}_k(.)$ into two parts, one due to sampling errors and the other due to model errors

$$Var(P - \widehat{\Phi}_k(.)) = Var\left((P - \Phi_k(.)) + \left(\Phi_k(.) - \widehat{\Phi}_k(.)\right)\right) = \Sigma_s + \Sigma_m \quad (5.2)$$

assuming that these two errors are uncorrelated with each other. The assumption that the model errors are uncorrelated with the sampling errors is rather standard in the statistics literature on

survey imputation (see, e.g., Rubin (1987) or Rao and Molina (2015)) or poverty mapping (see, e.g., Elbers *et al.* (2003).[25] The variance for the sampling errors Σ_s can be estimated using the bootstrap method.

Using the delta method, the variance for the model errors Σ_m can be written as

$$\sum_{m=1}^{k} \nabla_{\hat{\beta}_m}^{'} V(\hat{\beta}_m) \nabla_{\hat{\beta}_m} + \sum_{m=1}^{k} \nabla_{\hat{\sigma}_{\varepsilon_m}}^{'} V(\hat{\sigma}_{\varepsilon_m}) \nabla_{\hat{\sigma}_{\varepsilon_m}} + \sum_{m=1}^{k-1} \sum_{n=m+1}^{k} \nabla_{\hat{\rho}_{y_{im}y_{in},d}}^{'} V(\hat{\rho}_{y_{im}y_{in},d}) \nabla_{\hat{\rho}_{y_{im}y_{in},d}} \quad (5.3)$$

To make notations less cluttered, let $\beta_{(jxl)}$ represent the matrix of estimated coefficients obtained from Equations (1), $\Phi_k(.)$ the standard k-variate normal probability, and $\hat{a}_{dmj} = d_m \frac{z_m - \hat{\beta}_m{}' x_{ij}}{\hat{\sigma}_{\varepsilon_m}}$ and $\bar{\hat{a}}_{dmnj} = \frac{(\hat{\rho}_{dmq} - \hat{\rho}_{dnq}\hat{\rho}_{dmn})\hat{a}_{dmj} + (\hat{\rho}_{dnq} - \hat{\rho}_{dmq}\hat{\rho}_{dmn})\hat{a}_{dnj}}{\sqrt{1 - \hat{\rho}_{dmn}^2}}$ for *m, n, q= 1,…, k*, and $m \neq n \neq q$.

Also let $\hat{\Sigma}_{\rho(-m)}$ be the *(k-1)x(k-1)* partial correlation matrix given $\hat{\beta}_m$with the off-diagonal entries $\hat{\rho}_{dst.m} = \frac{\hat{\rho}_{dst} - \hat{\rho}_{dsm}\hat{\rho}_{dtm}}{\sqrt{1 - \hat{\rho}_{dsm}^2}\sqrt{1 - \hat{\rho}_{dtm}^2}}$for *s, t= 1,…, k* and $s, t \neq m$; similarly, let $\hat{\Sigma}_{\rho(-m,-n)}$ be the *(k-2)x(k-2)* partial correlation matrix given $\hat{\beta}_m$ and $\hat{\beta}_n$ with the off-diagonal entries$\hat{\rho}_{dst.mn} = \frac{\hat{\rho}_{dst.m} - \hat{\rho}_{dsn.m}\hat{\rho}_{dtn.m}}{\sqrt{1 - \hat{\rho}_{dsn.m}^2}\sqrt{1 - \hat{\rho}_{dtn.m}^2}}$ for *s, t= 1,…, k* and $s, t \neq m, n$. Applying the chain rule and taking the first partial derivative with regards to $\hat{\beta}_m$ and $\hat{\sigma}_{\varepsilon_m}$ (see, for example, Prekopa (1970)) and $\hat{\rho}_{y_{im}y_{in},d}$ (see, for example, Plackett (1954)), we can write out the derivatives of the terms in Equation (5.3) as follows. Note that $\varphi_2(.)$ stands for the standard bivariate normal probability density function (pdf).

$$\nabla_{\hat{\beta}_m} = d_m \left(\frac{-x_{ij}}{\hat{\sigma}_{\varepsilon_m}}\right) \varphi(\hat{a}_{dmj})\ \Phi_{k-1}\left(\frac{\hat{a}_{d1j} - \hat{\rho}_{dm1}\hat{a}_{dmj}}{\sqrt{1 - \hat{\rho}_{dm1}^2}}, \ldots, \frac{\hat{a}_{dkj} - \hat{\rho}_{dmk}\hat{a}_{dmj}}{\sqrt{1 - \hat{\rho}_{dmk}^2}}, \hat{\Sigma}_{\rho(-m)}\right) + \sum_{\substack{n=1 \\ n \neq m}}^{k} d_m d_n \frac{-var(x_{ij})\hat{\beta}_n}{\hat{\sigma}_{\varepsilon_m}\hat{\sigma}_{\varepsilon_n}}\ \varphi_2(\hat{a}_{dmj}, \hat{a}_{dnj}, \hat{\rho}_{dmn})\ \Phi_{k-2}(\hat{a}_{d1j} - \bar{\hat{a}}_{dm1j}, \ldots, \hat{a}_{dkj} - \bar{\hat{a}}_{dmkj}, \hat{\Sigma}_{\rho(-m,-n)}) \quad (5.4)$$

$$\nabla_{\hat{\sigma}_{\varepsilon_m}} = \left(\frac{-\hat{a}_{dmj}}{\hat{\sigma}_{\varepsilon_m}}\right) \varphi(\hat{a}_{dmj})\ \Phi_{k-1}\left(\frac{\hat{a}_{d1j} - \hat{\rho}_{dm1}\hat{a}_{dmj}}{\sqrt{1 - \hat{\rho}_{dm1}^2}}, \ldots, \frac{\hat{a}_{dkj} - \hat{\rho}_{dmk}\hat{a}_{dmj}}{\sqrt{1 - \hat{\rho}_{dmk}^2}}, \hat{\Sigma}_{\rho(-m)}\right) - \sum_{\substack{n=1 \\ n \neq m}}^{k} \left(d_m d_n \frac{\rho_{y_{im}y_{in}}\sqrt{var(y_{im})\, var(y_{in})} - \beta_m{}' var(x_i)\beta_n}{\hat{\sigma}_{\varepsilon_m}^2 \hat{\sigma}_{\varepsilon_n}}\right) \varphi_2(\hat{a}_{dmj}, \hat{a}_{dnj}, \hat{\rho}_{dmn}) * \Phi_{k-2}(\hat{a}_{d1j} - \bar{\hat{a}}_{dm1j}, \ldots, \hat{a}_{dkj} - \bar{\hat{a}}_{dmkj}, \hat{\Sigma}_{\rho(-m,-n)}) \quad (5.5)$$

$$\nabla_{\hat{\rho}_{y_{im}y_{in},d}} = \frac{d_m d_n}{\sqrt{1 - R_m^2}\sqrt{1 - R_n^2}} * \varphi_2(\hat{a}_{dmj}, \hat{a}_{dnj}, \hat{\rho}_{dmn}) * \Phi_{k-2}(\hat{a}_{d1j} - \bar{\hat{a}}_{dm1j}, \ldots, \hat{a}_{dkj} - \bar{\hat{a}}_{dmkj}, \hat{\Sigma}_{\rho(-m,-n)}) \quad (5.6)$$

[25] This assumption is also analogous to the standard variance decomposition formula where the unconditional variance is decomposed into the variance of the conditional expectation and the expectation of the conditional variance.

$V(\hat{\beta}_m)$ is the asymptotic covariance-variance matrix for the estimated coefficients obtained from the corresponding Equation (1). $V(\hat{\sigma}_{\varepsilon_m})$ can be approximated by $\frac{(8N-7)\hat{\sigma}^2_{\varepsilon m}}{(4N-3)^2}$ using the formula in Montgomery (2012, pp. 720) where N>25. And $V(\hat{\rho}_{y_{im}y_{in},d})$ can be estimated from Lemma 1.[26]

For the special case of two periods in Proposition 3, the formulae for the different variance terms are provided below

$$\nabla_{\hat{\beta}_m} = d_m\left(\frac{-x_{ij}}{\hat{\sigma}_{\varepsilon_m}}\right)\varphi\left(d_m\frac{z_m-\hat{\beta}_m{}'x_{ij}}{\hat{\sigma}_{\varepsilon_m}}\right)\Phi\left(\frac{d_n\frac{z_n-\hat{\beta}_n{}'x_{ij}}{\hat{\sigma}_{\varepsilon_n}}-\hat{\rho}_d d_m\frac{z_m-\hat{\beta}_m{}'x_{ij}}{\hat{\sigma}_{\varepsilon_m}}}{\sqrt{1-\hat{\rho}_d^2}}\right)$$
$$-\frac{d_m d_n\, var(x_{ij})\hat{\beta}_n}{\hat{\sigma}_{\varepsilon_m}\hat{\sigma}_{\varepsilon_n}}\varphi_2\left(d_m\frac{z_m-\hat{\beta}_m{}'x_{ij}}{\hat{\sigma}_{\varepsilon_m}}, d_n\frac{z_n-\hat{\beta}_n{}'x_{ij}}{\hat{\sigma}_{\varepsilon_n}}, \hat{\rho}_d\right) \quad (5.7)$$

$$\nabla_{\hat{\sigma}_{\varepsilon_m}} = \left(-d_m\frac{z_m-\hat{\beta}_m{}'x_{ij}}{\hat{\sigma}^2_{\varepsilon_m}}\right)\varphi\left(d_m\frac{z_m-\hat{\beta}_m{}'x_{ij}}{\hat{\sigma}_{\varepsilon_m}}\right)\Phi\left(\frac{d_n\frac{z_n-\hat{\beta}_n{}'x_{ij}}{\hat{\sigma}_{\varepsilon_n}}-\hat{\rho}_d d_m\frac{z_m-\hat{\beta}_m{}'x_{ij}}{\hat{\sigma}_{\varepsilon_m}}}{\sqrt{1-\hat{\rho}_d^2}}\right)$$
$$-\left(d_m d_n\frac{\rho_{y_{im}y_{in}}\sqrt{var(y_{im})\,var(y_{in})}-\beta_m{}'var(x_i)\beta_n}{\hat{\sigma}^2_{\varepsilon_m}\hat{\sigma}_{\varepsilon_n}}\right)\varphi_2\left(d_m\frac{z_m-\hat{\beta}_m{}'x_{ij}}{\hat{\sigma}_{\varepsilon_m}}, d_n\frac{z_n-\hat{\beta}_n{}'x_{ij}}{\hat{\sigma}_{\varepsilon_n}}, \hat{\rho}_d\right) \quad (5.8)$$

$$\nabla_{\hat{\rho}_{y_{i1}y_{i2},d}} = \frac{d_1 d_2\sqrt{var(y_{i1})\,var(y_{i2})}}{\hat{\sigma}_{\varepsilon_1}\hat{\sigma}_{\varepsilon_2}}\varphi_2\left(d_1\frac{z_1-\hat{\beta}_1{}'x_{ij}}{\hat{\sigma}_{\varepsilon_1}}, d_2\frac{z_2-\hat{\beta}_2{}'x_{ij}}{\hat{\sigma}_{\varepsilon_2}}, \hat{\rho}_d\right) \quad (5.9)$$

with $n = 3-m$. Similarly as with the general case, $V(\hat{\beta}_1)$ and $V(\hat{\beta}_2)$ being respectively the estimated asymptotic covariance-variance matrix for the estimated coefficients obtained from Equations (1), $V(\hat{\sigma}_{\varepsilon_m})$ being approximated by $\frac{(8N-7)\hat{\sigma}^2_{\varepsilon m}}{(4N-3)^2}$, and $V(\hat{\rho}_{y_{i1}y_{i2},d})$ the estimated asymptotic variance obtained from Lemma 1.

Corollary 3.1- Asymptotic results for point estimates of relative quantities of poverty dynamics for two periods

Given the same assumptions and notations in Proposition 3, let P_{i1} and $P_{i,12}$ respectively be the population parameters of interest in period j (j= 1, 2) and both periods (e.g., $P_{ij} = P(y_{ij} \leq z_j)$ and $P_{i,12} = P(y_{i1} \leq z_1 \text{ and } y_{i2} > z_2)$ for household i, i=1,..., N). And let the sample averaged estimated quantities of poverty dynamics represented by $\hat{\Phi}(.) = \Phi\left(d_j\frac{z_j-\hat{\beta}_j{}'x_{ij}}{\hat{\sigma}_{\varepsilon_j}}\right)$, our point estimates are distributed as

$$\sqrt{n}\left[\frac{P_{i,12}}{P_{ij}} - \frac{\Phi_2\left(d_1\frac{z_1-\hat{\beta}_1{}'x_{ij}}{\hat{\sigma}_{\varepsilon_1}}, d_2\frac{z_2-\hat{\beta}_2{}'x_{ij}}{\hat{\sigma}_{\varepsilon_2}}, \hat{\rho}_d\right)}{\Phi\left(d_j\frac{z_j-\hat{\beta}_j{}'x_{ij}}{\hat{\sigma}_{\varepsilon_j}}\right)}\right] \sim N(0,\ V_r) \quad (5.10)$$

The full formulae for covariance-variance matrix V_r is provided in the proof.

[26] See also Mullahy (2011) for a related derivation.

Proof

Note that since we have to estimate both the numerators and denominators (and their standard errors) in Equation (5.10), this would reduce the accuracy of our estimates compared to those for the absolute quantities of poverty dynamics provided in Proposition 3.[27]

Since $\widehat{\Phi}(.) = \Phi\left(d_j \frac{z_j - \widehat{\beta}_1 \prime x_{ij}}{\widehat{\sigma}_{\varepsilon_{ij}}}\right)$ is a consistent estimator of P_{ij} and $\widehat{\Phi}_2(.)$ is a consistent estimator of $P_{i,12}$ as discussed in the proof of Proposition 5 above, it follows that $\frac{\widehat{\Phi}_2(.)}{\widehat{\Phi}(.)}$ is a consistent estimator of $\frac{P_{i,12}}{P_{ij}}$. Then note that, since $\frac{\partial(\widehat{\Phi}_2(.)/\widehat{\Phi}(.))}{\partial \widehat{\Phi}_2(.)} = \frac{1}{\widehat{\Phi}(.)}$ and $\frac{\partial(\widehat{\Phi}_2(.)/\widehat{\Phi}(.))}{\partial \widehat{\Phi}(.)} = \frac{-\widehat{\Phi}_2(.)}{(\widehat{\Phi}(.))^2}$, using the delta method,[28] we have

$$\sqrt{n}\left[\frac{P_{i,12}}{P_{ij}} - \frac{\Phi_2\left(d_1 \frac{z_1 - \widehat{\beta}_1 \prime x_{ij}}{\widehat{\sigma}_{\varepsilon_{i1}}}, d_2 \frac{z_2 - \widehat{\beta}_2 \prime x_{ij}}{\widehat{\sigma}_{\varepsilon_{i2}}}, \widehat{\rho}_d\right)}{\Phi\left(d_j \frac{z_j - \widehat{\beta}_1 \prime x_{ij}}{\widehat{\sigma}_{\varepsilon_{ij}}}\right)}\right] \sim N\,(0,\ V_r) \tag{5.11}$$

The covariance-variance matrix V_r can be estimated as

$$V_r = \frac{1}{\left(\widehat{\Phi}(.)\right)^2} Var(\widehat{\Phi}_2(.)) + \frac{\left(\widehat{\Phi}_2(.)\right)^2}{\left(\widehat{\Phi}(.)\right)^4} Var(\widehat{\Phi}(.)) - 2\frac{\widehat{\Phi}_2(.)}{\left(\widehat{\Phi}(.)\right)^3} Cov(\widehat{\Phi}_2(.), \widehat{\Phi}(.))$$

$$= \left(\frac{\widehat{\Phi}_2(.)}{\widehat{\Phi}(.)}\right)^2 \left[\frac{Var(\widehat{\Phi}_2(.))}{(\widehat{\Phi}_2(.))^2} + \frac{Var(\widehat{\Phi}(.))}{(\widehat{\Phi}(.))^2} - 2\frac{Cov(\widehat{\Phi}_2(.), \widehat{\Phi}(.))}{\widehat{\Phi}_2(.)\widehat{\Phi}(.)}\right] \tag{5.12}$$

Similar to $Var(\widehat{\Phi}_2(.))$, $Var(\widehat{\Phi}(.))$ can be decomposed into a model error Σ_{jm} and a sampling error Σ_{js} assuming these two errors are uncorrelated.[29] The model error can be estimated as

$$\Sigma_{jm} = \nabla_{\widehat{\beta}_j}^{'} V(\hat{\beta}_j) \nabla_{\widehat{\beta}_j} + \nabla_{\widehat{\sigma}_{\varepsilon_j}}^{'} V(\hat{\sigma}_{\varepsilon_j}) \nabla_{\widehat{\sigma}_{\varepsilon_j}} \tag{5.13}$$

with $\nabla_{\widehat{\beta}_j} = d_j \left(\frac{-x_{ij}}{\widehat{\sigma}_{\varepsilon_j}}\right) \varphi\left(d_j \frac{z_j - \widehat{\beta}_j \prime x_{ij}}{\widehat{\sigma}_{\varepsilon_j}}\right)$ and $\nabla_{\widehat{\sigma}_{\varepsilon_j}} = -d_j \left(\frac{z_j - \widehat{\beta}_j \prime x_{ij}}{\widehat{\sigma}_{\varepsilon_j}^2}\right) \varphi\left(d_j \frac{z_j - \widehat{\beta}_j \prime x_{ij}}{\widehat{\sigma}_{\varepsilon_j}}\right)$.

Additional References

Anderson, Theodore W. (2003). "*An Introduction to Multivariate Statistical Analysis*". USA: John Wiley & Sons.

Casella, George and Roger L. Berger. (2002). *Statistical Inference*, 2nd Edition. California: Duxbury Press.

Dang, Hai-Anh H., and Peter F. Lanjouw. (2017). "Welfare dynamics measurement: Two definitions of a vulnerability line and their empirical application." *Review of Income and Wealth* 63(4): 633-660.

Montgomery, Douglas C. (2012). *Introduction to Statistical Quality Control*. USA: Wiley.

[27] We estimate $\Phi(.)$ using Equation (1) and its variance as discussed above, but do not use the corresponding sample-based statistics (i.e., poverty headcount ratio) to be consistent with the way we estimate $\Phi_2(.)$. If the model has good fit, $\Phi(.)$ would be very similar to the sample-based poverty headcount ratio but has much smaller variance. Another practical implication is that if we divide $\widehat{\Phi}_2(.)$ by the sample poverty rate instead of $\widehat{\Phi}(.)$, this ratio can be larger than 100 percent when we consider estimates for certain subpopulation groups.

[28] See, for example, theorem 5.5.28 in Casella and Berger (2002).

[29] Pham-Gia, Turkkan and Marchand (2006) offer an alternative expression of the density of a ratio of two normal random variables in terms of Hermite and confluent hypergeometric functions.

Mullahy, John. (2011). “Marginal Effects in Multivariate Probit and Kindred Discrete and Count Outcome Modes, ith Applications in Health Economics”. NBER Working paper 17588.
Pham-Gia, T, N. Turkkan, and E. Marchand (2006). “Density of the Ratio of Two Normal Random Variables and Applications”. *Communications in Statistics- Theory and Method*, 35(9): 1569-1591.
Plackett, R. L. (1954). “A Reduction Formula for Normal Multivariate Integrals”. *Biometrika,* 41:351-360.
Prekopa, Andras. (1970). “On Probabilistic Constrained Programming”. *In Proceedings of the Princeton Symposium on Mathematical Programming*. New Jersey: Princeton Press.
Rao, J. N. K. and Isabel Molina. (2015). *Small Area Estimation*, 2nd edition, New York: Wiley.
Rubin, Donald B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Appendix 2: Further Monte Carlo Simulation

We briefly describe the Monte Carlo simulation: i) draw the random variables x_i's and the error terms v_i's using the assumed distribution parameters for each given sample size, ii) calculate y_1 and y_2 using these values and the given parameters α's and β's, iii) estimate the "true" quantities of poverty mobility using y_1 and y_2, iv) treat y_1 and y_2 as if they came from two separate cross sections, estimate Equation (1) for each and obtain the predicted coefficients for α's and β's, v) estimate poverty mobility using the drawn values for x_i's, the predicted coefficients for α's and β's, and the corresponding ρ as in Equation (4), and vi) simulate steps 1 to 4 1,000 times, and finally take the averages for all the simulated data.

We use the typical setting where just a few variables are available (i.e., ρ=0.58) with a medium sample size in the following simulations.

2.1 Relaxing Assumption 1

Time-invariant household characteristics have different distributions

Assumption 1 requires that the time-invariant household characteristics x in the two survey rounds have the same distributions. As discussed earlier, this assumption may be violated if the underlying population changes over time for various reasons (e.g., the population's education achievement may increase over time). We can test this assumption by letting the distribution of some time-invariant characteristics change between over time. In particular, we keep the distribution of x_{i3} fixed in the first period (i.e., $x_{i31} \sim N(0, 6)$), and vary its distribution in the second period as $x_{i32} = x_{i31} + k * \sqrt{6}$, where k ranges from -0.12 to 0.12 in an incremental step of 0.3. Put differently, we let the distribution of x_{i32} change from that of x_{i31} by as much as 12 percent of the square root of its variance. Estimation results shown in Figure 2.1 below suggest that the estimated poverty rates fall within the true rates' 95 percent CIs, except for the extreme cases where k= -0.12 or -0.9. Even in these cases, only some of the estimated poverty rates fall outside the true rates' 95 percent CIs.

Correlated variables

Another interesting case that can happen in practice is that the variables x's can be correlated with each other, and we may only have data on some, but not all, of the x's. For example, age can be (positively) correlated with education achievement, and we may only have data on age but not education in the survey. We examine this situation by drawing x_{i1} and x_{i2} jointly from a bivariate normal distribution (with their same mean 0 and variances as before), but we now let their correlation range from 0.14 to 0.85 in an incremental step of 0.14. Figure 2.2 below suggests that, when x_{i2} is omitted, the stronger the correlation between x_{i1} and x_{i2}, the estimated poverty rates are, unsurprisingly, more likely to fall inside the true rates' 95 percent CIs.

2.2. Relaxing Assumption 2

Heterogeneity (or mis-measurement) of ρ

The simulation above relies on the true value for ρ. However, ρ may be heterogenous and change for different population groups (i.e., more educated households may likely exit poverty or have more poverty mobility over time). Furthermore, both our approximations for ρ are based on asymptotic theory, and given our reliance, in practical applications, on Lemma 1 to approximate

$\rho_{y_{i1}y_{i2}}$ with the synthetic panel cohort-level simple correlation coefficient $\rho_{y_{c1}y_{c2}}$, we may not be able to precisely estimate this parameter in practice. Would this heterogeneity (or mis-measurement) of ρ significantly affect estimation results?

Fixing the true value of ρ at 0.58, we then vary its value in each increment of 10 percent, within a range of [-40%, +40%], to mimic situations where this parameter is misestimated. We plot in Figure 2.2 the estimated poverty rates using these incorrect values for ρ against the true poverty rates for the mid-sized sample (N= 4,000). While estimates are, unsurprisingly, more inaccurate as ρ deviates more from its true value, estimates fall inside the 95 percent confidence interval of the true rates when ρ is misestimated within the range [-20%, 20%]. Estimates still remain largely inside the 95 percent confidence interval of the true rates when the measurement error increases to $\pm$30%, and only fall outside the 95 percent confidence interval for the middle part of the distribution when the measurement error increases to $\pm$40%. Overall, these simulation results indicate that our method has a reasonable performance under a theoretical setting where ρ can be mismeasured up to a certain extent (i.e., $\pm$30%). This range compares reasonably well with panel data. As an example, we estimate ρ for different population groups for three sets of panel data for Peru during 2004 and 2006 (Table 2.2), the difference in ρ for a given population group and for the whole population varies from -28 percent to 10 percent.

Heteroskedastic errors

Closely related to the situation above is one where the error terms are heteroskedastic, or vary with the values of (some of) the x_i's. We examine one form of heteroskedasticity where we add to the error term in each period an additional component $k * x_{i3}$, and we let k vary from 0.1 to 0.9 in an incremental step of 0.1. Estimation results shown in Figure 2.3 below suggest that they still regularly fall inside the true rates' 95 percent CIs for these cases.

Table 2.1: Model Parameters for Simulation

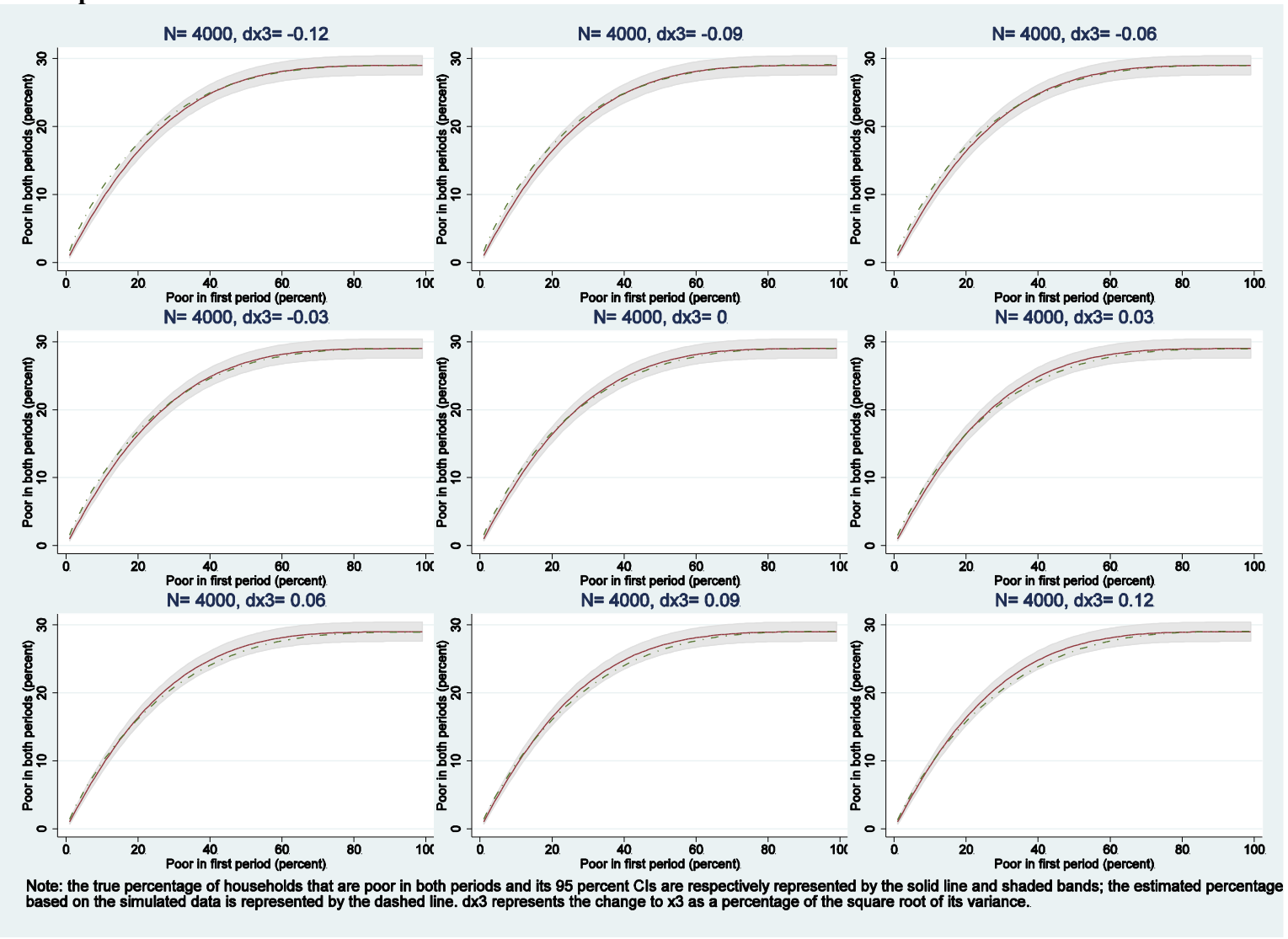
Model	X	$V(\varepsilon_1)$	R_1^2	$V(\varepsilon_2)$	R_2^2	$\boldsymbol{\rho}$
1	1	24.5	0.09	35.0	0.09	0.88
2	1, 2	19.5	0.28	28.9	0.25	0.85
3	1, 2, 3	17.5	0.35	22.3	0.42	0.71
4	1, 2, 3, 4	13.5	0.50	15.6	0.60	0.60
5	1, 2, 3, 4, 5	12.5	0.54	14.7	0.62	0.58
6	1, 2, 3, 4, 5, 6	9.5	0.65	10.8	0.72	0.43
7	1, 2, 3, 4, 5, 6, 7	7.5	0.72	6.9	0.82	0.22
8	1, 2, 3, 4, 5, 6, 7, 8	6.5	0.76	6.5	0.83	0.15

Note: The x_i's variables are added sequentially and cumulatively to each simulation model. For example, Model 1 includes only the intercept and x_{i1}, Model 2 includes the intercept, x_{i1} and x_{i2}, and so on. The error terms v_1 and v_2 are assumed to follow a standard bivariate normal distribution where $var(v_j)$= 6.5, for j= 1, 2, and $cov(v_1, v_2)$= 1. The vectors of coefficients are b_1= (1,1,1,1,1,1,1,1) and b_2= (1.2,1.1,1.05,1.3,0.9,1.15,1.4,0.6). The x_i's are assumed to follow normal distributions where the means are 0 and the variances are respectively (2.5,5,6,4,1,3,2,1) for x_{ik}, k= 1,…, 8.

Table 2.2: Estimated ρ from Actual Panels for Different Population Groups for Peru, 2004-2006

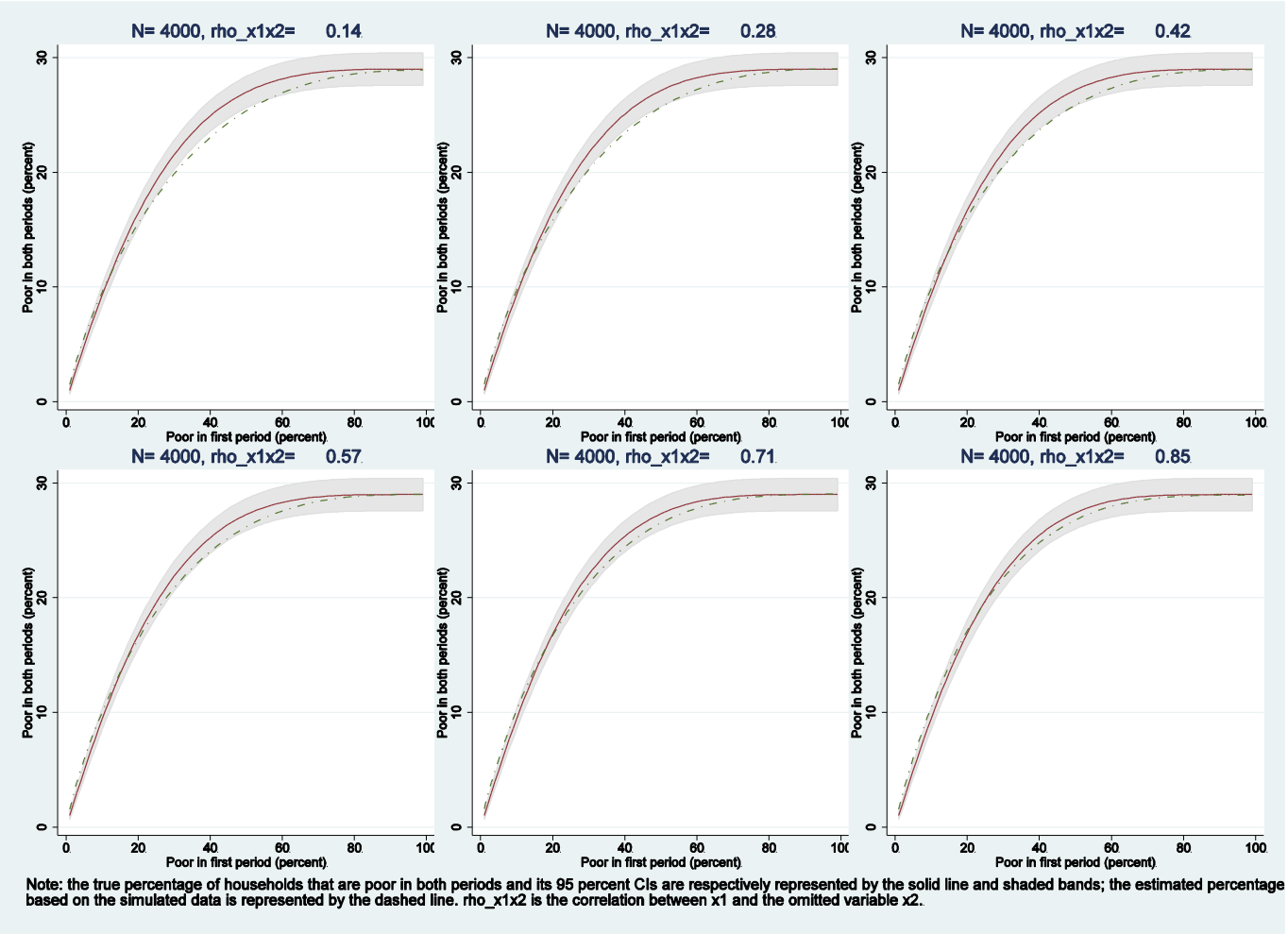
	2004-2005	2005-2006	2004-2006
Primary education or higher	0.68	0.70	0.65
Secondary education or higher	0.69	0.73	0.67
Ethnic minorities	0.59	0.56	0.66
Female head	0.46	0.63	0.57
Urban	0.67	0.69	0.64
Urban Costa	0.66	0.72	0.69
Rural Costa	0.56	0.71	0.56
Urban Sierra	0.66	0.64	0.57
Rural Sierra	0.58	0.50	0.68
Urban Selva	0.65	0.69	0.64
Rural Selva	0.59	0.59	0.57
Metro Lima	0.66	0.66	0.59
All population	**0.64**	**0.66**	**0.63**

Figure 2.1: Predicted Poverty Rates vs. True Poverty Rates for Two Periods Based on Simulated Data, with Relaxing Assumption 1



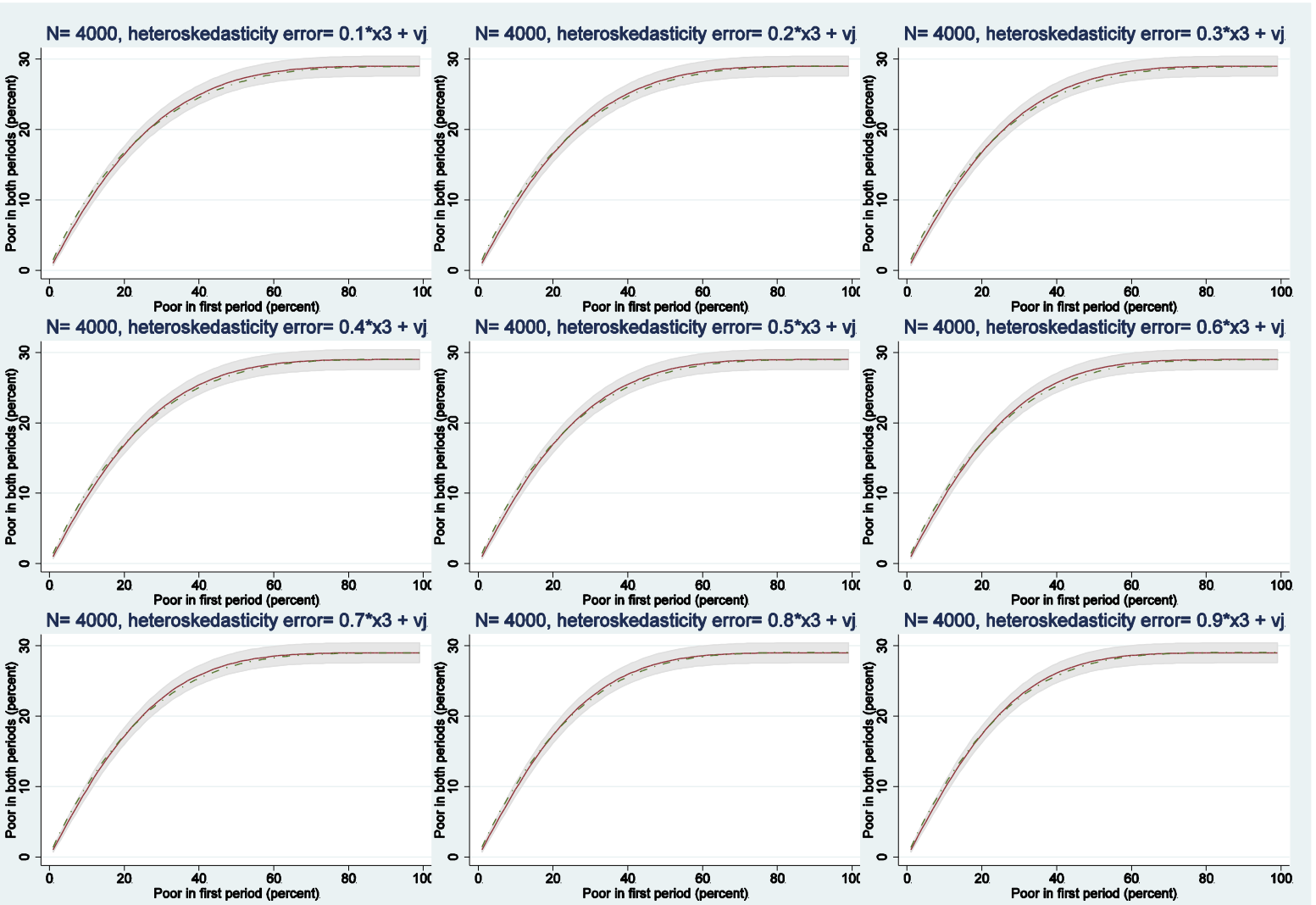
Note: the true percentage of households that are poor in both periods and its 95 percent CIs are respectively represented by the solid line and shaded bands; the estimated percentage based on the simulated data is represented by the dashed line. dx3 represents the change to x3 as a percentage of the square root of its variance.

Figure 2.2: Predicted Poverty Rates vs. True Poverty Rates for Two Periods Based on Simulated Data, with Omitted and Correlated Variables



Note: the true percentage of households that are poor in both periods and its 95 percent CIs are respectively represented by the solid line and shaded bands; the estimated percentage based on the simulated data is represented by the dashed line. rho_x1x2 is the correlation between x1 and the omitted variable x2.

Figure 2.3: Predicted Poverty Rates vs. True Poverty Rates for Two Periods Based on Simulated Data, with Heteroskedasticity Errors



Note: the true percentage of households that are poor in both periods and its 95 percent CIs are respectively represented by the solid line and shaded bands; the estimated percentage based on the simulated data is represented by the dashed line.

Appendix 3: Data Description, Robustness Checks and Extensions

Except for the PSID that is implemented by the University of Michigan, all the other surveys are nationally representative surveys implemented by each country's statistical agency, with previous or current technical assistance from international organizations (the World Bank with Peru and Vietnam), leading universities (University of Essex with Bosnia-Herzegovina) or statistical agencies in richer countries (Statistics Sweden with Lao PDR). Also except for the PSID, all the other surveys are similar to the LSMS-type (Living Standards Measurement Survey) surveys supported by the World Bank, and provide detailed information on household consumption and demographics, as well as schooling, health, employment, migration, and housing. The PSID has a more complex structure and provides similarly detailed, if not richer, information. All these surveys are widely used in academic studies (especially the PSID) as well as in poverty assessments undertaken by the respective governments, the policy-research community. We use the official poverty lines for Lao PDR, Peru, and Vietnam; for the USA, we use the poverty lines provided in the PSID data (which adjust for family size and demographics); for Bosnia-Herzegovina we use the 20^{th} percentile of the consumption distribution in 2001 as the poverty line. We use income, and household consumption, as a household welfare measure respectively for the US, and all the other countries.[30]

One particular feature with the LECSs, VHLSSs and ENAHOs is a rotating panel design, which collects panel data for a subset of each survey round between two adjacent years. Around one third and one half of the households in the first round are repeated in the next round for the LECs and VHLSSs respectively, and the corresponding repetition ratio for the ENAHOs is around one quarter. This combination of both cross-sectional data and panel data in one survey provides an appropriate setting for us to implement our procedures on the cross section components, and then validate our estimates against the true rates from the panel components for each country. For the BLSMSs and the PSIDs, there is no rotating panel design and so we use the panel halves, pretending that these are cross sectional data. To ensure comparability between estimates based on the panel and cross section components, we use household weights with our estimates for the ENAHOs and population weights for the remaining surveys.[31]

There are pros and cons with using panel data versus a rotating panel survey for validation purposes. On one hand, actual panels may offer smaller sample sizes and may not appear as "real" as rotating panel surveys that offer both panel and cross-sectional data. But on the other hand, actual panels by definition satisfy Assumption 1 on the same distribution of the time-invariant household characteristics in the two survey rounds. Thus validation exercises using rotating panel surveys, compared to actual panels, would additionally require that the data from the cross-section

[30] We build the data for the BSLM using the data from Demirguc-Kunt, Klapper and Panos (2011). For the PSID, we only consider the sample persons with non-zero longitudinal weights. The PSID also has some information on household consumption but this measure is not commonly used to measure poverty and is much less comprehensive than those for other surveys. Appendix 3 in Dang and Lanjouw (2013) provides a more detailed description of these surveys and other data quality checks.

[31] For example, the household-weighted headcount poverty rates based on the (actual) panel component for Peru are around 5 percent lower than those based on the cross-section component (and the population-weighted estimates are even more different). The Peruvian data are thus not perfect for validation purposes, but we believe it is still useful to show estimates for this country using household weights.

component be similar to those from the panel component. We employ a mix of actual panels and rotating panel surveys in our validation exercises to provide more robustness checks.

We would like to emphasize the importance of good-quality panel data for validation purposes with synthetic panels. In the worst-case scenario, if the actual panel data are of low quality (e.g., heavily affected by attrition issue), they can themselves provide biased estimates of the quantities of interest. In another (somewhat better) scenario, the actual panel data may not provide biased estimates, but can offer imprecise estimates with large CIs. This would in turn weaken both our goodness-of-fit measures for validation purposes (i.e., through higher probabilities of containing both the synthetic panel point estimates and CIs). Note that the actual panel data sets we analyze in this paper are generally recognized to be of reasonable quality as discussed in Section V. See also Dorfman (2011) for a related test for the bias of the point estimate in the context of small area estimation.

An alternative to assuming a bivariate normal distribution is to use a copula approach (see, e.g., Trivedi and Zimmer (2005) and Nelsen (2006)). As a robustness check on the (parametric) bivariate normality distribution, we also use a Gaussian copula that combines the empirical distribution of the error terms from the cross sections for Vietnam in 2006-2008 and Peru in 2005-2006. This estimation approach requires multiple simulations (we use 500 times), but offers a predicted error term $\hat{\varepsilon}_{ij}$ that we can add to the deterministic part $\hat{\beta}_j' x_{ij}$ in Equation (1) to obtain predicted household consumption. While the copula-based estimates are rather similar to our synthetic panel estimates for Vietnam (Appendix 3, Table 3.5), they are somewhat less accurate than those for Peru. We further pursue this approach in Bourguignon and Dang (2019).

In addition to these two assumptions, we also rely on the standard assumption that the sampling methodology is consistent and comparable over the periods (such that household consumption aggregates are consistently constructed and comparable over the two periods).

In addition, we can obtain good estimates of correlation at the cohort-level aggregated data if the individual data within a cohort show very similar values (or the intraclass correlation is close to 1 (Snijders and Bosker, 2011)). Furthermore, if these cohort dummy variables do not capture any variation in household consumption, the synthetic panel cohort-level simple correlation coefficient $\rho_{y_{c1}y_{c2}}$ would simply be 0. In the extreme case, consumption (or poverty) mobility can happen entirely within cohorts, but this case would be easily detected since it results in $\rho_{y_{c1}y_{c2}}$ being equal to 1 (i.e., since cohort means remain unchanged across the two survey rounds). We return to more discussion in the next section on Monte Carlo simulation.

Estimates using data in the first survey round as the base year are provided in Table 3.4, and show similarly encouraging results. Weighted regressions provide qualitatively similar but somewhat less accurate results, thus we use unweighted regressions. Estimates for the earlier survey rounds (x_{i1}) for the last three countries are provided in Dang and Lanjouw (2013) and offer qualitatively similar results. For comparison, we also calculate the bootstrap standard errors by bootstrapping (y_{ij}, $\boldsymbol{x}_{ij}$) from its empirical distribution function (1,000 times), adjusting for the complex survey design (including stratification, cluster sample, and population weights for all countries except for the PSID), and applying the estimated parameters for Equations (1) from the original samples. The bootstrap standard errors, shown in Table 3.7 in Appendix 3, are in fact

slightly smaller than those based on the analytical standard errors. As such, to be conservative, we work with the analytical standard errors.

We next examine in Table 3.8 the poverty mobility estimates for three periods using data from all three survey rounds for the US in 2005-2007, and 2009, Vietnam in 2004, 2006, and 2008, and Peru in 2004, 2005, and 2006, where there are 8 possible poverty categories that each household can fall in in these three periods (for the unconditional probabilities).[32] As discussed earlier, we should expect estimates to be less accurate than those for two periods; however, our proposed method turns out to work quite well with more than two thirds (i.e., 17 out of 24) of all the point estimates being contained in the 95 percent CIs around the true rates; the corresponding figure for the stricter test of one standard error is just one half (i.e., 12 out of 24). The efficiency test points to a coverage ranging from 62 percent to 76 percent, and almost two thirds (i.e., 15 out of 24) of the point estimates pass 100 percent inclusion mark.

Additional References

Bourguignon, Francois and Hai-Anh Dang. (2019). "Investigating Welfare Dynamics with Repeated Cross Sections: A Copula Approach". *Paper presented at the WB-IARIW conference*, Washington DC.

Cappellari, Lorenzo, and Stephen P. Jenkins. (2006). "Calculation of Multivariate Normal Probabilities by Simulation, with Applications to Maximum Simulated Likelihood Estimation". *Stata Journal*, 6(2): 156- 189.

Demirguc-Kunt, Asli, Leora F. Klapper, and Georgios A. Panos. (2011). "Entrepreneurship in Post-Conflict Transition: The Role of Informality and Access to Finance". *Economics of Transition*, 19(1): 27-78.

Dorfman, Alan H. (2011). "A Coverage Approach to Evaluating Mean Square Error". *Pakistan Journal of Statistics*, 27(4): 493-506.

Nelsen, Roger B. (2006). *An Introduction to Copulas*. 2nd Edition. New York: Springer.

Snijders, Tom AB and Roel Bosker. (2011). *Multilevel analysis: an introduction to basic and advanced multilevel modelling*. Sage Publications, London.

Trivedi, P. K. and D. M. Zimmer. (2005). "Copula Modeling: An Introduction for Practitioners". *Foundations and Trends in Econometrics*, 1(1): 1–111.

[32] Estimation results for the trivariate normal probabilities in this table are calculated using the Stata algorithm by Cappellari and Jenkins (2006) with 100 Halton draws.

Table 3.1: Estimated Parameters of Household Consumption for Each Year

	Bosnia-Herzegovina		Lao PDR		Peru				United States				Vietnam			
	2001	2004	2002/03	2007/08	2004-05		2005-06		2005-07		2007-09		2004-06		2006-08	
Age	0.006***	0.012***	0.004***	0.006***	0.010***	0.012***	0.012***	0.013***	0.008***	0.006***	0.011***	0.008***	0.009***	0.010***	0.011***	0.009***
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Female	0.190***	0.277***	0.086*	0.137***	0.166***	0.153***	0.144***	0.192***	-0.306***	-0.463***	-0.433***	-0.516***	0.133***	0.094***	0.084***	0.113***
	(0.041)	(0.043)	(0.048)	(0.041)	(0.022)	(0.016)	(0.016)	(0.016)	(0.016)	(0.020)	(0.020)	(0.024)	(0.023)	(0.021)	(0.022)	(0.022)
Years of schooling	0.035***	0.038***	0.032***	0.046***	0.064***	0.068***	0.068***	0.067***	0.419***	0.579***	0.573***	0.794***	0.051***	0.053***	0.053***	0.056***
	(0.005)	(0.005)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.022)	(0.028)	(0.028)	(0.036)	(0.003)	(0.002)	(0.003)	(0.003)
Bosnian	-0.227***	-0.042														
	(0.051)	(0.053)														
Serb	-0.128**	-0.068														
	(0.051)	(0.053)														
Ethnic majority group			0.239***	0.261***	0.209***	0.197***	0.188***	0.205***	0.150***	0.182***	0.200***	0.253***	0.393***	0.389***	0.361***	0.383***
			(0.021)	(0.022)	(0.025)	(0.018)	(0.018)	(0.017)	(0.016)	(0.020)	(0.020)	(0.024)	(0.027)	(0.025)	(0.026)	(0.026)
Urban	-0.151***	-0.020	0.132***	0.133***	0.352***	0.430***	0.439***	0.446***	0.004***	0.008***	0.008***	0.006***	0.529***	0.447***	0.433***	0.310***
	(0.030)	(0.031)	(0.026)	(0.024)	(0.027)	(0.020)	(0.020)	(0.019)	(0.001)	(0.002)	(0.002)	(0.002)	(0.026)	(0.024)	(0.024)	(0.023)
Constant	7.525***	7.022***	11.264***	11.658***	4.091***	3.928***	3.937***	3.946***	10.822***	10.538***	10.360***	10.086***	6.901***	7.192***	7.166***	7.492***
	(0.119)	(0.131)	(0.051)	(0.055)	(0.057)	(0.040)	(0.040)	(0.040)	(0.040)	(0.053)	(0.049)	(0.065)	(0.053)	(0.048)	(0.051)	(0.050)
σ_v	0.522	0.543	0.518	0.537	0.547	0.556	0.553	0.546	0.407	0.519	0.511	0.628	0.473	0.482	0.485	0.489
Adjusted R^2	0.077	0.077	0.157	0.218	0.407	0.443	0.443	0.463	0.293	0.328	0.335	0.340	0.454	0.421	0.407	0.370
N	1342	1342	3032	3215	4493	9169	8593	9084	3275	3275	3368	3368	3527	3674	3596	3701

Note: *p<0 .1, **p<0.05, ***p<0.01. Standard errors are in parentheses. Household heads' ages are restricted to between 25 and 55 for the first survey round and adjusted accordingly with the year difference for the second survey round. For the US, dummy variables for college degree and being white are used instead of years of schooling and ethnic majority group respectively. Other control variables used for the US include dummy variables indicating high school education and dummy variables indicating religion. Estimation is provided using the cross sections for Lao PDR, Peru, and Vietnam and from panel halves for Bosnia-Herzegovina and the US.

Table 3.2: Estimated Bounds on Poverty Dynamics Based on Synthetic Panel Data for Two Periods, Conditional Probabilities (Percentage)

Poverty Status	Bosnia- Herzegovina		Lao PDR		Peru		United States		Vietnam	
First Period--> Second Period	2001- 2004		2002/03- 2007/08		2005-06		2007-09		2006-08	
	Actual panel	Synthetic panel	Actual panel	Synthetic panel	Actual panel	Synthetic panel	Actual panel	Synthetic panel	Actual panel	Synthetic panel
Poor--> Poor	45.0	[22.4, 86.7]	49.0	[32.2, 91.3]	72.0	[53.6, 94.2]	61.2	[26.3, 93.5]	62.8	[32.6, 96.3]
	(4.6)		(3.0)		(1.9)		(2.2)		(2.8)	
Poor--> Nonpoor	55.0	[13.3, 77.6]	51.0	[8.7, 67.8]	28.0	[5.8, 46.4]	38.8	[6.5, 73.7]	37.2	[3.7, 67.4]
	(4.6)		(3.0)		(1.9)		(2.2)		(2.8)	
Nonpoor--> Poor	13.6	[2.8, 19.7]	15.2	[0.6, 21.8]	15.1	[0.4, 31.2]	5.0	[1.5, 8.5]	5.9	[0.7, 11.6]
	(1.8)		(1.3)		(1.3)		(0.4)		(0.6)	
Nonpoor--> Nonpoor	86.4	[80.3, 97.2]	84.8	[78.2, 99.4]	84.9	[68.8, 99.6]	95.0	[91.5, 98.5]	94.1	[88.4, 99.3]
	(1.8)		(1.3)		(1.3)		(0.4)		(0.6)	
N	1342	1342	1989	3215	2250	9084	3368	3368	2723	3701

Note: Synthetic panels are constructed from cross sections for Lao PDR, Peru, and Vietnam and from panel halves for Bosnia-Herzegovina and the US. Predictions are obtained using the estimated parameters from the first and second survey rounds on data in the second survey round. The estimated bounds are shown in brackets under the "Synthetic Panel" for each country. All numbers are weighted using household weights for Peru, and population weights for other countries. Poverty rates are in percent. Household heads' ages are restricted to between 25 and 55 for the first survey round and adjusted accordingly with the year difference for the second survey round.

Table 3.3: Poverty Dynamics Based on Synthetic Panel Data for Two Periods, Conditional Probabilities (Percentage)

Poverty Status	Bosnia- Herzegovina		Lao PDR		Peru		United States		Vietnam	
First Period--> Second Period	2001- 2004		2002/03- 2007/08		2005-06		2007-09		2006-08	
	Actual panel	Synthetic panel	Actual panel	Synthetic panel	Actual panel	Synthetic panel	Actual panel	Synthetic panel	Actual panel	Synthetic panel
Poor--> Poor	45.0	39.4	49.0	50.0	72.0	71.5	61.2	65.5	62.8	66.0
	(4.6)	(1.2)	(3.0)	(1.6)	(1.9)	(1.0)	(2.2)	(2.0)	(2.8)	(1.5)
Poor--> Nonpoor	55.0	60.6	51.0	50.0	28.0	28.5	38.8	34.5	37.2	34.0
	(4.6)	(1.7)	(3.0)	(1.1)	(1.9)	(0.3)	(2.2)	(0.9)	(2.8)	(0.6)
Nonpoor--> Poor	13.6	15.3	15.2	15.5	15.1	17.6	5.0	4.4	5.9	5.9
	(1.8)	(0.2)	(1.3)	(0.3)	(1.3)	(0.2)	(0.4)	(0.1)	(0.6)	(0.1)
Nonpoor--> Nonpoor	86.4	84.7	84.8	84.5	84.9	82.4	95.0	95.6	94.1	94.1
	(1.8)	(0.7)	(1.3)	(0.8)	(1.3)	(0.7)	(0.4)	(0.3)	(0.6)	(0.3)
Goodness-of-fit Tests										
Within 95% CI	4/4		4/4		4/4		2/4		4/4	
Within 1 standard error	2/4		4/4		2/4		0/4		2/4	
Mean coverage (percent)	100		100		79.5		66.6		96.8	
Coverage of 100%	4/4		4/4		2/4		2/4		3/4	
N	1342	1342	1989	3215	2250	9084	3368	3368	2723	3701

Note: Synthetic panels are constructed from cross sections for Lao PDR, Peru, and Vietnam and from panel halves for Bosnia-Herzegovina and the US. Predictions are obtained using the estimated parameters from the first and second survey rounds on data in the second survey round. Standard errors are obtained adjusting for complex survey design for all countries, except for the US PSID. All numbers are weighted using household weights for Peru, and population weights for other countries. Poverty rates are in percent. Household heads' ages are restricted to between 25 and 55 for the first survey round and adjusted accordingly with the year difference for the second survey round. The "Within 95% CI" row shows the number of times that the estimates based on the synthetic panels fall within the 95% confidence interval (CI) of the estimates based on the actual panels; the "Within 1 standard error" row shows a similar figure but using one standard error around the estimates based on the actual panels. The "Mean coverage (percent)" row shows the mean proportion of the 95% CI around the synthetic panel estimates that overlap with those based on the actual panels; the "Coverage of 100%" row shows a similar figure for the number of times that the former fall completely inside the latter.

Table 3.4: Poverty Dynamics Based on Synthetic Data for Two Periods, Using Data in the First Survey Round as the Base (Percentage)

Poverty Status	Bosnia- Herzegovina		Lao PDR		Peru		United States		Vietnam	
First Period & Second Period	2001- 2004		2002/03- 2007/08		2005-06		2007-09		2006-08	
	Actual panel	Synthetic panel	Actual panel	Synthetic panel	Actual Panel	Synthetic Panel	Actual Panel	Synthetic Panel	Actual Panel	Synthetic Panel
Poor, Poor	10.8	9.8	13.5	15.1	29.4	32.2	6.0	7.3	9.6	10.2
	(2.3)	(0.3)	(1.2)	(0.4)	(1.4)	(0.4)	(0.4)	(0.3)	(0.7)	(0.3)
Poor, Nonpoor	13.4	12.3	16.0	13.6	11.7	11.9	3.8	3.7	6.2	5.2
	(1.5)	(0.3)	(1.1)	(0.1)	(0.9)	(0.1)	(0.3)	(0.1)	(0.6)	(0.1)
Nonpoor, Poor	11.5	14.4	8.9	12.4	8.8	9.7	4.6	4.4	4.5	5.2
	(1.7)	(0.2)	(0.9)	(0.2)	(0.7)	(0.1)	(0.4)	(0.1)	(0.5)	(0.1)
Nonpoor, Nonpoor	64.3	63.5	61.7	59.0	50.1	46.2	85.7	84.6	79.7	79.4
	(2.7)	(0.7)	(1.6)	(0.6)	(1.6)	(0.4)	(0.6)	(0.4)	(1.0)	(0.4)
Goodness-of-fit Tests										
Within 95% CI	4/4		2/4		2/4		3/4		4/4	
Within 1 standard error	3/4		0/4		1/4		2/4		2/4	
Mean coverage (percent)	100		41.9		62.6		63.3		94.2	
Coverage of 100%	4/4		1/4		2/4		2/4		2/4	
N	1342	1342	1989	3032	2250	8593	3368	3368	2723	3596

Note: Synthetic panels are constructed from cross sections for Lao PDR, Peru, and Vietnam and from panel halves for Bosnia-Herzegovina and the US. Predictions are obtained using the estimated parameters from the first and second survey rounds on data in the second survey round. Standard errors are obtained adjusting for complex survey design for all countries, except for the US PSID. All numbers are weighted using household weights for Peru, and population weights for other countries. Poverty rates are in percent. Household heads' ages are restricted to between 25 and 55 for the first survey round and adjusted accordingly with the year difference for the second survey round. The "Within 95% CI" row shows the number of times that the estimates based on the synthetic panels fall within the 95% confidence interval (CI) of the estimates based on the actual panels; the "Within 1 standard error" row shows a similar figure but using one standard error around the estimates based on the actual panels. The "Mean coverage (percent)" row shows the mean proportion of the 95% CI around the synthetic panel estimates that overlap with those based on the actual panels; the "Coverage of 100%" row shows a similar figure for the number of times that the former fall completely inside the latter.

Table 3.5: Poverty Dynamics Based on Synthetic Data for Two Periods, Using Gaussian Copula (Percentage)

Poverty Status	Peru		Vietnam	
First Period & Second Period	2005-06		2006-08	
	Actual Panel	Synthetic Panel	Actual Panel	Synthetic Panel
Poor, Poor	29.9	31.2	9.9	9.7
	(1.3)	(0.6)	(0.8)	(0.6)
Poor, Nonpoor	11.6	12.6	5.9	5.2
	(0.9)	(0.5)	(0.5)	(0.4)
Nonpoor, Poor	8.9	10.1	4.9	5.0
	(0.8)	(0.4)	(0.5)	(0.4)
Nonpoor, Nonpoor	49.7	46.1	79.3	80.1
	(1.6)	(0.7)	(1.0)	(0.8)
Goodness-of-fit Tests				
Within 95% CI	3/4		4/4	
Within 1 standard error	0/4		3/4	
Mean coverage (percent)	71.2		91.1	
Coverage of 100%	1/4		2/4	
N	2250	8593	2723	3596

Note: Synthetic panels are constructed from cross sections for Peru and Vietnam. Predictions are obtained using the estimated parameters from the first and second survey rounds on data in the second survey round. Standard errors are obtained adjusting for complex survey design for all countries. All numbers are weighted using household weights for Peru, and population weights for other countries. Poverty rates are in percent. Household heads' ages are restricted to between 25 and 55 for the first survey round and adjusted accordingly with the year difference for the second survey round. The "Within 95% CI" row shows the number of times that the estimates based on the synthetic panels fall within the 95% confidence interval (CI) of the estimates based on the actual panels; the "Within 1 standard error" row shows a similar figure but using one standard error around the estimates based on the actual panels. The "Mean coverage (percent)" row shows the mean proportion of the 95% CI around the synthetic panel estimates that overlap with those based on the actual panels; the "Coverage of 100%" row shows a similar figure for the number of times that the former fall completely inside the latter.

Table 3.6: Poverty Dynamics Based on Synthetic Data for Two Periods, Using Data in the First Survey Round as the Base (Percentage)

Poverty Status	Bosnia- Herzegovina		Lao PDR		Peru		United States		Vietnam	
First Period & Second Period	2001- 2004		2002/03- 2007/08		2005-06		2007-09		2006-08	
	Actual panel	Synthetic panel	Actual panel	Synthetic panel	Actual Panel	Synthetic Panel	Actual Panel	Synthetic Panel	Actual Panel	Synthetic Panel
Poor, Poor	10.8	9.8	13.5	15.1	29.4	32.2	6.0	7.3	9.6	10.2
	(2.3)	(0.3)	(1.2)	(0.4)	(1.4)	(0.4)	(0.4)	(0.3)	(0.7)	(0.3)
Poor, Nonpoor	13.4	12.3	16.0	13.6	11.7	11.9	3.8	3.7	6.2	5.2
	(1.5)	(0.3)	(1.1)	(0.1)	(0.9)	(0.1)	(0.3)	(0.1)	(0.6)	(0.1)
Nonpoor, Poor	11.5	14.4	8.9	12.4	8.8	9.7	4.6	4.4	4.5	5.2
	(1.7)	(0.2)	(0.9)	(0.2)	(0.7)	(0.1)	(0.4)	(0.1)	(0.5)	(0.1)
Nonpoor, Nonpoor	64.3	63.5	61.7	59.0	50.1	46.2	85.7	84.6	79.7	79.4
	(2.7)	(0.7)	(1.6)	(0.6)	(1.6)	(0.4)	(0.6)	(0.4)	(1.0)	(0.4)
Goodness-of-fit Tests										
Within 95% CI	4/4		2/4		2/4		3/4		4/4	
Within 1 standard error	3/4		0/4		1/4		2/4		2/4	
Mean coverage (percent)	100		41.9		62.6		63.3		94.2	
Coverage of 100%	4/4		1/4		2/4		2/4		2/4	
N	1342	1342	1989	3032	2250	8593	3368	3368	2723	3596

Note: Synthetic panels are constructed from cross sections for Lao PDR, Peru, and Vietnam and from panel halves for Bosnia-Herzegovina and the US. Predictions are obtained using the estimated parameters from the first and second survey rounds on data in the second survey round. Standard errors are obtained adjusting for complex survey design for all countries, except for the US PSID. All numbers are weighted using household weights for Peru, and population weights for other countries. Poverty rates are in percent. Household heads' ages are restricted to between 25 and 55 for the first survey round and adjusted accordingly with the year difference for the second survey round. The "Within 95% CI" row shows the number of times that the estimates based on the synthetic panels fall within the 95% confidence interval (CI) of the estimates based on the actual panels; the "Within 1 standard error" row shows a similar figure but using one standard error around the estimates based on the actual panels. The "Mean coverage (percent)" row shows the mean proportion of the 95% CI around the synthetic panel estimates that overlap with those based on the actual panels; the "Coverage of 100%" row shows a similar figure for the number of times that the former fall completely inside the latter.

Table 3.7: Poverty Dynamics Based on Synthetic Data for Two Periods with Bootstrap Standard Errors, Joint Probabilities (Percentage)

Poverty Status	Bosnia- Herzegovina		Lao PDR		Peru		United States		Vietnam	
First Period & Second Period	2001- 2004		2002/03- 2007/08		2005-06		2007-09		2006-08	
	Actual panel	Synthetic panel	Actual panel	Synthetic panel	Actual Panel	Synthetic Panel	Actual Panel	Synthetic Panel	Actual Panel	Synthetic Panel
Poor, Poor	10.3	8.2	13.8	13.2	29.9	30.9	6.0	6.2	9.9	9.6
	(1.7)	(0.2)	(1.2)	(0.3)	(1.3)	(0.3)	(0.4)	(0.2)	(0.8)	(0.2)
Poor, Nonpoor	12.6	12.6	14.3	13.2	11.6	12.3	3.8	3.2	5.9	4.9
	(1.2)	(0.2)	(1.1)	(0.1)	(0.9)	(0.1)	(0.3)	(0.1)	(0.5)	(0.0)
Nonpoor, Poor	10.5	12.1	10.9	11.4	8.9	10.0	4.6	4.0	4.9	5.0
	(1.4)	(0.1)	(1.0)	(0.1)	(0.8)	(0.0)	(0.4)	(0.1)	(0.5)	(0.0)
Nonpoor, Nonpoor	66.5	67.2	61.0	62.2	49.7	46.8	85.7	86.6	79.3	80.4
	(2.2)	(0.4)	(1.6)	(0.5)	(1.6)	(0.3)	(0.6)	(0.3)	(1.0)	(0.3)
Goodness-of-fit Tests										
Within 95% CI	4/4		4/4		4/4		4/4		4/4	
Within 1 standard error	2/4		4/4		2/4		1/4		2/4	
Mean coverage (percent)	100		100		92.9		83.0		100	
Coverage of 100%	4/4		4/4		3/4		2/4		4/4	
N	1342	1342	1989	3215	2250	9084	3368	3368	2723	3701

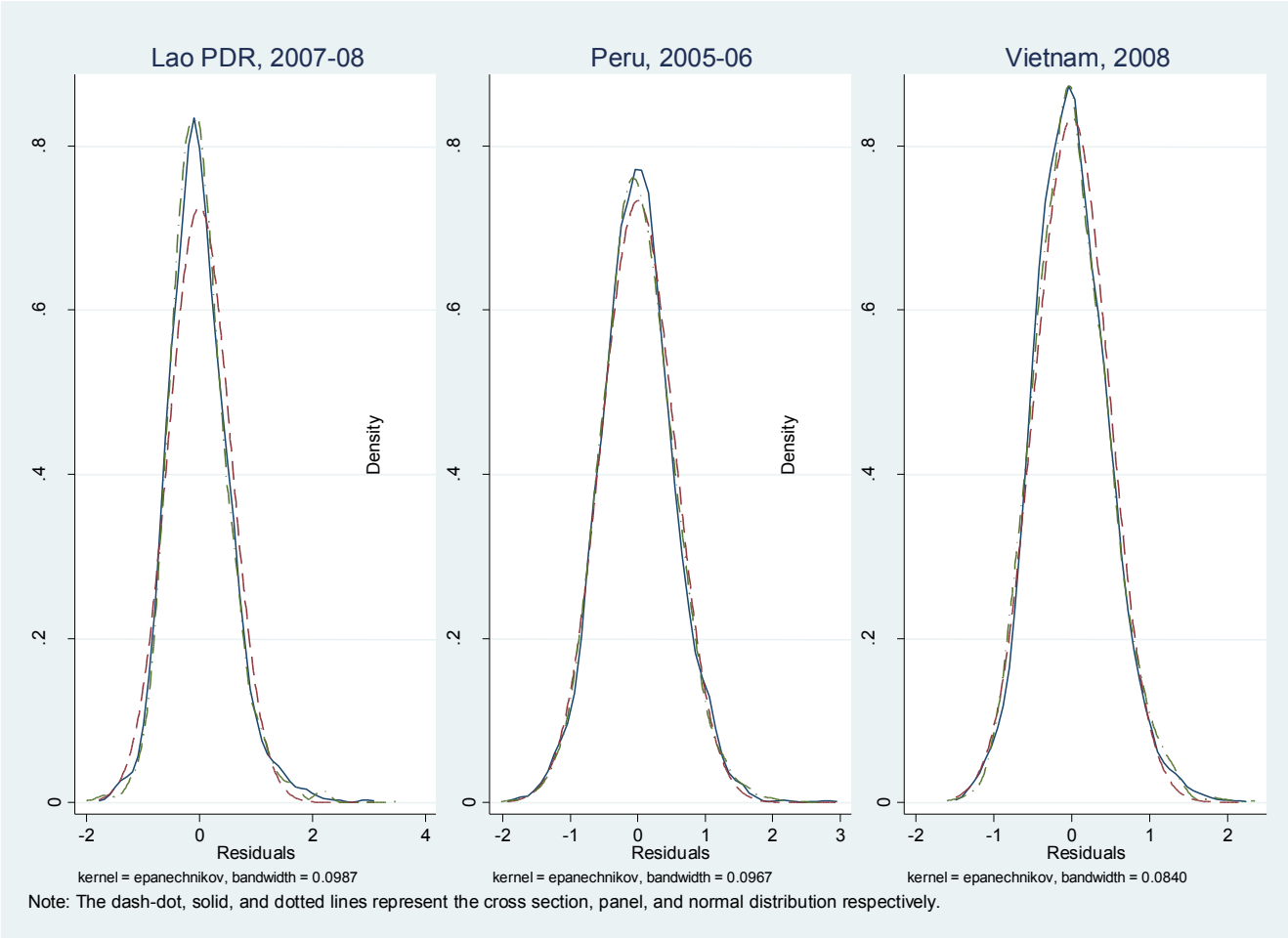
Note: Synthetic panels are constructed from cross sections for Lao PDR, Peru, and Vietnam and from panel halves for Bosnia-Herzegovina and the US. Predictions are obtained using the estimated parameters from the first and second survey rounds on data in the second survey round. Bootstrap standard errors are obtained adjusting for complex survey design for all countries, except for the US PSID. We use 1,000 bootstraps. All numbers are weighted using household weights for Peru, and population weights for other countries. Poverty rates are in percent. Household heads' ages are restricted to between 25 and 55 for the first survey round and adjusted accordingly with the year difference for the second survey round. The "Within 95% CI" row shows the number of times that the estimates based on the synthetic panels fall within the 95% confidence interval (CI) of the estimates based on the actual panels; the "Within 1 standard error" row shows a similar figure but using one standard error around the estimates based on the actual panels. The "Mean coverage (percent)" row shows the mean proportion of the 95% CI around the synthetic panel estimates that overlap with those based on the actual panels; the "Coverage of 100%" row shows a similar figure for the number of times that the former fall completely inside the latter.

Table 3.8: Poverty Dynamics Based on Synthetic Panel Data for Three Periods (Percentage)

Poverty Status	Peru		United States		Vietnam	
	2004-05-06		2005-07-09		2004-06-08	
First, Second & Third Period	Actual panel	Synthetic panel	Actual panel	Synthetic panel	Actual panel	Synthetic panel
Poor, Poor, Poor	26.6	24.0	4.0	4.0	8.1	7.6
	(1.4)	(0.4)	(0.4)	(0.2)	(1.0)	(0.3)
Poor, Poor, Nonpoor	6.9	7.1	1.4	2.0	3.1	2.8
	(0.7)	(0.1)	(0.2)	(0.0)	(0.6)	(0.0)
Poor, Nonpoor, Poor	4.4	3.3	1.0	0.5	2.3	2.9
	(0.6)	(0.0)	(0.2)	(0.0)	(0.5)	(0.1)
Poor, Nonpoor, Nonpoor	7.2	6.4	2.7	2.8	6.6	5.0
	(0.7)	(0.0)	(0.3)	(0.0)	(0.8)	(0.1)
Nonoor, Poor, Poor	3.9	6.1	1.8	1.7	0.8	1.1
	(0.5)	(0.1)	(0.2)	(0.1)	(0.2)	(0.0)
Nonpoor, Poor, Nonpoor	5.4	5.0	2.0	1.1	1.7	2.6
	(0.6)	(0.0)	(0.3)	(0.0)	(0.4)	(0.0)
Nonpoor, Nonpoor, Poor	4.6	6.4	3.1	3.2	2.9	2.7
	(0.6)	(0.0)	(0.3)	(0.1)	(0.5)	(0.0)
Nonpoor, Nonpoor, Nonpoor	41.0	41.6	84.0	84.6	74.5	75.3
	(1.7)	(0.4)	(0.7)	(0.4)	(1.5)	(0.4)
Goodness-of-fit Tests						
Within 95% CI	6/8		5/8		6/8	
Within 1 standard error	3/8		5/8		4/8	
Mean coverage (percent)	69.1		62.1		75.8	
Coverage of 100%	4/8		5/8		6/8	
N	1987	8608	3036	3036	1282	3808

Note: Synthetic panels are constructed from cross sections for Peru and Vietnam and from panel halves for the US. Predictions are obtained using the estimated parameters from the first, second, and third survey rounds on data in the third survey round. Standard errors are obtained adjusting for complex survey design for all countries, except for the US PSID. All numbers are weighted using household weights for Peru, and population weights for other countries. Poverty rates are in percent. Household heads' ages are restricted to between 25 and 55 for the first survey round and adjusted accordingly with the year difference for the second survey round. The "Within 95% CI" row shows the number of times that the estimates based on the synthetic panels fall within the 95% confidence interval (CI) of the estimates based on the actual panels; the "Within 1 standard error" row shows a similar figure but using one standard error around the estimates based on the actual panels. The "Mean coverage (percent)" row shows the mean proportion of the 95% CI around the synthetic panel estimates that overlap with those based on the actual panels; the "Coverage of 100%" row shows a similar figure for the number of times that the former fall completely inside the latter.

Figure 3.1: Plotting the Cross section and Panel Data vs. the Normal Distribution



Note: The dash-dot, solid, and dotted lines represent the cross section, panel, and normal distribution respectively.

Appendix 4: Estimation Procedures and Related Concerns

We propose the following steps to obtain poverty mobility for two periods:

Step 1: Using the data in survey round 1, estimate Equation (1) and obtain the predicted coefficients $\hat{\beta}_1'$, and the predicted standard error $\hat{\sigma}_{\varepsilon_{i1}}$ for the error term ε_{i1}. Similarly, using the data in survey round 2, estimate Equation (1) and obtain similar parameters $\hat{\beta}_2'$ and $\hat{\sigma}_{\varepsilon_{i2}}$.

Step 2a: Aggregate data in both survey rounds 1 and 2 by cohorts and obtain the estimated cohort-level simple correlation coefficient $\hat{\rho}_{y_{i1}y_{i2}}$. Calculate $\hat{\rho}$ using Proposition 1 (and check that $\hat{\rho}_{y_{i1}y_{i2}} \geq \hat{\rho}$).

Step 2b: (If relevant) Calculate $\hat{\rho}$ using Proposition 2 as a robustness check.

Step 3: For each household in survey round *j*, calculate the unconditional quantities of poverty mobility as $\Phi_2\left(d_1 \frac{z_1-\hat{\beta}_1'x_{ij}}{\hat{\sigma}_{\varepsilon_1}}, d_2 \frac{z_2-\hat{\beta}_2'x_{ij}}{\hat{\sigma}_{\varepsilon_2}}, \hat{\rho}_d\right)$, where d_j is an indicator function that equals 1 if the household is poor and equals -1 if the household is non-poor in period *j, j= 1, 2*, and $\hat{\rho}_d = d_1 d_2 \hat{\rho}$. Calculate the standard errors using Proposition 3. Make the appropriate adjustments to obtain population-level numbers.

Step 4: (If relevant) Calculate the conditional quantities of poverty mobility for period *j* as $\frac{\Phi_2\left(d_1 \frac{z_1-\hat{\beta}_1'x_{ij}}{\hat{\sigma}_{\varepsilon_1}}, d_2 \frac{z_2-\hat{\beta}_2'x_{ij}}{\hat{\sigma}_{\varepsilon_2}}, \hat{\rho}_d\right)}{\hat{P}_j}$, where d_j is an indicator function that equals 1 if the household is poor and equals -1 if the household is non-poor in period *j, j= 1, 2*, and $\hat{\rho}_d = d_1 d_2 \hat{\rho}$. Calculate the standard errors using Corollary 3.1. Make the appropriate adjustments to obtain population-level numbers.

The estimation procedures for three groups (or more) are similar, where the formulae for the quantities estimated in Steps 3 and 4 are given in Proposition 4. The formulae for three periods (or more) are given in Dang and Lanjouw (2013).[33] Compared to the previous case of two periods, the computation now becomes more involved since the number of integral dimensions corresponds to the number of survey rounds. Estimates will likely be less accurate for three (or more) periods than those for two periods due to increased layers of potential (modeling and sampling) errors.

As shown in Proposition 3 in Appendix 1, the standard errors for the synthetic panel estimates consist of two components, the model errors and the sampling errors, with the latter's variance expected to be larger than the former's variance when the regressions have good fits. This is indeed the case where (results not shown) the variances of the sampling errors are significantly larger than those for the model errors. Thus, since the sampling errors account for most of the errors with the synthetic estimates and the cross sections used for the synthetic estimates have larger sample sizes than panel data, the synthetic panel estimates unsurprisingly have smaller standard errors than those based on actual panel data. This is supported by the empirical estimates provided in Table 3.

A practical concern with estimation is whether or not Equation (1) should be estimated with household weights. There appear to be both advantages and disadvantages with both approaches. Weighted regressions are especially relevant when the provided household weights were constructed to account for non-response or attrition bias or were specifically based on the dependent variables (informative sampling); on the other hand, unweighted regressions are most

[33] Stata programs that implement these procedures are available upon request.

relevant when the proposed super-population (i.e., Equation (1)) model is correct and can provide some causal interpretation. Estimation without weights in the former case results in biased estimates, while estimation with weights in the latter case yields inefficiency (i.e., larger standard errors). Thus it seems advisable to estimate Equation (1) both with and without weights and compare results, particularly where there is limited information on how the weights have been constructed. Once the parameters from Equation (1) are obtained, the quantities of interest in Steps 3 and 4 should be estimated using weights as usual.[34]

For the test on Proposition 2, the variables used to construct cohorts should be correlated with household consumption. In practice, the test for Proposition 2 is simply the F test for the joint significance of the cohort dummy variables in a cohort fixed-effects model. The estimate for ρ can also be obtained directly from this regression. For example, this can be done in Stata by using the following command for fixed-effects regressions "*xtreg y, i(panid) fe*", where the panel id (*panid*) is represented by the cohorts formed by a combination of all the different values of the time-invariant variables in x_{ij}. (Note there is no need for a cohort time-invariant regressor in this command because all such time-invariant variables are washed out in the fixed-effects regression). The desired estimate for ρ is provided by the estimate for "*rho*" in the outputs of this regression.

As discussed in Section III.2, a larger sample size would reduce the sampling variance; thus, this points to the advantages of cross sections over panel data when the former have much larger sample sizes than the latter. A natural extension of this would be to pool estimates from the two cross sections for a larger sample size to reduce the sampling errors even more, where we can simply use the corresponding population weight for each cross section to estimate the means (assuming that the sample sizes of the cross sections are similar). See also Kish (1999, 2002) for overviews on combining surveys. On a related note, whether the model variance or the sampling variance is the dominant component would depend on the dynamics of the underlying regression relationship and the overall precision of our theoretical models. Our estimation results indicate that the sampling variance is significantly larger than the model variance, which is consistent with findings in the small area estimation literature (see, e.g., Rao (2003, p. 35)).

Additional References

Deaton, Angus. (1997). "*The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*." MD: The Johns Hopkins University Press.

Kish, Leslie. (1999). "Cumulating/ Combining Population Surveys". *Survey Methodology*, 25(2): 129- 138.

---. (2002). "New Paradigms (Models) for Probability Sampling". *Survey Methodology*, 28(1): 31-34.

Lorh, Sharon L. (2010). *Sampling, Design and Analysis*. Massachusetts: Duxbury Press.

Pfeffermann, Danny. (2011). "Modelling of Complex Survey Data: Why model? Why Is It a Problem? How Can We Approach It?" *Survey Methodology*, 37(2): 115- 136.

Rao, J. N. K. (2003). *Small Area Estimation*. New Jersey: Wiley.

[34] We briefly summarize here the arguments provided in Deaton (1997), Lorh (2010), and Pfeffermann (2011) in this section; see the cited studies for further discussion on this topic.