Policy Research Working Paper 8545

Lassoing Welfare Dynamics with Cross-Sectional Data

Leonardo Lucchetti
Paul Corral
Andrés Ham
Santiago Garriga

WORLD BANK GROUP
Poverty and Equity Global Practice
August 2018

Policy Research Working Paper 8545

Abstract

This paper introduces, validates, and applies a Least Absolute Shrinkage and Selection Operator with multiple imputation by Predictive Mean Matching (LASSO-PMM) method to estimate intra-generational welfare dynamics using cross-sectional data. Compared to previous welfare dynamic prediction methods, the LASSO-PMM makes fewer and less restrictive assumptions and allows estimating poverty transitions and income changes. We validate the method using 36 harmonized panel data sets in four Latin American countries and then apply it to cross-section data from 43 countries across the world. To the best of our knowledge, this is the first paper that uses these many datasets to validate and estimate welfare and mobility predictions. Validation results indicate that LASSO-PMM predictions are in general statistically indistinguishable from actual household poverty rates, mobility indicators, and income or consumption changes; results which are further supported by a series of sensitivity tests and robustness checks. These findings are sufficiently encouraging to suggest that estimating economic mobility using a LASSO-PMM approach may accurately approximate actual welfare dynamics in settings where panel data are unavailable. This application provides useful policy information on the dynamics of individual welfare in countries where two or more rounds of cross-section data are available.

This paper is a product of the Poverty and Equity Global Practice. 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://www.worldbank.org/research. The authors may be contacted at llucchetti@worldbank.org.

Lassoing welfare dynamics with cross-sectional data

Leonardo Lucchetti
The World Bank

Paul Corral
The World Bank

Andrés Ham
Universidad de los Andes

Santiago Garriga
Paris School of Economics

Updated December 16, 2020

Keywords: Poverty; Poverty transitions; LASSO; Machine learning; Multiple Imputation; Synthetic panels; Welfare dynamics.

JEL classification: C53, D31, O15, I32.
Sector Board: POV

Leonardo Lucchetti (llucchetti@worldbank.org) is a Senior Economist with the Poverty and Equity Global Practice, World Bank. Paul Corral (pcorralrodas@worldbank.org) is a Senior Economist with the HD Chief Economist Office, World Bank. Andres Ham (a.ham@uniandes.edu.co) is an Assistant Professor in the School of Government at Universidad de los Andes. Santiago Garriga (garrigasantiago@gmail.com) is a Ph.D candidate at the Paris School of Economics. We thank Aziz Atamanov, Eduard Bukin, Oscar Calvo, Andres Castañeda, Francis Addeah Darko, Carolina Diaz-Bonilla, Reno Dewina, Leonardo Gasparini, Gabriela Inchauste, Jose Montes, Laura Liliana Moreno, David Newhouse, Minh Cong Nguyen, Obert Pimhidzai, Carolina Sanchez-Paramo, Carlos Silva-Jauregui, Liliana D. Sousa, Shinya Takamatsu, Mariana Viollaz, Judy Yang, Nobuo Yoshida, Matthew Wai-Poi, Bagus Arya Wirapati, and Salman Zaidi. We also thank participants of the LACEA-LAMES 2018 meeting. Carlos Santiago Guzmán-Gutiérrez provided outstanding research assistance. Pablo Gluzmann provided outstanding support in the construction of all panel datasets. All remaining errors are ours.

1. Introduction

Many countries now gather survey microdata at the household and individual level. The rapid expansion of household surveys at frequent intervals that are comparable over time and across countries has facilitated poverty monitoring in the developing world: coverage increased from 13 countries in the 1990s to over 60 countries in 2011 (Serajuddin *et al*. 2015). One of the main uses for these surveys is to study poverty and welfare within countries and over time. However, most of the available micro data are cross-sectional (i.e., they do not track individuals and households over time and therefore only provide aggregate welfare trends). Panel datasets that follow individuals over several periods of time are less readily available, which limits the understanding of the factors that explain movements out of poverty, into poverty, and the duration of poverty (Dang et al. 2019).

This paper proposes a machine learning approach with a multiple imputation technique to overcome this empirical challenge. We propose using a Least Absolute Shrinkage and Selection Operator method with multiple imputation by Predictive Mean Matching (LASSO-PMM) to estimate intra-generational economic mobility using cross-sectional data.[1] The LASSO-PMM approach we propose estimates a model of household-level income[2] using a set of time-invariant and deterministic harmonized variables as controls to provide estimates of welfare dynamics. Compared to previous methods in the literature, LASSO-PMM makes fewer assumptions and estimates not only poverty transitions, but also income changes. We present, validate, and apply the LASSO-PMM method on harmonized microdata. Specifically, we validate the method using 36 panels in four Latin American countries (Argentina, Chile, Nicaragua, and Peru) and apply the

[1] We developed a Stata command to implement the LASSO-PMM, which can be accessed here. The command makes use of Wilbur Townsend's lassoregress Stata command.
[2] For simplicity we refer to income as the welfare measure throughout this paper.

procedure to cross-section data from 43 countries across the world. To the best of our knowledge, this is the first paper that uses this many datasets to validate and estimate welfare and mobility predictions.

Some previous studies have proposed methods that harness cross-section data to infer about dynamic welfare changes. "Synthetic Panels", developed by Dang, Lanjouw, Luoto, and McKenzie (2014), is the most recent.[3] The authors estimate a (log) income model in both the first and second rounds of cross-sectional data, including time-invariant, deterministic, and retrospective regressors. Parameters estimated in the first round are then used to predict unobserved first round income for all households interviewed in the second round. Depending on the assumptions on the correlation between error terms in the underlying regressions in both rounds, this "non-parametric" approach provides upper and lower bound estimates of poverty and mobility. The methodology has been validated using seven panels in Chile, Nicaragua, and Peru by Cruces et al. (2015), while Ferreira et al. (2013) applied it to predict intra-generational poverty mobility in 18 countries in Latin America and the Caribbean (LAC) using harmonized cross-sectional micro data.

Because this synthetic panel approach only provides bounds, Dang and Lanjouw (2013) proposed a method that obtains a point estimate of intra-generational poverty mobility by assuming that the error terms in the underlying regressions follow a binormal distribution and using the age-cohort correlation of residuals from cross-sections. This "parametric" method was validated by the authors on panel data from five countries and has been applied by Dang and Ianchovichina (2018) to study welfare dynamics in six Arab countries; by Dang and Dabalen (2018) to analyze whether growth has been pro-poor in 21 countries in Africa; and by Vakis, Rigolini, and Lucchetti (2016)

[3] The synthetic panel method builds on the poverty mapping technique developed by Elbers, Lanjouw, and Lanjouw (2003).

to analyze chronic poverty in 17 LAC countries for which harmonized cross-sectional micro data are available.

The LASSO-PMM approach introduced in this paper is more flexible and entails fewer assumptions compared to these two methods. As such, it does not require estimating any error correlation terms or assuming a binormal distribution of residuals in the underlying regressions as in Dang and Lanjouw (2013). In addition, the LASSO-PMM method not only provides transition probabilities, but also predicts income growth for every observation between two rounds of cross-sectional data as if we had actual panels. Finally, unlike Dang et al. (2014), the LASSO-PMM method obtains point estimates instead of bounds of welfare dynamics.

We use 36 panels from four countries to validate our estimates of mobility using harmonized variables frequently used in many regional and global studies, which facilitates implementation by researchers and policymakers in developing countries. The LASSO-PMM approach for predicting poverty and mobility approximates observed welfare changes. In general, validation results show that LASSO-PMM predictions are statistically indistinguishable from actual poverty rates, transitions, and income changes calculated using panel data. A series of validation exercises support the method's predictive performance. We divide these exercises into: i) *sensitivity tests*, in which the actual poverty rates and transitions do not change; and ii) *robustness exercises*, when we change the estimation procedure, sample, or other attributes that change the actual welfare estimates. Both of these validation exercises show that the LASSO-PMM approach for predicting poverty and mobility accurately approximates actual welfare changes. The two main conclusions from this validation are: i) changing parameters to make the method more computationally intensive provides only marginal changes in estimates; and ii) model specification matters. All these findings are sufficiently encouraging in general to suggest

that estimating economic mobility using LASSO-PMM may approximate actual welfare indicators in settings where cross-sections are routinely collected, but where panel data are unavailable.

Given these validation results, we apply the LASSO-PMM method to harmonized cross-section data from 43 countries across the world in circa 2010 and 2015. Compared to previous methods, the models we estimate have greater explanatory power, which increases our predictive capabilities. In general, aggregate poverty rates are accurately approximated when using cross-sections across a heterogeneous group of countries. Several additional messages emerge from the dynamic analysis. First, there has been considerable mobility in recent years; approximately 22% of individuals changed their economic status in all countries under analysis. Second, even in a period of generalized poverty reduction, about 9% of the population fell into poverty. Third, a large share of the population is immobile; about 24% of all individuals remained in chronic poverty. Fourth, initial average income for the chronic poor was slightly lower but grew significantly less than for those who moved out of poverty. Lastly, those who were not poor but fell into poverty had an income substantially lower and decreased more than those who were not poor and remained out of poverty.

The remainder of this paper is organized as follows. Section 2 summarizes existing cross-section welfare dynamic predicting methods and presents the LASSO-PMM approach. Section 3 describes the harmonized data sources used to validate and apply the method. Section 4 presents the results of estimating the LASSO-PMM approach and performs a wide array of validation tests to determine its sensitivity and robustness with 36 panel data sets. Section 5 applies the LASSO-PMM approach to cross-section data from 43 countries and discusses the findings. Section 6 concludes.

2. Welfare dynamic prediction methods[4]

2.1. Non-parametric synthetic panels

Assume two rounds of cross-sectional micro data are available: round 1 and round 2. These two rounds of survey are random samples of an underlying population and consist of a sample of N_1 and N_2 households, respectively. Let us define y_{it} as household's i log per capita income in moment t, x_{it} as a vector of household characteristics for household i in round t, and z as the poverty line. Variables included in x_{i2} refer to characteristics of household i in round 2 that are also observed (for different households) in round 1. These characteristics often include: i) time-invariant variables such as gender of the head of the household if his/her identity remains constant between the rounds of data; ii) deterministic variables such as age; and iii) retrospective variables such as whether a household surveyed in the second round had an asset in the first round (Cruces et al. 2015; Dang and Lanjouw 2018). The relationship between income and these characteristics can be expressed in regression form as:

$$y_{it} = \beta_t{}' x_{it} + \varepsilon_{it} \qquad \text{t = 1, 2} \tag{1}$$

where ε_{it} is an error term and x_{it} is a vector of K regressors whose first element is equal to one so that the first element of β_t is the intercept of the model.

We introduce superscripts to refer to observations surveyed at different times. Our objective is to calculate, for a household i interviewed in round 2, the income change across the two rounds of data: $\Delta y_{i1}^2 = y_{i2}^2 - y_{i1}^2$, where y_{i1}^2 and y_{i2}^2 are the first and second round incomes of household i surveyed in round 2, respectively. Therefore, because we observe the household in round 2, we need to estimate its income in round 1. Consequently, we can also estimate the

[4] This section draws from Dang and Lanjouw (2013), Dang et al. (2014), Cruces et al. (2015), Vakis et al. (2016), and Lucchetti (2017).

household's poverty dynamics: the joint probability that household *i* surveyed in round 2 escapes poverty between round 1 and round 2 ($\Pr(y_{i1}^2 < z \; and \; y_{i2}^2 > z)$), remains poor ($\Pr(y_{i1}^2 < z \; and \; y_{i2}^2 < z)$), becomes poor ($\Pr(y_{i1}^2 > z \; and \; y_{i2}^2 < z)$), and remains non-poor ($\Pr(y_{i1}^2 > z \; and \; y_{i2}^2 > z)$).[5,6]

These estimates can be easily calculated with panel data, since all households are interviewed in both rounds (i.e., y_{i1}^2 is known for every household *i* interviewed in round 2). However, these datasets are rarely available in developing countries and costly to collect. Alternatively, synthetic panels allow predicting the first round "unobserved" incomes of households surveyed in the second round by multiplying their time-invariant characteristics and the first-round Ordinary Least Squares (OLS) estimates of parameters $\hat{\beta}_1^{OLS}$ that solve the following optimization problem:

$$\hat{\beta}_1^{OLS} = \underset{\beta_1}{\operatorname{argmin}}\left[\sum_{i=1}^{N_1}(y_{i1}^1 - \beta_1{}' x_{i1}^1)^2\right] = \underset{\beta_1}{\operatorname{argmin}}[RSS_1^1] \qquad (2)$$

where y_{i1}^1 is the first-round log income of household *i* surveyed in round 1, N_1 indexes the number of observations in round 1, and RSS_1^1 refers to the residual sum of squares.

The two non-parametric synthetic panel approaches differ in the treatment given to the correlation between the error terms in the first and second round of cross-sectional data, which is likely to be non-negative according to Dang et al. (2014). Upper bound estimates assume no correlation between the first and second round error terms. The authors propose estimating the first-round incomes of households interviewed in the second round of data by drawing randomly

[5] For ease of exposition, we will only focus on the probability of escaping poverty in this section. However, the results generalize to the other three mobility situations (remaining poor, becoming poor, and remaining non-poor).
[6] Note that it is possible to carry out the exact inverse prediction (i.e., predict income of round 2 for those households interviewed in round 1). We will return to this point in the results section.

with replacement from the empirical distribution of first-round estimated residuals (denoted as $\tilde{\varepsilon}_{i1}^{2}$). In this case, the upper bound prediction of the first-round incomes for households surveyed in the second round is:

$$\hat{y}_{i1}^{2U} = \hat{y}_{i1}^{2} + \tilde{\varepsilon}_{i1}^{2} \tag{3}$$

where $\hat{y}_{i1}^{2}$ is the product between time-invariant characteristics and the first-round OLS estimates: $\hat{y}_{i1}^{2} = \hat{\beta}_{1}^{OLS\prime} x_{i1}^{2}$. Once incomes are predicted for the sample, we can calculate the joint probability of a household *i* surveyed in round 2 of being poor in round 1 and escaping poverty in round 2, $\Pr\left(\hat{y}_{i1}^{2U} < z \; and \; y_{i2}^{2} > z\right)$, as well as the income change between both periods $\Delta y_{i1}^{2U} = y_{i2}^{2} - \hat{y}_{i1}^{2U}$. Since predictions arise from a random draw of the empirical distribution of residuals, the method needs to be repeated *R* times and results shown are averaged over these *R* replications.

Lower bound estimates on the other hand assume perfect positive correlation between the first and second round error terms. The authors propose estimating first round incomes for households interviewed in the second round of data by using the estimates of the scaled residuals from the second-round regression (denoted as $\hat{\varepsilon}_{i2}^{2}$). The lower bound predictions are:

$$\hat{y}_{i1}^{2\mathrm{L}} = \hat{y}_{i1}^{2} + \frac{\hat{\sigma}_{\varepsilon_1}}{\hat{\sigma}_{\varepsilon_2}} \hat{\varepsilon}_{i2}^{2} \tag{4}$$

where $\hat{\sigma}_{\varepsilon_1}$ and $\hat{\sigma}_{\varepsilon_2}$are estimated standard errors for the two error terms ε_{i1} and ε_{i2}, respectively. The joint probability of a household *i* surveyed in round 2 of being poor in round 1 and escaping poverty in round 2 is given by $\Pr\left(\hat{y}_{i1}^{2\mathrm{L}} < z \; and \; y_{i2}^{2} > z\right)$, while the change in incomes between both periods is $\Delta y_{i1}^{2L} = y_{i2}^{2} - \hat{y}_{i1}^{2\mathrm{L}}$. Since the method is not randomly drawing from any the empirical distribution of residuals, there is no need to repeat the procedure *R* times.

The main limitation of the non-parametric approach is that the estimated bounds are often too wide.[7] As such, a parametric approach was developed to obtain a point estimate of poverty dynamics.

2.2. Parametric synthetic panels

Given that bounds of this non-parametric approach are often too wide, Dang and Lanjouw (2013) proposed a method to obtain a parametric point estimate of intra-generational poverty mobility. In order to accomplish this objective, the authors assume a bivariate normal distribution for the error terms with a non-negative correlation coefficient ρ. Thus, they can obtain a point estimate of the probability of escaping poverty by calculating:

$$\Pr(y_{i1}^2 < z \; and \; y_{i2}^2 > z) = \Phi\left(\frac{z - \hat{\beta}_1^{OLS\prime} x_{i1}^2}{\hat{\sigma}_{\varepsilon_1}}, \frac{z - \hat{\beta}_2^{OLS\prime} x_{i1}^2}{\hat{\sigma}_{\varepsilon_2}}, -\rho\right) \tag{6}$$

where $\hat{\beta}_2^{OLS}$ are the second-round OLS parameter estimates. A parametric lower bound estimate can be obtained by setting $\rho = 1$, while the upper bound estimate emerges from setting $\rho = 0$. The authors suggest estimating an age-cohort correlation of residuals using cross-section data to obtain an estimate of the unknown parameter ρ. Thus, the method requires an additional estimate.

Despite obtaining a parametric estimate of poverty mobility, there are several limitations of this approach. First, although it cannot be directly demonstrated from the results obtained in the

[7] Lucchetti (2017) proposes an adaptation of the lower and upper bound estimations to obtain non-parametric point estimates of welfare dynamics. The author suggests computing a weighted average of the residuals to get a point estimate of mobility. First round non-parametric predicted incomes are

$$\hat{y}_{i1}^{2NP} = \hat{y}_{i1}^2 + \left[(1-\gamma)\hat{\varepsilon}_{i1}^2 + \gamma \frac{\hat{\sigma}_{\varepsilon_1}}{\hat{\sigma}_{\varepsilon_2}} \hat{\varepsilon}_{i2}^2\right]$$

where $0 \leq \gamma \leq 1$. The joint probability of a household i surveyed in round 2 of being poor in round 1 and escape poverty in round 2 is given by $\Pr(\hat{y}_{i1}^{2NP} < z \; and \; y_{i2}^2 > z)$, while the change in incomes between both periods is $\Delta y_{i1}^{2NP} = y_{i2}^2 - \hat{y}_{i1}^{2\mathrm{NP}}$. Since upper bound residuals are used, the method needs to be repeated R times. The lower bound estimates can be obtained by setting $\gamma = 1$, while the upper bound estimates emerge from setting $\gamma = 0$. The author sets $\gamma = 0.5$ and test the sensitivity of results to changes in the value of the γ. The main limitation of this approach is that the selection of the value of γ can be considered arbitrary.

literature, the quality of the predictions might not be linked to the quality of the underlying welfare model used. For instance, Dang and Lanjouw (2013) obtain predictions that are statistically indistinguishable from actual household poverty transitions with a relatively low adjusted R^2=0.08 in the underlying welfare model.[8] The intuition is that a worse fit of the model makes the upper and lower bound estimates get far away from each other. However, the point estimate ultimately depends on the value of ρ selected and is unaffected by the value of the upper and lower bound predictions.[9]

Second, Herault and Jenkins (2019) show that the estimates of ρ are quite sensitive to the choice of the cohort, which (connected to the previous point) could generate inadequate estimates of ρ and subsequently the dynamics of poverty. Third, the method does not allow predicting income changes (i.e., $\Delta y_{i1}^2 = y_{i2}^2 - y_{i1}^2$); it only obtains poverty transitions like the one shown in equation (6). Finally, the assumption of normality of the error terms is rejected in Vietnam and Indonesia by Dang et al. (2014). In summary, obtaining a point estimate for welfare dynamics is feasible, but requires several restrictive assumptions and additional estimations.

2.3. A machine learning and multiple imputation approach (LASSO-PMM)

This paper proposes using a Least Absolute Shrinkage and Selection Operator complemented with multiple imputation by Predictive Mean Matching (LASSO-PMM) to estimate intra-generational poverty mobility and household-level income growth using cross-sectional data. The Least Absolute Shrinkage and Selection Operator (LASSO) procedure is one of the most popular

[8] Herault and Jenkins (2019) also obtain good results with a relatively low adjusted R^2 in the range of 0.098 to 0.226, while Dang and Ianchovichina (2018) obtain good predictions with the following adjusted R^2 coefficients: 0.16 for Jordan, 0.09 for Palestine, 0.05 for Syria, and 0.13 for Yemen.

[9] Lucchetti (2017) found a similar result in non-parametric estimates. Lower values of the adjusted R^2 make the lower and upper bounds get far away from each other, despite the fact that the non-parametric approach that uses a weighted average of the residuals to get a point estimate of mobility remains largely unaffected by the quality of the underlying welfare model.

machine learning methods among economists and consists of minimizing a quadratic loss function plus the sum of the absolute value of the coefficients (Mullainathan and Jann Spiess 2017). We propose estimating parameters in the first round of cross-sectional data by solving the following optimization problem:

$$\hat{\beta}_{1\lambda}^{LASSO} = \underset{\beta_1}{\operatorname{argmin}} \left[RSS_1^1 + \lambda \sum_{s=1}^{K} |\beta_{s1}| \right] \tag{7}$$

The estimates depend on the value of the "shrinkage" factor λ. Whenever $\lambda \to 0$, the objective function becomes the OLS objective function in (2) and $\hat{\beta}_{1\lambda}^{LASSO} \to \hat{\beta}_1^{OLS}$. The LASSO estimate will deviate from the OLS estimate for positive values of λ. Finally, $\hat{\beta}_{1\lambda}^{LASSO}$ will be shrunk to zero as $\lambda \to \infty$. Therefore, for values $\lambda > 0$, the LASSO attenuates towards zero if compared with OLS.

The factor λ is introduced for two reasons. First, the shrinkage penalty $\sum_{s=1}^{K} |\beta_{s1}|$ in LASSO provides corner solutions, which implies that some coefficients are forced to be zero. Therefore, the LASSO works well for model selection when the number of candidate variables *K* is large. Second, for appropriate values of λ, the bias introduced is compensated by a reduction of variance. Given that the main objective is to predict welfare, and not estimate causal effects of the explanatory variables, minimizing the bias-variance trade-off is essential for accurate predictions.

To avoid arbitrariness, the shrinkage factor λ is selected by means of a 10-fold cross-validation algorithm,[10] which is a method to test the out of sample fit of the income model.[11] The algorithm randomly divides the first-round of data into 10 equal sized folds. By leaving one fold out (the test fold), the model is fit in the other 9 folds (the training folds). Once the income model

[10] This is the first stage of the cross-validation process we employ in our validation exercise. A second stage is explained in section 3.
[11] Variables are standarized to have a mean of zero and standard deviation of one.

is estimated, the withheld fold is used to predict the model. This is repeated 10 times until all folds have been left out and all observations have a predicted value $\hat{y}_{i1}^{1}$. The value of λ is selected so that it minimizes the mean squared error (MSE) defined as $\sum_{i=1}^{N_1}(y_{i1}^{1} - \hat{y}_{i1}^{1})^2/N_1$.

Like in the case of the other synthetic panels presented in previous sections, we can proceed to compute two linear fits. The LASSO linear fit of the first-round incomes for all households surveyed in the first round is

$$\hat{y}_{i1\lambda}^{1LASSO} = \hat{\beta}_{1\lambda}^{LASSO\prime} x_{i1}^{1} \tag{8}$$

On the other hand, the LASSO first-round linear fit for all households surveyed in the second round is

$$\hat{y}_{i1\lambda}^{2LASSO} = \hat{\beta}_{1\lambda}^{LASSO\prime} x_{i1}^{2} \tag{9}$$

Like in the upper and lower bound non-parametric synthetic panels, the full welfare measure is completed by the addition of the residuals. The data generating process for the residuals is unknown under LASSO. This is one of the reasons why we simulated these unobservable using a predictive mean matching approach (PMM). This method is common in the multiple imputation literature (see Van Buuren, 2018), and always finds a value which has been observed in the data to produce a simulated value.

Once the linear fits in the first and second rounds are obtained using LASSO (equations (8) and (9)), they are compared with the aim of finding neighbors. For every observation in the second round of data, a set of neighbors (in terms of linear fit) are found in the first-round data by looking for observations with the smallest absolute difference between the two linear fits. Among the set of neighbors (the number of neighbors is determined by the researcher), one neighbor is selected at random. The observed welfare measure from that neighbor is then imputed to the corresponding observation in the second round, which represents the first-round income for that particular household surveyed in the second round. This imputation needs to be done for every

single observation in the second round. First-round imputed incomes are then compared with second-round actual incomes for every household surveyed in the second round of data in order to estimate income dynamics and poverty transitions.

Standard errors are obtained by relying on bootstrap samples of the first round data across simulations. For every bootstrap sample of the data, a new LASSO is fit and the whole welfare vector is constructed via PMM. This implies that it is possible that a variable which was used in one bootstrap sample may not be used in the next, this is because LASSO is executed for each realized data point and finds the MSE minimizing model.

The following steps describe in detail our approach if we want to impute first-round incomes to all households surveyed in the second round:

- Take a bootstrap sample of the first round cross-sectional data.
- Obtain LASSO estimates in the first round as in equation (7): $\hat{\beta}_{1\lambda}^{LASSO} = \underset{\beta_1}{\operatorname{argmin}}[RSS_1^1 + \lambda \sum_{s=1}^{K} |\beta_{s1}|]$.
- Obtain the LASSO prediction (linear fit) of the first-round incomes for all households surveyed in the first round bootstrapped sample as in equation (8): $\hat{y}_{i1\lambda}^{1LASSO} = \hat{\beta}_{1\lambda}^{LASSO\prime} x_{i1}^1$.
- Obtain the LASSO prediction (linear fit) of the first-round incomes for all households surveyed in the second round as in equation (9): $\hat{y}_{i1\lambda}^{2LASSO} = \hat{\beta}_{1\lambda}^{LASSO\prime} x_{i1}^2$
- For a particular observation *m* in the second round, obtain the closest *K* nearest neighbors by minimizing the difference of fit with every observation *i* in first round: $\left|\hat{y}_{i1\lambda}^{1LASSO} - \hat{y}_{m1\lambda}^{2LASSO}\right|, i = 1, 2, \ldots, N_1$.
- Randomly select one neighbor *s* from the list of *K* nearest neighbors of observation *m*.
- From this chosen neighbor *s* in the first round, select the observed log income (y_{s1}^1) and assign it to observation m surveyed in the second round. This is the same as adding the residual term

$\hat{\varepsilon}_{s1}^{1} + \epsilon$ of the *s* nearest neighbor to the linear fit $\hat{y}_{m1\lambda}^{2LASSO}$ of the observation *m*,[12] where ϵ is a real value which encompasses the difference between $\hat{y}_{s1\lambda}^{1LASSO} - \hat{y}_{m1\lambda}^{2LASSO}$

- Observation *m* would move out of poverty if $y_{s1}^{1} < z$ and $y_{m2}^{2} > z$
- Find nearest neighbors and compute income dynamics for the rest of the observations in the second round.
- Repeat all steps *R* times.

It is important to note that estimating LASSO with a multiple imputation approach has at least three advantages with respect to previous methods in the dynamic welfare prediction literature. First, the approach described in this paper does not require additional estimation of unknown parameters, such as the age-cohort correlation of residuals from cross-sections as in the parametric approach. Second, unlike the parametric approach, the method allows researchers to calculate household-level income changes and not just probabilities of poverty mobility.[13] Finally, given that the objective is prediction and not causal inference, the method takes advantage of machine learning techniques that minimize MSE (both bias and variance), to produce accurate welfare predictions outside the estimation sample. While previous studies minimize their discussion of model specification and explanatory power, the LASSO-PMM procedure we apply here selects the optimal set of variables in the underlying method by minimizing the MSE of the dependent variable (log of per capita household income) out of sample and with minimal

[12] As such, instead of randomly allocating first-round residuals as in the upper bound non-parametric synthetic panel, the LASSO-PMM allocates actual incomes from the first-round nearest neighbords, which is similar to allocating residuals based on the minimization of the distance of linear fits.

[13] Bourguignon and Moreno (2018) introduce in a working paper a modification of the parametric Synthetic panel approach that avoids its most arbitrary assumptions. Their method also allows to estimate income changes. However, their's method is more difficult to implement in our context; their work involves the calibration (instead of an estimation) of Synthetic panels within the scope of AR(1) processes, which makes its implementation in several countries at once a difficult task because the error formation process needs to be modeled.

assumptions. This is important because the quality of the predictions is directly linked to the quality of the underlying welfare model used under this approach.

3. Data and empirical approach

This paper uses harmonized microdata for several countries to validate and estimate economic mobility using LASSO-PMM. We first employ 36 panels from four selected countries to validate the method, performing a battery of sensitivity and robustness exercises in a controlled environment where we can observe household income in both periods. Then, we proceed to apply the method to harmonized cross-section data from 43 countries around the world, located in six broad regions.

3.1 Validation with harmonized panel data

For the validation of the LASSO-PMM method, we employ panel data from four Latin American countries: Argentina, Chile, Peru, and Nicaragua. These data are part of the SEDLAC and LABLAC projects.[14] The selection of these countries depends mainly on the availability of comparable panel data over time, but we highlight that these sources contain information for small and large economies; lower-middle, upper-middle, and high income Latin-American countries, and encompass short and long panels.

We construct 36 different panels for the four selected countries, which vary in the length of time in which households are followed. There are 23 one-year panels, 7 two-year panels, 5 three-year panels, and one four-year panel. To the best of our knowledge, this is the first paper that

[14] The SEDLAC project consists of more than 400 household surveys in more than 25 LAC countries to provide statistics on poverty and other distributional and social variables. The LABLAC project is a complementary source to SEDLAC, which monitors trends in labor market variables with the highest frequency information available in 13 LAC countries. Both projects are a joint venture between the World Bank and and the Center for Distributive, Labor, and Social Studies (CEDLAS, for its acronym in Spanish) at the *Universidad Nacional de La Plata* in Argentina. See Bourguignon (2015) for a detailed description of the SEDLAC project. Data for Argentina, Peru, and Nicaragua are part of the SEDLAC project, while data for Chile are obtained from LABLAC project.

uses this many panel data sets from multiple countries to validate cross-section poverty dynamics prediction methods.[15] Variation in poverty rates, aggregate conditions, and panel length provide an added value that allows investigating the performance, sensitivity, and robustness of the LASSO-PMM method. This knowledge is pivotal to understand whether this method can be applied more generally to monitor and predict household welfare dynamics using data sources from other regions and countries.

All 36 panels are harmonized to maximize comparability across countries and over time. The harmonization procedure follows the criteria established by SEDLAC (2018), which consists in using identical definitions in each data set. The advantage of this procedure lies in ensuring comparability across relevant dimensions, so that any observed differences across countries are due to specific changes in poverty dynamics, aggregate conditions, and panel length. The harmonization procedure requires cleaning and processing each panel to ensure that all samples, definitions, and variables are identical in each country and year. The full list of available covariates for each country and panel in the validation data are shown in Appendix Table A.1.

The decision to use harmonized panels is taken in order to test how the method would perform on cross-sectional data that applies some form of harmonization, such as SEDLAC, LABLAC, and other worldwide repositories commonly used to measure and monitor poverty, much like we use in our application in section 4. Satisfactory validation results would provide confidence to implement LASSO-PMM in environments where cross-sectional harmonized microdata exist to predict welfare dynamics. Moreover, the need to use harmonized data incorporates an additional restriction to the validation exercise, apart from the existing constraint

[15] For instance, Herault and Jenkins (2019) use 31 panels from two countries (Australia and the UK) in their validations; Dang and Lanjouw (2013) use 11 panels from 5 countries; Dang et al. (2014) use 8 panels from 6 countries; and Bourguignon and Moreno (2018) use one panel from Mexico.

of time invariance of the predictors used in the models. Therefore, the objective of the first part of this paper is to validate the LASSO-PMM method when we have comparable data across countries, weighting generalizability over specificity.

3.2 Application to harmonized cross-section data

After performing the validation on panel data, we apply the LASSO-PMM method to harmonized cross-section data. We employ the Global Monitoring Database (GMD) which gathers income and/or expenditure household surveys across the world. Raw surveys routinely conducted by national statistical offices in each country are compiled by the GMD project that is hosted and maintained by the World Bank. Most of the value added of this exercise, besides the compilation of a large number of household surveys, is the additional data processing and cleaning within the GMD initiative that harmonizes data sources to be comparable not only across countries but also within countries over time. The GMD database is the main input for the well-known *PovcalNet* online tool[16] to monitor global poverty and inequality.[17]

The benefits of using the GMD data in terms of population coverage are undeniable, since they contain more than 1,000 surveys in 156 countries.[18] However, with greater coverage, some specificity is lost. The number of surveys and variables that can be harmonized in order to be plausibly compared across and within countries is limited. Therefore, a set of dimensions have been selected and harmonized by the GMD project. These dimensions are: welfare (nominal, real, and in PPP terms); demographics (household size, gender, relation to household head, and marital status); education (school attendance, education level, and literacy); location (either subnational regions or rural-urban areas); some basic assets (landline phone, cellular phone, and computer);

[16] http://iresearch.worldbank.org/PovcalNet/povOnDemand.aspx
[17] At the time of writing, our results are drawn from the July 2020 version of the GMD database.
[18] In this paper we use surveys in 43 countries for which two round of cross-sectional data exist.

and access to water, sanitation, and electricity. The data also contain household and individual identifiers, as well as sampling weights. These dimensions contain most of the required time-invariant and deterministic variables necessary to apply the LASSO-PMM method for predicting welfare dynamics and poverty mobility.

This paper uses household surveys for 43 countries around the world. The selection of countries is based on the availability of at least two comparable data points and on the availability of the necessary variables to conduct the estimation. We use the most recent surveys for each country at the time or writing, trying to cover a period of five years between both rounds of cross-sectional data (i.e., circa 2010-2015). Our sample includes countries in the six geographical regions defined by the World Bank: East Asia and the Pacific (EAP), Europe and Central Asia (ECA), Latin America and the Caribbean (LAC), Middle East and North Africa (MNA), South Asia (SAR) and Sub-Saharan Africa (SSA). While the selected surveys do not cover all countries in the world, they do encompass a large set of nations with different levels of development and welfare. Appendix Table A.2 summarizes the countries and years, as well as the specific household surveys in the GMD that we use in the analysis.

3.3 Empirical strategy

Estimations are carried out for household-level income changes as well as for poverty dynamics defined as the proportion of individuals with a harmonized per capita income lower than a US$5.5 per person per day poverty line, both in 2011 purchasing power parity (PPP) per day.[19] The main difference between the validation and application of the LASSO-PMM method is that we can

[19] The US$5.5 per person per day in 2011 PPP is the World Bank poverty line used for upper-middle-income countries (UMICs). This line is the median value of UMIC national poverty lines (Jolliffe and Prydz 2016, Ferrerira et al 2016, Castañeda et al. 2018). The only exception is Chile in the validation analysis where we define a relative poverty line equal to 0.5 of median per capita income, because the poverty rate using the US$5.5 threshold is too low in this particular case.

observe actual welfare and mobility for households in the former case, but not the latter. Following previous studies (Cruces et al. 2005; Dang et al. 2014; Bourguignon and Moreno 2018), the key time invariance assumption is maintained by considering only those households whose heads are between 25 and 65 years of age in all estimates so that life cycle events are avoided in general.

We follow previous literature that estimates poverty and welfare by including time invariant and deterministic regressors in the models. The harmonized data used in this paper allow estimating poverty transitions and welfare changes with variables frequently associated with well-being (e.g., Ferreira et al. 2012; Vakis et al. 2016), specifically harmonized variables commonly found in household surveys that are comparable within and between countries and over time. The models consider the log of per capita household income in 2011 PPP/day as the left-hand side variable in the underlying regressions and the following regressors: [1] the household head's characteristics (age, age squared,[20] gender, and years of education), household size, and household size squared;[21] [2] regional fixed effects;[22] and [3] the interaction between the first and the second set of covariates. The selected specification follows previous studies that define a core set of time-invariant variables and correlates of welfare (Cruces et al. 2005; Dang et al. 2014; Bourguignon and Moreno 2018).

All these covariates are natural candidates as time invariant and deterministic regressors. However, it is important to bear in mind that all variables considered vary to some extent over time. The age, gender, and education of the head of the household can change each time the head

[20] Age is a deterministic function of time. Therefore, we subtract the number of years betwen rounds of data from the household head's age when predicting first round incomes for those surveyed in the second round. On the other hand, we add the number of years betwen rounds to the household head's age when changing the forecasting direction. This procedure is commonly applied in the Synthetic Panel's literature. See Dang and Lanjouw (2013) and Dang et al. (2014) for specific details and a more comprehensive explanation.

[21] In the validation, we also include area of residence (i.e., urban vs. rural) in Chile, since it is the only country for which this variable is available.

[22] We consider all regions for which surveys are representative in each country. In the application, we use area of residence (i.e., urban vs. rural) when regional fixed effects are not available.

of the household changes between survey rounds (e.g., death, migration, and divorce, among others, can force changes in the leadership of the household). Household size can also change due to births, deaths, divorces, among others.[23] The area of residence and regional fixed effects can change between rounds due to migration. On the other hand, other variables are easily discarded from the specifications, since they are clearly time-variant (e.g., employment status, sector of employment, etc.).[24]

The LASSO-PMM approach takes the included variables and selects the relevant covariates in each of the countries and years by minimizing the MSE. The penalty parameter in the LASSO, λ, is chosen using a 10-fold cross validation exercise. The PMM method selects the household's nearest neighbor. This procedure is repeated 20 times. The results provide first round predicted welfare for each household surveyed in the second round. These predictions are then compared to observed household's welfare in the second round.

In the particular case of the validation, we carry out several robustness and sensitivity analyses to determine the stability of the predictions from the LASSO-PMM method. For instance, we change model parameters, estimation procedure, the direction of the prediction, sample, included variables, weighting structures, consider different ages, and use different poverty lines in our validation exercises in the next section. Given that we have panel data in this exercise, we follow Cruces et al. (2015) and perform a second stage cross-validation that randomly splits the

[23] Household size was not considered in any of the studies with the exception of Bourguignon and Moreno (2018). However, household size is an important variable to be included for at least two reasons. First, the dependent variable (total household income) of the underlying regressions is in per capita terms. If we take two households that are identical in terms of head of household's characteristics and have the same total income, but the first one is only one individual while the second one is a married couple, the second one would be twice as poor as the first one. Household size will be capturing this difference. Second, we are using household level data but individual weights (household weights times household size), and therefore household size also affects results through weights.

[24] Bourguignon and Moreno (2018) present a thorough discussion of what variables may be considered as time-invariant in poverty prediction approaches.

panels into two subsamples and treats each of these subsamples as a cross-section. Therefore, the coefficients are estimated in one of these subsamples in the first round of data and applied to the second subsample in the second round. By treating each subsample of the panel as a cross-section, this second stage cross-validation avoids any bias that might arise from using the panel dataset to validate the method.

4. Validation results

4.1 Predicted poverty rates

We apply the LASSO-PMM approach to predict the log of per capita household income in 2011 PPP/day in each country at the household-level. Covariates are harmonized across countries and years, and include the demographic, educational, and regional variables described in the previous section. The number of regressors varies from 27 in Nicaragua to 103 in Chile (see Appendix Table A.1).[25]

The objective is to predict first round per capita household incomes for every household we observe in the second round.[26] Based on the estimated LASSO-PMM predictions, a first step of the intra-generational mobility analysis is to compare actual poverty rates in round 1 with the predicted poverty headcounts that emerge when applying the LASSO-PMM approach to those households surveyed in the second round. Estimates for all 36 panels are shown in Figure 1 and for the last available panel in each country in Table 1. The table presents both point estimates and 95 percent confidence intervals for actual and predicted poverty headcount rates.

[25] The panel for Peru covering 2009-2012 was excluded due to low number of observations.
[26] We also test the robustness of predictions to changes in the forecasting direction. That is, we predict second round incomes for every household observed in round one. We find qualitatively similar results when changing this direction. The full set of results are omitted due to space, but are available upon request.

Overall, the LASSO-PMM method performs well, since the predicted poverty rate estimates are close to the actual poverty headcounts calculated from the panel (or in general lie within their 95% confidence intervals). The figure shows predictions on the y-axis and actual poverty headcounts on the x-axis. Perfect correspondence between both is represented by the 45-degree line. Many of the predictions are either on the line or in close proximity. The table shows this overlap more clearly. For instance, in Argentina and Nicaragua, predicted and actual poverty rate estimates are identical. In Chile and Peru, while the point estimates are different, the 95% confidence intervals overlap. Therefore, predicted poverty rates from the LASSO-PMM are statistically indistinguishable from actual poverty rates calculated from panel data.

4.2 Predicted poverty transitions

The LASSO-PMM method also allows estimating poverty mobility patterns (i.e., transitions into and out of poverty). Columns 2 to 5 in Table 1 show the point estimates and 95% confidence intervals for actual mobility from panel data and predictions obtained by means of the LASSO-PMM approach using the latest panel available in each country. We report four joint probabilities: the probability that a household is poor in both rounds of data (*Poor, poor*), escapes poverty (*Poor, non-poor*), becomes poor (*Non-poor, poor*), and the likelihood that a household remains non-poor in both periods (*Non-poor, non-poor*).

Figure 2 summarizes the results for poverty transitions for all available years.[27] With few exceptions, many of the point estimates arising from the LASSO-PMM approach fall within the 95 percent confidence interval of actual poverty transitions. The confidence intervals in Table 1 show that in many countries and years, the LASSO-PMM approach accurately predicts mobility patterns. There is variation across panels, with some predictions being more optimistic, while

[27] The results for Peru using 2- and 3-year panels are shown in Appendix Figure A.1.

others more pessimistic. For instance, the method underestimates persistent poverty and overestimates upward mobility in Argentina. Conversely, the method is pessimistic in other cases, since downward mobility is overestimated in Argentina and Peru. However, the results indicate that the LASSO-PMM predictions approximate actual poverty transitions with a reasonable degree of precision.

The results in Table 1 and Appendix Figure A.1 also suggest that the method performs well irrespective of the length of the prediction window, given that we have one-year, two-year, three-year, and four-year panels. We find suggestive evidence that the method performs slightly better when the length of the panel is longer, perhaps due to less risk of idiosyncratic and unexpected aggregate shocks. This can be seen by comparing the results for Peru, where we can construct 1-year, 2-year, and 3-year panels. As the length of time between panels increases, mobility is better approximated by LASSO-PMM both in terms of the point estimates and their respective confidence intervals.

4.3 Changes in income and non-anonymous growth incidence curves

One additional contribution of the LASSO-PMM procedure compared to the parametric synthetic panel is that it also predicts income changes apart from just poverty transitions. Figure 3 graphs estimated household per capita income growth for two population groups defined by: i) the dynamic poverty transitions (poor, poor; poor, non-poor; non-poor, poor; and non-poor, non-poor) and ii) the quintiles of the income distribution in the first round—i.e., the non-anonymous growth incidence curves (GIC).[28] The figure presents the point estimate and the 95 percent confidence interval for the last available panel in each country. All estimates from the LASSO-PMM approach

[28] The anonymous GIC refers to quantile-level (or any othe percentile) growth of mean incomes by quantile (or any other percentile) of the income distribution (Ravallion and Chen 2003). On the other hand, the non-anonymous GIC refers to mean individual-level growth of incomes by quantile (or any other percentile) of the income distribution.

are compared with the actual income growth from panel data. Following the literature using GIC, we change the forecasting direction. That is, we predict second round household per capita incomes for every household interviewed in the first round of data.[29]

We find encouraging results for the prediction of income changes from the LASSO-PMM method. The predictions tend to trace out the actual income changes across the distribution. Therefore, in comparison to the parametric synthetic panel prediction method, the LASSO-PMM provides additional information on income growth, that approximates actual changes quite well. With some exceptions, our predictions overlap with the 95% confidence intervals for all sub-groups of poverty dynamics and quintiles of the income distribution. In some cases, the LASSO-PMM approach is optimistic, predicting higher income growth for the bottom two quintiles in certain countries. Once again, there is suggestive evidence that results are better when the length of the panel is longer.

4.4 Validation: sensitivity and robustness tests

These findings suggest that the LASSO-PMM approach is well-suited to predict aggregate poverty rates, transitions, and welfare changes with few assumptions and simple models that use harmonized variables. While our empirical approach is fairly general, we now proceed to test whether departures from the main assumptions and changes in how the method is applied affect the precision of the resulting welfare predictions. The main objective of these validations is to test the sensitivity and robustness of the LASSO-PMM approach to different scenarios that may be encountered in practice by other researchers that only have access to cross-section data.

[29] We also test the robustness of predictions to changes in the forecasting direction and find similar results.

Similar to other cross-section poverty prediction methods, applying the LASSO-PMM approach requires making several empirical decisions. We test the extent to which these decisions affect the sensitivity and robustness of the LASSO-PMM approach to obtain welfare dynamic predictions. For simplicity, we divide these exercises into two groups: i) *sensitivity tests*, in which the actual poverty rates and transitions do not change; and ii) *robustness exercises*, when we change the estimation procedure, sample, or other attributes that may change the actual welfare estimates. We compare predictions from the LASSO-PMM approach with the actual estimate for the poverty headcount and mobility for the latest available panel in each country.[30] Tables 2 to 7 show the results for our validation exercises by country. The upper panel presents sensitivity analyses, and the lower panel shows robustness tests, with LASSO-PMM estimates from Table 1 in the first row.

We begin with sensitivity analyses. These simulations change parameters that affect the prediction procedure but do not affect the actual poverty rate or poverty transitions. First, we increase the number of neighbors used in the PMM matching procedure from one to five. Second, we increase the number of PMM repetitions from 20 to 40. Last, we reduce the number of PMM repetitions from 20 to 10. The results for all countries suggest that varying these parameters does not affect predicted poverty rates and confidence intervals significantly.

Results for mobility patterns paint a similar picture. The full results for each of the four possible poverty transitions are shown in the last columns of Tables 2 to 7. In general, the confidence intervals of the LASSO-PMM method overlap with actual mobility estimates. These results can be interpreted as evidence that the method is stable to variation in the chosen parameters.

[30] Sensitivity and robustness results for all panels are omitted due to space restrictions but are available upon request.

We also test whether the accuracy of LASSO-PMM changes when considering different sets of variables. In particular, we begin with a model that includes household head gender, age, age squared, and years of schooling (model 1); then add household size and its square (model 2), and then include regional variables (model 3). The model tested so far is model 3 with interactions between the regional variables and all included covariates. Figure 4 shows predictions from these simplified models and compares the results to actual poverty transitions. Overall, we see an improvement in a model's predictions when we include more time-invariant and deterministic covariates. The models that include household size and its square and region fixed effects are closer to actual poverty transitions than just including household head attributes. Together with the results from our preferred model, these findings suggest that model specification matters.

We now turn to robustness exercises. These simulations involve making changes to the sample, covariates, and weighting structure that changes actual poverty and transition estimates. The lower panels of Tables 2 to 7 present the results for the most recent panel in each country.

First, we estimate household per capita income using Ordinary Least Squares (OLS) with PMM instead of LASSO. OLS sets $\lambda = 0$, using all included variables instead of selecting those that minimize the mean squared error. The results between both estimation procedures yield similar predictions within the confidence interval of true poverty rates and transitions. One plausible explanation is that, if the underlying model exhibits an adequate fit, OLS and LASSO find suitable neighbors throughout the original welfare distribution in a similar way. However, OLS and LASSO might arrive to different predictions in situations in which the underlying model does not have a good fit.

Second, we include household heads aged between 25 and 75 years old in the estimations, increasing the upper limit by ten years. This affects the actual poverty rate because we use a larger

sample. The findings suggest that the predictions remain accurate when including household heads older than 65 years of age.

Third, we consider what happens if the model had perfect foresight with respect to the included covariates. Given that we have panel data, we can observe all included covariates in both time periods. Instead of using cross-section controls, we make use of the actual panel structure to include actual variables for the period in which we predict poverty, in our case retrospective variables. Given that we estimate the model for households observed in round 2 and predict their income in round 1, this is feasible in our setup. As already mentioned, all variables are time-variant to some extent. Therefore, this exercise intends to correct for any changes in age, years of education, household size, and migration from households in each of the four countries. In general, using retrospective covariates does not affect performance substantially.

Last, we consider population changes by means of different weighting structures. As mentioned in Section 3, we weight the estimates using the survey weights in each of our data sources. This is common practice in household surveys to carry out inference on the entire population. There are two factors that may affect the performance of LASSO-PMM. On the one hand, statistics institutes in some countries vary their sampling structure from survey to survey. On the other hand, immigration shocks can dramatically alter the population in a country if many people enter or leave the country, which can affect welfare and poverty rates. If either of these events occur, they may affect welfare predictions from LASSO-PMM and other methods because they rely on the assumption that the population is relatively stable between periods.

The baseline LASSO-PMM approach estimates individual-level poverty using household level data and therefore multiplies household-level weights by household size. We try two different

alternatives. On one hand, we predict household-level poverty using household-level weights. On the other hand, we predict household-level poverty assuming equal weights across households.

The results in the tables show that changes in weighting structure do not significantly alter the predictions from the LASSO-PMM approach; in general, predictions are statistically indistinguishable.

In unreported results available upon request, we also test whether changing the direction of the prediction affects the estimates. Changing the direction so we predict second round "unobserved" income for households observed in period 1 leads to similar results across the board.

While not shown here, we also conduct these validation exercises on estimated growth incidence curves, finding that LASSO-PMM approximates actual income changes accurately. We also test whether the LASSO-PMM approach is sensitive to changes in the poverty line. We have used the $5.5 a day line throughout this paper, but also consider how the method fares with thresholds ranging from $5.5 to $13 a day.[31] Results show that irrespective of how we define the poverty threshold, poverty rates, transition, and income change predictions from the LASSO-PMM approach are in general statistically similar to the actual welfare estimates using panel data.

In addition to considering poor and non-poor individuals, we also test whether the method can predict welfare not only in two states, but across more than two. This extension is relevant when there is a fraction of "vulnerable" population that are more likely to enter poverty. Previous studies have identified these households as important to policy (Cruces et al. 2012), and we test whether the method is able to predict household defined as poor (income < $5.5 a day), vulnerable (income between $5.5-$11 a day), and middle class (income > $11 a day). The results, available upon request, indicate that expanding the number of population groups does not affect the

[31] In Chile, we test lines from $6 a day to $14.

method's overall performance. The confidence intervals for LASSO-PMM predictions and actual classifications of households into these categories overlap in general. These tests further support the accuracy of LASSO-PMM to predict welfare dynamics using cross-sections.

Overall, the sensitivity analyses and robustness exercises show that the LASSO-PMM approach for predicting poverty and mobility approximates actual welfare changes. The method performs well considering changes in parameters, estimation procedure, direction of the prediction, sample, covariates, and weighting schemes. The two main conclusions from this validation are: i) changing parameters to make the method more computationally intensive provides only marginal changes; and ii) models with greater explanatory power may improve LASSO-PMM's predictions. While we focus on generalizability by using a reduced set of harmonized controls to ensure comparability across countries and years, the determinants of poverty vary across countries and over time. This provides researchers in specific countries more leeway since they can use other variables in the models, such as ethnicity, and other retrospective and time-invariant correlates of welfare. As previous studies have shown, machine learning methods tend to work best when the model has high predictive power. While we show that the LASSO-PMM approach works well with a core-set of harmonized explanatory variables, our results suggest that poverty and mobility predictions may improve if the model is chosen adequately. An underlying model with a better fit will produce more accurate neighbors along a larger proportion of the welfare distribution and, therefore, the original distribution will be better replicated. Given these results, we now apply the procedure to harmonized cross-section data to predict poverty, mobility, and income changes in 43 countries across the world.

5. Application to cross-section data

We employ the basic LASSO-PMM specification used throughout the paper to estimate welfare dynamics using cross-section data in circa 2010 and 2015, considering one nearest neighbor and 20 PMM repetitions. We take all households in the first round cross-section survey and predict their income in the second round to obtain our main results. The welfare variable used in each of the 43 countries is shown in Appendix Table A.2. Given the number of countries and our harmonized approach to welfare prediction, we only include common variables in the model, which are described in Appendix Table A.3. We begin by predicting aggregate poverty rates, then study predicted mobility transitions, and finish by plotting predicted income changes across the distribution in growth incidence curves.

5.1 Predicted poverty rates across the world

We plot the poverty rate (using a $5.5 a day poverty line) that results from the cross-section data in the second round and the poverty rate that arises using the predicted income in the second round for those households surveyed the first round (Figure 5). First, as expected from the exercise with panel data, predicted and actual poverty rates are similar, given their proximity to the 45-degree line. Second, there is large heterogeneity in poverty across the world. While some countries have low poverty rates, such as Poland, there are nations where most of the population is below the poverty line, as in Rwanda. Despite the large differences across countries, LASSO-PMM approximates aggregate poverty rates well.[32]

[32] We originally had cross-sectional data for 49 countries. However, countries Bangladesh, India, Pakistan from SAR and Indonesia, Mongolia, and Phillipines from EAP were removed from the analysis because they presented substantial differences between actual and predicted poverty rates.

In the previous section, we highlighted that in order to predict well, the model matters. While we use a harmonized approach that includes a restricted set of variables, the models we fit have better predictive power than in previous studies. The second panel of Figure 5 presents the cumulative distribution of R-squared coefficients from the core papers in the welfare dynamics prediction literature and compares them with our estimations.[33] The figure shows that the Cumulative Distribution Function (*cdf*) in our application is shifted to the right, which implies that the models we estimate have better predictive power. Roughly 80 percent of our estimations have an R-squared greater than 0.30. This means that despite using a restricted set of harmonized variables, the estimated models have good predictive power. In country-specific applications that allow the inclusion of additional predictors of welfare, the explanatory power of implementing LASSO-PMM may rise further, potentially leading to greater precision in welfare prediction.[34]

5.2 Predicted poverty transitions across the world

We proceed to study predicted poverty transitions. We use the same four categories as in the previous section, but for simplicity refer to them as *chronic poor* (Poor, poor), *downward mobile* (Non-poor, poor), *upward mobile* (Poor, non-poor), and *never poor* (Non-poor, non-poor).

Figure 6 presents the predicted share of individuals in each of these categories by region. For our sample of 43 countries, the figure suggests that there has been considerable mobility; 22 percent of individuals changed their economic status. Moreover, the figure shows that, even in a period of poverty reduction (i.e., poverty decreased by 5 percentage points for this set of countries), about 9 percent of the population fell into poverty. Finally, the figure suggest that a large part of the population is immobile, about 24 percent of all individuals remained chronic poor.

[33] The papers considered are: Hérault and Jenkins (2019), Dang and Dabalen (2018), Dang and Lanjouw (2013), Dang, Lanjouw, Luoto and McKenzie (2014), Dang and Ianchovichina (2018) and Lucchetti (2017).
[34] We get similar results when predicting first round income for those surveyed in the second round. These results are available upon request.

This picture does not account for heterogeneity across and within regions. The results suggest that there are three types of regions. The first group has low chronic poverty levels and includes Europe and Central Asia (ECA), Latin America and the Caribbean (LAC), and East Asia and the Pacific (EAP). The second group, composed by South Asia (SAR) and Middle East and North Africa (MNA) lies on middle ground, where roughly one quarter to one third of the population seems to be chronically poor. The last group is composed by the Sub-Saharan Africa (SSA), where almost 80 percent of the population is chronically poor. Appendix Figure A.2 presents transition categories by country, sorted in increasing order by their chronic poverty level. Uruguay presents, for instance, one of the lowest chronic poverty levels, while on the other hand, Rwanda has one of the highest values. Interestingly, there are also large heterogeneities within regions. For instance, Kyrgyz Republic has a relatively large chronic poverty rate, probably comparable to African countries, but it is geographically surrounded by a group of countries with low chronic poverty.

These predictions suggest that a sizable fraction of individuals across the world remain trapped in chronic poverty, although this share varies by region. These predictions can help governments in different regions and countries to target policies that mitigate welfare losses for some households and that promote upward mobility at all levels.

5.3 Changes in income and growth incidence curves across the world

As mentioned throughout this paper, one of the advantages of the LASSO-PMM approach to welfare prediction is that it allows producing a point estimate of unobserved household income in any given period. This feature is particularly interesting because it allows looking at income changes at individual level (i.e., non-anonymized). We can therefore analyze what drives, or not, movements into the different categories: initial income or income changes.

Welfare changes depend on the initial level of welfare and the magnitude of changes in welfare. For instance, people can remain poor in both periods either because they start with an income far below the poverty line or because their income growth over time is not sufficient to lift them out of poverty.

To study whether predicted transitions are due to initial welfare or its respective changes, Figure 7 presents the initial and final welfare by region and the aggregate set of 43 countries. These welfare changes are presented for the same four groups: chronic poor, downward mobile, upward mobile, and never poor. Various findings are drawn from this figure. First, for the 43 countries, initial average income for the chronic poor ($2.8) was slightly lower than for the upward mobile ($3.8). Second, average income grew significantly for the upward mobile (to $11.0), while it remained almost constant for the chronic poor (final income was $2.9 on average for this group). Third, the never poor had an initial income substantially higher on average ($18.1) than those who were non-poor but move into poverty ($11.0). Finally, the average income of those who remained out of poverty almost did not change (final income was $17.2 on average for this group), while it decreased considerably for the downward mobile (final income was $3.9 on average for this group).

These magnitudes vary by region, although the overall message remains the same. For instance, the largest income losses for the downward mobile occur in Sub-Saharan Africa and Latin America and the Caribbean. Welfare gains are also highest in those two regions for the upward mobile. The remaining regions see lower changes, with losses for the downward mobile of around 55 percent and increases for the upward mobile close to 140 percent.[35]

[35] Appendix Figure A.3 presents the same results at the country-level.

We also estimate a growth incidence curve for the 43 available countries to determine whether predicted welfare changes are pro-poor or not. One important advantage of the LASSO-PMM is that, unlike traditional growth incidence curves developed by Ravallion and Chen (2003), we do not need to rely on the rank preservation assumption. On the contrary, for each household surveyed in the first round, we can estimate the annualized income growth at individual level, using the predicted income the second round. Afterwards, we construct centiles (deciles) at world (regional) level using the baseline income and we compute the average of the individual income growth for each group. Pooling all countries together, we observe in Figure 7 that annualized income growth is positive until the 68th percentile and becomes negative afterwards, suggesting that predicted income growth is pro-poor. Observing differences by region highlights some heterogeneity in these predictions. For example, income growth in the SAR region is always positive until the 8th decile, while in ECA the lower half of the distribution is predicted to grow, while the top half is not.

6. Conclusion

This is the first paper, to the best of our knowledge, that uses supervised machine learning and multiple imputation techniques to estimate welfare dynamics in the absence of panel datasets. The LASSO-PMM approach proposed in this paper estimates household-level welfare using a set of deterministic and time-invariant harmonized variables as controls and employs Predictive Mean Matching (PMM) with a household's nearest neighbor to provide estimates of poverty rates, transitions, and income changes. Compared to previous poverty prediction methods in the literature, LASSO-PMM makes fewer assumptions and estimates both poverty transitions and income changes. This research presents, validates, and applies the method using harmonized data

from 36 panels in four Latin American countries and cross-section data from 43 countries across the world.

In general, validation results indicate that LASSO-PMM predictions are statistically indistinguishable from actual household level poverty rates, transitions, and income changes. A series of validation exercises support the method's overall performance. We divide these exercises into two groups: i) *sensitivity tests*, in which the actual poverty rates and transitions do not change; and ii) *robustness exercises*, when we change the estimation procedure, sample, or other attributes that may change the actual welfare estimates. The validation exercises show that the LASSO-PMM approach for predicting poverty and mobility approximates actual welfare changes. The two main conclusions from this validation are: i) changing parameters to make the method more computationally intensive provides only marginal changes; and ii) models with greater explanatory power improve LASSO-PMM's predictions. These findings are sufficiently encouraging to suggest that estimating welfare dynamics using LASSO-PMM may accurately approximate actual welfare indicators in settings where panel data are unavailable. Interestingly, the robustness exercises show that OLS and LASSO yield similar results. A plausible explanation for this is that, if the underlying model exhibits an adequate fit, the two estimation methods find suitable neighbors throughout the original welfare distribution in a similar way. However, the two estimation methods might arrive to different predictions in situations in which the underlying model does not have a good fit.

We then apply the LASSO-PMM method to harmonized cross-section data. Aggregate poverty rates are accurately approximated when using cross-sections across a heterogeneous group of countries. Mobility probabilities are also estimated at the world, regional, and country-levels. Worldwide, we predict that 24 percent of all households are in chronic poverty, 9 percent are

downward mobile, 13 percent are upward mobile, and 53 percent are never poor. The method estimates that downward mobile households see their welfare fall by 65 percent on average, while upward mobile households escape poverty with a mean welfare increase of 190 percent. We also predict growth incidence curves and find evidence of a marked pro-poor growth.

This paper contributes to two strands of literature. First, it contributes to the methodological discussion on welfare dynamic prediction methods when panel data are unavailable. The LASSO-PMM approach presented in this paper embraces machine learning methods whose primary goal is accurate prediction, makes less restrictive assumptions than previous synthetic panel methods, and provides a simple framework that is easier to replicate in practice. The paper also highlights that the quality of the predictions is directly linked to the quality of the underlying welfare model used. Previous papers present predictions that are statistically indistinguishable from actual household poverty transitions with a relatively low fit. This is not the case for predictions in this paper; the quality of the underlying welfare model is key. Although the paper is restricted by a limited set of harmonized variables, the models we estimate have greater explanatory power compared to previous methods. Second, we contribute to the discussion of how the results from welfare prediction methods can be used to guide policy. The application of LASSO-PMM is carried out on 43 cross section datasets. To our knowledge, no other paper in the literature has used this amount of microdata to estimate welfare dynamics using cross-sections. The results from applying LASSO-PMM to cross-section data yield not only predicted poverty rates in the aggregate, but also the share of individuals who are upward and downward mobile, a key population for targeting anti-poverty policies. The results also provide insight on the magnitude of income changes that drive individuals both into and out of poverty, which is also useful knowledge for development efforts.

The LASSO-PMM approach described in this paper presents several advantages compared to other methods. For instance, unlike the parametric synthetic panel, the LASSO-PMM allows to calculate household-level income changes instead of just probabilities of poverty transitions. In addition, unlike the parametric approach, the LASSO-PMM does not require additional estimation of unknown parameters, such as the age-cohort correlation of residuals from cross-sections. Finally, instead of allocating residuals from the first-round regression to estimate first-round incomes for households interviewed in the second round, as in the upper bound non-parametric synthetic panel method, the LASSO-PMM directly imputes the observed total income from a first-round near neighbor observations. Therefore, all first-round sources of incomes (e.g., labor income, public social transfers, private transfers, capital income, rent, etc.) from the same neighbors can also be imputed in the second round. As such, future research could then analyze the contribution of each source of income to poverty and welfare changes between two moments in time.

References

Bourguignon, François. 2015. “Appraising Income Inequality Databases in Latin America.” *Journal of Economic Inequality* 13(4): 557–78.

Bourguignon, F. and A. Hector Moreno M. 2018. On synthetic income panels. PSE Working Papers N°2018-63. <halshs-01988068>.

Andrés Castañeda, Leonardo Gasparini, Santiago Garriga, Leonardo Lucchetti, and Daniel Valderrama, 2018. "How Sensitive Is Regional Poverty Measurement in Latin America to the Value of the Poverty Line?," *Economía Journal*, The Latin American and Caribbean Economic Association - LACEA, vol. 0, pages 33-58, November.

Cruces, Guillermo, Marcelo Bérgolo, and Andrés Ham. 2012. “Assessing the Predictive Power of Vulnerability Measures: Evidence from Panel Data for Argentina and Chile”. *Journal of Income Distribution*, 21(1):28-64.

Cruces, Guillermo, Peter Lanjouw, Leonardo Lucchetti, Elizaveta Perova, Renos Vakis, and Mariana Viollaz. 2015. “Intra-Generational Mobility and Repeated Cross-Sections: A Three- Country Validation Exercise.” *Journal of Economic Inequality* 13 (2): 161–79.

Dang, Hai-Anh and Ianchovichina, Elena (2018). “Welfare Dynamics with Synthetic Panels: The Case of the Arab World in Transition”. *Review of Income and Wealth*.

Dang, Hai-Anh and Peter Lanjouw 2018. Poverty dynamics in India between 2004-2012: Insights from longitudinal analysis using synthetic panel data. *Economic Development and Cultural Change*.

Dang, Hai-Anh and Andrew Dabalen. 2018. “Is Poverty in Africa Mostly Chronic or Transient? Evidence from Synthetic Panel Data.” *The Journal of Development Studies*. DOI: 10.1080/00220388.2017.1417585

Dang, Hai-Anh, Peter Lanjouw, Jill Luoto, and David McKenzie. 2014. “Using Repeated Cross-Sections to Explore Movements into and out of Poverty”. *Journal of Development Economics*. 107, 112–128.

Dang, Hai-Anh and Peter Lanjouw. 2013. “Measuring Poverty Dynamics with Synthetic Panels Based on Cross-Sections.” World Bank Policy Research Working Paper 6540.

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

Elbers, C., Lanjouw, J.O., and Lanjouw, P. 2003. Micro-level estimation of poverty and inequality. *Econometrica* 71(1), 355–364

Ferreira, Francisco H. G., Julian Messina, Jamele Rigolini, Luis-Felipe López-Calva, Maria Ana Lugo, and Renos Vakis. 2013. *Economic Mobility and the Rise of the Latin American Middle Class*. Washington, DC: World Bank.

Ferreira, F.H.G., Chen, S., Dabalen, A. et al. 2016. A global count of the extreme poor in 2012: data issues, methodology and initial results. *The Journal of Economic Inequality* 14, 141–172.

Hérault, Nicolas, and Stephen P. Jenkins. (2019) "How valid are synthetic panel estimates of poverty dynamics?" *The Journal of Economic Inequality.* 17(1): 51-76.

Jolliffe, D., Prydz, E.B. 2016. Estimating international poverty lines from comparable national thresholds. *The Journal of Economic Inequality* 14, 185–198

LABLAC (2018). Labor Database for Latin America and The Caribbean (CEDLAS and The World Bank).

Lucchetti, Leonardo. 2017. "Who Escaped Poverty and Who Was Left Behind? A Non-Parametric Approach to Explore Welfare Dynamics Using Cross-Sections." World Bank Policy Research Working Paper No. 8220.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives,* 31 (2): 87-106.

SEDLAC (2018). Socio-Economic Database for Latin America and the Caribbean (CEDLAS and The World Bank).

Ravallion, M., and S. Chen. 2003. "Measuring Propoor Growth." *Economics Letters* 78 (1): 93–99.

Serajuddin, Umar; Uematsu, Hiroki; Wieser, Christina; Yoshida, Nobuo; Dabalen, Andrew L. 2015. "Data deprivation: another deprivation to end." Policy Research working paper; no. WPS 7252. Washington, D.C.: World Bank Group.

Van Buuren, S. (2018). *Flexible imputation of missing data*. Chapman and Hall/CRC.

Vakis, Renos; Jamele Rigolini; and Leonardo Lucchetti. 2016. *Left behind: chronic poverty in Latin America and the Caribbean*. Washington, DC; World Bank Group.

World Bank (2018). *Poverty and shared prosperity 2018: Piecing together the poverty puzzle.* Washington, DC: World Bank. License: Creative Commons Attribu- tion CC BY 3.0 IGO

Tables and figures

Table 1. Actual and predicted poverty headcounts and transitions

Country	Years	Poverty rate		Poor, poor		Poor, non-poor		Non-poor, poor		Non-poor, non-poor		Obs.
		Actual	Pred.	Actual	Pred.	Actual	Pred.	Actual	Pred.	Actual	Pred.	
Argentina	2013 - 2014	7	7	3	1	4	6	4	6	89	87	3,329
		(6, 7)	(4, 9)	(2, 3)	(0, 1)	(3, 5)	(4, 9)	(3, 5)	(5, 7)	(88, 90)	(85, 90)	
Chile	2015 - 2016	13	11	2	1	11	10	2	3	85	86	4,052
		(12, 14)	(9, 13)	(1, 2)	(0, 1)	(10, 12)	(9, 12)	(2, 2)	(3, 4)	(84, 86)	(84, 88)	
Peru	2015 - 2016	26	25	17	13	9	12	8	12	66	63	3,179
		(24, 27)	(21, 28)	(15, 18)	(11, 15)	(8, 10)	(10, 14)	(7, 9)	(10, 14)	(64, 68)	(61, 66)	
	2014 - 2016	25	23	15	11	10	12	8	12	67	65	1,163
		(22, 27)	(19, 28)	(12, 17)	(8, 14)	(9, 12)	(9, 16)	(6, 10)	(9, 14)	(64, 70)	(61, 69)	
	2013 - 2016	30	29	17	13	13	16	8	12	62	60	1,058
		(27, 33)	(23, 34)	(15, 19)	(10, 16)	(11, 15)	(11, 20)	(6, 9)	(9, 15)	(59, 65)	(54, 65)	
Nicaragua	2001 - 2005	60	60	41	36	19	24	10	15	30	24	907
		(57, 63)	(54, 67)	(38, 44)	(31, 41)	(16, 21)	(20, 29)	(8, 12)	(11, 19)	(27, 33)	(20, 29)	

Data source: SEDLAC and LABLAC data (CEDLAS and the World Bank). Note: Results are constrained to the panel sample of households whose heads are between 25 and 65 years old. Results in column [Actual] show actual panel poverty estimates. [Pred.] shows LASSO-PMM estimates. Poor are those individuals with a per capita income lower than $5.5 per person per day, except in Chile where we define a relative poverty line equal to 0.5 of median per capita income. Poverty lines and incomes are expressed in 2011 $PPP/day. 95% confidence intervals between parenthesis. Predictions are obtained with 20 repetitions and 1 neighbor of PMM. The column Pred. presents first round income predictions for all individuals surveyed in the second round. Results are based on the following datasets: Encuesta Permanente de Hogares-Contina in Argentina; Nueva Encuesta Nacional de Empleo in Chile; Encuesta Nacional de Hogares sobre Medición de Nivel de Vida in Nicaragua; and Encuesta Nacional de Hogares in Peru.

Table 2. Sensitivity and robustness of LASSO-PMM predictions, Argentina (2013-2014)

	Poverty rate		Poor, poor		Poor, non-poor		Non-poor, poor		Non-poor, non-poor	
	Actual	Pred.	Actual	Pred.	Actual	Pred.	Actual	Pred.	Actual	Pred.
LASSO-PMM	7	7	3	1	4	6	4	6	89	87
	(6, 7)	(4, 9)	(2, 3)	(0, 1)	(3, 5)	(4, 9)	(3, 5)	(5, 7)	(88, 90)	(85, 90)
Sensitivity										
5 neighbors	-	7	-	1	-	7	-	6	-	87
		(5, 10)		(0, 2)		(4, 9)		(5, 7)		(84, 90)
40 repetitions	-	7	-	1	-	6	-	6	-	87
		(4, 10)		(0, 1)		(4, 9)		(5, 7)		(84, 90)
10 repetitions	-	7	-	1	-	6	-	6	-	87
		(4, 10)		(0, 1)		(3, 10)		(5, 7)		(84, 90)
Robustness										
OLS	7	7	3	1	4	7	4	6	89	87
	(6, 7)	(5, 10)	(2, 3)	(0, 2)	(3, 5)	(4, 9)	(3, 5)	(5, 7)	(88, 90)	(85, 89)
Ages 25-75	7	5	2	1	5	5	4	5	89	89
	(7, 8)	(3, 8)	(2, 3)	(0, 1)	(5, 6)	(2, 7)	(3, 4)	(4, 6)	(88, 90)	(87, 92)
Retrospective X's	8	6	3	1	5	5	5	6	87	87
	(7, 9)	(4, 8)	(2, 3)	(0, 2)	(4, 6)	(3, 7)	(4, 5)	(5, 8)	(86, 89)	(85, 90)
HH weights	5	5	2	0	3	4	3	4	92	91
	(4, 6)	(3, 6)	(1, 2)	(0, 1)	(2, 4)	(3, 6)	(3, 4)	(4, 5)	(91, 93)	(89, 92)
Equal weights	6	5	2	1	4	4	3	4	92	91
	(5, 6)	(4, 6)	(1, 2)	(0, 1)	(3, 4)	(3, 5)	(2, 3)	(3, 5)	(91, 93)	(90, 92)

Data source: SEDLAC and LABLAC data (CEDLAS and the World Bank). Note: Results in column [Actual] show actual panel poverty estimates. [Pred.] shows predicted estimates. Poor are those individuals with a per capita income lower than $5.5 per person per day. Poverty lines and incomes are expressed in 2011 $PPP/day. 95% confidence intervals between parenthesis. Results in first row are constrained to the panel sample of households whose heads are between 25 and 65 years old and are obtained with 20 repetitions and 1 neighbor of PMM. We present the following validation exercises (see Section 4 for details): i) increasing the number of neighbors to 5, ii) performing 40 repetitions of PMM, iii) performing 10 repetitions of PMM, iv) using OLS instead of LASSO, v) expanding the sample to include household heads between 25 and 75 years old, vi) using retrospective X's from the panel, vii) using household-level weights, and viii) weighting each household equally. The column Pred. presents first round income predictions for all individuals surveyed in the second round. Results are based on the Encuesta Permanente de Hogares-Contina.

Table 3. Sensitivity and robustness of LASSO-PMM predictions, Chile (2015-2016)

	Poverty rate		Poor, poor		Poor, non-poor		Non-poor, poor		Non-poor, non-poor	
	Actual	Pred.	Actual	Pred.	Actual	Pred.	Actual	Pred.	Actual	Pred.
LASSO-PMM	13	11	2	1	11	10	2	3	85	86
	(12, 14)	(9, 13)	(1, 2)	(0, 1)	(10, 12)	(9, 12)	(2, 2)	(3, 4)	(84, 86)	(84, 88)
Sensitivity										
5 neighbors	-	11	-	1	-	10	-	3	-	86
		(9, 13)		(0, 1)		(8, 12)		(2, 4)		(84, 88)
40 repetitions	-	11	-	1	-	10	-	3	-	86
		(9, 13)		(0, 1)		(8, 12)		(2, 4)		(84, 88)
10 repetitions	-	11	-	1	-	10	-	3	-	86
		(9, 13)		(0, 1)		(8, 12)		(3, 4)		(84, 88)
Robustness										
OLS	13	11	2	1	11	10	2	3	85	86
	(12, 14)	(8, 13)	(1, 2)	(0, 1)	(10, 12)	(8, 13)	(2, 2)	(2, 4)	(84, 86)	(84, 89)
Ages 25-75	12	10	1	0	11	10	2	2	87	88
	(11, 13)	(8, 12)	(1, 1)	(0, 1)	(10, 11)	(7, 12)	(1, 2)	(2, 3)	(86, 87)	(85, 90)
Retrospective X's	13	11	2	1	11	11	2	3	85	86
	(12, 14)	(9, 13)	(1, 2)	(0, 1)	(10, 12)	(8, 13)	(1, 2)	(2, 4)	(84, 86)	(84, 88)
HH weights	11	9	2	0	10	9	2	3	87	88
	(10, 12)	(7, 11)	(1, 2)	(0, 1)	(9, 11)	(7, 11)	(2, 2)	(3, 4)	(86, 88)	(86, 90)
Equal weights	11	9	2	0	9	9	2	4	87	87
	(10, 12)	(7, 11)	(1, 2)	(0, 1)	(8, 10)	(7, 11)	(2, 3)	(3, 4)	(86, 88)	(85, 89)

Data source: SEDLAC and LABLAC data (CEDLAS and the World Bank). Note: Results in column [Actual] show actual panel poverty estimates. [Pred.] shows predicted estimates. Poor are those individuals with a per capita income lower than the relative poverty line equal to 0.5 of median per capita income. Poverty lines and incomes are expressed in 2011 $PPP/day. 95% confidence intervals between parenthesis. Results in first row are constrained to the panel sample of households whose heads are between 25 and 65 years old and are obtained with 20 repetitions and 1 neighbor of PMM. We present the following validation exercises (see Section 4 for details): i) increasing the number of neighbors to 5, ii) performing 40 repetitions of PMM, iii) performing 10 repetitions of PMM, iv) using OLS instead of LASSO, v) expanding the sample to include household heads between 25 and 75 years old, vi) using retrospective X's from the panel, vii) using household-level weights, and viii) weighting each household equally. The column Pred. presents first round income predictions for all individuals surveyed in the second round. Results are based on the Nueva Encuesta Nacional de Empleo.

Table 4. Sensitivity and robustness of LASSO-PMM predictions, Peru (2015-2016)

	Poverty rate		Poor, poor		Poor, non-poor		Non-poor, poor		Non-poor, non-poor	
	Actual	Pred.	Actual	Pred.	Actual	Pred.	Actual	Pred.	Actual	Pred.
LASSO-PMM	26	25	17	13	9	12	8	12	66	63
	(24, 27)	(21, 28)	(15, 18)	(11, 15)	(8, 10)	(10, 14)	(7, 9)	(10, 14)	(64, 68)	(61, 66)
Sensitivity										
5 neighbors	-	25	-	13	-	13	-	12	-	63
		(22, 28)		(11, 14)		(10, 15)		(10, 14)		(60, 65)
40 repetitions	-	25	-	13	-	12	-	12	-	63
		(22, 28)		(11, 14)		(10, 14)		(10, 14)		(61, 66)
10 repetitions	-	25	-	13	-	12	-	12	-	63
		(22, 28)		(11, 15)		(10, 14)		(10, 14)		(61, 66)
Robustness										
OLS	26	25	17	12	9	13	8	12	66	63
	(24, 27)	(22, 28)	(15, 18)	(11, 14)	(8, 10)	(10, 15)	(7, 9)	(10, 14)	(64, 68)	(60, 65)
Ages 25-75	25	24	15	11	9	13	8	12	68	64
	(23, 26)	(21, 26)	(14, 17)	(10, 13)	(8, 10)	(10, 15)	(7, 9)	(11, 13)	(66, 69)	(62, 67)
Retrospective X's	28	26	17	12	10	13	7	12	65	62
	(26, 29)	(23, 29)	(16, 19)	(11, 14)	(9, 11)	(11, 16)	(6, 8)	(10, 14)	(64, 67)	(59, 65)
HH weights	23	22	14	10	9	11	8	12	69	67
	(21, 24)	(19, 24)	(13, 15)	(9, 12)	(8, 10)	(9, 13)	(7, 9)	(10, 13)	(67, 71)	(64, 69)
Equal weights	29	28	19	14	10	14	9	14	62	59
	(27, 30)	(25, 30)	(17, 20)	(12, 16)	(9, 11)	(12, 15)	(8, 10)	(12, 15)	(60, 64)	(56, 61)

Data source: SEDLAC and LABLAC data (CEDLAS and the World Bank). Note: Results in column [Actual] show actual panel poverty estimates. [Pred.] shows predicted estimates. Poor are those individuals with a per capita income lower than $5.5 per person per day. Poverty lines and incomes are expressed in 2011 $PPP/day. 95% confidence intervals between parenthesis. Results in first row are constrained to the panel sample of households whose heads are between 25 and 65 years old and are obtained with 20 repetitions and 1 neighbor of PMM. We present the following validation exercises (see Section 4 for details): i) increasing the number of neighbors to 5, ii) performing 40 repetitions of PMM, iii) performing 10 repetitions of PMM, iv) using OLS instead of LASSO, v) expanding the sample to include household heads between 25 and 75 years old, vi) using retrospective X's from the panel, vii) using household-level weights, and viii) weighting each household equally. The column Pred. presents first round income predictions for all individuals surveyed in the second round. Results are based on the Encuesta Nacional de Hogares.

Table 5. Sensitivity and robustness of LASSO-PMM predictions, Peru (2014-2016)

	Poverty rate		Poor, poor		Poor, non-poor		Non-poor, poor		Non-poor, non-poor	
	Actual	Pred.	Actual	Pred.	Actual	Pred.	Actual	Pred.	Actual	Pred.
LASSO-PMM	25	23	15	11	10	12	8	12	67	65
	(22, 27)	(19, 28)	(12, 17)	(8, 14)	(9, 12)	(9, 16)	(6, 10)	(9, 14)	(64, 70)	(61, 69)
Sensitivity										
5 neighbors	-	22	-	11	-	12	-	12	-	66
		(19, 26)		(8, 13)		(9, 14)		(9, 14)		(62, 69)
40 repetitions	-	23	-	11	-	12	-	12	-	65
		(19, 28)		(8, 14)		(9, 16)		(9, 15)		(61, 69)
10 repetitions	-	24	-	11	-	13	-	12	-	65
		(19, 28)		(8, 14)		(9, 16)		(9, 14)		(61, 69)
Robustness										
OLS	25	23	15	11	10	13	8	12	67	65
	(22, 27)	(19, 28)	(12, 17)	(8, 13)	(9, 12)	(9, 17)	(6, 10)	(9, 15)	(64, 70)	(60, 69)
Ages 25-75	27	24	16	11	11	13	8	13	65	63
	(25, 29)	(19, 29)	(14, 18)	(9, 13)	(9, 12)	(9, 17)	(6, 9)	(10, 15)	(63, 68)	(59, 68)
Retrospective X's	25	26	15	11	10	14	8	12	67	63
	(22, 27)	(20, 31)	(13, 17)	(9, 14)	(8, 12)	(9, 19)	(7, 10)	(9, 14)	(64, 70)	(58, 68)
HH weights	22	20	11	8	10	12	7	10	71	70
	(19, 24)	(16, 24)	(10, 13)	(6, 11)	(9, 12)	(8, 15)	(6, 9)	(8, 13)	(68, 74)	(66, 73)
Equal weights	28	26	16	12	12	14	9	13	64	62
	(25, 30)	(22, 30)	(14, 18)	(10, 14)	(10, 13)	(10, 17)	(7, 10)	(10, 15)	(61, 67)	(58, 65)

Data source: SEDLAC and LABLAC data (CEDLAS and the World Bank). Note: Results in column [Actual] show actual panel poverty estimates. [Pred.] shows predicted estimates. Poor are those individuals with a per capita income lower than $5.5 per person per day. Poverty lines and incomes are expressed in 2011 $PPP/day. 95% confidence intervals between parenthesis. Results in first row are constrained to the panel sample of households whose heads are between 25 and 65 years old and are obtained with 20 repetitions and 1 neighbor of PMM. We present the following validation exercises (see Section 4 for details): i) increasing the number of neighbors to 5, ii) performing 40 repetitions of PMM, iii) performing 10 repetitions of PMM, iv) using OLS instead of LASSO, v) expanding the sample to include household heads between 25 and 75 years old, vi) using retrospective X's from the panel, vii) using household-level weights, and viii) weighting each household equally. The column Pred. presents first round income predictions for all individuals surveyed in the second round. Results are based on the Encuesta Nacional de Hogares.

Table 6. Sensitivity and robustness of LASSO-PMM predictions, Peru (2013-2016)

	Poverty rate		Poor, poor		Poor, non-poor		Non-poor, poor		Non-poor, non-poor	
	Actual	Pred.	Actual	Pred.	Actual	Pred.	Actual	Pred.	Actual	Pred.
LASSO-PMM	30	29	17	13	13	16	8	12	62	60
	(27, 33)	(23, 34)	(15, 19)	(10, 16)	(11, 15)	(11, 20)	(6, 9)	(9, 15)	(59, 65)	(54, 65)
Sensitivity										
5 neighbors	-	28	-	13	-	15	-	12	-	60
		(24, 32)		(10, 16)		(11, 19)		(9, 14)		(56, 65)
40 repetitions	-	28	-	13	-	15	-	12	-	60
		(23, 34)		(10, 16)		(11, 20)		(9, 15)		(55, 65)
10 repetitions	-	28	-	13	-	15	-	12	-	60
		(22, 35)		(10, 16)		(10, 21)		(9, 15)		(54, 66)
Robustness										
OLS	30	28	17	12	13	16	8	12	62	60
	(27, 33)	(22, 34)	(15, 19)	(9, 16)	(11, 15)	(11, 20)	(6, 9)	(9, 16)	(59, 65)	(55, 64)
Ages 25-75	27	24	17	12	11	13	7	12	66	64
	(25, 30)	(19, 30)	(15, 19)	(9, 15)	(9, 12)	(9, 16)	(5, 8)	(9, 15)	(63, 69)	(60, 68)
Retrospective X's	29	30	17	14	11	16	7	11	64	59
	(26, 31)	(25, 35)	(15, 20)	(10, 17)	(9, 13)	(12, 20)	(6, 9)	(8, 14)	(61, 67)	(55, 64)
HH weights	28	25	15	10	13	14	7	12	65	63
	(25, 31)	(20, 29)	(13, 17)	(8, 13)	(11, 15)	(10, 18)	(6, 9)	(9, 15)	(62, 68)	(59, 68)
Equal weights	32	30	19	14	13	16	8	13	59	57
	(30, 35)	(26, 34)	(17, 22)	(12, 17)	(11, 15)	(12, 19)	(7, 10)	(10, 16)	(56, 62)	(53, 61)

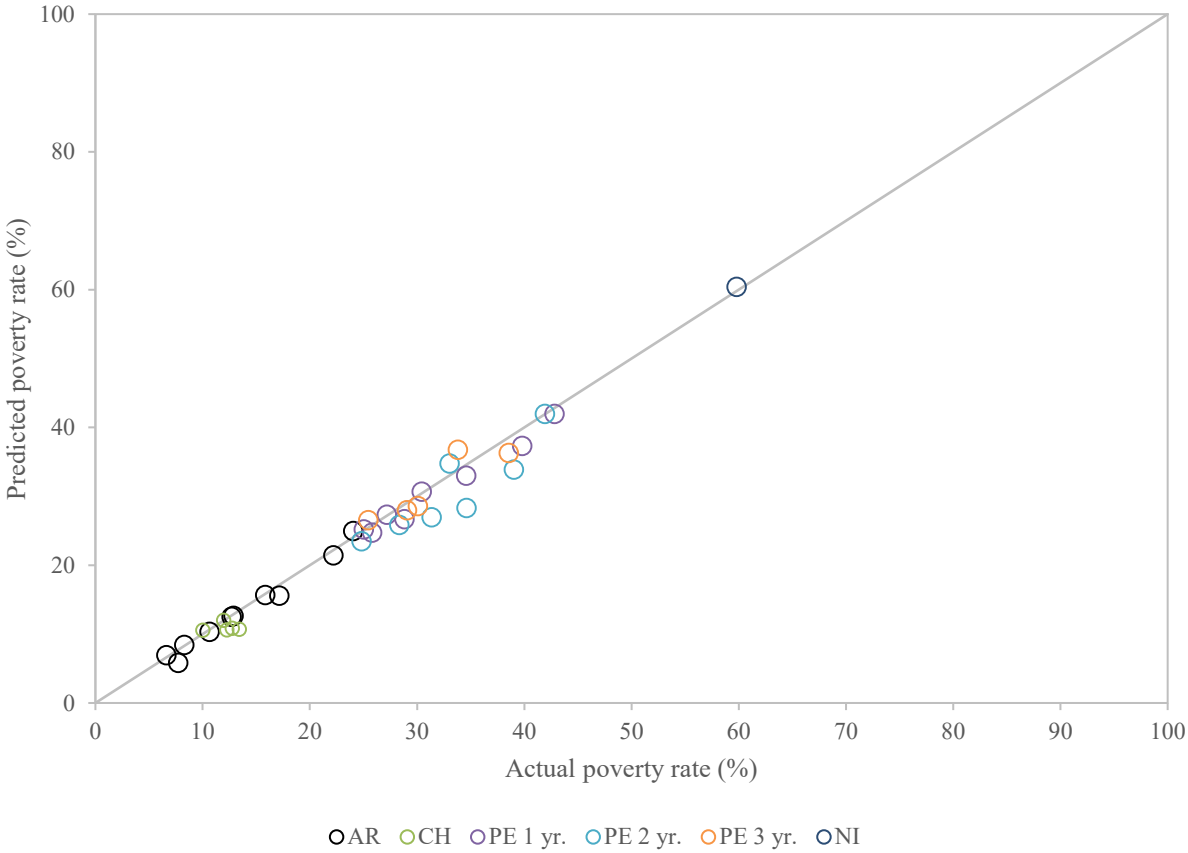
Data source: SEDLAC and LABLAC data (CEDLAS and the World Bank). Note: Results in column [Actual] show actual panel poverty estimates. [Pred.] shows predicted estimates. Poor are those individuals with a per capita income lower than $5.5 per person per day. Poverty lines and incomes are expressed in 2011 $PPP/day. 95% confidence intervals between parenthesis. Results in first row are constrained to the panel sample of households whose heads are between 25 and 65 years old and are obtained with 20 repetitions and 1 neighbor of PMM. We present the following validation exercises (see Section 4 for details): i) increasing the number of neighbors to 5, ii) performing 40 repetitions of PMM, iii) performing 10 repetitions of PMM, iv) using OLS instead of LASSO, v) expanding the sample to include household heads between 25 and 75 years old, vi) using retrospective X's from the panel, vii) using household-level weights, and viii) weighting each household equally. The column Pred. presents first round income predictions for all individuals surveyed in the second round. Results are based on the Encuesta Nacional de Hogares.

Table 7. Sensitivity and robustness of LASSO-PMM predictions, Nicaragua (2001-2005)

	Poverty rate		Poor, poor		Poor, non-poor		Non-poor, poor		Non-poor, non-poor	
	Actual	Pred.	Actual	Pred.	Actual	Pred.	Actual	Pred.	Actual	Pred.
LASSO-PMM	60	60	41	36	19	24	10	15	30	24
	(57, 63)	(54, 67)	(38, 44)	(31, 41)	(16, 21)	(20, 29)	(8, 12)	(11, 19)	(27, 33)	(20, 29)
Sensitivity										
5 neighbors	-	61	-	36	-	25	-	15	-	24
		(55, 66)		(32, 40)		(20, 29)		(12, 19)		(20, 28)
40 repetitions	-	60	-	36	-	24	-	15	-	25
		(54, 66)		(31, 41)		(20, 28)		(11, 20)		(20, 29)
10 repetitions	-	60	-	36	-	24	-	16	-	25
		(52, 68)		(31, 41)		(19, 29)		(11, 20)		(19, 30)
Robustness										
OLS	60	60	41	36	19	25	10	16	30	24
	(57, 63)	(54, 67)	(38, 44)	(31, 40)	(16, 21)	(20, 29)	(8, 12)	(11, 20)	(27, 33)	(20, 28)
Ages 25-75	61	59	42	36	19	24	10	16	29	25
	(58, 64)	(54, 65)	(39, 45)	(32, 40)	(17, 22)	(20, 27)	(8, 12)	(13, 19)	(26, 32)	(21, 28)
Retrospective X's	60	65	40	37	20	28	9	12	31	23
	(57, 63)	(59, 70)	(37, 43)	(33, 41)	(17, 23)	(24, 32)	(7, 11)	(9, 16)	(28, 34)	(19, 27)
HH weights	53	55	34	29	19	25	10	14	37	32
	(50, 56)	(48, 61)	(30, 37)	(25, 34)	(17, 22)	(21, 30)	(8, 12)	(10, 18)	(34, 41)	(27, 36)
Equal weights	58	59	38	34	19	25	9	14	33	28
	(55, 61)	(53, 64)	(35, 42)	(30, 38)	(17, 22)	(21, 29)	(7, 11)	(10, 17)	(30, 36)	(23, 32)

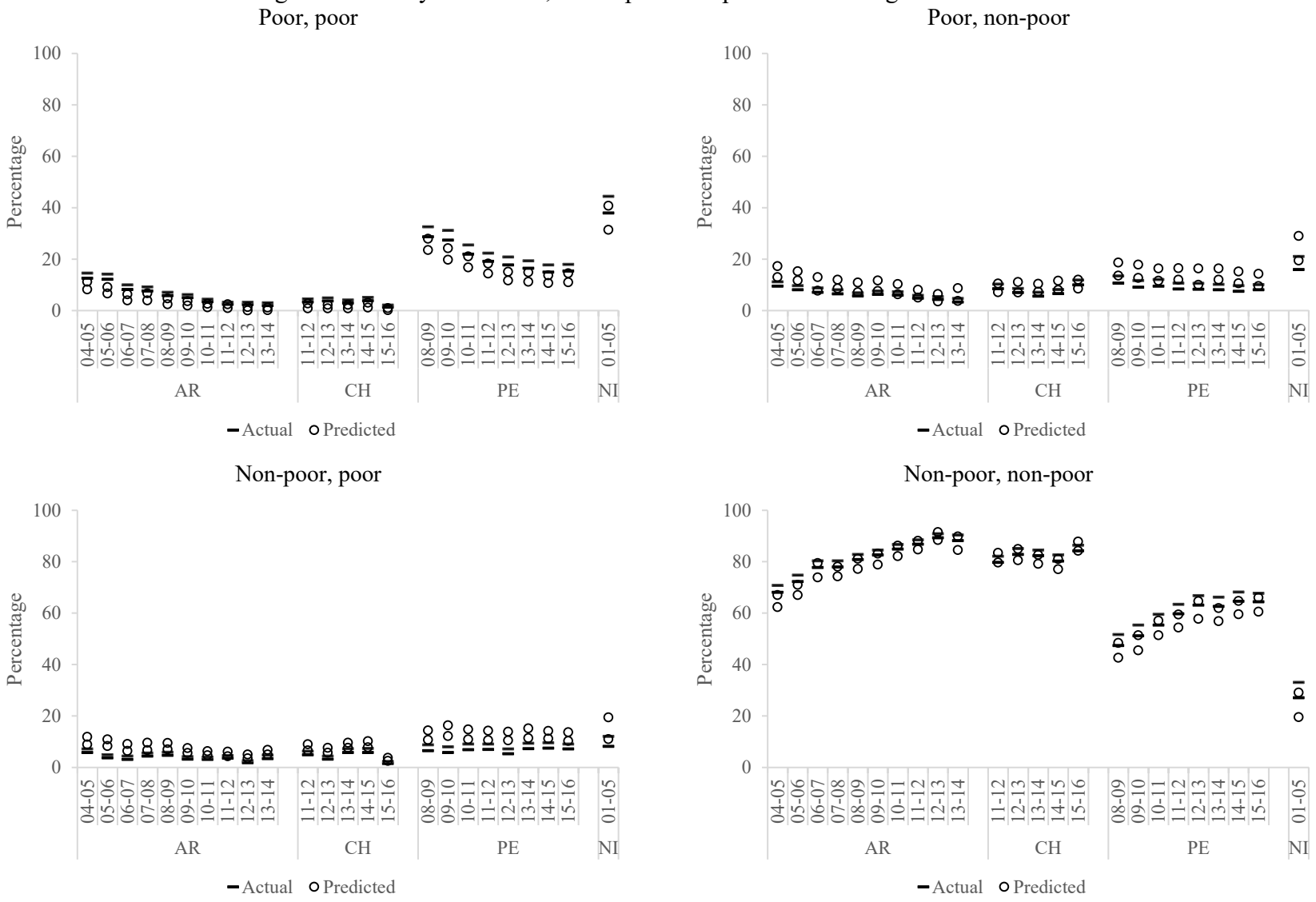
Data source: SEDLAC and LABLAC data (CEDLAS and the World Bank). Note: Results in column [Actual] show actual panel poverty estimates. [Pred.] shows predicted estimates. Poor are those individuals with a per capita income lower than $5.5 per person per day. Poverty lines and incomes are expressed in 2011 $PPP/day. 95% confidence intervals between parenthesis. Results in first row are constrained to the panel sample of households whose heads are between 25 and 65 years old and are obtained with 20 repetitions and 1 neighbor of PMM. We present the following validation exercises (see Section 4 for details): i) increasing the number of neighbors to 5, ii) performing 40 repetitions of PMM, iii) performing 10 repetitions of PMM, iv) using OLS instead of LASSO, v) expanding the sample to include household heads between 25 and 75 years old, vi) using retrospective X's from the panel, vii) using household-level weights, and viii) weighting each household equally. The column Pred. presents first round income predictions for all individuals surveyed in the second round. Results are based on the Encuesta Nacional de Hogares sobre Medición de Nivel de Vida.

Figure 1. Actual and predicted poverty headcount in first round using second round observations



Data source: SEDLAC and LABLAC data (CEDLAS and the World Bank). Note: Results are constrained to the panel sample of households whose heads are between 25 and 65 years old. Results on the x-axis [Actual] show actual panel poverty estimates. The y-axis [Predicted] shows LASSO-PMM estimates. The 45-degree line represents points where actual and predicted values are equal. Poor are those individuals with a per capita income lower than $5.5 per person per day, except in Chile where we define a relative poverty line equal to 0.5 of median per capita income. Poverty lines and incomes are expressed in 2011 $PPP/day. Predictions are obtained with 20 repetitions and 1 neighbor of PMM. The vertical axis presents first round income predictions for all individuals surveyed in the second round. The horizontal axis presents actual poverty in the first round. Results are based on the following datasets: Encuesta Permanente de Hogares-Contina in Argentina; Nueva Encuesta Nacional de Empleo in Chile; Encuesta Nacional de Hogares sobre Medición de Nivel de Vida in Nicaragua; and Encuesta Nacional de Hogares in Peru. AR refers to Argentina; CH to Chile; PE to Peru, and NI to Nicaragua.

Figure 2. Poverty transitions, actual panel vs. predictions using LASSO-PMM



Data source: SEDLAC and LABLAC data (CEDLAS and the World Bank). Note: The figure shows confidence intervals for actual poverty transitions (lines) and predictions obtained using the LASSO-PMM approach (circles). AR refers to Argentina; CH to Chile; PE to Peru, and NI to Nicaragua. See Table 1 for additional notes.

Figure 3. Household-level income change by groups of mobility transition and quintiles of the income distribution

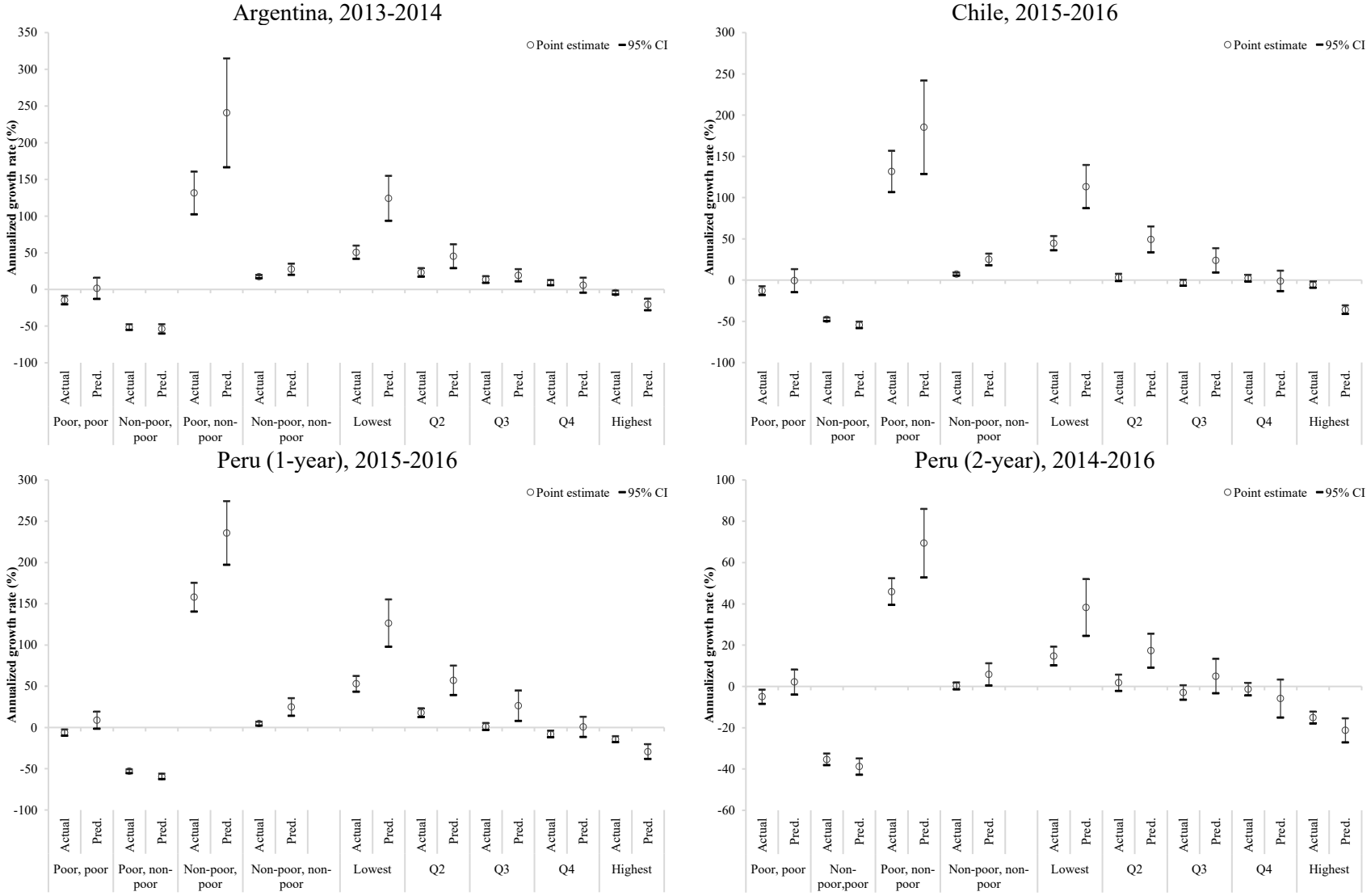
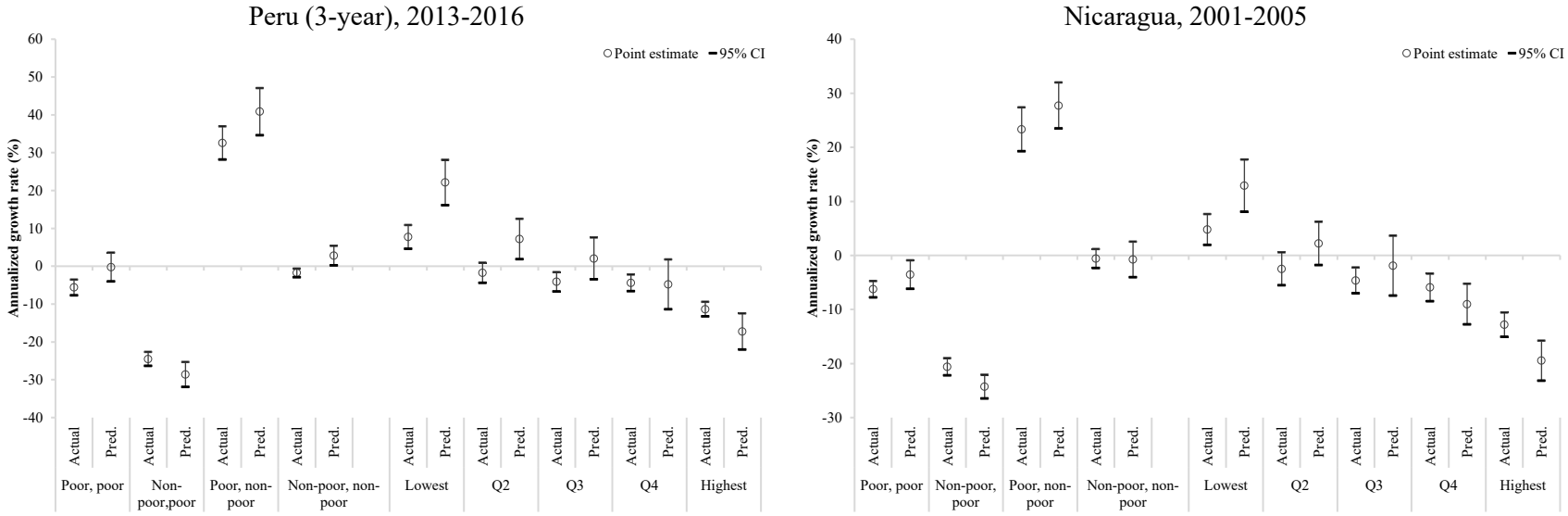


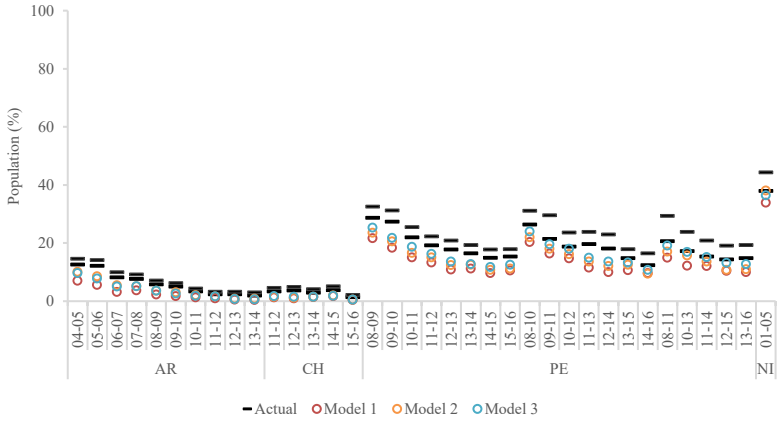
Figure 3 (continued)



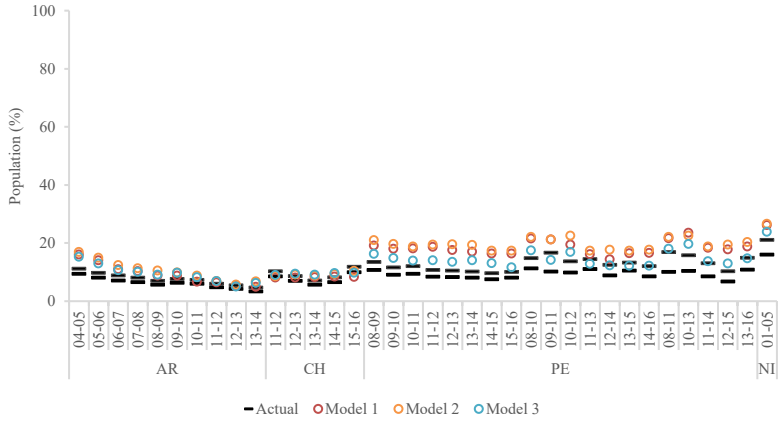
Data source: SEDLAC data (CEDLAS and the World Bank). Note: The figure presents actual and predicted income changes for households by their transition categories and their period 1 income quintile. That is, predictions for round 2 are obtained for households observed in round 1. See Table 1 for additional notes.

Figure 4. Poverty transitions, actual panel vs. predictions using different covariates

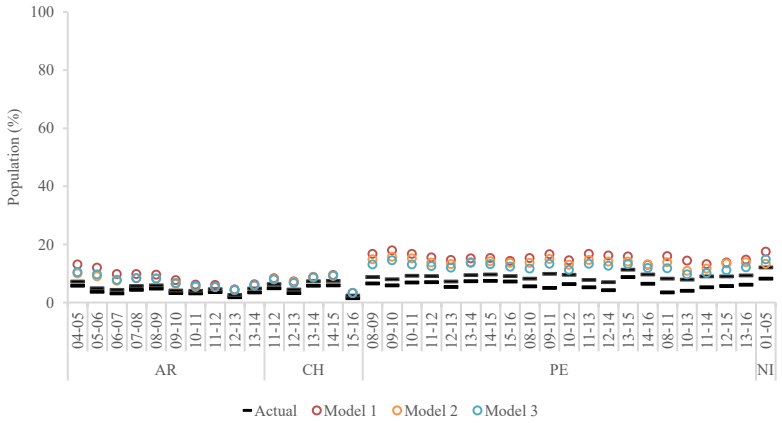
Poor, poor



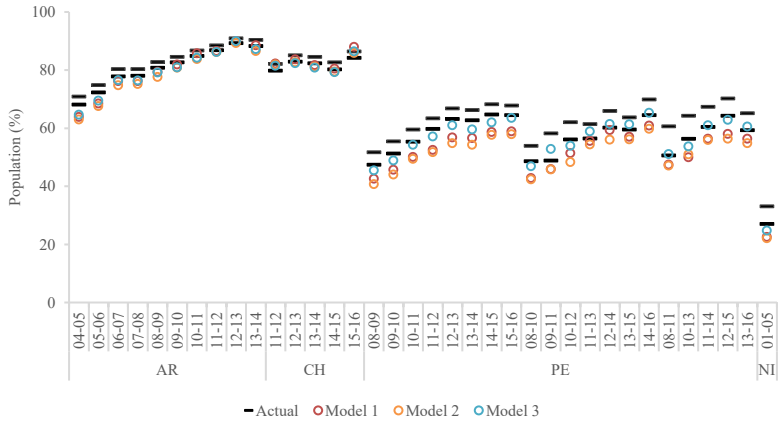
Poor, non-poor



Non-poor, poor



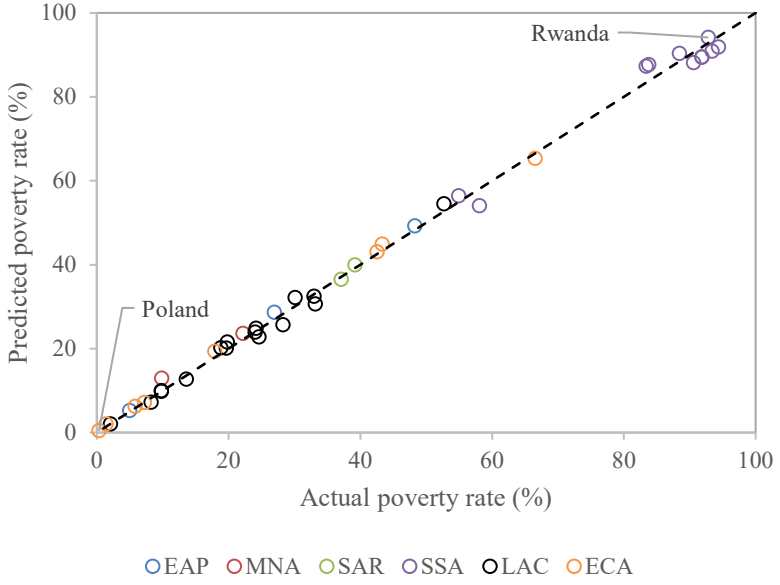
Non-poor, non-poor



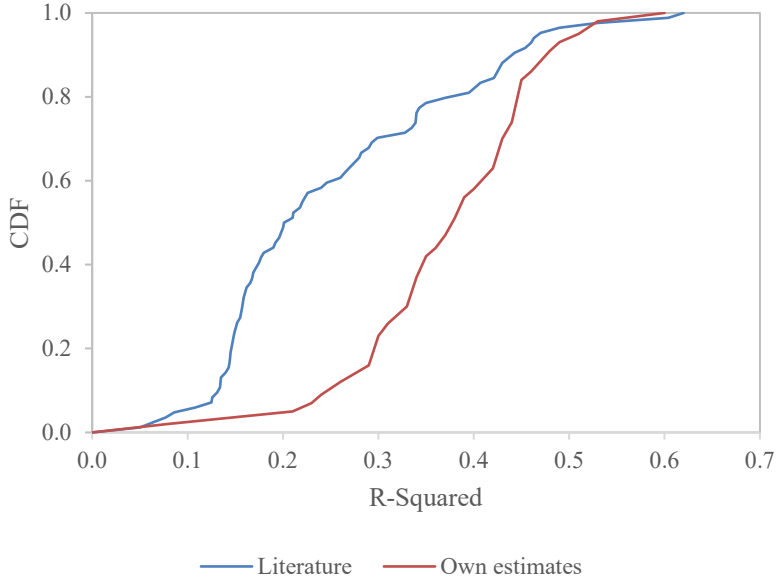
Data source: SEDLAC and LABLAC data (CEDLAS and the World Bank). Note: The figure shows confidence intervals for actual poverty transitions (lines) and predictions obtained using the LASSO-PMM approach with different model specifications. Model 1 includes household head characteristics (gender, age and its square, and years of schooling); Model 2 includes the variables in model 1 plus household size and its square; Model 3 includes the variables in model 2 plus region fixed effects. AR refers to Argentina; CH to Chile; PE to Peru, and NI to Nicaragua. Models present results based on first round income predictions for all individuals surveyed in the second round. See Table 1 for additional notes.

Figure 5. Poverty prediction results from cross-section data around the world

Panel a. Actual and predicted poverty rates

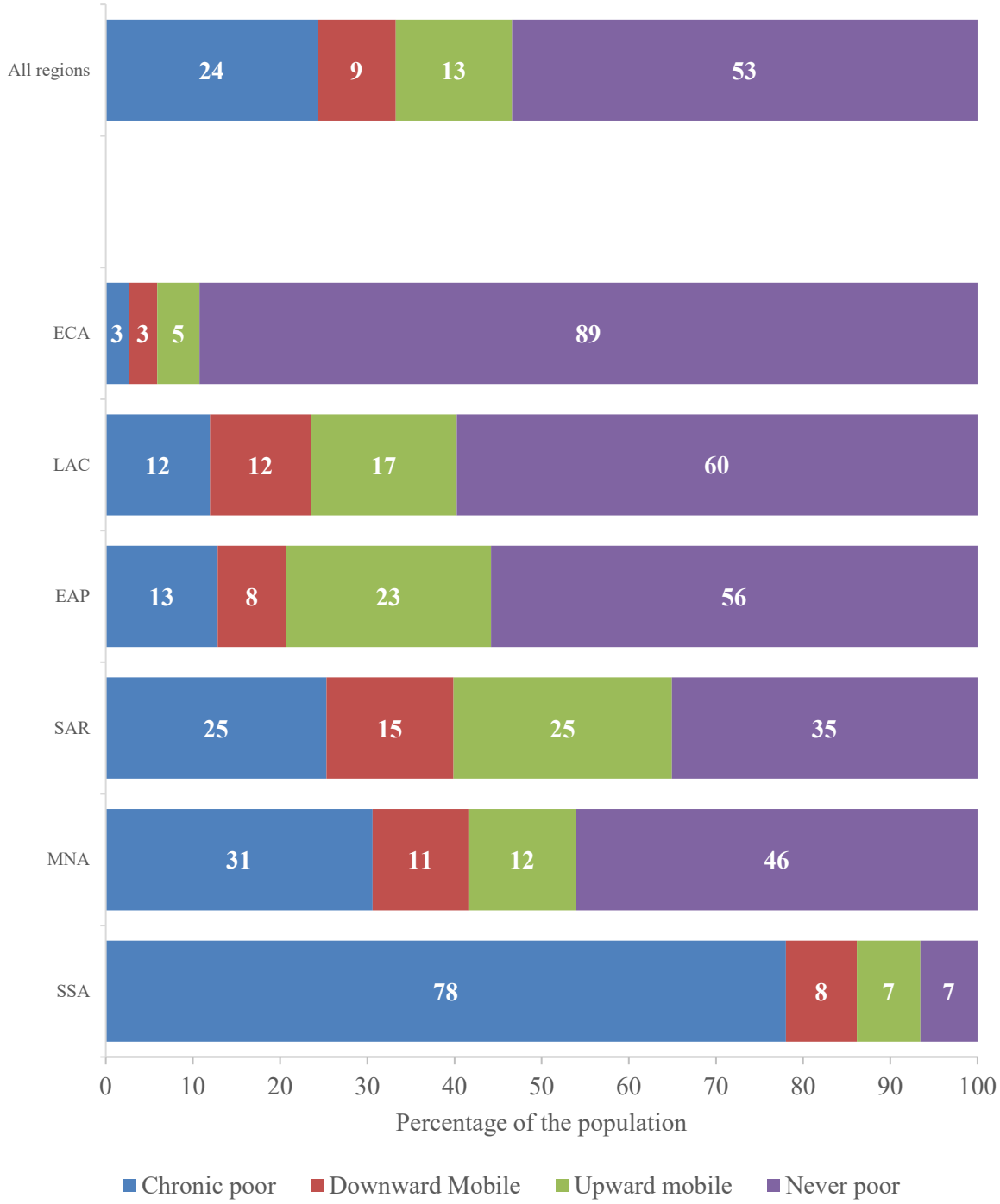


Panel b. Cummulative density function of R-Squareds



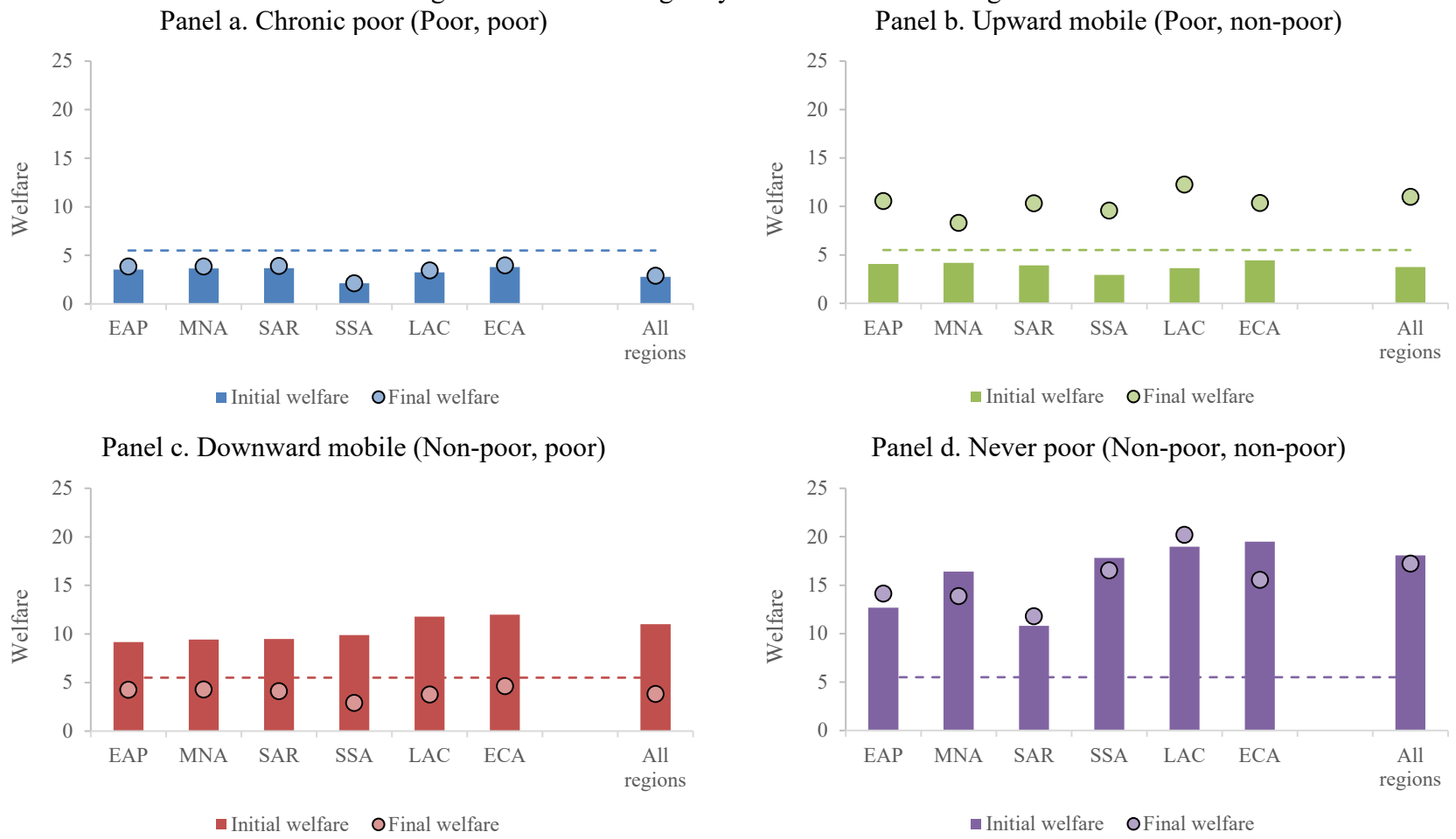
Data source: Own calculations based on GMD, World Bank. Note: Panel a presents poverty rates calculated using the cross-section data (horizontal axis) and the predicted income (vertical axis) using the $5.5 per person per day international poverty line. Our sample includes countries in the six geographical regions defined by the World Bank: East Asia and the Pacific (EAP), Europe and Central Asia (ECA), Latin America and the Caribbean (LAC), Middle East and North Africa (MNA), South Asia (SAR) and Sub-Saharan Africa (SSA). Results are obtained with 20 repetitions and 1 neighbor of PMM. The vertical axis presents results based on second round income predictions for all individuals surveyed in the first round. The horizontal axis presents actual poverty in the second round. Panel b presents the cumulative density function of R-squared in the existing literature (in blue) and the R-squared that results from our estimations. The papers considered are: Hérault N. and Stephen P. Jenkins (2018), Hai-Anh H. Dang and Andrew L. Dabalen (2018), Hai-Anh Dang and Peter Lanjouw (2013), Hai-Anh Dang, Peter Lanjouw, Jill Luoto and David McKenzie (2014), Hai-Anh Dang and Elena Ianchovichina (2018), Cruces et al. (2015), Lucchetti L. (2017), and Bourguignon and Moreno (2019). Poverty rates in this figure do not necessarily match country-specific poverty rates since our estimates are based on the LASSO-PMM approach which uses a sub-population of households with household heads who are between 25 and 65 years of age.

Figure 6. Predicted welfare transitions across the world



Data source: Own calculations based on GMD, World Bank. Note: This figure presents poverty transitions using the $5.5 per person per day international poverty line; transitions are calculated using the base income and the simulated one (circa 2010 and 2015). Chronic poor refers to those who are poor in both periods, downward (upward) mobile to those that fall (escape) from poverty in the second period, while never poor to those above the poverty line. "All regions" include the 43 countries that we used from the GMD data that are listed in Table A.2; the six geographical regions are those defined by the World Bank: East Asia and the Pacific (EAP), Europe and Central Asia (ECA), Latin America and the Caribbean (LAC), Middle East and North Africa (MNA), South Asia (SAR) and Sub-Saharan Africa (SSA). Results are obtained with 20 repetitions and 1 neighbor of PMM. Poverty rates in this figure do not necessarily match regional-specific poverty rates since our estimates are based on the LASSO-PMM approach which uses a sub-population of households with household heads who are between 25 and 65 years of age. Results based on second round income predictions for all individuals surveyed in the first round.

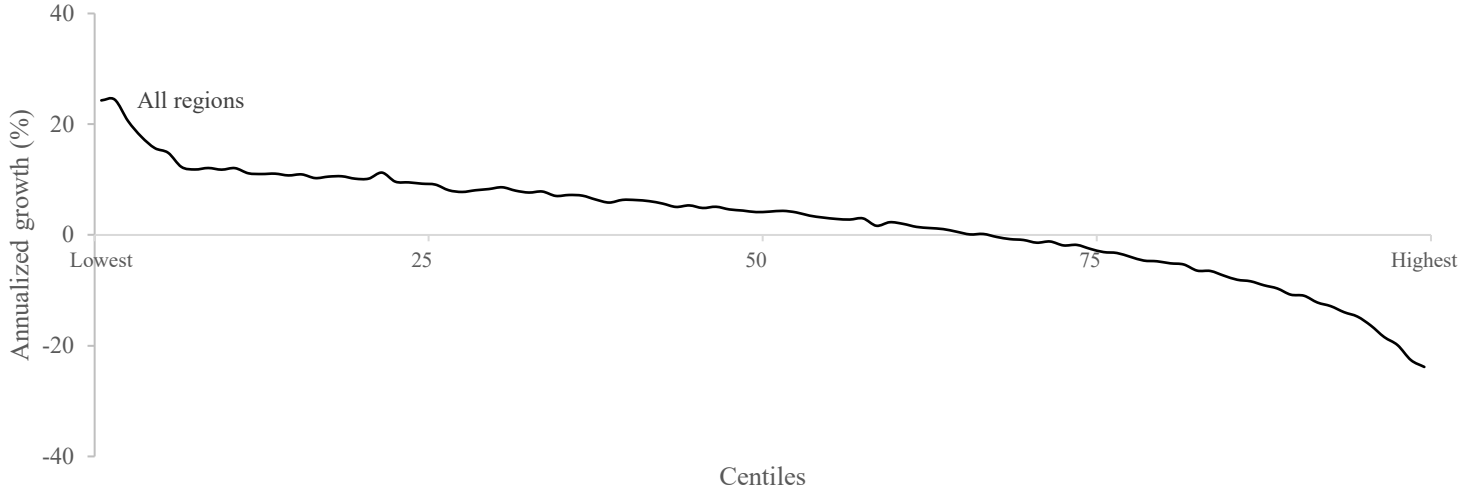
Figure 7. Income changes by welfare transition categories



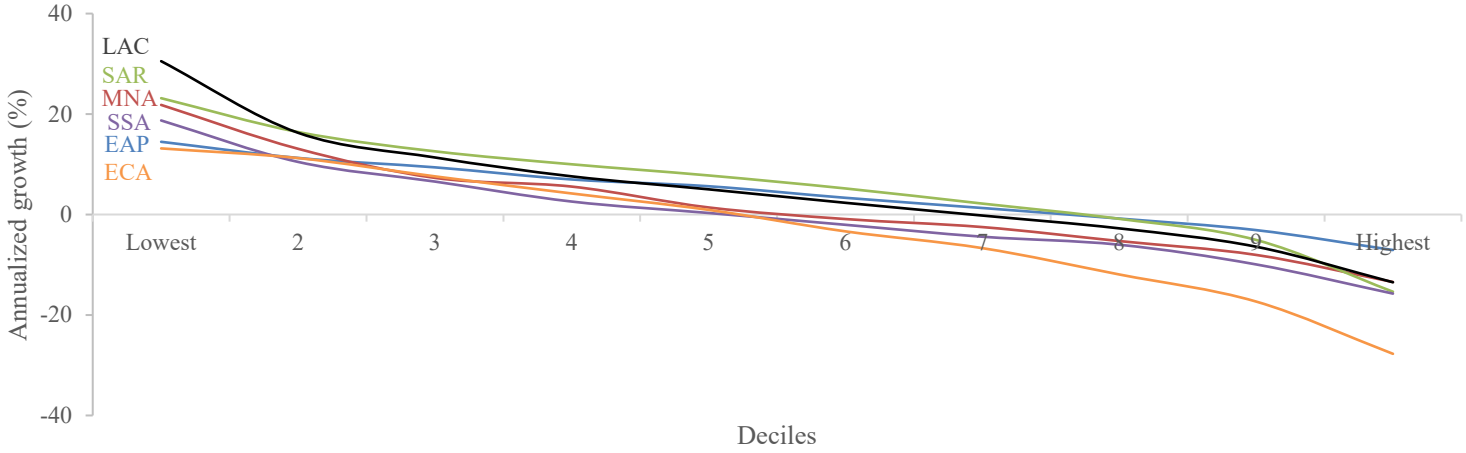
Data source: Own calculations based on GMD, World Bank. Note: These figures present baseline and final income by welfare transition category and region (circa 2010 and 2015). Categories are defined based on the $ 5.5 per person per day international poverty line (dashed line). "All regions" include the 43 countries that we used from the GMD data that are listed in Table A.2; the six geographical regions are those defined by the World Bank: East Asia and the Pacific (EAP), Europe and Central Asia (ECA), Latin America and the Caribbean (LAC), Middle East and North Africa (MNA), South Asia (SAR) and Sub-Saharan Africa (SSA). Results are obtained with 20 repetitions and 1 neighbor of PMM. Results based on second round income predictions for all individuals surveyed in the first round.

Figure 8. Predicted non-anonymous growth incidence curves

Panel a. All regions together



Panel b. By region



Data source: Own calculations based on GMD, World Bank. Note: These figures present the non-anonymous annualized income growth by centiles (deciles) for all countries together (each regions) in circa 2010 and 2015. “All regions” include the 43 countries that we used from the GMD data that are listed in Table A.2; the six geographical regions are those defined by the World Bank: East Asia and the Pacific (EAP), Europe and Central Asia (ECA), Latin America and the Caribbean (LAC), Middle East and North Africa (MNA), South Asia (SAR) and Sub-Saharan Africa (SSA). Results based on second round income predictions for all individuals surveyed in the first round.

Appendix

Table A.1. Available countries, panels, and variables used for LASSO-PMM validation

Country	Year 1	Year 2	Demographics	HH. head's Education	Household size	Area of residence	Regions	Included variables
Argentina	2004	2005	Yes	Yes	Yes	No	6	41
Argentina	2005	2006	Yes	Yes	Yes	No	6	41
Argentina	2006	2007	Yes	Yes	Yes	No	6	41
Argentina	2007	2008	Yes	Yes	Yes	No	6	41
Argentina	2008	2009	Yes	Yes	Yes	No	6	41
Argentina	2009	2010	Yes	Yes	Yes	No	6	41
Argentina	2010	2011	Yes	Yes	Yes	No	6	41
Argentina	2011	2012	Yes	Yes	Yes	No	6	41
Argentina	2012	2013	Yes	Yes	Yes	No	6	41
Argentina	2013	2014	Yes	Yes	Yes	No	6	41
Chile	2011	2012	Yes	Yes	Yes	Yes	13	103
Chile	2012	2013	Yes	Yes	Yes	Yes	13	103
Chile	2013	2014	Yes	Yes	Yes	Yes	13	103
Chile	2014	2015	Yes	Yes	Yes	Yes	13	103
Chile	2015	2016	Yes	Yes	Yes	Yes	13	103
Peru	2008	2009	Yes	Yes	Yes	No	7	48
Peru	2009	2010	Yes	Yes	Yes	No	7	48
Peru	2010	2011	Yes	Yes	Yes	No	7	48
Peru	2011	2012	Yes	Yes	Yes	No	7	48
Peru	2012	2013	Yes	Yes	Yes	No	7	48
Peru	2013	2014	Yes	Yes	Yes	No	7	48
Peru	2014	2015	Yes	Yes	Yes	No	7	48
Peru	2015	2016	Yes	Yes	Yes	No	7	48
Peru	2008	2010	Yes	Yes	Yes	No	7	48
Peru	2009	2011	Yes	Yes	Yes	No	7	48
Peru	2010	2012	Yes	Yes	Yes	No	7	48
Peru	2011	2013	Yes	Yes	Yes	No	7	48
Peru	2012	2014	Yes	Yes	Yes	No	7	48
Peru	2013	2015	Yes	Yes	Yes	No	7	48
Peru	2014	2016	Yes	Yes	Yes	No	7	48
Peru	2008	2011	Yes	Yes	Yes	No	7	48
Peru	2010	2013	Yes	Yes	Yes	No	7	48
Peru	2011	2014	Yes	Yes	Yes	No	7	48
Peru	2012	2015	Yes	Yes	Yes	No	7	48
Peru	2013	2016	Yes	Yes	Yes	No	7	48
Nicaragua	2001	2005	Yes	Yes	Yes	No	4	27

Data source: SEDLAC and LABLAC data (CEDLAS and the World Bank). Note: Demographics includes gender and age of the household head and household size. Area of residence refers to urban or rural areas, while regions refer to the regions for which the surveys are representative. Results are based on the following datasets: Encuesta Permanente de Hogares-Contina in Argentina; Nueva Encuesta Nacional de Empleo in Chile; Encuesta Nacional de Hogares sobre Medición de Nivel de Vida in Nicaragua; and Encuesta Nacional de Hogares in Peru.

Table A.2. Available cross-section surveys for LASSO-PMM application

Region	Country	ISO code	Year	Survey	Welfare aggregate
East Asia and Pacific (EAP)	Fiji	FJI	2008 & 2013	HIES	Consumption
	Thailand	THA	2009 & 2015	SES	Consumption
	Vietnam	VNM	2010 & 2016	VHLSS	Consumption
Europe and Central Asia (ECA)	Armenia	ARM	2011 & 2016	ILCS	Consumption
	Georgia	GEO	2012 & 2016	HIS	Consumption
	Kazakhstan	KAZ	2010 & 2015	HBS	Consumption
	Kyrgyz Republic	KGZ	2012 & 2016	KIHS	Consumption
	Poland	POL	2011 & 2016	HBS	Consumption
	Romania	ROU	2008 & 2016	HBS	Consumption
	Russian Federation	RUS	2012 & 2015	HBS	Consumption
	Serbia	SRB	2009 & 2015	HBS	Consumption
Latin America and the Caribbean (LAC)	Argentina	ARG	2011 & 2016	EPHC-S2	Income
	Bolivia	BOL	2011 & 2016	EH	Income
	Brazil	BRA	2011 & 2015	PNAD	Income
	Chile	CHL	2011 & 2015	CASEN	Income
	Colombia	COL	2011 & 2016	GEIH	Income
	Costa Rica	CRI	2011 & 2016	ENAHO	Income
	Dominican Republic	DOM	2011 & 2016	ENFT	Income
	Ecuador	ECU	2011 & 2016	ENEMDU	Income
	Honduras	HND	2011 & 2016	EPHPM	Income
	Mexico	MEX	2010 & 2014	ENIGH	Income
	Nicaragua	NIC	2009 & 2014	EMNV	Income
	Panama	PAN	2011 & 2016	EH	Income
	Peru	PER	2011 & 2016	ENAHO	Income
	Paraguay	PRY	2011 & 2016	EPH	Income
	El Salvador	SLV	2011 & 2016	EHPM	Income
	Uruguay	URY	2011 & 2016	ECH	Income
Middle East and North Africa (MNA)	Egypt, Arab Rep.	EGY	2010 &2012	HIECS	Consumption
	Iran, Islamic Rep.	IRN	2009 & 2014	HEIS	Consumption
	West Bank and Gaza	PSE	2011 & 2016	PECS	Consumption
South Asia (SAR)	Bhutan	BTN	2012 & 2017	BLSS	Consumption
	Sri Lanka	LKA	2012 & 2016	HIES	Consumption
Sub-Saharan Africa (SSA)	Burkina Faso	BFA	2009 & 2014	ECVM; EMC	Consumption
	Cote d'Ivore	CIV	2008 & 2015	ENV	Consumption
	Ethiopia	ETH	2010 & 2015	HICES	Consumption
	Mozambique	MOZ	2008 & 2014	IOF	Consumption
	Mauritania	MRT	2008 & 2014	EPCV	Consumption
	Niger	NER	2011 & 2014	ECVMA	Consumption
	Rwanda	RWA	2010 & 2013	EICV-III; EICV-IV	Consumption
	Togo	TGO	2011 & 2015	QUIBB	Consumption
	Uganda	UGA	2012 & 2016	UNHS	Consumption
	South Africa	ZAF	2010 & 2014	IES; LCS	Consumption
	Zambia	ZMB	2010&2015	LCMS-VI; LCMS-VII	consumption

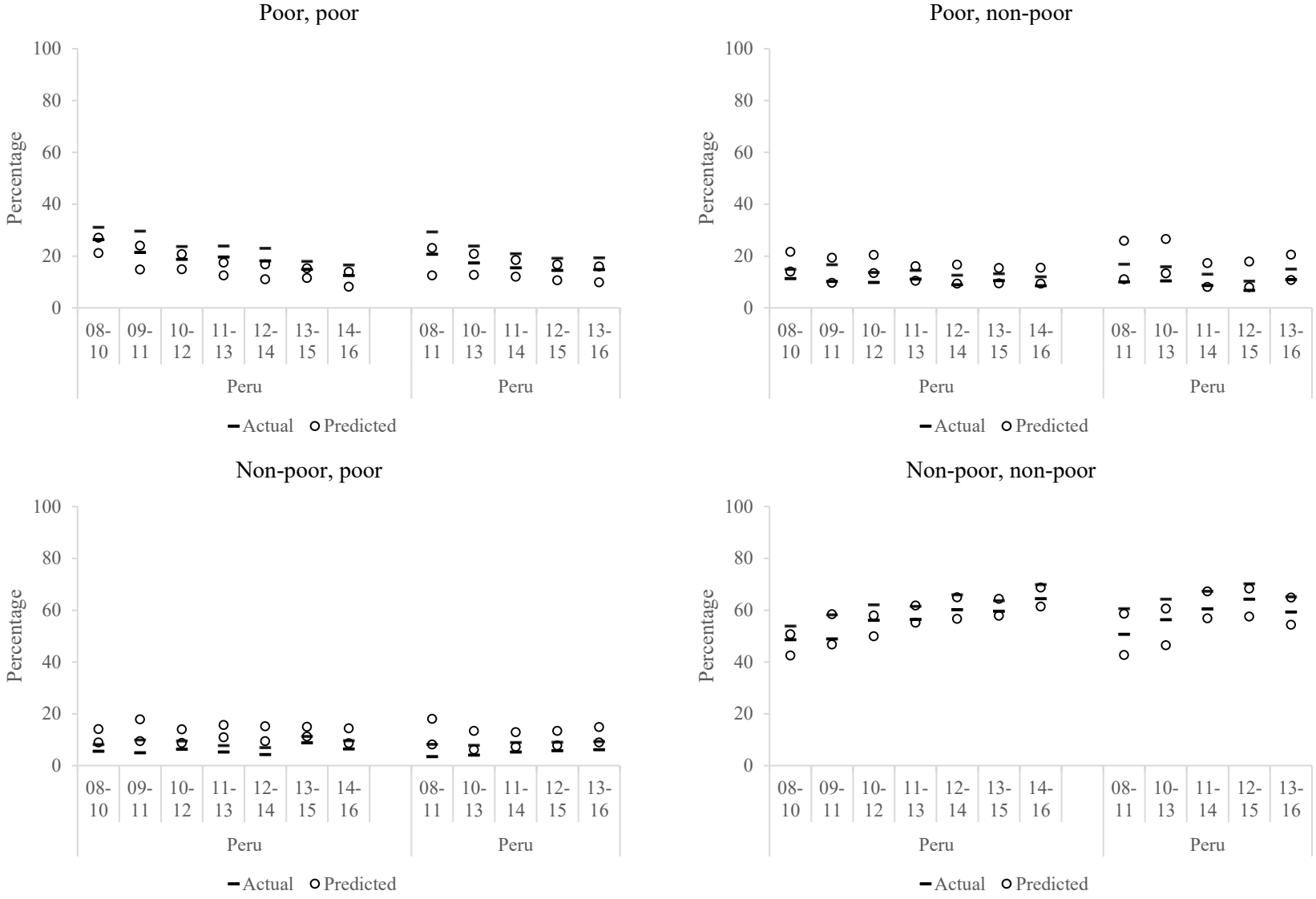
Data source: GMD, World Bank.

Table A.3. Harmonized variables used for LASSO-PMM application

Variable	Definition
Welfare aggregate	Welfare variable used for international poverty estimations. It is expressed on a per capita daily basis and deflated using 2011 PPPs
Gender	Gender of the household head
Age	Age of the household head (level and squared)
Education level	Education of the household head in four education levels: (i) No education (ii) Primary (complete or incomplete) (iii) Secondary (complete or incomplete) (iv) Tertiary (complete or incomplete)
Household size	Number of members within the dwelling (level and squared)
Region fixed effects	Depending on the country, we use either regional or urban rural fixed effects
Weight	Survey weights

Data source: GMD, World Bank.

Figure A.1. Poverty transitions, actual panel vs. predictions using LASSO-PMM (Peru 2- and 3-year panels)



Data source: SEDLAC and LABLAC data (CEDLAS and the World Bank). Note: The figure shows confidence intervals for actual poverty transitions (lines) and predictions obtained using the LASSO-PMM approach (circles). See Table 1 for additional notes.

Figure A.2. Predicted welfare transitions across the world, by country

Data source: GMD, World Bank. Note: This figure presents poverty transitions using the $5.5 per person per day international poverty line; transitions are calculated using the base income and the simulated one (circa 2010 and 2015). Chronic poor refers to those who are poor in both periods, downward (upward) mobile to those that fall (escape) from poverty in the second period, while never poor to those above the poverty line. Results are obtained with 20 repetitions and 1 neighbor of PMM. Poverty rates in this figure do not necessarily match country-specific poverty rates since our estimates are based on the LASSO-PMM approach which uses a sub-population of households with household heads who are between 25 and 65 years of age. Results based on second round income predictions for all individuals surveyed in the first round.

Figure A.3 Income changes by welfare transition category and country

Panel a. Chronic poor (Poor, poor)

Panel b. Upward mobile (Poor, non-poor)

Panel c. Downward mobile (Non-poor, poor)

Panel d. Never poor (Non-poor, non-poor)

Data source: GMD, World Bank. Note: These figures present baseline and final income by welfare transition category and country (circa 2010 and 2015). Categories are defined based on the $ 5.5 international poverty line (dashed line). Results are obtained with 20 repetitions and 1 neighbor of PMM. Results based on second round income predictions for all individuals surveyed in the first round.