Public Disclosure Authorized

Public Disclosure Authorized

WPS6504

Policy Research Working Paper 6504

Measuring Poverty Dynamics with Synthetic Panels Based on Cross-Sections

Hai-Anh Dang
Peter Lanjouw

The World Bank
Development Research Group
Poverty and Inequality Team
June 2013



Policy Research Working Paper 6504

Abstract

Panel data conventionally underpin the analysis of poverty mobility over time. However, such data are not readily available for most developing countries. Far more common are the "snap-shots" of welfare captured by cross-section surveys. This paper proposes a method to construct synthetic panel data from cross sections which can provide point estimates of poverty mobility. In contrast to traditional pseudo-panel methods that require multiple rounds of cross-sectional data to study poverty at the cohort level, the proposed method can be applied to settings with as few as two survey rounds and also permits investigation at the more disaggregated household level. The procedure is implemented using cross-section survey data from several countries, spanning different income levels and geographical regions. Estimates fall within the 95 percent confidence interval—or even one standard error in many cases—of those based on actual panel data. The method is not only restricted to studying poverty mobility but can also accommodate investigation of other welfare outcome dynamics.

This paper is a product of the Poverty and Inequality Team, Development Research Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The authors may be contacted at hdang@worldbank.org and planjouw@worldbank.org.

The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

Produced by the Research Support Team

Measuring Poverty Dynamics with Synthetic Panels Based on Cross-Sections

Hai-Anh Dang and Peter Lanjouw[*]

World Bank

Keywords: Transitory and Chronic poverty; Synthetic panels; Mobility

JEL Codes: O15, I32, C53.

Sector Board: POV

[*]Dang (hdang@worldbank.org) and Lanjouw (planjouw@worldbank.org) are respectively Economist and Research Manager with the Poverty and Inequality Unit, Development Research Group, World Bank. We thank Chris Elbers, Francisco Ferreira, Paul Glewwe, Bill Greene, Dean Jolliffe, Michael Lokshin, David McKenzie, Tuoc Van Phan, Sergiy Radyakin, Carolina Sanchez-Paramo, Renos Vakis, Roy van der Weide, Nobuo Yoshida, and participants at seminars at IFPRI and the World Bank for comments on previous versions of this paper, and Alan H. Dorfman for referring us to a recent paper of his. We also thank Reema Nayar and Yue Man Lee for their support and encouragement for application of our method to the World Bank 2012 Job Flagship report for South Asia, and we further thank Renos Vakis and Leonardo Lucchetti for their help with the Peruvian data. The findings and interpretations in this paper do not necessarily reflect the views of the World Bank, its affiliated institutions, or its Executive Directors.

I. Introduction

To effectively reduce poverty, we want to understand the factors that help households escape poverty as well as those that induce them to remain in or fall back into poverty. It is commonly claimed that we need panel data to answer these questions, especially at the household or individual level. However, for most developing countries, cross-sectional data are far more common than panel data. This can be for a variety of reasons. Panel data collection can be very costly, for example, and can also pose a variety of logistical and capacity-related challenges. For whatever reason, the scarcity of panel data has rendered the analysis of welfare dynamics difficult, if not impossible, in most developing country settings.

To overcome the non-availability of panel data, there have been a number of studies that develop pseudo-panels (or synthetic panels) out of multiple rounds of cross-sectional data. Following the seminal contributions of Deaton (1985), synthetic panels based on age cohorts have been widely used to investigate income and consumption over time (e.g., Deaton and Paxson, 1994; Banks, Blundell, and Brugiavini, 2001; and Pencavel, 2007). Other outcomes have also been analyzed with synthetic panels; these include, for example, labor responses to tax reforms (Blundell, Duncan, and Meghir, 1988), the returns to academic and vocational qualifications (Mcintosh, 2006), and household demand for private medical insurance (Propper, Rees, and Green, 2001). In particular, since cross-section samples are typically refreshed each time that the surveys are fielded, synthetic panels are possibly also less exposed to concerns about attrition and measurement error often leveled at

true panel data relating.[1] Thus, unsurprisingly, the econometrics of pseudo-panel data is a rapidly growing field of research.[2]

Perhaps because of their emphasis on cohorts rather than the household or individual, pseudo panel methods have not been widely applied to the analysis of poverty 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.

Against this background, a recent paper by Dang, Lanjouw, Luoto, and McKenzie (2011) (hereafter referred to as DLLM) proposes both parametric and non-parametric approaches to construct synthetic panels at the household level from two rounds of cross sections with rather parsimonious assumptions. These synthetic panels can then be used to predict lower-bound and upper-bound estimates of household poverty dynamics. [3]

The DLLM method is applied in two empirical settings: Vietnam and Indonesia. Drawing on both cross sectional data and genuine panel data, the authors compare mobility estimates based on synthetic panels to those that would be obtained from actual panel data.

[1] See, for example, Glewwe and Jacoby (2000) and Kalton (2009) for recent overviews of the advantages and disadvantages of cross sections and panel data in both developing and richer country contexts.
[2] See, for example, Inoue (2008) for a recent development and a brief review of this literature.
[3] This method, and its non-parametric variant in particular, is related to the small-area estimation approach developed by Elbers, Lanjouw, and Lanjouw (2002, 2003). See, for example, Agostini and Brown (2010), Elbers et al. (2007), and Demombynes and Ozler (2005) for recent applications of the poverty mapping method. 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 literature on imputing missing data (see, e.g., Little and Rubin, 2002). See also Ridder and Moffitt (2007) for a recent review on the econometrics of data combination.

In their validation exercise the authors find that the "true" estimate of mobility (as revealed by the actual panel data) is generally sandwiched between the upper-bound and lower-bound assessments derived from the DLLM method. In particular, the interval between the bound estimates can be narrowed if an appropriate range for the correlation between the error terms can be postulated. DLLM propose that panel data, where available from other sources, such as other countries with similar characteristics, might be scrutinized to identify such a narrower range. Despite its infancy, applications of the DLLM method in various settings have been yielding encouraging results.[4]

In this paper we generalize the method introduced by DLLM in several important aspects. First, by proposing a method to calculate the appropriate correlation term and its upper theoretical bound using each country's *own* cross sectional surveys, we overcome the limitations of having to rely on external estimates from true panels in similar settings. Notably, such "similar" true panels might not be available (we will discuss other more technical issues in a later section). Thus with the current more general framework, we can apply our method to study poverty dynamics with any two rounds of cross sections where our (rather standard) assumptions are satisfied.

Second, the identification of point values for the correlation term allows us to move from providing *bound* estimates to *point* estimates of poverty mobility. This advance renders estimation more accurate and more easily interpreted. In particular, we can investigate different measures of poverty dynamics, such as the population shares in different poverty statuses in both survey periods considered together (i.e., absolute measure or joint probability) or the population shares in different poverty statuses in one period given

[4] For example, recent applications/ validations of the DLLM method against true panel data include Cruces et al., (2011) for Chile, Nicaragua, and Peru, and Bierbaum and Gassmann (2012) for the Kyrgyz Republic.

their welfare status in the other period (i.e., relative measure or conditional probability). The former measure provides one way to define chronic poverty rate,[5] and the latter includes the percentage of the poor in the first period that escape poverty in the second period.

Third, also by providing point estimates of poverty mobility, we can conveniently generalize the method introduced in DLLM to settings where more than two rounds of data are available. This potentially useful since very few panel datasets in developing countries span more than two periods; and in those cases where they do, the datasets are likely to suffer heavily from attrition problems. By considering poverty mobility in more than two periods, we can investigate richer inter-temporal profiles of movement into and out of poverty. Fourth, and finally, we also provide standard errors and discuss the asymptotic distributions for the point estimates, which is not a focus in the bound estimates offered by DLLM.

We validate our estimates by using both cross sectional and actual panel survey data from high-income and developing countries including Bosnia-Herzegovina, Lao PDR, Peru, the United States, and Vietnam. We find that our synthetic panel estimates are close to—and mostly lying within the 95 percent confidence intervals or even one standard error in many cases—of those of actual panel data. Assuming our model is valid, the standard errors on our model-based synthetic panel estimates are also found to be smaller than the sampling-based standard errors of actual panel data.

This paper consists of six sections. Given that our paper builds on the parametric approach introduced in DLLM to obtain bound estimates on poverty mobility, we start, in the next section, with a brief description of this method. Our generalization of this method

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

and estimation procedures are then introduced in Section III. We describe the data in Section IV before applying our method to investigate poverty dynamics in Section V and Section VI concludes.

II. Bound Estimates on Poverty Mobility

Let x_{ij} be a vector of household characteristics observed in survey round *j* (j= 1 or 2) that are also observed in the other survey round for household i, i= 1,…, N. Subject to data availability, these household characteristics can include such time-invariant variables as ethnicity, religion, language and if the household heads remain the same across survey rounds, variables such as household heads' age, sex, education, place of birth, and parental 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.

Then let y_{ij} represent household consumption or income in survey round j, j= 1 or 2. The linear projection of household consumption (or income) on household characteristics for each survey round is given by

$$y_{i1} = \beta_1' x_{i1} + \varepsilon_{i1} \quad (1)$$

$$y_{i2} = \beta_2' x_{i2} + \varepsilon_{i2} \quad (2)$$

Let z_j be the poverty line in period j, j= 1 or 2. We are interested in knowing such quantities as

$$P(y_{i1} < z_1 \text{ and } y_{i2} > z_2) \quad (3a)$$

which represents the percentage of households that are poor in the first period but nonpoor in the second period (considered together for two periods), or

$$P(y_{i2} > z_2 \mid y_{i1} < z_1) \quad (3b)$$

which represents the percentage of poor households in the first period that escape poverty in the second period.

If panel data are available, we can easily calculate the quantities in (3a) and (3b); otherwise, we have to rely on synthetic panels for this purpose. Assume that the underlying population being sampled in survey rounds 1 and 2 are the same, or more specifically, $x_{i1} = x_{i2}$, and $y_{i1}|x_{i1}$ and $y_{i1}|x_{i2}$ have identical distributions (Assumption 1 in DLLM). We can rely on the time-invariant variables x_{ij} that are collected in both survey rounds to construct the consumptions in period 1 for households interviewed in period 2, and vice versa.[6] In particular, assume that ε_{i1} and ε_{i2} have a bivariate normal distribution with correlation coefficient ρ and standard deviations σ_{ε_1} and σ_{ε_2} respectively (Assumption 2'). If ρ is known, DLLM propose to estimate quantity (3a) by

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

where $\Phi_2(.)$ stands for the bivariate normal cumulative distribution function (cdf)) (and $\phi_2(.)$ stands for the bivariate normal probability density function (pdf)).

However, since ρ is usually unknown in most contexts, assume also that it is bounded by the interval [0, 1] (Assumption 2''),[7] since for any x, y, and ρ, $\frac{\partial \Phi_2(x, y, \rho)}{\partial \rho} = \phi_2(x, y, \rho) > 0$ (Sungur, 1990). Equation (4) indicates that a lower (higher) value of ρ means a higher (lower) probability of being poor in the first period but non-poor in the second period. Thus the lower bound and upper bound estimates of mobility can be established by identifying the

[6] In other words, this assumption implies that households in period 2 that have similar characteristics to those of households in period 1 would have achieved the same consumption levels in period 1 or vice versa.

[7] DLLM provide several reasons why this assumption can be expected to hold and show this is the case using household survey data from several countries. Also see DLLM for more discussion on the implications for these assumptions and proofs for the bound estimates.

appropriate range of values for the correlation term ρ. Absent any other information, DLLM indicate one can start by assuming that ρ is either 0 or 1. However, by examining empirical estimates from actual panel data for other countries, DLLM propose a narrower range for Indonesia during 1997-2000 of [0.3, 0.7]. This in turn yields a lower bound and upper bound interval of [8.1, 11.8] for the proportion of households that were poor in 1997 but non-poor in 2000. DLLM show that this interval encompasses the poverty rate of 10.1 percent based on true panel data.

III. Point Estimates on Poverty Mobility in a Generalized Framework

Despite its relevance for identifying bound estimates on poverty dynamics at the household level, the method introduced in DLLM suffers from a couple of drawbacks. First, unless one is content to work with the extreme case of ρ in the [0,1] range, a group of countries with actual panel data that is comparable to the country under investigation must be found so that a more reasonable empirical range of values for ρ can be identified. This task would require a certain degree of homogeneity for these countries, since ρ may vary depending on a host of factors caused by any difference ranging from economic structures to modeling methods and survey designs. Complicating this issue even further, for the same country, ρ is likely to be different for different household welfare outcomes; thus an appropriate range of correlation term for, say, household food consumption must be estimated separately from that for household non-food consumption. Over time, ρ might also change.

Second, this method only provides bound estimates rather than point estimates. While some bound estimates are certainly useful in the absence of true panel data, they also leave

room for accuracy improvement, since there always exists a tradeoff between accuracy and encompassment with the bound approach: a larger bound interval is more likely to encompass the true rates but will be less accurate, and vice versa.

Thus in this section we generalize the method introduced in DLLM by offering a theoretical framework to obtain estimates for ρ using a country's own cross sectional surveys and discuss the asymptotic properties of the new point estimators before we extend it to settings with three survey rounds or more. We maintain the DLLM framework discussed in the previous section.

III.1. Theoretical Estimates for ρ

We offer the following proposition to obtain the simple correlation coefficient between household consumption in two survey rounds $\rho_{y_{i1}y_{i2}}$, which is closely related to ρ.

Proposition 1- Approximate estimation of $\rho_{y_{i1}y_{i2}}$

Assume household consumption follows a simple linear dynamic data-generating process given by $y_{i2} = \alpha + \delta' y_{i1} + \eta_{i2}$ *(*), where* η_{i2} *is the random error term. Also assume that the sample size of each household survey round is large enough (or* $N \to \infty$*), the number of age cohorts (C) constructed from the survey data is fixed, and the age cohort dummy variables satisfy the relevance and exogeneity criteria for instrumental variables for* y_{i1} *in (*). 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 age cohorts constructed from the household survey data.*

Proof

See Appendix 1.

In the absence of true panel data, we do not observe y_{i1} for the same household with household consumption in period 2, but we can predict it by projecting household consumption in period 1 on the cohort dummy variables; this process is equivalent to an instrumental variables (IV) estimation where the instrumental variables are the age cohorts,

and practically results in taking period-by-period sample averages within the head of household i's cohort (Verbeek, 2008). Thus a consistent estimate for $\rho_{y_{i1}y_{i2}}$ (and δ) requires that the instruments age cohort dummy variables are relevant (i.e., statistically significant in the regression of household consumption on themselves) and exogenous (i.e., being uncorrelated with the error term η_{i2}).

While the former assumption can be easily checked given the available data, the latter needs to be assumed and depends on our assumption or prior knowledge about the specific data under consideration. For example, while we expect the latter assumption holds in most contexts, it would be violated if there are cohort effects in the random error term η_{i2} (see, for example, Moffitt (1993) or McKenzie (2004)).[8] Furthermore, in addition to the relevance and exogeneity conditions, good instruments also need to be strong for unbiased estimates (Stock and Yogo (2005)), which effectively requires age cohort dummy variables to be strongly correlated with household consumption; fortunately, as with the relevance condition, this additional condition can be easily checked using the cross sections.

Note that the assumptions stated in Proposition 1 on large cohort sizes are standard in the traditional pseudo-panel literature (e.g., type 1 asymptotics in Verbeek (2008)) and helps preclude measurement errors with cohort means. Clearly, the implicit assumption underlying pseudo-panel analysis is that the cohort means of household consumptions change across cohorts as well as over time, or put differently, the cohort means of household consumption are expected to capture poverty mobility at the household level in our contexts. It is theoretically possible that poverty mobility can happen mostly within cohorts and thus

[8] On a related note, surveys that focus on a particular age group of the population (e.g., youth surveys) or with particular designs (e.g., oversampling certain age cohorts) are unlikely to provide consistent estimates for the whole population.

results in underestimation of mobility since cohort means capture little of poverty mobility at the household level; however, whether this hypothesis holds true is an empirical issue that can easily be checked in our subsequent validation of synthetic panel estimates against those based on true panel data.[9]

While it appears reasonable to assume that N tending to infinity given the usually large number of households interviewed with current household surveys, there seems to be no current consensus in the literature on how large n_c should be.[10] Thus Proposition 1 provides an approximation of $\rho_{y_{i1}y_{i2}}$ based on asymptotic theory, and how well this approximation turns out to be in practice is an empirical issue. We will see later in our empirical estimates using household surveys that estimation results are rather encouraging. Furthermore, Corollary 2.1 to be discussed below will provide a lower value for $\rho_{y_{i1}y_{i2}}$ as a check on our cohort estimate.

Armed with an estimate for $\rho_{y_{i1}y_{i2}}$, we can then proceed to propose an estimate for the partial correlation coefficient ρ, which in turn helps provide the point estimate for poverty mobility. We also provide an upper value on ρ as a robustness check on our estimates for this parameter.

Proposition 2- Point estimate of ρ

Let R_j^2, for j= 1, 2, respectively represent the coefficients of determination obtained from estimating equations (1) and (2), and x_i represent the vector of household time-invariant characteristics. The partial correlation coefficient ρ can be estimated by

[9] In the extreme case, poverty mobility can happen entirely within cohorts and thus does not change cohort means in the aggregate. But we would expect this phenomenon to be rare in practice. Furthermore, this case would be easily detected since it will result in the synthetic panel cohort-level simple correlation coefficient $\rho_{y_{c1}y_{c2}}$ being equal to 0 (i.e., since cohort means remain unchanged over time). We would like to thank David McKenzie for suggesting this point to us.

[10] Monte Carlo simulations by Verbeek and Nijman (1992) suggests that cohort sizes of 100 to 200 are sufficient, but a recent study by Devereux (2007) suggests that n_c should be as large as 2000 or even more.

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

Corollary 2.1- Another approximation of ρ

If $\beta_1 \approx \beta_2$, the partial correlation coefficient ρ can also be estimated by

$$\rho = \frac{\rho_{y_{i1}y_{i2}} - \sqrt{R_1^2 R_2^2}}{\sqrt{1-R_1^2}\sqrt{1-R_2^2}} \tag{6}$$

Corollary 2.2- Upper value of ρ

Assume that the error terms ε_{ij} in equations (1) and (2) follows the traditional household effects model and can be broken down as $\varepsilon_{ij} = u_i + v_{ij}$ where conditional on the observed household characteristics, the unobserved household effects u_i has a normal distribution with mean 0 and variance σ_u^2, the idiosyncratic error terms v_{i1} and v_{i2} both have a normal distribution with means 0's and variance σ_v^2, and the covariance between v_{i1} and v_{i2} is 0. An upper value for the partial correlation coefficient ρ is given by the simple correlation coefficient $\rho_{y_{i1}y_{i2}}$.

Corollary 2.3- Lower value of $\rho_{y_{i1}y_{i2}}$

The simple correlation coefficient $\rho_{y_{i1}y_{i2}}$ for household consumption between the two survey rounds is greater than or equal to its lower value

i) $$\frac{\beta_1' \text{var}(x_i)\beta_2}{\sqrt{\text{var}(y_{i1})\,\text{var}(y_{i2})}} \text{ or} \tag{7}$$

ii) $$\sqrt{R_1^2 R_2^2} \text{ if } \beta_1 \approx \beta_2 \tag{8}$$

with equality occurring when the estimation model fully captures all the variations in the dependent variable (i.e., all the error terms are zero).

Proof

See Appendix 1.

It is useful to make several remarks here. First, while the first way of estimating ρ given in (5) naturally follows from our framework provided by equations (1) and (2) and provides more accurate results, the second way of estimating ρ in (6) is somewhat neater and perhaps more amenable to interpretation. It suggests that, given the estimated parameters in equations (1) and (2) 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. In our validation exercise to be discussed later, these two formulae give very similar estimates for ρ.[11]

Second, the variance-covariance matrix of the time-invariant household characteristics $\text{var}(x_i)$ in expression (5) is the same for each round of true panel data, but can vary for the cross sectional surveys. It thus may be useful to check one's data to make sure these matrices are similar; otherwise, one may 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 make almost no difference to our estimate whether we use the one from the first survey round or from the second survey round.[12]

Third, the assumption on the traditional household random effects in Corollary 2.2 is rather standard and should be satisfied as long as the variances of ε_{ij} are similar. As long as there are unobserved household characteristics that are not controlled for in the regression, σ_u^2 will be positive. However, the more prediction power the model has due to inclusion of previously unobserved household time-invariant characteristics (e.g., through better data

[11] Strictly speaking, we only require $\beta_1\beta_2^{'} \approx \beta_2\beta_1^{'}$ instead of assuming $\beta_1 \approx \beta_2$ for Corollary 2.1, but we make the above assumption for convenience. Note that for the three-variable case, the two formulae in (5) and (6) are identical. Also note that 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 much more calculations than the given formulae above, thus we do not discuss it further.

[12] We abuse the notations var(x) and var(y) 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.

collection), the less variance these unobserved household characteristics have or the smaller σ_u^2 is.

Fourth, we do not estimate ρ the same way we do with $\rho_{y_{i1}y_{i2}}$ as in Proposition 1, but need go through one more step with Proposition 2. The reason is straightforward once we recall that the cohort aggregation method in Proposition 1 is akin to an instrumental variable method where the cohort dummy (or age) variables work as the instruments (Moffitt, 1993). Thus, since the predicted error terms obtained from equations (1) and (2) are netted of age (and other time-invariant characteristics), when these error terms are aggregated by cohorts again, they would tend to zero.[13] On the other hand, we do not estimate ρ using the same procedures in Proposition 1, by, say, leaving out the age variables in estimating equations (1) and (2) since ρ obtained this way is different by construction from (and will overestimate) the partial correlation coefficient we are interested in.

Finally, it is rather straightforward to see that, given our assumption that ρ is non-negative, the numerators in equations (5) and (6) are also non-negative, thus leading to Corollary 2.3.[14] Since the lower values of $\rho_{y_{i1}y_{i2}}$ provided by Corollary 2.3 are derived in a different way, these represent a robustness check on our estimate based on the cohort

[13] In fact, an informal check on a lower value for the partial correlation coefficient can be done by just implementing the same procedures in Proposition 1, where the predicted error terms are obtained from estimating equations (1) and (2) (including the age variable). However, as discussed above, while this estimated partial correlation can provide some value for checking purposes, there is no guarantee that it will be statistically different from 0. See Lanjouw and Ruiz (2012) for an extension of the non-parametric approach introduced in DLLM to construct synthetic panels on data from EU countries that calculates ρ in this way.

[14] Note that the seemingly intuitive result that the partial correlation coefficient should be less than or equal to the simple correlation coefficient (e.g., since the former results when all possible correlation with other household characteristics are removed from the latter) stated in Corollary 2.2 may only hold in our context of the multiple correlation between household consumption and household characteristics, but not in general. For example, where the R^2's in expression (6) are not multiple correlation coefficients but just bivariate correlation coefficients, they can take on negative values that will invalidate this equality. This is the well-known suppression problem in the statistics literature; see, for example, Friedman and Wall (2005) for a recent discussion.

analysis in Proposition 1 above. While these "tests" may not be powerful in the sense that they can give a tight estimate just under $\rho_{y_{i1}y_{i2}}$, at least they can provide some assurance that our cohort estimate of $\rho_{y_{i1}y_{i2}}$ should satisfy a lower bound estimate and provide a positive estimate for ρ. Also a practical use of expression (8) is, given the R^2's from two cross sections, we can use their geometric mean as a lower value to quickly gauge the strength of the simple correlation between household consumption in the two periods. A similar use of Corollary 2.1 when we know $\rho_{y_{i1}y_{i2}}$ provides a shorthand calculation for the partial correlation ρ.

III.2. Further Discussion on Point Estimates and Poverty Dynamics

To fully characterize the distribution for the point estimates in (4), we provide below its asymptotics.

Proposition 3- Asymptotic results for point estimates for two periods
Assuming that household consumption can be explained by household characteristics as stated in equations (1) and (2) and all the standard regularity conditions are satisfied for each equation (i.e., $X'\varepsilon / N \xrightarrow{p} 0$ *and* $X'X / N \xrightarrow{p} M$ *finite and positive definite).*[15] *Let P represent household i's (i=1,…, N) quantity of poverty dynamics (e.g.,* $P = P(y_{i1} < z_1 \text{ and } y_{i2} > z_2)$*), 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 $\rho_{y_{i1}y_{i2},d} = d_1 d_2 \rho_{y_{i1}y_{i2}}$, *our point estimates are distributed as*

$$\sqrt{n}\left[P - \hat{\Phi}_2\left(d_1 \frac{z_1 - \hat{\beta}_1' x_{ij}}{\hat{\sigma}_{\varepsilon_1}}, d_2 \frac{z_2 - \hat{\beta}_2' x_{ij}}{\hat{\sigma}_{\varepsilon_2}}, \hat{\rho}_d \right) \right] \sim N(0, V) \qquad (9)$$

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 $V = \Sigma_s + \Sigma_m$.

[15] 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.

The first component Σ_s *is due to the sampling errors and can be estimated using the bootstrap method. The second component* Σ_m *is due to the model errors and can be estimated as* $\sum_{m=1}^{2}\nabla'_{\hat{\beta}_m}V(\hat{\beta}_m)\nabla_{\hat{\beta}_m}+\sum_{m=1}^{2}\nabla'_{\hat{\sigma}_{\varepsilon_m}}V(\hat{\sigma}_{\varepsilon_m})\nabla_{\hat{\sigma}_{\varepsilon_m}}+\nabla'_{\hat{\rho}_{y_{i1}y_{i2},d}}V(\hat{\rho}_{y_{i1}y_{i2},d})\nabla_{\hat{\rho}_{y_{i1}y_{i2},d}}$ *where*

$$\nabla_{\hat{\beta}_m}=d_m\left(\frac{-x_{ij}}{\hat{\sigma}_{\varepsilon_m}}\right)\phi\left(d_m\frac{z_m-\hat{\beta}_m'x_{ij}}{\hat{\sigma}_{\varepsilon_m}}\right)\Phi\left(\frac{d_n\frac{z_n-\hat{\beta}_n'x_{ij}}{\hat{\sigma}_{\varepsilon_n}}-\hat{\rho}_d d_m\frac{z_m-\hat{\beta}_m'x_{ij}}{\hat{\sigma}_{\varepsilon_m}}}{\sqrt{1-\hat{\rho}_d^2}}\right)$$
$$-\frac{d_m d_n\,\mathrm{var}(x_{ij})\hat{\beta}_n}{\hat{\sigma}_{\varepsilon_m}\hat{\sigma}_{\varepsilon_n}}\phi_2\left(d_m\frac{z_m-\hat{\beta}_m'x_{ij}}{\hat{\sigma}_{\varepsilon_m}},d_n\frac{z_n-\hat{\beta}_n'x_{ij}}{\hat{\sigma}_{\varepsilon_n}},\hat{\rho}_d\right) \tag{10}$$

$$\nabla_{\hat{\sigma}_{\varepsilon_m}}=\left(-d_m\frac{z_m-\hat{\beta}_m'x_{ij}}{\hat{\sigma}^2_{\varepsilon_m}}\right)\phi\left(d_m\frac{z_m-\hat{\beta}_m'x_{ij}}{\hat{\sigma}_{\varepsilon_m}}\right)\Phi\left(\frac{d_n\frac{z_n-\hat{\beta}_n'x_{ij}}{\hat{\sigma}_{\varepsilon_n}}-\hat{\rho}_d d_m\frac{z_m-\hat{\beta}_m'x_{ij}}{\hat{\sigma}_{\varepsilon_m}}}{\sqrt{1-\hat{\rho}_d^2}}\right)$$
$$-\left(d_m d_n\frac{\rho_{y_{im}y_{in}}\sqrt{\mathrm{var}(y_{im})\,\mathrm{var}(y_{in})}-\beta_m'\mathrm{var}(x_i)\beta_n}{\hat{\sigma}^2_{\varepsilon_m}\hat{\sigma}_{\varepsilon_n}}\right)\phi_2\left(d_m\frac{z_m-\hat{\beta}_m'x_{ij}}{\hat{\sigma}_{\varepsilon_m}},d_n\frac{z_n-\hat{\beta}_n'x_{ij}}{\hat{\sigma}_{\varepsilon_n}},\hat{\rho}_d\right) \tag{11}$$

$$\nabla_{\hat{\rho}_{y_{i1}y_{i2},d}}=\frac{d_1d_2\sqrt{\mathrm{var}(y_{i1})\,\mathrm{var}(y_{i2})}}{\hat{\sigma}_{\varepsilon_1}\hat{\sigma}_{\varepsilon_2}}\phi_2\left(d_1\frac{z_1-\hat{\beta}_1'x_{ij}}{\hat{\sigma}_{\varepsilon_1}},d_2\frac{z_2-\hat{\beta}_2'x_{ij}}{\hat{\sigma}_{\varepsilon_2}},\hat{\rho}_d\right) \tag{12}$$

with $n=3-m$, $V(\hat{\beta}_1)$ *and* $V(\hat{\beta}_2)$ *being respectively the estimated asymptotic covariance-variance matrix for the estimated coefficients obtained from equations (1) and (2),* $V(\hat{\sigma}_{\varepsilon_m})$ *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 Proposition 1.*

Proof
See Appendix 1.

A couple of remarks are in order here. First, this decomposition of the error terms is similar in spirit to the decomposition in Elbers, Lanjouw, and Lanjouw (ELL) (2003), which is related to the familiar formula in sampling statistics that the mean squared error of the

estimate is composed of the variance of the estimate plus its bias squared. However, the key difference between that study and ours is that ELL impute data from a survey into a census, but we are imputing data from a survey to a survey. Thus while ELL do not need to take into account the sampling errors, we have to model these sampling errors explicitly. An alternative to this analytical error would be the bootstrap error.

Second, the model variance will be smaller the better fit we have for our regressions in equations (1) and (2). And the larger the sample sizes used for prediction, the further the sampling variance can be reduced; thus, this points to the advantages of cross sections over panel data when the former has 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.[16] 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. In our validation experience, the sampling variance is significantly larger than the model variance, which is consistent with practical experience in the small area estimation literature (see, for example, Rao (2003, p. 35)).

Finally, the formulae in Proposition 3 are general and can be used to obtain our estimates using data either from the first or the second survey round, where the subscript *j* in x_{ij} should be adjusted accordingly to indicate data for the corresponding survey round. Estimates using either survey rounds are theoretically equivalent since the following identity always holds $P(y_{i1} < z_1 \; and \; y_{i2} > z_2) \equiv P(y_{i2} > z_2 \; and \; y_{i1} < z_1)$. Also note that while the

[16] Since we are mostly interested in estimating the means, assuming that the sample sizes of the cross sections are similar, we can simply use the corresponding population weight for each cross section. For estimates of other quantities such as totals, the population weights when pooling two cross sections can be adjusted by dividing by half; see, for example, Botman and Jack (1995) for a related discussion with the National Health Interview Surveys, and Kish (1999, 2002) for overviews on combining surveys.

number of households is the same for true panel data across survey rounds, it can vary for cross sectional data, thus the number of observations (N) in the variance approximation for $V(\hat{\sigma}_{\varepsilon_m})$ should be adjusted accordingly.

Effective policy reduction strategies require a good understanding of the proportions of the population that remains in certain poverty statuses in both periods, as well as the proportions of the population that move into or out of poverty in one period given their poverty status in the other period. Roughly speaking, the former outcomes include absolute numbers of poverty proportions (or joint probabilities) such as chronic poverty rates, while the latter outcomes include relative numbers of poverty proportions (or conditional probabilities) such as the proportion of the poor in the first period that exit poverty in the second period. The former outcomes thus provide a simultaneous view of poverty dynamics over time, but the latter outcomes emphasize its sequential nature, and both measures combined would provide a rich picture of poverty dynamics. We thus provide the following Corollary to Proposition 3 to provide the asymptotic results for such cases.

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

Given the same assumptions in Proposition 3, let P_{i1} and $P_{i,12}$ respectively represent household i's (i=1,…, N) quantities of poverty dynamics in period j (j= 1, 2) and both periods (e.g., $P_{ij} = P(y_{ij} < z_j)$ and $P_{i,12} = P(y_{i1} < z_1 \text{ and } y_{i2} > z_2)$), 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 $\rho_d = d_1 d_2 \rho$. And let the sampled averaged estimated quantities of poverty dynamics represented by $\hat{\bar{\Phi}}_2(.) = \frac{1}{N}\sum_{i=1}^{N} \Phi_2\left(d_1 \frac{z_1 - \hat{\beta}_1' x_{ij}}{\hat{\sigma}_{\varepsilon_1}}, d_2 \frac{z_2 - \hat{\beta}_2' x_{ij}}{\hat{\sigma}_{\varepsilon_2}}, \hat{\rho}_d \right)$,

$\hat{\bar{\Phi}}(.) = \frac{1}{N}\sum_{i=1}^{N} \hat{\Phi}\left(d_j \frac{z_j - \hat{\beta}_j' x_{ij}}{\hat{\sigma}_{\varepsilon_j}} \right)$, *our point estimates are distributed as*

$$\sqrt{n}\left[\frac{P_{i,12}}{P_{ij}}-\frac{\hat{\bar{\Phi}}_2\left(d_1\frac{z_1-\hat{\beta}_1'x_{ij}}{\hat{\sigma}_{\varepsilon_1}},d_2\frac{z_2-\hat{\beta}_2'x_{ij}}{\hat{\sigma}_{\varepsilon_2}},\hat{\rho}_d\right)}{\hat{\bar{\Phi}}\left(d_j\frac{z_j-\hat{\beta}_j'x_{ij}}{\hat{\sigma}_{\varepsilon_j}}\right)}\right]\sim N(0,V_r) \qquad (13)$$

where the covariance-variance matrix V_r can be estimated as

$$V_r=\left(\frac{\hat{\bar{\Phi}}_2(.)}{\hat{\bar{\Phi}}(.)}\right)^2\left[\frac{Var(\hat{\Phi}_2(.))}{\left(\hat{\bar{\Phi}}_2(.)\right)^2}+\frac{Var(\hat{\Phi}(.))}{\left(\hat{\bar{\Phi}}(.)\right)^2}-2\frac{Cov(\hat{\Phi}_2(.),\hat{\Phi}(.))}{\hat{\bar{\Phi}}_2(.)\hat{\bar{\Phi}}(.)}\right] \qquad (14)$$

where $Var(\hat{\Phi}(.))$ can be decomposed into a model error Σ_{jm} and a sampling error Σ_{js}. The model error can be estimated as $\Sigma_{jm}=\nabla_{\hat{\beta}_j}^{'}V(\hat{\beta}_j)\nabla_{\hat{\beta}_j}+\nabla_{\hat{\sigma}_{\varepsilon_j}}^{'}V(\hat{\sigma}_{\varepsilon_j})\nabla_{\hat{\sigma}_{\varepsilon_j}}$ with

$\nabla_{\hat{\beta}_j}=d_j\left(\frac{-x_{ij}}{\hat{\sigma}_{\varepsilon_j}}\right)\phi\left(d_j\frac{z_j-\hat{\beta}_j'x_{ij}}{\hat{\sigma}_{\varepsilon_j}}\right)$ *and* $\nabla_{\hat{\sigma}_{\varepsilon_j}}=-d_j\left(\frac{z_j-\hat{\beta}_j'x_{ij}}{\hat{\sigma}_{\varepsilon_j}^2}\right)\phi\left(d_j\frac{z_j-\hat{\beta}_j'x_{ij}}{\hat{\sigma}_{\varepsilon_j}}\right)$.

Proof

See Appendix 1.

Note that we estimate $\bar{\Phi}(.)$ using equations (1) or (2) 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 $\bar{\Phi}_2(.)$. If the model has good fit, $\bar{\Phi}(.)$ would be very similar to the sample-based poverty headcount ratio but has much smaller variance.[17] However, since we have to estimate both the numerators and denominators in the ratios (and their standard errors), this would reduce the accuracy of our estimates compared to those for the absolute quantities of poverty dynamics provided in Proposition 3.

III.3. Method for Three Periods or More

We now generalize this method to the general setting where there are three or even more rounds of survey data. More generally, assume that there are *k* rounds of survey data and

[17] Another practical implication is that if we divide $\hat{\bar{\Phi}}_2(.)$ by the sample poverty rate instead of $\hat{\bar{\Phi}}(.)$, this ratio can be larger than 100 percent when we consider estimates for certain subpopulation groups.

household consumption levels can be explained by household characteristics for survey round j^{th} in the following equations

$$y_{ij} = \beta_j' x_{ij} + \varepsilon_{ij} \tag{15}$$

where j= 1,…, k. We are interested in knowing such quantities as

$$P(y_{i1} \sim z_1 \; and \; y_{i2} \sim z_2, y_{i3} \sim z_3, ..., y_{ik} \sim z_k) \tag{16}$$

where z_j is the poverty line in period j and the relation sign ($\sim$) indicates either the larger sign (>) or smaller sign (<).

It is rather straightforward to see that the formula to calculate such quantities in (16) on data from the j^{th} survey round is the generalized version of (4) as follows

$$P(y_{i1} \sim z_1 \; and \; y_{i2} \sim z_2, ..., y_{ik} \sim z_k) = \Phi_k\left(d_1 \frac{z_1 - \beta_1' x_{ij}}{\sigma_{\varepsilon_{i1}}}, d_2 \frac{z_2 - \beta_2' x_{ij}}{\sigma_{\varepsilon_{i2}}}, ..., d_k \frac{z_k - \beta_k' x_{ij}}{\sigma_{\varepsilon_{ik}}}, \Sigma_\rho \right) \tag{17}$$

where $\Phi_k(.)$ stands for the k-variate normal cumulative distribution function, and d_j, with *j= 1,…., k* an indicator variable that equals 1 when y_{ij} is smaller than the corresponding poverty line in the same period (i.e., household *i* is poor in period *j*), and equals -1 otherwise.

Note that the matrix Σ_ρ of partial correlation coefficient is symmetric and is represented by

$$\Sigma_\rho = \begin{pmatrix} 1 & . & . & . & . & . \\ \rho_{d12} & 1 & . & . & . & . \\ \rho_{d13} & \rho_{d23} & 1 & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \rho_{d1k} & . & . & . & . & 1 \end{pmatrix}_{kxk}$$ where the subscripts *jj* indicates the particular two survey rounds under consideration and all the elements on the diagonal are 1s. ρ_{djl} stands for the

correlation coefficient between equation j and l and equals $d_j * d_l * \rho_{jl}$; there are such $\frac{k(k-1)}{2}$ correlation coefficients in the correlation coefficient matrix Σ_ρ.

However, 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 be likely to be less accurate for three periods, and longer periods in general, than those for two periods due to increased layers of (modeling and sampling) errors.

We can then generalize Proposition 3 to any setting with two periods or more. Estimates on the relative quantity of poverty dynamics can be obtained by extending the result in Corollary 3.1 to more periods, but again, note that estimates will be likely to be less accurate the more periods we consider.

Proposition 4- Asymptotic results for point estimates for k periods

Assuming that household consumption can be explained by household characteristics as stated in the following equations

$$y_{ij} = \beta_j' x_{ij} + \varepsilon_{ij} \qquad (18)$$

, j= 1,…, k and all the standard regularity conditions are satisfied for each equation (i.e., $X'\varepsilon / N \xrightarrow{p} 0$ *and* $X'X / N \xrightarrow{p} M$ *finite and positive definite). Let P represent household i's (i=1,…, N) quantity of poverty dynamics (e.g.,* $P(y_{i1} \sim z_1, y_{i2} \sim z_2, y_{i3} \sim z_3, ..., y_{ik} \sim z_k)$*), 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*

$$\sqrt{n}\left[P - \Phi_k\left(d_1 \frac{z_1 - \hat{\beta}_1' x_{ij}}{\hat{\sigma}_{\varepsilon_1}}, ..., d_k \frac{z_k - \hat{\beta}_k' x_{ij}}{\hat{\sigma}_{\varepsilon_k}}, \hat{\Sigma}_\rho \right) \right] \sim N(0, V) \qquad (19)$$

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.*

To make notations less cluttered, let $\beta_{(jxl)}$ *represent the matrix of estimated coefficients obtained from (18),* $\Phi_k(.)$ *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* $\bar{\hat{a}}_{dmnj} = \frac{(\hat{\rho}_{dmq} - \hat{\rho}_{dnq}\hat{\rho}_{dmn})\hat{a}_{dmj} + (\hat{\rho}_{dnq} - \hat{\rho}_{dmq}\hat{\rho}_{dmn})\hat{a}_{dnj}}{\sqrt{1-\hat{\rho}^2_{dmn}}}$ *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* $\hat{\rho}_{dst.m} = \frac{\hat{\rho}_{dst} - \hat{\rho}_{dsm}\hat{\rho}_{dtm}}{\sqrt{1-\hat{\rho}^2_{dsm}}\sqrt{1-\hat{\rho}^2_{dtm}}}$ *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* $\hat{\rho}_{dst.mn} = \frac{\hat{\rho}_{dst.m} - \hat{\rho}_{dsn.m}\hat{\rho}_{dtn.m}}{\sqrt{1-\hat{\rho}^2_{dsn.m}}\sqrt{1-\hat{\rho}^2_{dtn.m}}}$ *for s, t= 1,…, k and* $s,t \neq m,n$ *. The second component* Σ_m *is due to the model errors and can be estimated as*

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

where

$$\nabla_{\hat{\beta}_m} = d_m \left(\frac{-x_{ij}}{\hat{\sigma}_{\varepsilon_m}} \right) \phi(\hat{a}_{dmj}) \Phi_{k-1} \left(\frac{\hat{a}_{d1j} - \hat{\rho}_{dm1}\hat{a}_{dmj}}{\sqrt{1-\hat{\rho}^2_{dm1}}}, \ldots, \frac{\hat{a}_{dkj} - \hat{\rho}_{dmk}\hat{a}_{dmj}}{\sqrt{1-\hat{\rho}^2_{dmk}}}, \hat{\Sigma}_{\rho(-m)} \right) +$$

$$+ \sum_{\substack{n=1 \\ n \neq m}}^{k} d_m d_n \frac{-\mathrm{var}(x_{ij})\hat{\beta}_n}{\hat{\sigma}_{\varepsilon_m}\hat{\sigma}_{\varepsilon_n}} \phi_2(\hat{a}_{dmj}, \hat{a}_{dnj}, \hat{\rho}_{dmn}) \Phi_{k-2}(\hat{a}_{d1j} - \bar{\hat{a}}_{dm1j}, \ldots, \hat{a}_{dkj} - \bar{\hat{a}}_{dmkj}, \hat{\Sigma}_{\rho(-m,-n)}) \quad (20)$$

$$\nabla_{\hat{\sigma}_{\varepsilon_m}} = \left(\frac{-\hat{a}_{dmj}}{\hat{\sigma}_{\varepsilon_m}} \right) \phi(\hat{a}_{dmj}) \Phi_{k-1} \left(\frac{\hat{a}_{d1j} - \hat{\rho}_{dm1}\hat{a}_{dmj}}{\sqrt{1-\hat{\rho}^2_{dm1}}}, \ldots, \frac{\hat{a}_{dkj} - \hat{\rho}_{dmk}\hat{a}_{dmj}}{\sqrt{1-\hat{\rho}^2_{dmk}}}, \hat{\Sigma}_{\rho(-m)} \right) -$$

$$- \sum_{\substack{n=1 \\ n \neq m}}^{k} \left(d_m d_n \frac{\rho_{y_{im}y_{in}} \sqrt{\mathrm{var}(y_{im})\,\mathrm{var}(y_{in})} - \beta_m' \mathrm{var}(x_i) \beta_n}{\hat{\sigma}^2_{\varepsilon_m}\hat{\sigma}_{\varepsilon_n}} \right) \phi_2(\hat{a}_{dmj}, \hat{a}_{dnj}, \hat{\rho}_{dmn}) * \quad (21)$$

$$* \Phi_{k-2}(\hat{a}_{d1j} - \bar{\hat{a}}_{dm1j}, \ldots, \hat{a}_{dkj} - \bar{\hat{a}}_{dmkj}, \hat{\Sigma}_{\rho(-m,-n)})$$

$$\nabla_{\hat{\rho}_{y_{im}y_{in},d}} = \frac{d_m d_n}{\sqrt{1-R^2_m}\sqrt{1-R^2_n}} * \phi_2(\hat{a}_{dmj}, \hat{a}_{dnj}, \hat{\rho}_{dmn}) * \Phi_{k-2}(\hat{a}_{d1j} - \bar{\hat{a}}_{dm1j}, \ldots, \hat{a}_{dkj} - \bar{\hat{a}}_{dmkj}, \hat{\Sigma}_{\rho(-m,-n)}) \quad (22)$$

with $V(\hat{\beta}_m)$ *being the asymptotic covariance-variance matrix for the estimated coefficients obtained from the corresponding equation in (18) and* $V(\hat{\sigma}_{\varepsilon_m})$ *being approximated by* $\frac{(8N-7)\hat{\sigma}^2_{\varepsilon_m}}{(4N-3)^2}$.

III.4. Estimation Procedures

A practical concern that we did not yet discuss is whether or not equations (1) and (2) 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 specifically based on the dependent variables (informative sampling); on the other hand, unweighted regressions are most relevant when the proposed super-population (i.e., equations (1) and (2)) 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).[18] Thus it seems advisable to estimate models both with and without weights and compare results, particularly where there is limited information on how the weights have been constructed.

Given the framework discussed above, 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}. Using the data in survey round 2, estimate equation (2) and obtain similar parameters $\hat{\beta}_2'$ and $\hat{\sigma}_{\varepsilon_{i2}}$.

Step 2: Aggregate data in both survey rounds 1 and 2 by age cohort and obtain the estimated cohort-level simple correlation coefficient $\hat{\rho}_{y_{i1}y_{i2}}$. Calculate $\hat{\rho}$ using Proposition 2, and check that $\hat{\rho}_{y_{i1}y_{i2}} \geq \hat{\rho}$ (and also $\hat{\rho}_{y_{i1}y_{i2}} \geq \frac{\hat{\beta}_1' \text{var}(x_i) \hat{\beta}_2}{\sqrt{\text{var}(y_{i1})\,\text{var}(y_{i2})}}$).

Step 3: For each household in survey round j, calculate absolute quantities of poverty mobility as $\Phi_2\left(d_1 \frac{z_1 - \hat{\beta}_1' x_{ij}}{\hat{\sigma}_{\varepsilon_{i1}}}, d_2 \frac{z_2 - \hat{\beta}_2' x_{ij}}{\hat{\sigma}_{\varepsilon_{i2}}}, \hat{\rho}_d \right)$, where d_j is an indicator function that

[18] See also Deaton (1997), Lorh (2010), and Pfeffermann (2011) for further discussion on this topic.

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 population-level relative quantities of poverty mobility for period j as $\frac{\hat{\Phi}_2\left(d_1 \frac{z_1 - \hat{\beta}_1' x_{ij}}{\hat{\sigma}_{\varepsilon_{i1}}}, d_2 \frac{z_2 - \hat{\beta}_2' x_{ij}}{\hat{\sigma}_{\varepsilon_{i2}}}, \hat{\rho}_d\right)}{\hat{P}_j}$, 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.

The estimation procedures for three periods or more are similar, with poverty mobility rates and standard errors estimated in Steps 3 and 4 using Proposition 4 instead of Proposition 3.[19]

IV. Data

To validate our method, we analyze household survey data from Bosnia-Herzegovina (Bosnia-Herzegovina Living Standards Measurement Survey, BLSMS), Lao PDR (Expenditure and Consumption Survey, LECS), the United States (Panel Study of Income Dynamics, PSID), Peru (Peruvian National Household Survey, ENAHO), and Vietnam (Vietnam Household Living Standards Survey, VHLSS). We use two rounds from the first two surveys and three rounds from the last three surveys, with data from the BLSMS in 2001-2004,[20] the LECS in 2002/03-2007/08, the PSIDs in 2005, 2007, and 2009, the ENAHOs in 2004, 2005, and 2006, and the VHLSSs in 2004, 2006, and 2008. The number of households hovers around 2,376 households for Bosnia-Herzegovina, 6,500 households

[19] We are working on a Stata program to calculate poverty mobility based on synthetic panels and will make it publicly available soon.

[20] We build this data based on the data from Demirguc-Kunt, Klapper and Panos (2011).

for the LECS, 9,189 households for each round of the VHLSSs, more than 5,000 households for the PSIDs,[21] and almost 20,000 households for the ENAHOs.

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 agencies, 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 in a number of developing countries 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 poverty assessment by the government and the donor 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 20th percentile of the consumption distribution in 2001 as the poverty line.

One particular feature the LECSs, VHLSSs and ENAHOs share 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

[21] We only consider the sample persons in the PSID with non-zero longitudinal weights.

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 thus 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.[22] We use income and household consumption as a household welfare measure respectively for the US and all the other countries.[23]

Appendix 3 provides a more detailed description of these surveys and other data quality checks.

V. Estimates on Poverty Dynamics

V.1. Estimates for ρ

Consistent with the literature on pseudo-panel data, we restrict household heads' age range to 25-55 for the first survey round and adjust this appropriately for later survey rounds (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.

[22] There can be both pros and cons with using true panel data versus a survey with a rotating panel design for validation purposes. On one hand, a rotating panel design may be more suitable since actual panel surveys usually have a smaller sample size than those of cross sections, and a reasonably large sample size is required to obtain accurate estimates for the cohort-level simple correlation coefficient as well as for our asymptotic results. On the other hand, an important requirement for using rotating panel surveys is that the data from the cross section component be similar to those from the panel component. For Peru, the household-weighted headcount poverty rates based on the actual panel component are around 5 percent lower than those based on the cross section component, and the population-weighted estimates are even more different. Thus while the Peruvian data are not perfect for validation purposes, we believe it is still useful to show estimates for this country using household weights. But note that this difference will introduce more noise into our estimates.

[23] The PSID also has some information on household consumption but this measure is not commonly used to measure poverty and much less comprehensive than those for other surveys.

After obtaining an estimate for $\rho_{y_{i1}y_{i2}}$ (which are all highly statistically significant with p-values less than 0.01) we calculate ρ in two different ways using Proposition 2 and Corollary 2.1., but estimation results are very similar with the differences being at most 0.01; thus, we only show the estimates based on Proposition 2 in Table 1. Overall, the absolute difference for ρ ranges from 0.01 (Bosnia-Herzegovina) to 0.18 (the US during 2005-2007), which corresponds to a range of relative differences of 2 to 28 percent.

Nevertheless, it is reassuring to see that estimates for ρ are always less than those for $\rho_{y_{i1}y_{i2}}$, which is consistent with the hypothesis in Corollary 2.2; similarly, estimates for $\rho_{y_{i1}y_{i2}}$ are also larger than its lower value estimates as can be estimated in Corollary 2.3(ii) (not shown). These results are very encouraging and suggest that our framework could well be applied to these cross sections to obtain ρ in the absence of actual panel data.

V.2. Overall Poverty Mobility

To save space, we show estimates for poverty dynamics using the latest survey rounds available, that is for Bosnia-Herzegovina during 2001-2004, Lao PDR during 2002/03-2007/08, Peru during 2005- 2006, the US during 2007-2009, and Vietnam during 2006-2008 in Table 2.[24] Estimates for earlier survey rounds for the last three countries are provided in Appendix 2, Table 2.2.

Our estimation results using true panel data and synthetic panel data are respectively displayed under the columns "True Panel" and "Synthetic Panel".[25] Results appear very encouraging with the synthetic panel point estimates being close to the true point estimates,

[24] Unless otherwise noted, we use data in the second survey round (x_{i2}) for predictions. Estimates corresponding to those in Table 2 but use data in the first survey round are provided in Appendix 2, Table 2.3. Weighted regressions provide qualitatively similar but somewhat less accurate results, thus we use unweighted regressions.

[25] The full estimated parameters for equations (1) and (2) are provided in Appendix 2, Table 2.1.

with the former lying roughly within one standard error of the latter in two fifths of the cases (i.e., 8 out of 20). If we look at the 95 percent confidence intervals around the true estimates for these two countries, more than four fifths (i.e., 17 out of 20) of the synthetic estimates would be contained in these intervals.[26] Furthermore, the difference between the true rates and our estimates seems negligible for certain mobility categories; for example, the percentage of those remaining poor in both periods for Peru during 2005-2006 were 29.9 percent and 30.9 percent respectively for the true rates and our estimates.

As discussed above, 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.[27] 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 estimates unsurprisingly have smaller standard errors than those based on true panel data. Table 2 shows that when the ratio of the sample size for the cross sections over that of the panels increases from around 1.5 for Lao PRD and Vietnam to around 4 for Peru, the standard errors for the synthetic estimates shrink even further by more than half for certain cases (e.g., for those being poor in 2005 but non-poor in 2006). For the US and Bosnia-Herzegovina, the sample size is the same for both true

[26] See also Dorfman (2011) for a formal discussion of this type of test in the context of small area estimates.
[27] We also calculate the bootstrap standard errors by bootstrapping $(y_{ij}, \mathbf{x}_{ij})$ from its empirical distribution function (1,000 times) and applying the estimated parameters for equations (1) and (2) from the original samples. Estimation results are very similar to the analytical standard errors.

panel and synthetic panel estimates, but the standard errors for the latter are smaller than those for the former.[28]

An alternative way to evaluating the synthetic panel point estimates is to use the confidence intervals for a "coverage test", which considers the share of the overlap between the 95 percent confidence intervals of the synthetic panel estimates and the true estimates over the former's 95 percent confidence interval. Instead of testing for the bias of the point estimate as in Dorfman (2011), this test focuses on the proportional contribution of its variance in the total mean squared error and scores the accuracy of each estimate on a 100 percent scale. The results for this test for two pairs of survey years for each country are shown in Table 3, where more than four fifths (i.e., 17 out of 20) and three fourths (i.e., 15 out of 20) of the synthetic panel estimates pass the 47 percent mark and the 90 percent mark respectively.

Table 4 provides the proportions of the population that exit or fall into poverty in the second period given their poverty status in the first period, using the results derived in Corollary 3.1. Estimation results are, not surprisingly, less accurate than those in Table 2 since both the numerators and denominators in the ratios in Corollary 3.1 are estimated. Due to one additional layer of estimates, now only just almost one half and three fourths of the synthetic estimates are respectively within one standard error and the 95 percent confidence intervals of the true rates. Interestingly, estimates for Lao PDR and Vietnam all respectively fall within one standard error and the 95 percent confidence intervals of the true rates.

However, for these remaining cases of inaccurate estimates for Bosnia, Peru, and the US, the relative differences with the true rates (i.e., the difference between the synthetic

[28] More generally, this is consistent with the well-known result in survey sampling that the model-based variances (synthetic panel estimates in our case) are usually smaller than the design-based variances (weighted estimates based on true panel data). See, e.g, Binder and Roberts (2009) for a recent review on this topic.

estimate and the true rate divided by the true rate) range from 3 percent (for the proportion of the non-poor in the first period that remains non-poor in the second period for Peru in 2005-2006) to 17 percent (for the proportion of the poor in the first period that escapes poverty in the second period for Peru in 2005-2006). These differences are not small, but if we lower our expectations to just provide some ballpark figures for the true rates for these cases, we believe these estimates are still helpful.

V.3. Poverty Mobility for Population Sub-Groups

As argued in DLLM, it is important to investigate poverty dynamics for population sub-groups 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 actual panel data, and thus the larger sample sizes the former has, the more accurate estimates it can bring.

We estimate and plot the estimated rates with their 95 percent confidence intervals for the absolute and relative 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) respectively for Peru in Figures 1 to 4 and Vietnam in Figures 5 to 8.[29] Clearly, these categorizations can overlap but they can provide a first cut at profiling poverty mobility for different groups, and we would expect an overlap between the synthetic

[29] We do not show all eight graphs for one country or for other countries to save space, but results are qualitatively similar.

panel estimates and the true rates for the groups whose heterogeneity mimics that of the whole country.

Not surprisingly, the 95 percent confidence intervals for synthetic panels estimates for both Peru and Vietnam 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, especially for those who are poor in either period or both periods. Except for a few cases (e.g., households where heads only achieve primary education or living in urban Selva in Figure 1, or households living in urban areas in Figure 6), there is much overlap between the true rates and our estimated rates.

V.4. Poverty Mobility in Three Periods

We turn next in Table 5 to examining our estimates on poverty mobility 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 absolute measures of poverty dynamics).[30] As discussed above, we should expect estimates to be less accurate than those for two periods; however, our proposed method turns out to work quite well for data from Vietnam with three fourths of all the point estimates being contained in the 95 percent confidence intervals around the true rates. For the US and Peru, this rate is lower with respectively more than half and exactly half of the point estimates lying in the 95 percent confidence intervals around the true rates.

[30] 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.

VI. Conclusion

For effective poverty reduction policies to take place, we need to understand well poverty mobility over time. In the absence of panel data, it has historically been difficult to study poverty mobility in developing countries. However, there are by now at least two rounds of cross-sectional household survey data available in a large majority of developing countries. Our proposed method, which generalizes that initiated by the DLLM paper, offers a means to convert these cross-sectional survey data into synthetic panel data.

In particular, by moving away from the bound estimates in the DLLM paper to point estimates, we show that our estimates are quite accurate and do not depend on additional information from ancillary data. Our method would thus seem applicable in most settings where two cross sections are available. We find that estimation results are good not only for the general population but for smaller population groups as well, and are associated with much tighter confidence intervals than even direct, panel-data based estimates in those settings where the sample sizes for the cross sections are large enough. Our estimates are validated against true panel data spanning different income levels and geographical regions. In addition, our method can be readily extended to settings with three survey rounds or more, although predictions are more accurate for shorter periods.

It should be noted that our method needs not be restricted only to the analysis of poverty mobility. It can in principle also be employed to study dynamics more generally. Possible application can involve dynamics in the areas of labor (e.g., what is the percentage of those who move from the rural farm sector to the non-farm sector?), health (e.g., what is the percentage of those who have medical insurance in the first period but do not in the second period?) or finance (e.g., what is the percentage of borrowers who do not default on loans in the first period but do in the second period?).

We come away with a growing sense that this basic methodology offers significant potential towards a better understanding of poverty dynamics in settings where panel data are absent, and can serve as a rather promising avenue for further research. For example, since measurement errors may be larger with richer households' consumption, future research may investigate the heterogeneity of the error terms for different types of households. Another potential direction is to examine the extent that the estimates offered by our synthetic panels can improve on and correct for "bad" panel data estimates resulting from serious attrition problems.

References

Agostini, Claudio A. and Philip H. Brown. (2010). "Local Distributional Effects of Government Cash Transfers in Chile". *Review of Income and Wealth*, 56(2): 366- 388.

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

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

Banks, James, Richard Blundell, and Agar Brugiavini. (2001). "Risk Pooling, Precautionary Saving and Consumption Growth". *Review of Economic Studies*, 68(4): 757-779.

Bierbaum, Mira and Franziska Gassmann. (2012). "Chronic and Transitory Poverty in the Kyrgyz Republic: What Can Synthetic Panels Tell Us?" *UNU-MERIT Working Paper #2012-064.*

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.

Botman, Steven L. and Susan S. Jack. (1995). "Combining National Health Interview Survey Datasets: Issues and Approaches". *Statistics in Medicine*, 14: 669-677.

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.

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.

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.

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

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

Cruces, Guillermo, Peter Lanjouw, Leonardo Lucchetti, Elizaveta Perova, Renos Vakis, and Mariana Viollaz. (2011). "Intra-generational Mobility and Repeated Cross-Sections: A

Three-country Validation Exercise". *World Bank Policy Research Working Paper No. 5916*. Washington DC: The World Bank.

Dang, Hai-Anh, Peter Lanjouw, Jill Luoto, and David McKenzie. (2011). "Using Repeated Cross-Sections to Explore Movements in and out of Poverty". *World Bank Policy Research Working Paper No. 5550*. Washington DC: The World Bank.

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

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

Deaton, Angus and Christina Paxson. (1994). "Intertemporal Choice and Inequality". *Journal of Political Economy*, 102(3): 437- 467.

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.

Demombynes, Gabriel and Berk Özler. (2005). "Crime and Local Inequality in South Africa," *Journal of Development Economics,* 76(2): 265–292.

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

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

Elbers, Chris, Jean O. Lanjouw, and Peter Lanjouw. (2002) "Micro-Level Estimation of Welfare". *World Bank Policy Research Working Paper # 2911*.

---. (2003). "Micro-level Estimation of Poverty and Inequality". *Econometrica*, 71(1): 355-364.

Elbers, Chris, Tomoki Fujii, Peter Lanjouw, Berk Özler, and Wesley Yin. (2007). "Poverty Alleviation Through Geographic Targeting: How Much Does Disaggregation Help?" *Journal of Development Economics*, 83: 198–213.

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.

Friedman, Lynn and Melanie Wall. (2005). "Graphical Views of Suppression and Multicollinearity in Multiple Linear Regression". *American Statistician*, 59(2): 127-136.

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.

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

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.

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.

Lanjouw, Peter and Nicolas Ruiz. (2012). "*Temporal Mapping of Poverty Using Synthetic Panel Data*". Working paper.

Little, Roderick J. A. and Donald B. Rubin. (2002). *Statistical Analysis with Missing Data*. 2nd Edition. New Jersey: Wiley.

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

Mcintosh, Steven. (2006). "Further Analysis of the Returns to Academic and Vocational Qualifications". *Oxford Bulletin of Economics and Statistics*, 68(2): 225- 251.

McKenzie, David. (2004). "Asymptotic Theory for Heterogeneous Dynamic Pseudo-Panels. *Journal of Econometrics*, 120, 235–262.

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

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

Mullahy, John. (2011). "Marginal Effects in Multivariate Probit and Kindred Discrete and Count Outcome Modes, with Applications in Health Economics". *NBER Working paper* 17588.

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.

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.

Peruvian Statistics Bureau (INEI). http://www.inei.gob.pe/srienaho/Consulta_por_Encuesta.asp Accessed October 2012.

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.

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.

PSID Main Interview User Manual: Release 2012.1. Institute for Social Research, University of Michigan, January 23, 2012. Available on the Internet at http://psidonline.isr.umich.edu/data/Documentation/UserGuide2009.pdf Accessed October 2012.

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

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

Stock, James H. and Motohiro Yogo. (2005). " Testing for Weak Instruments in Linear IV Regression." In D. W. K. Andrews and J. H. Stock. (Eds.) *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg.* Cambridge: Cambridge University Press.

Sungur, Engin A. (1990). "Dependence Information in Parameterized Copulas". *Communications in Statistics- Simulation and Computation*, 19(4): 1339-1360.

Tung, Phung Duc and Nguyen Phong. (undated). "*Vietnam Household Living Standards Surveys (VHLSSs) in 2002 and 2004- Basic Information"*. Available on the World Bank's LSMS website at http://siteresources.worldbank.org/INTLSMS/Resources/3358986-1181743055198/3877319-1207074161131/BINFO_VHLSS_02_04.pdf Accessed October 2012.

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

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

Country	Survey Year	True panel data		Synthetic panel estimates	
		P_{yi1yi2}	ρ	P_{yi1yi2}	ρ
Bosnia-Herzegovina	2001 2004	0.48	0.45	0.43	0.40
Lao PDR	2002-03 2007-08	0.51	0.43	0.56	0.46
Peru	2004 2005	0.82	0.64	0.82	0.69
	2005 2006	0.82	0.66	0.80	0.63
	2004 2006	0.79	0.63	0.73	0.51
Vietnam	2004 2006	0.81	0.66	0.85	0.73
	2006 2008	0.78	0.62	0.85	0.76
	2004 2008	0.75	0.58	0.84	0.74
United States	2005 2007	0.76	0.66	0.89	0.84
	2007 2009	0.82	0.70	0.86	0.79
	2005 2009	0.72	0.57	0.71	0.59

Note: 1. The synthetic panels estimates in panel A are based on cross sectional data, and the synthetic panel estimates in panel B are based on two rounds of true panel data.

2. Pyi1yi2 is the simple correlation across two survey rounds for household consumption for all countries except for the US; for the US it is the correlation for household income. ρ is the partial correlation between the residuals of the regression of household consumption on household head's gender, years of schooling, ethnicity, and residence areas.

3. All estimates for Pyi1yi2 are significant at the 0.01 level.

4. Household heads' ages are restricted to between 25 and 55 in the first survey round.

Table 2: 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	
	True panel	Synthetic panel	True panel	Synthetic panel	True Panel	Synthetic Panel	True Panel	Synthetic Panel	True Panel	Synthetic Panel
Poor, Poor	10.8	8.2	13.8	13.2	29.9	30.9	6.0	6.2	9.9	9.6
	(0.9)	(0.1)	(0.8)	(0.2)	(1.0)	(0.2)	(0.4)	(0.2)	(0.6)	(0.2)
Poor, Nonpoor	13.1	12.6	14.3	13.2	11.6	12.3	3.8	3.2	5.9	4.9
	(0.9)	(0.1)	(0.8)	(0.1)	(0.7)	(0.0)	(0.3)	(0.1)	(0.4)	(0.1)
Nonpoor, Poor	10.9	12.1	10.9	11.4	8.9	10.0	4.6	4.0	4.9	5.0
	(0.9)	(0.1)	(0.7)	(0.1)	(0.6)	(0.0)	(0.4)	(0.1)	(0.4)	(0.1)
Nonpoor, Nonpoor	69.2	67.2	61.0	62.2	49.7	46.8	85.7	86.6	79.3	80.4
	(1.3)	(0.3)	(1.1)	(0.3)	(1.1)	(0.3)	(0.6)	(0.3)	(0.8)	(0.3)
N	1342	1342	1989	3215	2250	9084	3368	3368	2723	3701

Note: 1. For each country, synthetic panels poverty rates are calculated using the cross section component except for Bosnia-Herzegovina and the USA. Predictions are obtained based on data in the second survey round. We use 500 bootstraps in calculating standard errors.

2. All numbers are weighted using household weights for Peru, and population weights for other countries. Poverty rates are in percent.

3. 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: Coverage Test for Poverty Dynamics Based on Synthetic Panel Data for Two Periods (Percentage)

	Bosnia-Herzegovina	Lao PDR	Peru	United States	Vietnam
Poverty status	**2001-04**	**2005-06**	**2005-06**	**2007-09**	**2006-08**
Poor, Poor	0.0	100	100	100	100
Poor, Nonpoor	100	100	100	46.9	0.0
Nonpoor, Poor	100	100	100	96.9	100
Nonpoor, Nonpoor	96.6	100	0.0	73.5	89.8

Note: 1. Coverage tests are calculated based on estimates from Table 2 and Appendix 2, Table 2.2.

Table 4: 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	
	True panel	Synthetic panel	True panel	Synthetic panel	True panel	Synthetic panel	True panel	Synthetic panel	True panel	Synthetic panel
Poor--> Poor	46.8	39.4	49.0	50.0	72.0	71.5	61.2	66.0	62.8	66.2
	(3.1)	(0.5)	(2.1)	(0.3)	(1.4)	(0.2)	(2.2)	(1.8)	(2.4)	(0.5)
Poor--> Nonpoor	57.2	60.6	51.0	50.0	28.0	28.5	38.8	34.0	37.2	33.8
	(3.1)	(0.8)	(2.1)	(0.4)	(1.4)	(0.2)	(2.2)	(0.8)	(2.4)	(0.4)
Nonpoor--> Poor	14.2	15.3	15.2	15.5	15.1	17.6	5.0	4.4	5.9	5.9
	(1.1)	(0.2)	(0.9)	(0.2)	(1.0)	(0.1)	(0.4)	(0.1)	(0.5)	(0.1)
Nonpoor--> Nonpoor	89.8	84.7	84.8	84.5	84.9	82.4	95.0	95.6	94.1	94.1
	(1.1)	(0.2)	(0.9)	(0.2)	(1.0)	(0.2)	(0.4)	(0.3)	(0.5)	(0.1)
N	1342	1342	1989	3215	2250	9084	3368	3368	2723	3701

Note: 1. For each country, synthetic panels poverty rates are calculated using the cross section component except for Bosnia-Herzegovina and the USA. Predictions are obtained based on data in the second survey round. We use 500 bootstraps in calculating standard errors.

2. All numbers are weighted using household weights for Peru, and population weights for other countries. Poverty rates are in percent.

3. 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. year difference for the second survey round.

Table 5: 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	True panel	Synthetic panel	True panel	Synthetic panel	True panel	Synthetic panel
Poor, Poor, Poor	26.6	24.0	4.0	4.0	8.1	7.6
	(1.0)	(0.2)	(0.4)	(0.2)	(0.8)	(0.2)
Poor, Poor, Nonpoor	6.9	7.1	1.4	2.0	3.1	2.8
	(0.6)	(0.0)	(0.2)	(0.0)	(0.5)	(0.0)
Poor, Nonpoor, Poor	4.4	3.3	1.0	0.5	2.3	2.9
	(0.5)	(0.0)	(0.2)	(0.0)	(0.4)	(0.0)
Poor, Nonpoor, Nonpoor	7.2	6.4	2.7	2.8	6.6	5.0
	(0.6)	(0.0)	(0.3)	(0.0)	(0.7)	(0.1)
Nonoor, Poor, Poor	3.9	6.1	1.8	1.7	0.8	1.1
	(0.4)	(0.0)	(0.2)	(0.1)	(0.2)	(0.0)
Nonpoor, Poor, Nonpoor	5.4	5.0	2.0	1.1	1.7	2.6
	(0.5)	(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
	(0.5)	(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
	(1.1)	(0.3)	(0.7)	(0.4)	(1.2)	(0.3)
N	1987	8608	3036	3036	1282	3808

Note: Poverty rates are calculated using the predicted parameters from the first, second, and third survey rounds on the data in the third survey round. We use 500 bootstraps in calculating standard errors. All numbers are weighted using population weights for the third survey round with standard errors in parentheses. Household heads' ages are restricted to between 25 and 55 in the first survey round and adjusted accordingly for other years.

Figure 1: Profiles for Those Who Remained Poor in Both Periods, Peru 2005- 2006

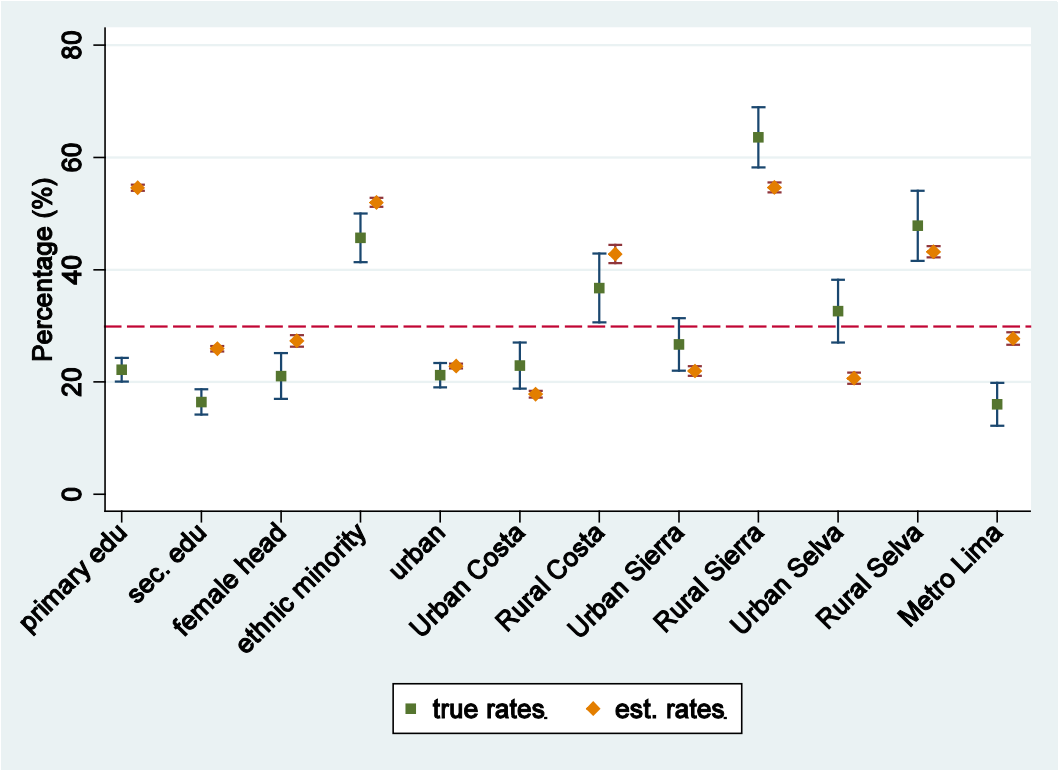


Figure 2: Profiles for Those Who Were Poor in the First Period but Non-poor in the Second Period, Peru 2005- 2006

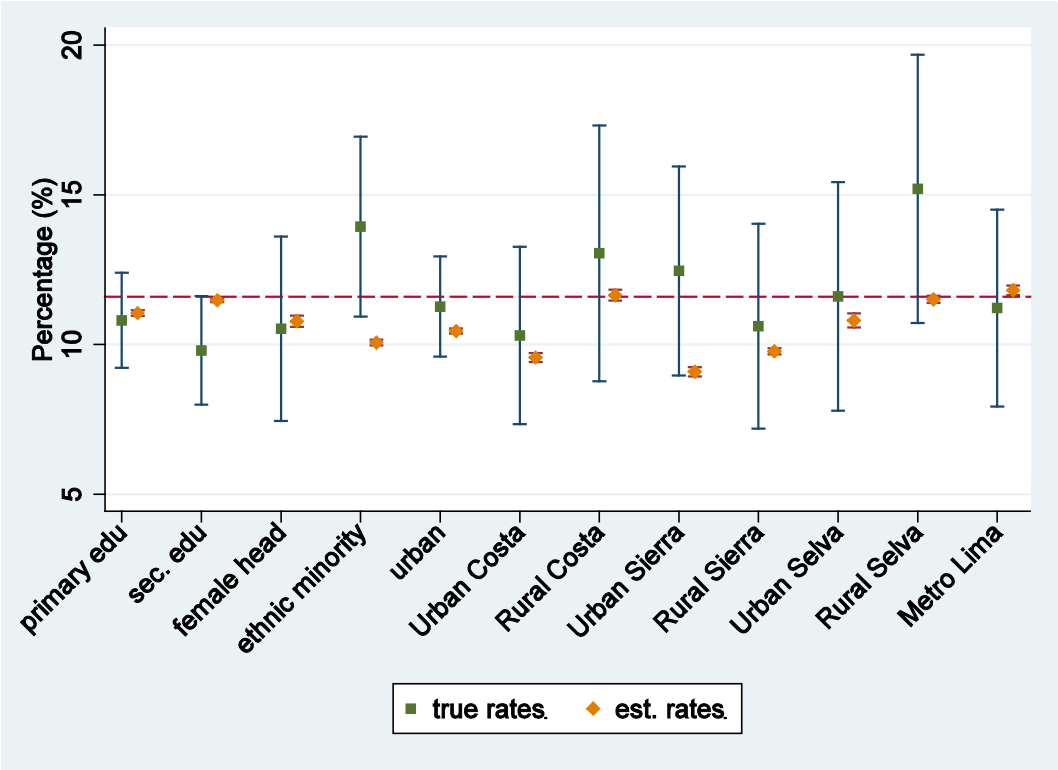


Figure 3: Profiles for Those Who Were Non-poor in the First Period but Poor in the Second Period, Peru 2005- 2006

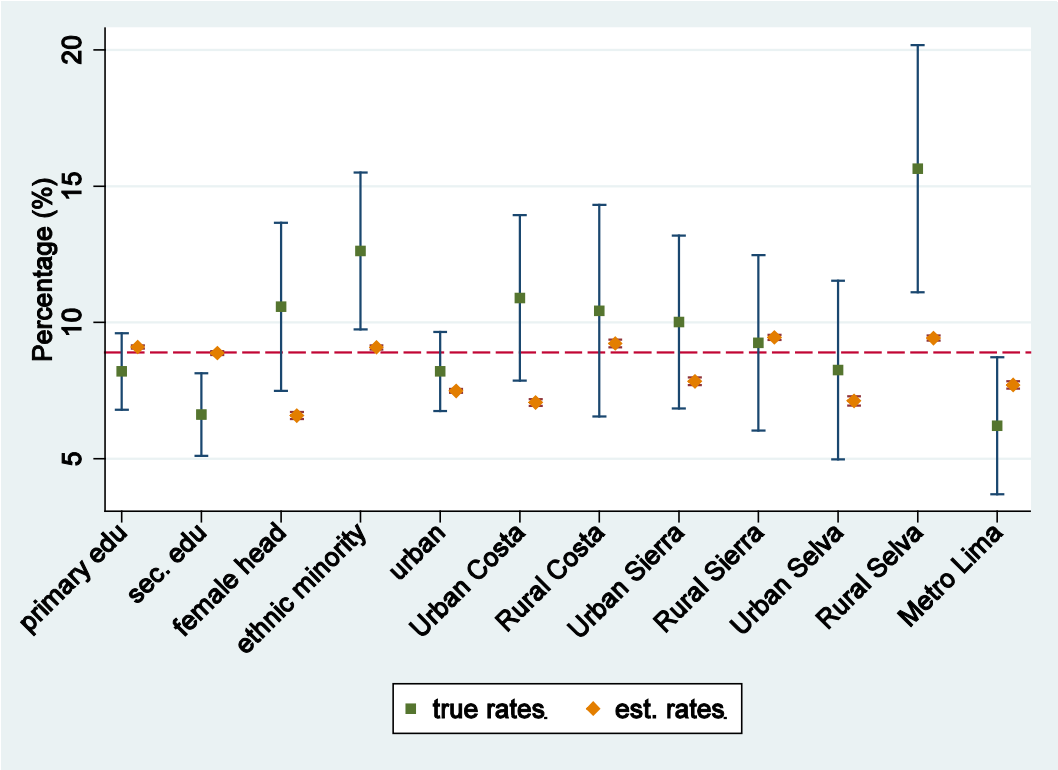


Figure 4: Profiles for Those Who Remained Non-poor in Both Periods, Peru 2005- 2006

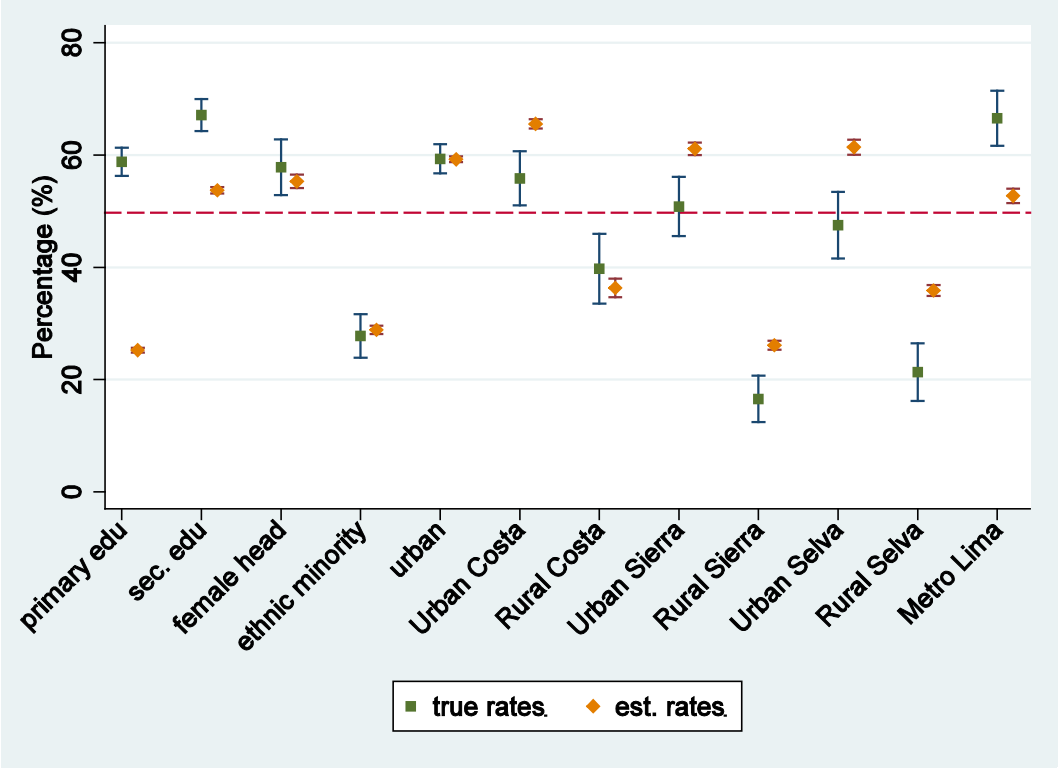


Figure 5: Profiles for the Proportion of the Population Who Were Poor in the Second Period Given that They Were Poor in the First Period, Vietnam 2006- 2008

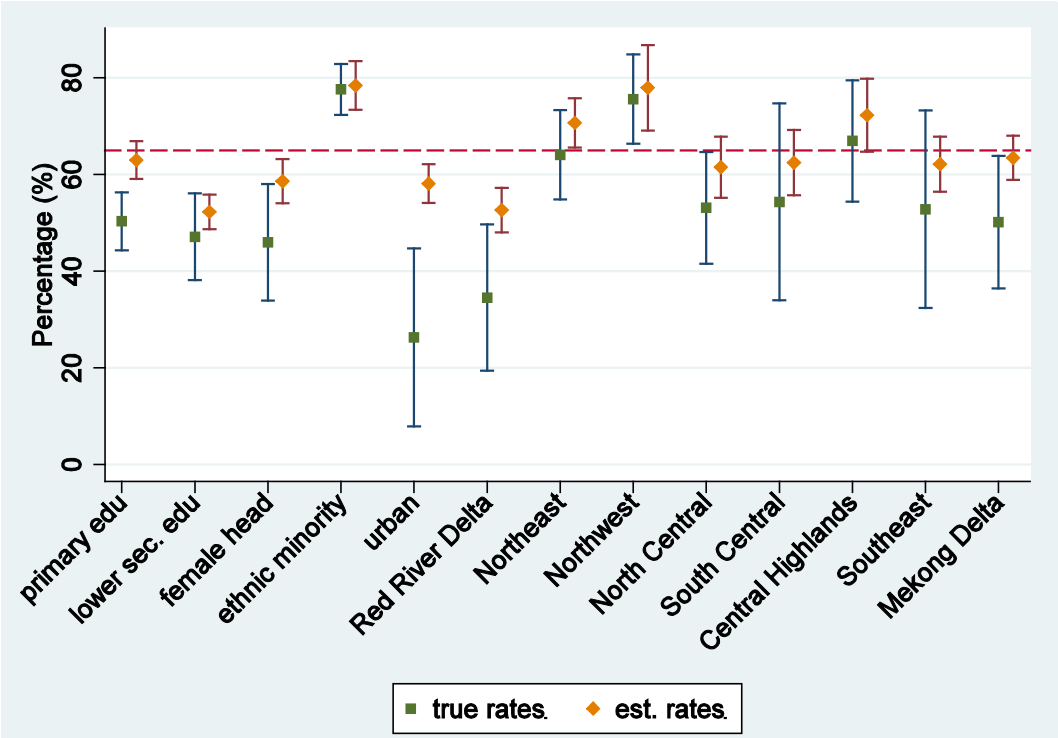


Figure 6: Profiles for the Proportion of the Population Who Were Non-poor in the Second Period Given that They Were Poor in the First Period, Vietnam 2006- 2008

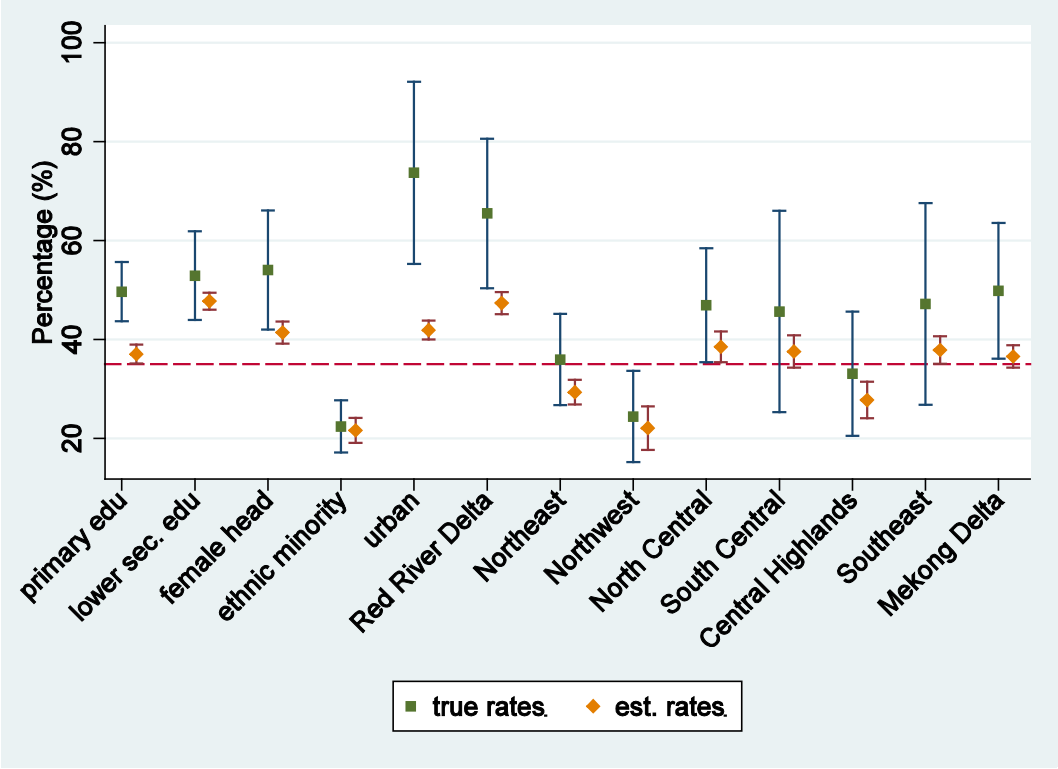


Figure 7: Profiles for the Proportion of the Population Who Were Poor in the Second Period Given that They Were Non-poor in the First Period, Vietnam 2006- 2008

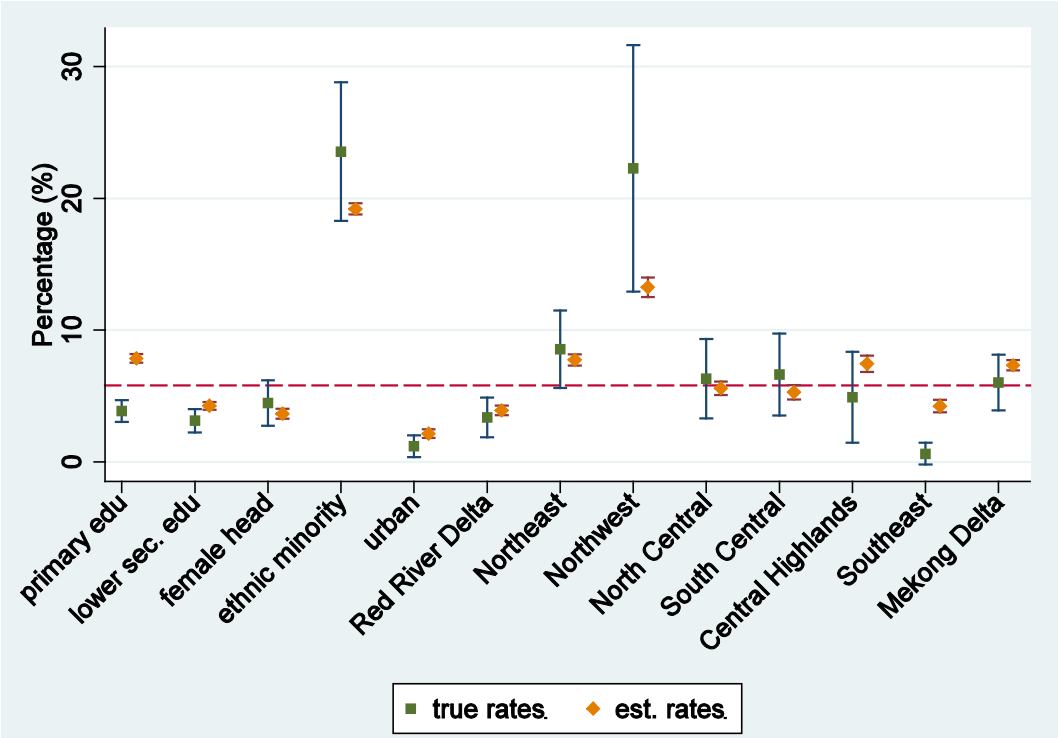
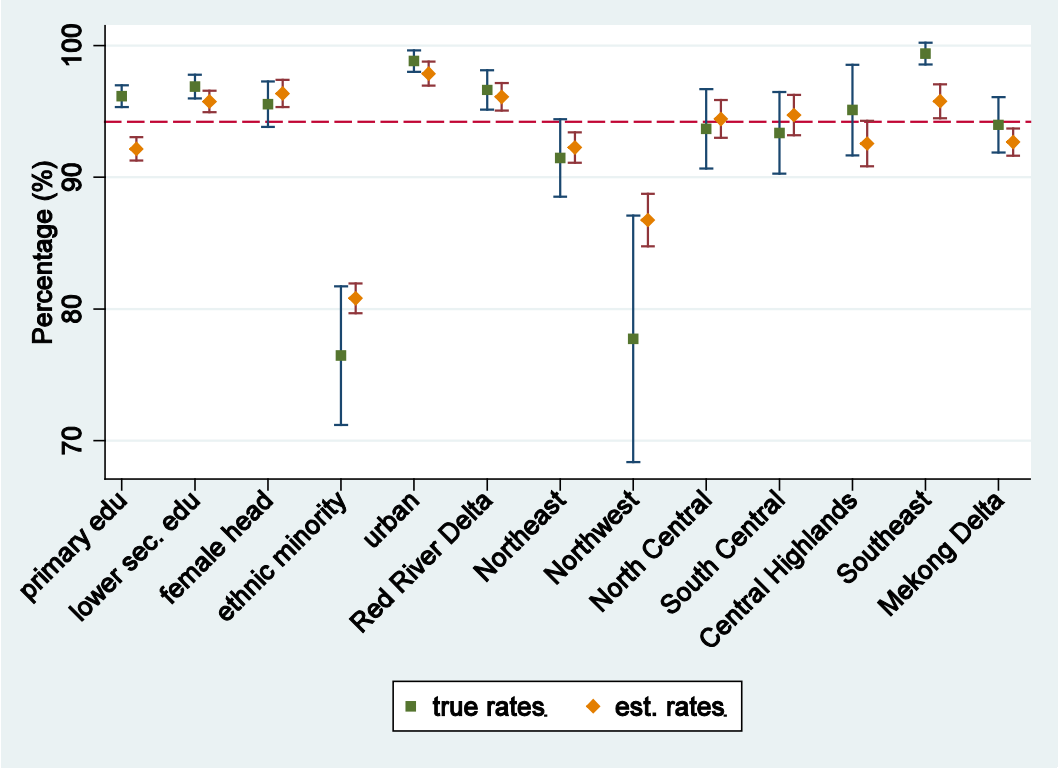


Figure 8: Profiles for the Proportion of the Population Who Were Non-poor in the Second Period Given that They Were Non-poor in the First Period, Vietnam 2006- 2008



Appendix 1: Proofs

Proof for Proposition 1

Consider a simple liner dynamic data-generating process for household consumption given by

$$y_{i2} = \alpha + \delta' y_{i1} + \eta_{i2} \qquad (1.1)$$

where y_{it} is household i's consumption in period t, t= 1, 2, and η_{i2} is a random error term. Note that in the absence of true 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 $\rho_{y_{i1}y_{i2}} = \frac{\text{cov}(y_{i1}, y_{i2})}{\sqrt{\text{var}(y_{i1})\,\text{var}(y_{i2})}} = \sqrt{\frac{\text{var}(y_{i1})}{\text{var}(y_{i2})}}\delta$. A consistent estimate for δ can be obtained by instrumenting for it with the age cohort dummy variables, as long as these instrumental variables are relevant and exogenous. Thus estimation of (1.1) this way is identical to applying OLS to the same model where all variables are aggregated to the cohort level (Verbeek, 2008)

$$\bar{y}_{c2} = \delta' \bar{y}_{c1} + \bar{\eta}_{c2} \qquad (1.2)$$

Thus from (1.2) we can consistently estimate δ, and $\rho_{y_{i1}y_{i2}}$ as $\rho_{y_{c1}y_{c2}} = \frac{\text{cov}(\bar{y}_{c1}, \bar{y}_{c2})}{\sqrt{\text{var}(\bar{y}_{c1})\,\text{var}(\bar{y}_{c2})}}$.

Proof for Proposition 2

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

$$\rho_{y_{i1}y_{i2}} = \frac{\text{cov}(y_{i1}, y_{i2})}{\sqrt{\text{var}(y_{i1})\,\text{var}(y_{i2})}} = \frac{\text{cov}(\beta_1' x_{i1} + \varepsilon_{i1}, \beta_2' x_{i2} + \varepsilon_{i2})}{\sqrt{\text{var}(y_{i1})\,\text{var}(y_{i2})}}$$

$$= \frac{\beta_1' \text{var}(x_i)\beta_2 + \rho\sqrt{\sigma_{\varepsilon_1}^2 \sigma_{\varepsilon_2}^2}}{\sqrt{\text{var}(y_{i1})\,\text{var}(y_{i2})}}$$

where the second line follows from Assumption 2 in DLLM. The third line follows from Assumption 1 that the underlying population being sampled in survey rounds 1 and 2 are the same, thus the time-invariant household characteristics x_{i1} and x_{i2} are replaced with x_i. Solving for ρ from the above equality, we have $\rho = \frac{\rho_{y_{i1}y_{i2}}\sqrt{\text{var}(y_{i1})\,\text{var}(y_{i2})} - \beta_1' \text{var}(x_i)\beta_2}{\sigma_{\varepsilon_1}\sigma_{\varepsilon_2}}$.

Proof for Corollary 2.1

If $\beta_1 \approx \beta_2$, we have

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

$$= \frac{\rho_{y_{i1}y_{i2}} - \dfrac{\beta_1'\text{var}(x_i)\beta_2}{\sqrt{\text{var}(y_{i1})\,\text{var}(y_{i2})}}}{\sqrt{1-R_1^2}\sqrt{1-R_2^2}}$$

$$= \frac{\rho_{y_{i1}y_{i2}} - \sqrt{\dfrac{\beta_1'\text{var}(x_i)\beta_2\beta_1'\text{var}(x_i)\beta_2}{\text{var}(y_{i1})\,\text{var}(y_{i2})}}}{\sqrt{1-R_1^2}\sqrt{1-R_2^2}}$$

$$\approx \frac{\rho_{y_{i1}y_{i2}} - \sqrt{\dfrac{\beta_1'\text{var}(x_i)\beta_1\beta_2'\text{var}(x_i)\beta_2}{\text{var}(y_{i1})\,\text{var}(y_{i2})}}}{\sqrt{1-R_1^2}\sqrt{1-R_2^2}}$$

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

where the denominator in the second row follows from the definition for R^2, and the last equality follows from the definition of R^2.

In fact, given that $\beta_1 \approx \beta_2$, another perhaps more intuitive way to prove Corollary 2.1 is for us to think about the systemic part of predicted household consumption as a single variable (i.e., $\hat{\beta}_1' x_{i1} \approx \hat{\beta}_2' x_{i2}$). By definition, the correlation between household consumption and this predicted variable is the multiple correlation coefficient R_j^2. Thus using the familiar expression that links the simple and partial correlation coefficients for bivariate normal variables[31]

$\rho_{12.3} = \dfrac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{1-\rho_{13}^2}\sqrt{1-\rho_{23}^2}}$ and replacing ρ_{12} with $\rho_{y_{i1}y_{i2}}$ and the simple correlation coefficients ρ_{13} and ρ_{23} respectively with R_1^2 and R_2^2, we have the expression in (6).

Proof for Corollary 2.2

The partial correlation coefficient between the error terms is then given as usual $\rho = \dfrac{\text{cov}(\varepsilon_{i1}, \varepsilon_{i2})}{\sqrt{\text{var}(\varepsilon_{i1})\,\text{var}(\varepsilon_{i2})}} = \dfrac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2}$, which implies that ρ is non-negative and is consistent of our Assumption 2. If we hold σ_v^2 fixed, it is straightforward to show that ρ is an increasing and concave function of σ_u^2. Thus ρ reaches its maximum when σ_u^2 reaches its maximum value. When σ_u^2 reaches its maximum value σ_u^{2+} (or $R_1^2 = R_2^2 = 0$), the estimation equations (1) and (2) have zero predictive power or all the terms with the estimated coefficients β's are equal to 0 and will drop out. Thus we have

[31] See, for example, equation (36) in Ridder and Moffitt (2007) or equation (20) in Anderson (2003, p. 39).

$$\rho \le \rho(\sigma_u^{2+}) = \frac{\sigma_u^{2+}}{\sigma_u^{2+} + \sigma_v^2} = \frac{\text{cov}(y_{i1}, y_{i2})}{\sqrt{\text{var}(y_{i1})\,\text{var}(y_{i2})}} = \rho_{y_{i1}y_{i2}}$$

with equality occurring only when the model has zero predictive power. Then the simple correlation coefficient $\rho_{y_{i1}y_{i2}}$ would be identical to the partial correlation coefficient ρ.

If $\beta_1 \approx \beta_2$, we can use Corollary 2.1 for an alternative proof. Holding R_2^2 fixed in the expression in (6), taking the first derivative of ρ with regards to R_1^2, we have

$$\frac{\partial \rho}{R_1^2} = \frac{1}{\sqrt{1-R_2^2}} \left(\frac{-\dfrac{R_2^2}{2\sqrt{R_1^2 R_2^2 (1-R_1^2)}} - \dfrac{\rho_{y_{i1}y_{i2}} - \sqrt{R_1^2 R_2^2}}{2\sqrt{(1-R_1^2)^3}}}{1-R_1^2} \right)$$

By definition, both R_1^2 and R_2^2 are bounded by [0, 1], and both two terms in the numerators are non-positive,[32] we have $\frac{\partial \rho}{R_1^2} \le 0$. Thus given any value for R_2^2, ρ is a decreasing function of R_1^2 and ρ reaches its maximum value when R_1^2 equals 0. By a similar argument, given any value for R_1^2, we can show that ρ is a decreasing function of R_2^2, and ρ reaches its maximum value when R_2^2 also equals 0. Thus ρ is less than or equal to $\rho_{y_{i1}y_{i2}}$, with equality occurring when both R_1^2 and R_2^2 equal 0.

Proof for Corollary 2.3

Corollary 2.3 directly follows from our assumption that ρ is non-negative. From the derivations in the proofs for Proposition 2 and Corollary 2.1 above, we see that

i) $\rho_{y_{i1}y_{i2}} = \dfrac{\beta_1' \text{var}(x_i)\beta_2}{\sqrt{\text{var}(y_{i1})\,\text{var}(y_{i2})}}$ when $\rho\sqrt{\sigma_{\varepsilon_1}^2 \sigma_{\varepsilon_2}^2} = 0$ or

ii) $\rho_{y_{i1}y_{i2}} - \sqrt{R_1^2 R_2^2}$ when the estimation model fully captures all the variations in the dependent variable (i.e., all the error terms are zero).

Proof for Propositions 3 and 4

To save space, we only show the proof for Proposition 4 since the two survey rounds scenario in Proposition 3 is a special case of the k survey rounds case in Proposition 4.

Given that household consumption can be explained by household characteristics in equations (1) and (2) and the standard regularity conditions are satisfied, our estimator $\hat{\Phi}_2(.)$ 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,

[32] Note that by Assumption 2, ρ is non-negative thus $\rho_{y_{i1}y_{i2}} - \sqrt{R_1^2 R_2^2}$ is non-negative.

and $j \neq m, n$, which are consistent estimators of the parameters. Thus $\hat{\Phi}_2(.)$ is a consistent estimator of $\Phi_2(.)$.

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

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

assuming that these two errors are uncorrelated with each other.

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

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

where 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 have the stated results.[33] Note that the approximation formula for $V(\hat{\sigma}_{\varepsilon_m}) = \frac{(8N-7)\hat{\sigma}^2_{\varepsilon_m}}{(4N-3)^2}$ is based on Montgomery (2012, pp. 720) where N>25.

Proof for Corollary 3.1

Since $\hat{\bar{\Phi}}(.) = \frac{1}{N}\sum_{i=1}^{N} \hat{\Phi}\left(d_j \frac{z_j - \hat{\beta}_1' x_{ij}}{\hat{\sigma}_{\varepsilon_{ij}}}\right)$ is a consistent estimator of P_{ij} and $\hat{\Phi}_2(.)$ is a consistent estimator of $P_{i,12}$ as discussed in the proof for Proposition 3 above, it follows that $\frac{\hat{\bar{\Phi}}_2(.)}{\hat{\bar{\Phi}}(.)}$ is a consistent estimator of $\frac{P_{i,12}}{P_{ij}}$. Then note that, since $\frac{\partial\left(\hat{\bar{\Phi}}_2(.)/\hat{\bar{\Phi}}(.)\right)}{\partial\hat{\bar{\Phi}}_2(.)} = \frac{1}{\hat{\bar{\Phi}}(.)}$ and $\frac{\partial\left(\hat{\bar{\Phi}}_2(.)/\hat{\bar{\Phi}}(.)\right)}{\partial\hat{\bar{\Phi}}(.)} = \frac{-\hat{\bar{\Phi}}_2(.)}{\left(\hat{\bar{\Phi}}(.)\right)^2}$, using the delta method[34] we have

$$\sqrt{n}\left[\frac{P_{i,12}}{P_{ij}} - \frac{\hat{\bar{\Phi}}_2\left(d_1 \frac{z_1 - \hat{\beta}_1' x_{ij}}{\hat{\sigma}_{\varepsilon_{i1}}}, d_2 \frac{z_2 - \hat{\beta}_2' x_{ij}}{\hat{\sigma}_{\varepsilon_{i2}}}, \hat{\rho}_d\right)}{\hat{\bar{\Phi}}\left(d_j \frac{z_j - \hat{\beta}_1' x_{ij}}{\hat{\sigma}_{\varepsilon_{ij}}}\right)}\right] \sim N(0, V_r)$$ where the covariance-variance

[33] See also Mullahy (2011) for a related derivation.
[34] See, for example, theorem 5.5.28 in Casella and Berger (2002).

matrix V_r can be estimated as

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

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

where similar to $Var(\hat{\Phi}_2(.))$, $Var(\hat{\Phi}(.))$ can be decomposed into a model error Σ_{jm} and a sampling error Σ_{js} assuming these two errors are uncorrelated.[35] The model error can be estimated as $\Sigma_{jm} = \nabla'_{\hat{\beta}_j} V(\hat{\beta}_j) \nabla_{\hat{\beta}_j} + \nabla'_{\hat{\sigma}_{\varepsilon_j}} V(\hat{\sigma}_{\varepsilon_j}) \nabla_{\hat{\sigma}_{\varepsilon_j}}$ with $\nabla_{\hat{\beta}_j} = d_j \left(\frac{-x_{ij}}{\hat{\sigma}_{\varepsilon_j}}\right) \phi\left(d_j \frac{z_j - \hat{\beta}_j' x_{ij}}{\hat{\sigma}_{\varepsilon_j}}\right)$ and

$\nabla_{\hat{\sigma}_{\varepsilon j}} = -d_j \left(\frac{z_j - \hat{\beta}_j' x_{ij}}{\hat{\sigma}^2_{\varepsilon_j}}\right) \phi\left(d_j \frac{z_j - \hat{\beta}_j' x_{ij}}{\hat{\sigma}_{\varepsilon_j}}\right)$.

[35] 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.

Appendix 2: Additional Tables

Table 2.1: Estimated Parameters of Household Consumption Using Cross Sections

	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
R^2	0.081	0.081	0.159	0.219	0.408	0.443	0.443	0.463	0.295	0.329	0.337	0.341	0.454	0.422	0.408	0.371
N	1342	1342	3032	3215	4493	9169	8593	9084	3275	3275	3368	3368	3527	3674	3596	3701

Note: 1. *p<0 .1, **p<0.05, ***p<0.01.

2. Household heads' ages are restricted to between 25 and 55 in the first survey round.

3. 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.

Table 2.2: Poverty Dynamics Based on Synthetic Data for Two Periods Using Earlier Survey Rounds (Percentage)

Poverty Status	Peru		United States		Vietnam	
First Period & Second Period	2004-05		2005-07		2004-06	
	True Panel	Synthetic Panel	True Panel	Synthetic Panel	True Panel	Synthetic Panel
Poor, Poor	32.4	32.7	5.6	7.2	11.3	11.0
	(1.0)	(0.2)	(0.4)	(0.2)	(0.6)	(0.3)
Poor, Nonpoor	9.8	9.7	3.9	3.8	9.2	7.8
	(0.7)	(0.0)	(0.3)	(0.1)	(0.6)	(0.1)
Nonpoor, Poor	9.7	11.2	4.0	3.1	4.0	3.9
	(0.6)	(0.0)	(0.3)	(0.1)	(0.4)	(0.1)
Nonpoor, Nonpoor	48.1	46.4	86.5	85.8	75.5	77.3
	(1.1)	(0.2)	(0.6)	(0.3)	(0.8)	(0.4)
N	2087	9169	3275	3275	2703	3674

Note: 1. For each country, synthetic panels poverty rates are calculated using the cross section component, and predictions obtained using data in the second survey round. We use 500 bootstraps in calculating standard errors.

2. All numbers are weighted using population weights for each survey round. Poverty rates are in percent.

3. 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 2.3: Poverty Dynamics Based on Synthetic Data for Two Periods, Using Data in the First Survey Round as the Base (Percentage)

Poverty Status	Lao PDR		Peru		Vietnam	
First Period & Second Period	2002/03- 2007/08		2005-06		2006-08	
	True panel	Synthetic panel	True Panel	Synthetic Panel	True Panel	Synthetic Panel
Poor, Poor	13.5	15.1	29.4	32.2	9.6	10.2
	(0.8)	(0.2)	(1.0)	(0.2)	(0.6)	(0.3)
Poor, Nonpoor	16.0	13.6	11.7	11.9	6.2	5.2
	(0.8)	(0.1)	(0.7)	(0.0)	(0.5)	(0.1)
Nonpoor, Poor	8.9	12.4	8.8	9.7	4.5	5.2
	(0.6)	(0.1)	(0.6)	(0.0)	(0.4)	(0.1)
Nonpoor, Nonpoor	61.7	59.0	50.1	46.2	79.7	79.4
	(1.1)	(0.4)	(1.1)	(0.3)	(0.8)	(0.4)
N	1989	3032	2250	8593	2723	3596

Note: 1. For each country, synthetic panels poverty rates are calculated using the cross section component except for Bosnia-Herzegovina and the USA. Predictions are obtained based on data in the first survey round. We use 500 bootstraps in calculating standard errors.

2. All numbers are weighted using household weights for Peru, and population weights for other countries. Poverty rates are in percent.

3. 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. year difference for the second survey round.

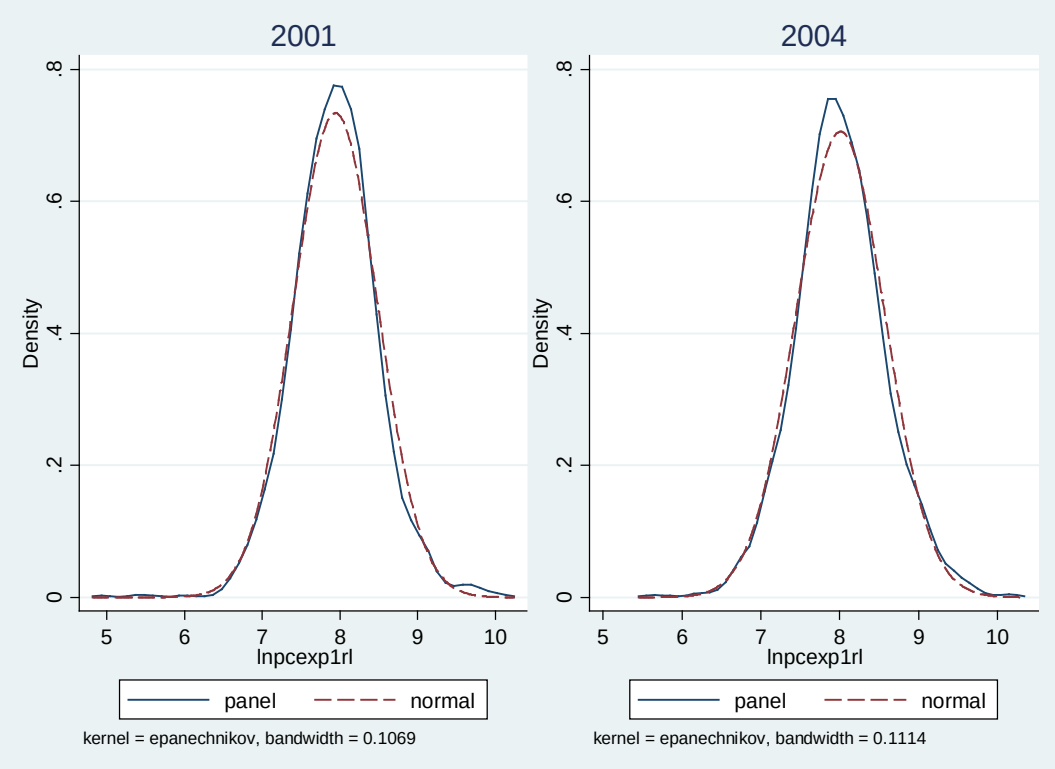
Appendix 3: Data Appendix

Bosnia-Herzegovina

We use two rounds of the panel data for the Bosnia-Herzegovina Living Standards Measurement Survey (also known as Living in BiH Survey) in 2001 and 2004, which is publicly available on the World Bank LSMS website. We build our data based on the files made available by Demirguc-Kunt, Klapper and Panos (2011).

There are 2,376 panel households between 2001 and 2004. After restricting household heads' age to between 25 and 55 for the first survey round and adjust accordingly for the second round, we are left with 1,353 panel households for analysis. We implement our analysis on the two halves of these panel data pretending that they are two cross sections. Figure 3.1 below provides the density graphs for household consumptions for Bosnia-Herzegovina in 2001 and 2004.

Figure 3.1. Log of Consumptions for Panel Data, Bosnia-Herzegovina 2001-2004



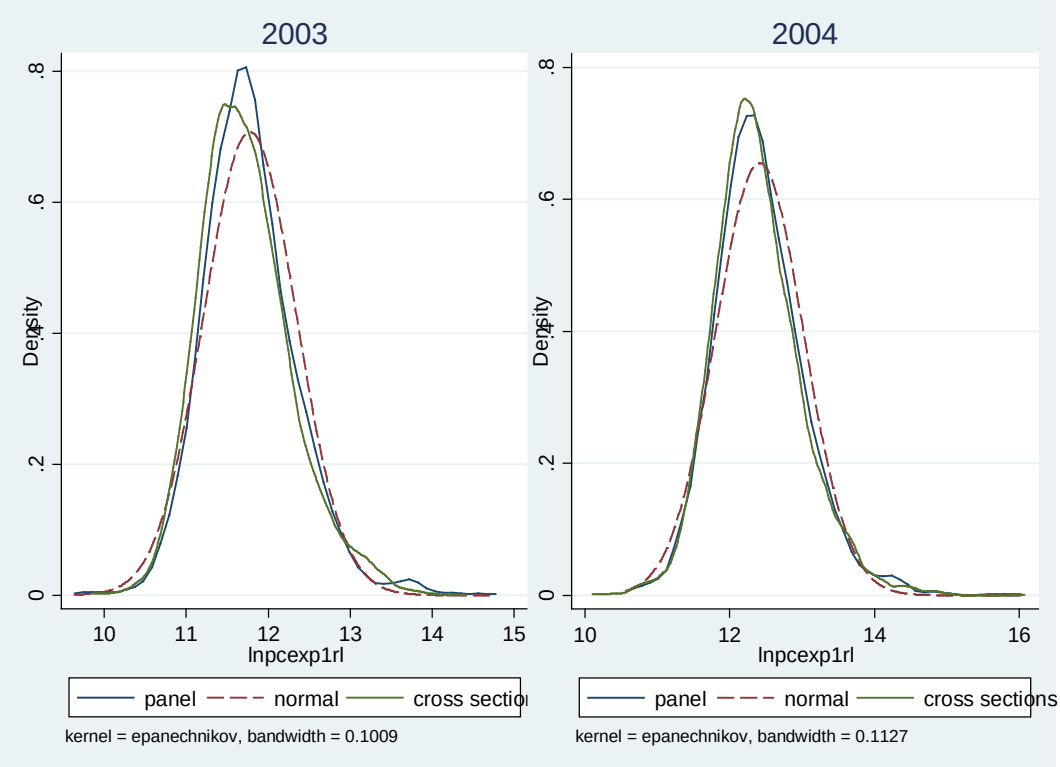
Lao PDR

We use two rounds of the Lao Expenditure and Consumption Survey (LECS) in 2002/03 and 2007/08, which is provided to us by the World Bank office in Lao PDR.

There are 2,357 panel households between 2002/03 and 2007/08. After restricting household heads' age to between 25 and 55 for the first survey round and adjust accordingly for the second round (i.e., increasing this age range to 30-60), we are left with 1,989 panel households for analysis. The corresponding numbers of cross sectional households we analyze are 3,223 and 3,225 respectively for 2002/03 and 2007/08.

Two sample t-tests with unequal variances show that household consumptions in the panel component are statistically but negligibly higher than those in cross section component for the 2002/03 round (e.g., with a difference of 0.05 between the two means of 11.78 and 11.73 respectively for the panel and cross section); however, these consumptions are not statistically different at the 5% level for the 2007/08 round. Figure 3.2 below provides as an example the density graphs for household consumptions in the panel and cross section components for Lao PDR in 2002/03 and 2007/08.

Figure 3.2. Log of Consumptions for Panel and Cross Section Components, Lao PDR 2002/03- 2007/08



Peru

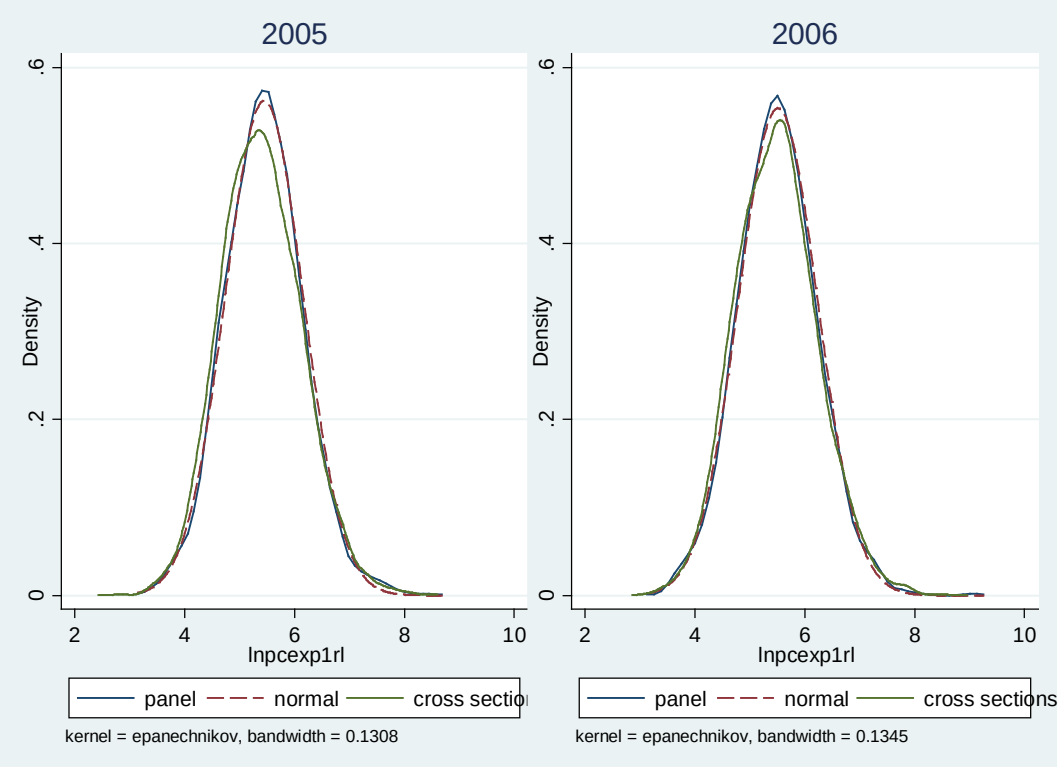
We use three rounds of the Peruvian National Household Survey (ENAHO) in 2004, 2005, and 2006, which is publicly available on the Peruvian Statistics Bureau (INEI)'s website. Both the panel and cross sectional households constructed from the ENAHOs are graciously provided to us by Renos Vakis and Leonardo Lucchetti based on their paper (Cruces et al., 2011).

There are 3,247 and 3,559 panel households respectively between 2004-2005 and 2005-2006. After restricting household heads' age to between 25 and 55 for the first survey round and adjust accordingly for the second round (i.e., increasing this age range to 26-56), we are left with 2,087 and 2,250 panel households for analysis. The corresponding numbers of cross sectional households we analyze are 4,493 and 9,169 respectively for 2004 and 2005 for the survey pair 2004-2005, and 8,593 and 9.084 respectively for 2005 and 2006 for the survey pair 2005-2006.

Two sample t-tests with unequal variances show that household consumptions in the panel component are statistically but negligibly higher than those in cross section component for the survey pair 2005-2006 (e.g., with a difference of 0.05 between the two means of 5.39 and 5.44 respectively for the panel and cross section); however, these consumptions are not statistically different for the survey pair 2004-2005. Figure 3.3 below provides as an example the density graphs for household consumptions in the panel and cross section components for Peru in 2006.

For all three years 2004-2005-2006, we have 2,668 panel households which is reduced to 1,987 panel households after a similar restriction on heads' age ranges. The corresponding number of cross sectional households we analyze is 8,608 for 2006.

Figure 3.3. Log of Consumptions for Panel and Cross Section Components, Peru 2005-2006



United States

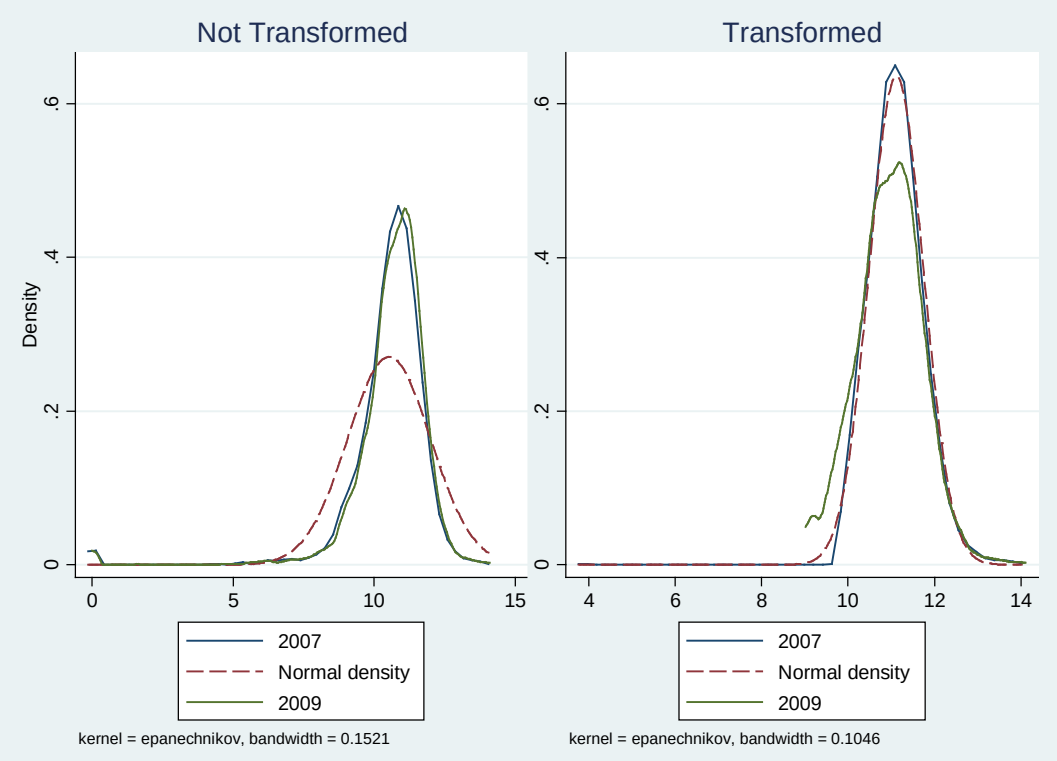
We use the three most recent rounds of the Panel Study of Income Dynamics (PSIDs) in 2005, 2007, and 2009, which is publicly available on the University of Michigan Institute for Social Research's website. The PSID started in 1968 and is the longest-running panel household survey implemented in the United States. The PSID was implemented annually between 1968 and 1997, and biannually after 1997. A useful documentation is provided in the PSID Main Interviewer User Manual Release 2012.1.

We use the sample persons in the PSID (i.e., those with a positive longitudinal weight), and after restricting household heads' age to between 25 and 55 for the first survey round and adjust accordingly for the second round (i.e., increasing this age range to 27-57) and keeping

the age difference across the two survey rounds between one and three,[36] we are left with 3,275 and 3,368 panel households respectively in 2005-2007 and 2007-2009 for analysis. For all three years 2005-2007-2009, we have 3,036 panel households after the restriction on heads' age range.

Since no comprehensive consumption aggregates are available in the PSID, we use income as a measure of households' welfare. Different from consumption measures, there are two potential issues with income measures: one is that the latter can be zero or negative (even though there are generally less than one percent of households at this welfare level in the PSID), which will translate into missing observations when we take the logarithm; the other is that income measures can have a lop-sided distribution that not as close to the normal distribution as consumption measures. We thus deal with both of these issues by implementing a Box-Cox transformation on the income variables using the *lnskew0* command in Stata, which effectively adds a positive constant *k* to the income before taking logarithm to minimize the skewness of the income variables. This constant *k* is 39,077, 19,727, and 8,279 respectively for incomes in 2005, 2007, and 2009.[37] Figure 3.4 below provides as an example the density graphs for log of incomes in 2007 and 2009 before and after the Box-Cox transformation.

Figure 3.4. Log of Incomes Before and After Box-Cox Transformation, USA 2007-2009



Vietnam

We use three rounds of the Vietnam Household Living Standards Surveys (VHLSSs) in 2004, 2006, and 2008, which is provided to us by Vietnam's General Statistical Office (GSO)

[36] This helps ensure the household heads remain the same across the two surveys.

[37] Since the number of panel observations change slightly between the two pairs of survey years, this constant *k* also changes slightly for 2007.

and the World Bank office in Vietnam. The VHLSSs have a rotating panel design with around one half of the data for the 2004 round of the VHLSS is repeated in the 2006 round, and one half of the 2006 round consisting of one half of the 2004-2006 panel and one half of the 2006 cross section is repeated in the 2008 round. An introduction to the VHLSSs (with a focus on the years 2002 and 2004) is provided by Tung and Phong (undated).

We construct panel data for the VHLSSs using household identification codes. Where we suspect mismatching between panel households due to incorrect identification codes, we double check and correct these cases with household heads' names. As a result, we could match 4,276 panel households between 2004 and 2006 out of 9,189 households for each year. After restricting heads' age to between 25 and 55 for 2004 and 27 and 57 for 2006 and keeping the age difference across the two survey rounds between one and three, we are left with 2,723 panel households for analysis.

Following a similar procedure, we could match 4,120 panel households between 2004 and 2006 out of 9,189 households for each year. After placing the restriction on heads' age range and difference, we are left with 2,703 panel households for analysis. The corresponding numbers of cross sectional households we analyze are 3,527 and 3,674 respectively for 2004 and 2006 for the survey pair 2004-2006, and 3,596 and 3,701 respectively for 2006 and 2008 for the survey pair 2006-2008.

Two sample t-tests with unequal variances show that household consumptions in the panel component are statistically but negligibly less than those in cross section component in the first round for each of the survey pairs 2004-2006 and 2006-2008 (e.g., with a difference of 0.04 between the two means of 8.44 and 8.48 respectively for the panel and cross section); however, these consumptions are not statistically different for the second rounds. Figure 3.5 below provides as an example the density graphs for household consumptions in the panel and cross section components for Vietnam in 2008.

The numbers for the panel households we could match between 2004, 2006, and 2008 are 1,848 (before restriction) and 1,282 (after restriction). The corresponding number of cross sectional households we analyze is 3,808 for 2008.

Figure 3.5. Log of Consumptions for Panel and Cross Section Components, Vietnam 2006-2008

