Estimating Poverty in India without Expenditure Data

A Survey-to-Survey Imputation Approach

David Newhouse * Pallavi Vyas †

December 31, 2020

Abstract

This paper applies an innovative method to estimate poverty in India in the absence of recent expenditure data. The method utilizes expenditure data from 2004-05, 2009-10, and 2011-12 to impute household expenditure into a survey of durable goods expenditure conducted in 2014-15. At the $1.90 per day international poverty line, the preferred model predicts a 2014-15 headcount poverty rate of 10.4 percent in urban areas and 13.8 percent in rural areas, implying a poverty rate of 12.7 percent nationally. The model's predictions are comparable to the World Bank's current adjustment method for the rural areas but imply a slower rate of poverty reduction for urban areas. In two validation tests, using past data, three alternative model specifications perform worse than the preferred model. The analysis indicates that survey-to-survey imputation, when feasible, is a preferable alternative to the current method of adjusting survey-based poverty estimates to later years.

Keywords: Poverty, India, Nowcasting, Survey-to-Survey Imputation
JEL classification: I32

*David Newhouse (corresponding author) is a Senior Economist at the World Bank Group, Washington DC, and Research Fellow at IZA., Bonn. His email address is dnewhouse@worldbank.org.

†Pallavi Vyas is an Associate Professor at Ahmedabad University, Ahmedabad and a Consultant at the World Bank Group, Washington DC. Her email address is pallavi.vyas@ahduni.edu.in. The authors thank Joao Pedro Azevedo, Benu Bidani, Urmila Chatterjee, Paul Corral, Francisco Ferreira, Haishan Fu, Dean Jolliffe, Christoph Lakner, Daniel Mahler, Rinku Murgai, Minh Nguyen, Sutirtha Roy, Carolina Sanchez-Paramo, Benjamin Stewart, Roy Van der Weide, and Nobuo Yoshida for invaluable suggestions, help, and support. This paper was previously published as a working paper entitled "Nowcasting poverty in India for 2014-15", which is available at: http://documents.worldbank.org/curated/en/294251537365339600/Nowcasting-Poverty-in-India-for-2014-15-A-Survey-to-Survey-Imputation-Approach

I Introduction

This paper explains the methodology and results of a survey-to-survey imputation exercise that was used to generate 2015 headcount poverty estimates for India at the World Bank. India was selected as a pilot case because of its size and importance, as well as the lack of recent data on household living standards. The latest survey-based estimates of poverty are from 2011-12. The most recent nationally representative survey that could be used to impute poverty into a later year is the NSSO expenditure on services and durables survey of 2014-15. The resulting poverty estimates from the imputation informed the 2015 poverty estimates, which are the most recent that are currently available, marking the first time that the World Bank used this type of imputation method as an input into its global and regional poverty estimates.

Understanding trends in poverty in India is not only vital to Indian policy makers to measure changes in the welfare of the poor, but also has major implications for regional and global estimates of the prevalence of poverty. In 2013, an estimate of the number of extreme poor in India, defined as those consuming less than $1.90 per day per person, numbered 250 million [1]. This means that India accounted for over a quarter of the global total of 783 million extreme poor. Although Nigeria has now likely passed India as the nation with the largest number of extremely poor persons, accurately monitoring India's progress in the fight against extreme poverty remains critical to assess progress towards the goal of reducing poverty to less than 3% by 2030.

The second motivation for undertaking a modeling exercise is that the typical method used by the World Bank to adjust poverty to a later year tends to overestimate the pace of poverty decline. When estimating global and regional poverty rates, the World Bank usually adjusts or "lines up" consumption survey data from a variety of years to a common year, by scaling up the measured levels of household per capita consumption to account for growth in the intervening years. For example, to "line-up" a survey from 2012 to 2015, each household's per capita consumption, measured in 2012, would be assumed to grow at the same rate. In most countries, this common rate of growth is assumed to be equal to the growth in household final consumption expenditure

[1] Estimate is based on the traditional line-up method explained below.

(HFCE), the value of expenditure incurred by resident households on goods and services in the national accounts, during the intervening period.[2] Because it is a component of the national accounts, HFCE is published annually and can therefore be used to line-up the survey means to any later year.

Unfortunately, however, growth in HFCE tends to substantially exceed mean consumption growth in household surveys.[3] One study, for example, concludes that on average only about half of the growth in HFCE was reflected in growth in survey consumption.[4] This is partly because official surveys face severe challenges obtaining accurate measures of welfare for the very wealthy. Because there are few very wealthy households, they may not be included in the sample. Wealthy households are also more likely than poor households to refuse to participate in the survey, and tend to underreport their income or consumption. The underrepresentation of wealthy households may lead growth in mean household expenditure to be underestimated in surveys, especially in cases where inequality appears to be rising. However, there are also good reasons why national accounts data may overestimate growth in mean income or consumption, including the use of outdated 'rates and ratios' based on old survey data.[5]

Because of the systematic discrepancy between growth in national accounts data and household surveys, using HFCE to adjust survey data is prone to overstating poverty reduction for countries that lack a recent survey such as India. In fact, past analysis from India demonstrates how the typical line-up adjustment procedure can greatly overestimate poverty decline, using data from 2004-05 and 2009-10.[6] Estimates are generated for 2009-10 by lining up the 2004-05 data to 2009-10, applying the HFCE growth rate during that period. These line-up estimates for 2009-10 are then compared with measurements from actual survey data from that year. The line-up procedure generated an estimated reduction in poverty from 41.6 to 13.4 percent, whereas the actual data from 2009-10 yielded a poverty rate of 32.7%.[7] In other words, the line-up method overstated

[2] In some countries, growth in per capita GDP is used instead of final household consumption expenditure.

[3] See Edward and Sumner (2013), Hillebrand (2008), Korinek, Mistiaen, and Ravallion (2006), Deaton and Kozel (2005), Ravallion (2003) and Minhas (1988).

[4] Ravallion (2003).

[5] Deaton and Kozel (2005).

[6] Jolliffe (2014) p.253.

[7] These figures are based on the older, $1.25 poverty line (in 2005 PPP).

the decline in poverty by 19 percentage points. However, virtually all of the discrepancy between the line-up estimate and the actual data resulted from consumption expenditure growth in the national accounts exceeding mean welfare growth in the survey. In contrast, assuming that the 2004-05 welfare distribution remained constant except for a scale factor created negligible error. In this case, scaling up measured welfare in the initial year by the wrong factor led to less accurate predictions than by assuming distributionally neutral changes.

The concern that growth in the national accounts exceeds growth in survey mean welfare motivates the use of a survey-to-survey imputation method to determine an appropriate scale factor or "pass-through rate" to apply to HFCE. This paper estimates poverty at the $1.90 line in 2014-15, using an econometric model specified at the household level. Urban and rural sectors are modeled separately, because they differ in fundamental ways.[8] The model draws on four rounds of past Indian National Sample Survey data, namely the 61st, 66th, 68th, and 72nd rounds, which were fielded in 2004-05, 2009-10, 2011-12, and 2014-15 respectively.[9] The 2014-15 survey, unlike the three previous rounds used in this exercise, did not collect information on aggregate household consumption and therefore cannot generate direct estimates of poverty. The 2014-15 survey, however, contains information on several household characteristics that are also present in past rounds and are reasonably well-correlated with per capita consumption.[10] These common characteristics include household age, size, caste, religion, a few labor market variables, and expenditure on three categories of services that are asked in the same way as previous rounds.

The analysis estimates a regression model that predicts household per capita consumption using these common household characteristics, plus contemporaneous rainfall shocks and a linear time trend. Most coefficients are allowed to vary over time in a linear way, slightly relaxing the usual strong assumption that the relationship between the explanatory variables and log welfare remains constant over time.[11] The model is estimated using data from the first three rounds of the survey.

[8] Also, India is one of three countries for which separate poverty estimates are presented by the World Bank for urban and rural areas.

[9] Unfortunately, data between 1993 and 2004-05 were unavailable because the welfare aggregate in the 1999-2000 round was not comparable.

[10] In the pooled regression including 2004-05, 2009-10, and 2011-12 data, the R^2s are 0.57 for urban areas and 0.46 for rural areas (Table 14 (Appendix).)

[11] Nguyen and van der Weide (2018).

The coefficients from this regression model are then applied to measured values of the same common characteristics in the 2014-15 survey, as well as 2014-15 rainfall levels, to repeatedly simulate levels of welfare and poverty. The results of these simulations are aggregated to generate estimates of poverty headcount rates and their standard errors.

This exercise is similar in spirit to the adjustments for non-comparability in the consumption aggregate that led to "the great Indian poverty debate".[12] Since then, this type of nowcasting exercise has also been carried out in several other countries.[13] While this method has not yet been successfully applied to estimate changes in inequality, a significant body of evidence now exists where out of sample predictions from models estimated on past data generate credible estimates, in a variety of country contexts, of poverty headcount and gaps.

The preferred model generates an estimate of 10.4% in urban areas and 13.8% in rural areas, which implies an overall poverty level of 12.7% for 2014-15 (Table 1 and Figure 1). The estimate is consistent with poverty reduction from 2004-05 to 2011-12 in rural areas and suggests a slight deceleration in poverty reduction in urban areas (Table 2). This in turn implies that the elasticity of poverty with respect to growth of per capita GDP was -2.8, which is within the range of past experience (Table 10). Reassuringly, when looking at the state level, predicted poverty reduction is greater in states with higher rates of GDP growth (Figure 2).

We use two main tests to validate the survey-to-survey imputation models. The first estimates a model with the same specification, using data from 2004-05 and 2009-10, to predict into 2011-12. These predictions are then compared with actual measured levels of poverty from the 2011-12 national sample survey. In a similar, vein, the second validation test uses data from 2009-10 and 2011-12 to predict backwards to 2004-05. This is a much more challenging exercise because data from two years apart are being used to extrapolate five years in the past.

While the traditional line-up method performs well when projecting forward into 2011-12, it gives implausible estimates when projecting back to 2004-05. To be specific, the standard line-up method

[12] Deaton and Dreze (2002) and Kijima and Lanjouw (2003) Deaton and Kozel (2005).

[13] See, for example, Christiaensen and Stifel (2007) in Kenya, Newhouse, Shivakumaran, Takamatsu, and Yoshida (2014) in Sri Lanka, Douidich, Ezzrari, Van der Weide, and Verme (2016) in Morocco, and Dang, Lanjouw, and Serajuddin (2017) in Jordan.

is off by only by 1.8 percentage points in 2011-12 but exaggerates poverty by approximately 10 percentage points when predicting back to 2004-05 (See Tables 6 and 7). Our preferred model performs much better on average, as there is no discrepancy when projecting into 2011-12 and about 3 percentage points for the 2004-05 projection.

The traditional line-up method happens to do well when projecting forward because growth in mean reported consumption in the survey was unusually high between 2009-10 and 2011-12 and nearly matched growth in HFCE. This is a sharp break from the pattern observed during the previous period, and in most other contexts. It remains to be seen whether this correspondence between HFCE and survey mean growth will be sustained, but there are at least two reasons to speculate it may not be. The first is that survey-based consumption may be more responsive than national account data to weather shocks. India experienced significant droughts during 2004 and 2009, followed by above-average rainfall in the 2011 season. To the extent that survey-based consumption is more sensitive than national accounts to agricultural income, favorable rainfall in 2011 would cause a larger jump in survey consumption than HFCE.[14]The second factor is the recovery from the 2008 financial crisis, which is distinctly visible in the 2009 national accounts data. Survey-based consumption, however, may have taken longer to recover. Both factors could have contributed to the unusually large increase in mean consumption observed between 2009-10 and 2011-12, which counteracted the usual tendency for HFCE growth to exceed growth in mean consumption. We know of no reason to believe, however, that growth in HFCE will accurately reflect growth in mean consumption going forward.

Turning back to the model predictions, we consider three alternative specifications that impose more restrictive assumptions. These estimate a slower poverty decline, with estimated headcount rates ranging from 15.4 to 18.8 percent (Table 5). Similar to our model, the traditional line-up method shows an overall decline in poverty, from 21.1% in 2011-12 to 12.1% in 2014-15 (Table 8). However, the line-up method indicates a sharper decline in urban poverty compared to our model (7.7 vs. 10.4) and a slightly lower decline in poverty in the rural areas (14.2 vs. 13.8). The different predictions indicate that the choice of modeling assumptions matters, at least in this case.

[14]According to the Economic Survey 2017-18 of the Ministry of Finance in India, agriculture employs over half of Indian workers but only contributes 17 to 18 percent of India's GDP.

The rest of this paper is organized as follows. The next section reviews the data used for the analysis. Section III turns to the methodology. It discusses the estimation method and the process used to select the set of predictor variables, and examines the results of four alternative specifications of the model. Section IV compares the preferred model's predictions for 2014-15 to the usual method used to project poverty forward, and investigates which specific variables are accounting for the change in average welfare predicted by the model. Section V shows that the estimated poverty rate implies a reasonable elasticity of poverty reduction with respect to growth, verifies that implied poverty reduction is greater in fast-growing states, and examines the model's predictions at higher poverty lines, and Section VI concludes.

II Data

In order to obtain estimates of welfare for 2014-15, we first identify variables that are common both to the three earlier consumption surveys and the 2014-15 round. The surveys have five demographic variables in common. They are household size, age of household members, gender, religion, and caste contained in the "household characteristics" section of each survey. The three common labor market variables are the household's principal industry, principal occupation and principal means of livelihood. [15] To ensure consistency, the occupation categories are harmonized across survey years.[16] We further group households into high skill, middle skill and low skill occupation categories.[17] The household's principal means of livelihood is determined from the "household type" variable. The categories are self-employed, regular wage/salary earning, casual

[15]The principal industry of a household is determined according to the National Industrial Classification (NIC) 2008 code for the NSSO72, NSSO68 and NSSO66 surveys and the NIC 1998 code for the NSSO61 survey. Households whose principal industry is agriculture, forestry or fishing are classified into the agricultural sector. Those that work principally in mining, manufacturing, construction or for the utilities are industrial sector households. Lastly, service sector households are those that work in wholesale or retail trade, food and accomodation, transportation and storage, information and communication, finance and insurance, real estate, professional and scientific, administrative and support, public administration and defense, education, health or arts and recreation.

[16]Regarding the principal occupation, the NSSO72, NSSO68 and NSSO66 surveys categorize households according to the National Classification of Occupations (NCO) 2004 code and the NSS061 survey according to the NCO 1968 code.

[17]Legislators, senior officials, managers, professionals, technicians and associate professionals are in the high skill category. Clerks, service workers, shop and market sales workers are classified as middle skill. Workers in agriculture and fishery, craft and related trades, plant and machine operators, assemblers and those in "elementary" occupations are classified as low skill.

labor and "other".[18]

In addition, the surveys ask questions regarding expenses on miscellaneous services such as household service (domestic help, barber, beauty, laundry, priest, grinding, tailor), recreation (cinema/theater, fairs/picnics, club fees, photography, VCD/DVD on hire) and transport (conveyance related expenses). We include that portion of transportation expenditures that makes it comparable to the previous consumption expenditure surveys. Therefore, expenditures on fuel and bus transportation are excluded. We also include a district-level rainfall variable that measures standardized deviations from the historical mean. The period used to calculate the historical mean is 1981-2017.[19]

Second, we ensure that the wording and recall periods on the questions are comparable over time. Both different wording and/or recall periods would change the interpretation of the relevant variables and thus affect the predictability of our model. We find for all the variables that we include, the questions are similar across surveys. On the other hand, there are questions related to expenditure on consumer durables in both surveys that we do not include for two reasons. First, there are discrepancies in relation to the wording of the questions between the NSSO72 and earlier surveys.[20] Second, the recording of possession and expenditures was different between surveys.[21]

[18]For rural areas, each of the categories are reported as agriculture and non-agriculture sub-classifications that we group together to make them comparable to the urban areas.

[19]Rainfall data are taken from the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS) data set on precipitation (Funk, Verdin, Michaelsen, Peterson, Pedreros, and Husak (2015)).

[20]The NSSO61, NSSO66 and NSSO68 ask about "expenditure for purchase and construction (including repair and maintenance) of durable goods for domestic use" and the NSSO72 asks about "expenditure on durable goods acquired during the last *365 days* other than those used exclusively for entrepreneurial activity". Second, the NSSO61, NSSO66 and NSSO68 (type 1 survey) has questions for a 30 day and a 365 day recall period while the question on the NSSO72 only asks about expenditure for the 365 recall period. Third, the NSSO61, NSSO66 and NSSO68 asks about expenditure for "purchase and construction", while the NSSO72 asks about "acquisition" rather than explicitly about purchase. Fourth, the NSSO61, NSSO66 and NSSO68 ask about purchase for "domestic use" while the NSSO72 asks about goods "other than those used exclusively for entrepreneurial activity". This includes the total value of raw materials, services and/or labor charges and any other charges. There is no explicit mention of "construction" or "repair and maintenance" in the NSSO72, but there is a value of components column that most likely includes raw materials used in construction. Fifth, in the NSSO68, there is a separate question on whether the good was bought for "hire purchase". The surveyor is asked to make the distinction between "hire purchase" and a loan. The NSSO72 instruction manual does not mention anything about goods bought on "hire purchase".

[21]We observed the discrepancies in the manuals provided to the surveyors.First, if an asset was bought and sold during the reference period, it is recorded in the earlier NSSO surveys but not in NSSO72. Second, in the earlier surveys if the item has been purchased but not yet in the household's possession, the expenditure is recorded by the surveyor. However, in the NSSO72 a durable good that is not in the household's possession even

Lastly, we confirm that the sampling frame is similar across all surveys. Like the earlier rounds, the NSSO72 is a multi-stage stratified survey of all states/union territories in India. To more accurately represent population density, for the 72nd round the primary sampling units (PSUs) are selected from the 2011 census list of villages in the rural sector while for the earlier rounds they are drawn from the 2001 census list of villages. Also, the number of PSUs are slightly higher, 14,088, in the 72nd round. For each of the earlier surveys, 12,784 PSUs are selected.[22] For the urban sector, in all surveys the PSUs are sampled from the Urban Frame Survey (UFS) blocks. For all surveys, the methodology to select the PSUs, the strata, sub-strata and the ultimate stage units (USUs) or households remains the same.[23]

A Descriptive Statistics

Tables 3 and 4 report the mean values of the variables used in the model by year, for urban and rural areas. The descriptive statistics generally reflect India's rapid economic development. Average log per capita consumption, in real terms increased from 4.5 to 4.74 between 2004-05 and 2011-12. When expressed in levels, per capita consumption grew approximately 24 percent over the seven-year period, or 3.1% a year. Household size steadily fell during this time. In the decade between the first and last survey, the share of the population living in households with 4 or fewer persons rose from 39 to 47 percent in urban areas, and from 29 to 35 percent in rural areas. A byproduct of smaller households is a drop in the share of household members that are children in the age group 0 to 14, which fell by 5 percentage points in both urban and rural areas. Regular wage work became slightly more prevalent in urban areas, and the share of rural workers working in agriculture declined 3 percentage points, as industrial work became more common in rural areas. Expenditure on household miscellaneous services, recreational spending and transport showed a continual strong increase from 2004-05 to 2014-15.[24]

if full payment has been made is not included. Third, if the durable good is in possession but has not been paid for, it is not included in the earlier NSSO surveys but is included in the NSSO72 (Instructions to Field Staff, Chapter Four).

[22]The NSSO refers to PSUs as First Stage Units (FSUs). The term 'village' is Panchayat wards for Kerala.

[23]In the event that the population is greater than 1,200 in a PSU, the PSU is divided into "hamlet groups" "sub-blocks" in the rural and urban PSUs, respectively.

[24]While the recall periods are the same and most of the categories of transportation expenditure are similar across surveys, in the NSSO72 the surveyors were asked to exclude expenditure on fuel for one's own trans-

The final rows of Tables 3 and 4 report the population-weighted mean rainfall, across districts, in terms of deviation from historical district means. Monsoons in India typically occur between June and October, making the third quarter rainfall particularly important for agricultural production. Table 4 indicates that in rural areas, rainfall was below average in the second half of 2004, the third quarter of 2009 and the first two quarters of 2010. On the other hand, rainfall was substantially above average in the third quarter of 2011, and close to exactly average in the third quarter of 2014. As mentioned above, favorable rainfall may have contributed to the strong growth in survey consumption between 2009-10 and 2011-12, but did not continue in 2014.

III Empirical Methodology

A Econometric Model

The specification of the model is based on the small area estimation (SAE) methodology originally proposed by Elbers, Lanjouw, and Lanjouw (2003). This methodology generates a joint distribution of economic welfare and independent variables in the target data set, where a welfare measure is not available. The process consists of two steps. The first involves estimating the relationship between economic welfare and explanatory variables in the source data set, in which a welfare measure is available. The parameters are then used to simulate economic welfare into the target data set.

The first step in this analysis is to estimate the relationship between household per capita consumption expenditure (the measure of economic welfare) and other explanatory variables, in the three available source data sets with data on total household consumption. The candidate explanatory variables are those that are available in both the source and target data sets. The regression is a random effect model relating log per capita consumption expenditure to household and regional

port. The consumption expenditure surveys do include expenditure on petrol and diesel for vehicles. Also, in the NSSO72 survey questions regarding transportation expenditure were asked separately for overnight and non-overnight journeys.

variables using the 2004-05, 2009-10 and 2011-12 surveys (equation 1).[25]

$$ln(y_{cht}) = X^h_{cht}\beta_1 + Z^h_{cht}t\beta_2 + X^d_{ct}\beta_3 + Z^d_{ct}t\beta_4 + \beta_5 t + u_{cht} \quad (1)$$

The welfare measure is y_{cht}, which is household per capita expenditure for a household in cluster c and household h, interviewed in survey round t. A cluster is a district as defined in the NSSO surveys. For all surveys, y_{cht} is measured in 2011 rupees using separate CPI indices for urban and rural areas.[26] The X^h_{cht} vector in equation (1) consists of an intercept plus household level demographic variables, labor market variables and expenses on miscellaneous services for household h in cluster c in time t. Expenses on miscellaneous services for all survey years are measured in 2011 rupees.[27] The X^d_{ct} vector consists of district means of household characteristics associated with cluster c in time t. This vector also includes a measure of rainfall shocks, which is the district's deviation from mean historical rainfall. District level characteristics are included both to improve the accuracy of model predictions, and to generate more accurate estimates of the standard error.[28] We chose to include district means instead of PSU means because of concerns about measurement error arising from insufficient numbers of observations per PSU. The linear time trend is the variable t. The vector Z^h is a subset of the X^h variables whose coefficients are allowed to vary linearly over time (Nguyen and van der Weide (2018)). Similarly, the vector Z^d is a group of X^d variables allowed to vary linearly with time.

The variables ultimately included in equation (1) are selected from the candidate variables using the Least Absolute Shrinkage and Selection Operator (LASSO) method (Tibshirani (1996)). They consist of all variables common to both surveys listed in Tables 3 and 4. LASSO roughly equalizes in and out of sample R^2, and therefore avoids "overfitting" the model to the observed data, which is important in this case because the goal of the exercise is to predict welfare out of sample. We utilize a procedure known as "Post-Lasso" where variables selected using the LASSO are used

[25]NSSO61, NSSO66 AND NSSO68. The coefficients in this semi-log model are interpreted as a relative change in welfare as a result of an absolute change in the explanatory variables.

[26]For rural areas, we use the CPI for Agricultural and Rural Laborers and for urban areas the CPI for Industrial workers. Source: CEIC (from Labor Bureau Government of India).

[27]The CPI indices are the same as those used for the welfare measure, y_{cht}.

[28]Elbers, Lanjouw, and Leite (2008).

in a subsequent linear model (Belloni and Chernozhukov (2013)). Post-lasso, as compared with standard LASSO, has the crucial advantage of producing unbiased predictions of log welfare[29].

In the pool of candidate variables for the model, we also include interactions of all the candidate variables with a linear time trend, interaction and the district means of each variables, and the district means interacted with a linear time trend. The resulting LASSO specification selects a γ vector to minimize

$$\underset{\gamma}{\text{M}in} \quad \frac{1}{N}\left[ln(y_{cht}) - x^h_{cht}\gamma_1 - x^h_{cht}t\gamma_2 - x^d_{ct}\gamma_3 - x^d_{ct}t\gamma_4 - \gamma_5 t\right]^2 + \lambda\sum_{j=1}^{5}|\gamma_j| \tag{2}$$

As in equation 1, the x^h_{cht} vector in equation (2) consists of an intercept plus household level demographic variables, labor market variables and expenses on miscellaneous services for household h in cluster c in time t. The vector of household characteristics interacted with the time trend is $x^h_{cht}t$. The x^d_{ct} vector consists of district means of household characteristics associated with cluster c in time t as described above for equation (1). The variables in x^d_{ct} interacted with a time trend constitute the $x^d_{ct}t$ vector. The tuning parameter, λ, is determined through the process of cross validation.[30] The optimization procedure from equation (2) sets the coefficients on several variable to zero. They are then dropped to estimate equation (1).[31]

The regressions are weighted using population weights.[32] In addition, because of important differences in consumption patterns in urban versus rural areas, and because poverty rates are reported separately for urban and rural India, separate models are estimated for each sector.

Taking X as the vector of all variables and β as the vector of coefficients, equation (1) can be rewritten as:

[29]The coefficients from the standard lasso procedure are biased towards zero, which would systematically underestimate poverty rates in this case.

[30]The data is divided into training and validation subsamples. The estimator finally selected is the one with the smallest out-of-sample MSE. The training data is used to estimate the model parameters of each of the competing estimators. The out-of-sample Mean Squared Error (MSE) is calculated for the predictions produced by each competing estimator.

[31]We do not estimate a model that interacts each variable with a time trend. This is because for the validation exercise, only two rounds of data are available to estimate the model, leaving insufficient degrees of freedom to estimate polynomial time trend interactions.

[32]Household weights are multiplied by household size to calculate population weights.

$$ln(y_{cht}) = X'\beta + u_{cht} \tag{3}$$

An initial estimate of β is obtained using weighted Ordinary Least Squares (OLS) (equation 2). The consumption spending of households within a cluster are assumed to be correlated. To allow for this within-cluster correlation, the random disturbance term u_{cht} has a cluster, η_{ct}, and a household component ϵ_{cht}. Therefore, the random disturbance term can be written as,

$$u_{cht} = \eta_{ct} + \epsilon_{cht} \tag{4}$$

The variables η_{ct} and ϵ_{cht} are assumed to be independent of each other and uncorrelated with the explanatory variables. We do not allow for heteroskedasticity in the cluster component of u_{cht}, due to the small number of clusters.[33] However, the household error term, ϵ_{cht}, is assumed to be heteroskedastic reflecting unequal variances in the error terms across households.

Because of heteroskedasticity in the error term and spatial correlation from the introduction of η_{ct}, we re-estimate equation (2) using Generalized Least Squares (GLS). The weights for the GLS specification are the predicted variances of the error terms from the OLS regression. We estimate the variance of the error term parametrically as specified in Elbers, Lanjouw, and Lanjouw (2003) and Nguyen, Corral, Azevedo, and Zhao (2017). The variance model has a logistic form with (ϵ^2_{cht}) as the dependent variable (equation 4). In equation 5, the dependent variable is the estimated variance, derived from the residuals of the first OLS regression. The explanatory variables of the variance model are also chosen using the LASSO method.[34] The estimates from the LASSO regression and the residuals are then used to predict the variance of (ϵ_{ch}) (equation 6).

$$E[\epsilon^2{}_{cht}] = \sigma^2{}_{\epsilon_{cht}} = \left[\frac{Ae^{Z'\alpha} + B}{1 + e^{Z'\alpha}}\right] \tag{5}$$

[33] Elbers, Lanjouw, and Lanjouw (2003).

[34] In this method all district and household characteristics used in step 1 are included on the right- hand side. The resulting estimates include zero coefficients for several variables, thereby selecting the remaining variables with non-zero coefficients for the variance model.

$$ln\left[\frac{\epsilon_{ch}^2}{A-\epsilon_{ch}^2}\right] = Z'_{cht}\alpha + r_{cht} \tag{6}$$

$$\hat{\sigma}^2_{\varepsilon_{cht}} \approx \left[\frac{Ae^{Z'\alpha}}{1+e^{Z'\alpha}}\right] + \frac{1}{2}\widehat{Var}(r)\left[\frac{Ae^{Z'\alpha}(1-e^{Z'\alpha})}{(1+e^{Z'\alpha})^3}\right] \tag{7}$$

The predicted variances are used to re-estimate equation (2) using GLS (equation (7)).

$$ln(y_{cht}) = X'\beta_{GLS} + u_{cht} \tag{8}$$

Next, we estimate predicted welfare using estimates from the analysis in the first step. We would like to generate a joint distribution of welfare, and not solely the expected value. Therefore, we use Monte-Carlo simulations to generate a vector of error terms that can be used to calculate the measure of welfare for each household.

Following Elbers, Lanjouw, and Lanjouw (2003), we make the following assumptions for the simulations. The vector $\tilde{\beta}$ is drawn from a normal distribution with mean $\hat{\beta}_{GLS}$ and variance Var($\hat{\beta}_{GLS}$).

$$\tilde{\beta} \sim N(\hat{\beta}_{GLS}, Var(\hat{\beta}_{GLS}))$$

A hundred simulations of β_{GLS}, gives a consistent estimate of β. Using the estimated β one can estimate $X_h^T\beta_{GLS}$, the expected value, for each household. The residuals η_{ct} and ϵ_{cht} are then calculated for each household using Monte Carlo simulations.

The cluster component of the error term, η_{ct}, is drawn from a normal distribution with mean 0 and variance, $\hat{\sigma}_\eta^2$. The data generating process of the variance term is assumed to follow a gamma distribution.

$$\tilde{\eta}_{ct} \sim N(0, \hat{\sigma}_\eta^2)$$

$$\hat{\sigma}_\eta^2 \sim Gamma(\bar{\sigma}_\eta^2, Var(\hat{\sigma}_\eta^2))$$

There is no defined distribution for the household component of the error term, ϵ_{cht}. The error terms are drawn from the empirical distribution of the household residuals, thus avoiding the need to assume a parametric distribution.

A simulated consumption level, $\tilde{y}_{cht}$, is calculated for each household as indicated in equation (8) below.

$$\tilde{y}_{cht} = x'\tilde{\beta} + \tilde{\eta}_{ct} + \tilde{\epsilon}_{cht} \tag{9}$$

The joint distribution of the welfare measure, $\tilde{y}_{cht}$, and other variables is determined by 100 simulations of $\tilde{y}_{cht}$. The mean of the 100 simulations gives the point estimate of household expenditure. The standard error is calculated using Rubin's rules (Rubin (2004)) which takes into account the variation both within and across households so that the variance is a weighted average of the two. Once the distribution of consumption expenditure is estimated, we predict headcount poverty rates according to the $1.90 per day international poverty line as well as the other two thresholds of $3.10 and $5.50 per day. We use the first version of the STATA package SAE developed by (Nguyen, Corral, Azevedo, and Zhao (2017)) to carry out the estimation and the Monte Carlo simulations.

B Alternative Model Specifications

Besides the primary model that was ultimately used for the estimation, we consider three alternative specifications. In one we interact each district level variable with a linear time trend. We refer to this model as the "District dummies*Time Trend" model. The district dummies subsume the district mean characteristics, which are dropped from this specification. The inclusion of district time trends in the model has the advantage of accounting for unobserved characteristics of the district whose effects vary over time in a linear way. The potential downside of the district time trends model is that it extrapolates from past trends to predict district mean welfare in 2014-15 rather than using actual data on district characteristics in 2014-15. While this is more likely to give a rate of poverty reduction in line with past trends, it will ignore the signal from district aggregate indicators in characteristics.

The second alternative model tested includes dummy variables for the three types of expenditure instead of the actual expenditure amounts. The dummy variables are equal to one if the household incurred any expenditure for the particular item. We refer to this model as "Expenditures at the Extensive Margin". This has the advantage of potentially being more robust to outliers in consumption, as well as changes in the way that the question was asked. However, it potentially ignores valuable information on the level of positive expenditures.

Lastly, we use only the 2011-12 data to predict into 2014-15. Because only one round of data is used to estimate the model, no variables are interacted with a time trend. We refer to this as the "Constant Coefficient" model, because the coefficients are fixed over time. This method has been utilized in existing studies that apply survey-to-survey imputation.

We report predicted poverty rates of the four models in Table 5. We validate each model's predictive quality by conducting two tests. First, for each specification we use data from 2004-05 and 2009-10 to predict poverty in 2011-12. We refer to this as forward projection. In the second analysis, we reverse project poverty in 2004-05 using the 2009-10 and 2011-12 data. In the forward and reverse projections, we compare the actual versus predicted headcount of poverty, in 2011-12 and 2004-5 respectively.[35] We finally choose the primary model as our preferred specification because it utilizes all the available data, including levels of expenditure, imposes the least restrictive modeling assumptions on the data, and generates the most accurate out-of-sample predictions, on average, in both tests.

B.1 Forward projection into 2011-12

Table 6 compares actual poverty in 2011-12 with the predictions of the four models. The preferred model is referred to as model 1 in these tables. For urban areas, the projection of model 3 (15.9%) is closest to the actual poverty number (13.4%) followed by model 1. Model 4, the constant coefficient model, gives predictions that are closet to the actual poverty number for rural areas (25.7%), followed by model 1 (23.2%). Therefore, models 3 or 4 could be possibilities for the final

[35] Using the $1.90 per day threshold.

choice based on the forward projection test with model 1 following closely. Therefore, models 3 or 4 perform well on the forward projection test with model 1 following closely behind.

B.2 Reverse projection into 2004-05

Table 7 compares model predictions with measured poverty in 2004-05. For urban areas, the prediction generated by model 1 (30.4%) is closest to the actual poverty number (25.4%). For rural areas, model 1 outperforms all other models. The predicted poverty rate of 45.8% is approximately 2 percentage points different from the actual rate of 43.4%. Therefore, according to the reverse projection test, model 1 is the most accurate overall specification.

The model finally selected is that which has the most accurate closest overall prediction. Based on the forward projection, model 1, 3 or 4 all perform well. However, for the 2004-05 projection, model 3 fares worse than model 1 and is far off the mark, by almost 30 percentage points, in rural areas. Therefore, we select model 1 as the primary specification because of its relatively accurate performance on the forward and reverse projections in both urban and rural areas.

IV Predicted Poverty Rates in 2014-15

A Comparison to the Line-Up Method

Table 8 compares the predictions for headcount poverty from the preferred model to that of the typical line-up method. The model gives an estimated 10.4% poverty rate in urban India, which represents a fall of 3 percentage points from 2011-12. The 95 percent confidence interval for urban areas ranges from 9.1 to 11.7 percent. The predicted headcount rate for rural areas is 13.8, with a 95 percent confidence interval ranging from 12.3 to 15.4 percent. This represents an 11 percentage point reduction from 2011-12. The national estimate is obtained by taking a weighted average of the urban and rural estimates, weighting by the proportion of population in 2011-12. This gives an estimated headcount rate of 12.7 percent, an 8.4 percentage point reduction from 2011-12.

The typical line-up method, gives a much lower prediction for headcount poverty in urban areas, namely 7.1%. In rural areas the prediction is slightly higher at 14.2%. The urban poverty estimate

implies a poverty reduction from 2011-12 at a rate that is approximately half as fast as the line-up method. A comparison of implied elasticities with respect to growth suggests that the typical line-up method, based on growth in HFCE, would have overestimated poverty reduction in India (Table 10).

B Variables that Account for Growth in Mean Welfare

This section considers which variables in the model account for the bulk of the predicted change in poverty between 2011-12 and 2014-15. While this question is difficult to answer directly, indirect evidence is available by decomposing the growth in log per capita consumption predicted by the model. A natural framework for better understanding of the contribution of individual variables is the classic Oaxaca-Blinder decomposition (Oaxaca (1973), Blinder (1973)). This technique decomposes the mean difference across two groups into a portion explained by differences in endowments and a portion due to differences in returns. In this case, the two groups are the households in the 2011-12 survey and the households in the 2014-15 survey. We decompose the difference between the model's mean predicted per capita consumption in 2014-15 and 2011-12. Because the means from each year are generated by predictions from the same model, none of the difference is attributable to changes in the coefficients (returns), and all of the change is due to the mean of the predictor variables (endowments). These changes can easily be decomposed into the portion due to each individual predictor variable, which helps to identify the variables that account for the largest changes in the model.

Table 9 displays the results. The cells report the change in mean log per capita consumption attributable to each set of variables, holding the others constant. Specifically, each cell is the difference between predicted log welfare, multiplied by 100, when the specified set of variables is set to its average in 2014-15, while holding other variables constant at their 2011-12 levels. Results are based on a linear regression, estimated in a dataset combining the two years. The dependent variable is the log welfare in 2011-12, and the predicted log welfare (in 2014-15).

Much of the change in average welfare, especially in rural areas, is attributable to changes in the three included expenditure variables. Expenditure on services, when combining the household and

district mean variables, accounts for an increase of approximately 11 percent in welfare, which amounts to roughly 70 percent of the total change. Expenditures on recreation and transportation account for an additional 10 percent of the total average change in welfare in rural areas. Changes in expenditures play a smaller role in urban areas, where the six expenditure variables account for an approximate 4.1 percent change, slightly less than half of the total. In urban areas, change in the age structure and a decline in the share of the low-caste population had large effects, combining to account for 89 percent of the average increase in welfare in urban areas. Declines in household size also contributed to the predicted rise in welfare in both rural and urban areas.

V Robustness Checks

We conduct three checks to assess the robustness of the predictions generated by the model. First, we calculate elasticities and semi-elasticities to see what the models' predictions imply about the change in extreme poverty with respect to real GDP growth. Second, we examine predicted headcount rates at the state level to test whether the model generates plausible predictions. Lastly, we examine estimated poverty rates at higher poverty lines to ensure that they are reasonable. Ultimately, only the estimated urban and rural poverty rates at the $1.90 line are used to determine the official poverty estimates.

A Elasticities and Semi-Elasticities

Table 10 shows the elasticity and semi-elasticity of the predictions, compared with past experience. As mentioned above, India experienced a rapid fall in poverty despite moderate growth between 2009-10 and 2011-12. This is reflected in a large swing in the poverty-growth elasticity, from about -0.6 between 2004-05 and 2009-10, to -3 between 2009-10 and 2011-12. This swing is also reflected in the semi-elasticity, which jumped in magnitude from about -21 to -85. The model predictions for 2014-15 imply an elasticity and semi-elasticity of -2.8 and -50.6. Both are within the range of the two previous measurements, with the predicted elasticity closer to the measure from 2009-10 to 2010-11, and the semi-elasticity closer to the 2004-05 to 2009-10 period.

As a point of comparison, the typical method used by the World Bank would scale up the 2011-12 survey welfare measure by the growth rate in Household Final Consumption Expenditure (HFCE). That method would predict a poverty rate of 12.1% in 2014-15, which would imply a larger elasticity of -3.4 and a semi-elasticity of -59.4.

B Implied State Level Results

One would expect that states with greater GDP growth would see larger reductions in extreme poverty. State per capita GDP growth does not enter into the model, and the model includes only one global time trend rather than state-specific time trends. Therefore comparing state GDP growth with predicted poverty reductions therefore reveals the extent to which differences in state level per capita GDP growth are reflected in the predictors included in the model.

Figure 2 displays, for each state, change in headcount poverty between 2011-12 and 2014-15, as predicted by the model, on the y axis. The x axis represents the real annual state GDP growth during that period. Goa, which suffered a sharp decline in growth during this period, is a clear outlier. Whether Goa is included or excluded, there is a clear negative correlation between state GDP growth and poverty reduction, as would be expected. Figure 3 shows a comparable plot for the period from 2004-05 to 2011-2. With all states included, the relationship between state GDP growth and poverty reduction between 2011-12 and 2014-15 is remarkably similar to the actual relationship between state GDP growth and poverty reduction between 2004-05 and 2011-12. In particular, both show that a growth rate of 5% per year is associated with about a one percentage point per year reduction in poverty, while a growth rate of 10% is associated with a reduction of about two percentage points per year. Excluding Goa from the more recent period makes the relationship between state GDP growth and poverty reduction somewhat stronger, underscoring the point that the model predicts more rapid poverty reduction in faster-growing states.

C Higher Poverty Lines

Tables 11 and 12 and Figures 4 and 5 report the predicted poverty rates at the lower middle-income line of $3.20 per day per person, and the upper middle- income line of $5.50 per day per

person. The predicted 2014-15 poverty rates for the $3.20 line are 49.4% in rural areas, 33.4% in urban areas, and 44.2% nationally. This is a substantial reduction from the 2011-12 estimate of nearly 16 percentage points, or 27 percent at the $3.20 line. The predicted poverty rates at the $5.50 line also suggest a substantial decline of 12 percentage points, from 89.7% in 2011-12 to a predicted rate of 77.3% in 2014-15 (Table 12). While this reflects that larger numbers of people near the $3.20 line were being pushed out of poverty as the economy improves, there was considerable movement out of the upper middle income levels of poverty as well. Overall, these predicted values for higher poverty lines appear to be consistent with past trends.

VI Conclusion

This analysis was motivated by concerns that the most recent available poverty data from India, were collected three and a half years before the 2014-15 target estimate. The concerns were heightened by the fact that the standard method of applying growth in HFCE to the most recent survey measurement of per capita consumption would overstate poverty reduction in India, and therefore the world. This paper describes an alternative method for estimating poverty, that utilizes a more recent nationally representative survey from 2014-15 containing some of the same demographic and socioeconomic characteristics collected in previous expenditure surveys. Importantly, the 2014-15 survey contains limited expenditure information on three types of services that are comparable to past surveys. The previous surveys are used to estimate a model to predict expenditure and poverty into the 2014-15 survey. This type of imputation method appears to work well in India, based on two validation exercises and a variety of robustness checks. Unlike most other previous applications of survey-to-survey imputation in other countries, the method allows for coefficients to vary linearly over time. The model's predictions imply an elasticity with respect to growth that is in line with past experience. The model also estimates plausible poverty rates at higher poverty lines, and predicts greater poverty reduction in faster growing states. The evidence thus suggests that a survey-to-survey approach can generate reliable estimates in the absence of complete expenditure data.

Table 1: International Poverty Rates: 2004-05 to 2014-15

	2004-05	2009-10	2011-12	2014-15
National				
Poverty rate	38.9	31.7	21.6	12.7
Standard Error	0.4	0.4	0.4	0.7
Urban				
Poverty rate	25.4	19.8	13.4	10.4
Standard Error	0.6	0.5	0.4	0.7
Rural				
Poverty rate	43.4	36.1	24.8	13.8
Standard Error	0.4	0.5	0.5	0.8

Sources: India National Sample Survey Office (NSSO) Surveys and staff estimates.

Figure 1: Historical Poverty Rates: 2004-05 to 2014-15

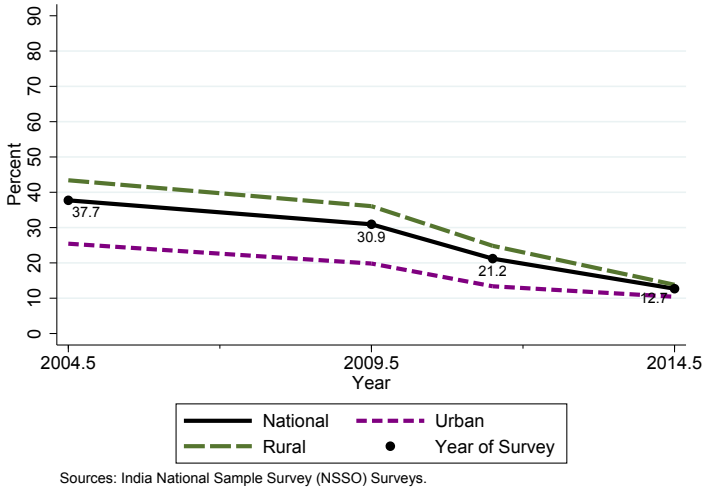


Table 2: Average Annual Change in Poverty Rates

	2004-05 to 2011-12	2011-12 to 2014-15
National	-2.4	-2.8
Urban	-1.7	-1.0
Rural	-2.7	-3.7

Changes in percentage points.

Source: India National Sample Survey Office (NSSO) Surveys.

Table 3: Descriptive Statistics: Urban model

	2004-05	**2009-10**	**2011-12**	**2014-15**
Log HH per capita expenditure				
Mean	4.50	4.61	4.74	.
HH size				
1 or 2	0.07	0.08	0.08	0.09
3	0.10	0.11	0.12	0.13
4	0.22	0.24	0.24	0.25
5	0.21	0.20	0.20	0.20
District HH size				
Share 1 or 2	0.06	0.07	0.08	0.08
Share 3	0.09	0.10	0.11	0.12
Share 4	0.20	0.23	0.23	0.24
Share 5	0.20	0.21	0.21	0.20
HH age structure				
Share 0-14	0.31	0.29	0.27	0.26
Share 15-24	0.19	0.18	0.18	0.18
Share 25-34	0.17	0.17	0.17	0.17
Share 35-49	0.20	0.20	0.21	0.22
Share 50-64	0.10	0.11	0.11	0.12
District avg HH age structure				
Share 0-14	0.33	0.30	0.29	0.28
Share 15-24	0.17	0.17	0.17	0.17
Share 25-34	0.16	0.17	0.17	0.17
Share 35-49	0.19	0.20	0.20	0.21
Share 50-64	0.10	0.11	0.11	0.12
Religion and social group				
Hindu	0.78	0.78	0.77	0.76
Share Hindu in district	0.81	0.81	0.81	0.79
Share sched caste in district	0.63	0.65	0.67	0.68
Household type				
Self-employed	0.43	0.42	0.41	0.37
Share self-employed in district	0.48	0.45	0.46	0.44
Casual laborer	0.12	0.14	0.13	0.15
Share casual labor in district	0.04	0.06	0.05	0.06
Regular wage worker	0.39	0.37	0.40	0.41
Share reg wage worker in dist.	0.20	0.19	0.21	0.23
Principal industry				
Agriculture	0.06	0.06	0.06	0.06
Industry	0.31	0.31	0.31	0.31
Share industry in district	0.24	0.25	0.26	0.26
HH expenditure				
Avg misc services in district	2265	2430	2904	4015
Misc services	2629	2903	3427	4840
Avg rec services in district	247	232	345	591
Recreational services	328	317	433	710
Avg transport*	1102	1093	1641	1964
Transport*	1458	1378	1968	2316
District rainfall shock				
July-September	-0.20	-0.04	0.50	0.02
July-September (squared)	0.34	0.17	0.44	0.28
October-December	-0.26	0.32	-0.51	-0.07
October-December (squared)	0.21	0.55	0.42	0.17
January-March	0.13	-0.25	-0.33	0.62
January-March (squared)	0.29	0.18	0.37	1.14
April-June	-0.04	-0.07	-0.18	0.42
April-June (squared)	0.28	0.35	0.19	0.40

Does not include expenditure on fuel or bus.

Table 4: Descriptive Statistics: Rural model

	2004-05	**2009-10**	**2011-12**	**2014-15**
Log HH per capita expenditure				
Mean	4.18	4.25	4.40	.
HH size				
1 or 2	0.05	0.06	0.06	0.06
3	0.08	0.09	0.09	0.09
4	0.16	0.19	0.19	0.20
5	0.19	0.20	0.21	0.21
District HH size				
Share 1 or 2	0.05	0.06	0.06	0.06
Share 3	0.08	0.09	0.10	0.10
Share 4	0.17	0.19	0.20	0.21
Share 5	0.20	0.20	0.21	0.21
HH age structure				
Share 0-14	0.38	0.35	0.34	0.33
Share 15-24	0.16	0.16	0.16	0.16
Share 25-34	0.15	0.15	0.15	0.15
Share 35-49	0.17	0.19	0.19	0.20
Share 50-64	0.10	0.11	0.11	0.11
District avg HH age structure				
Share 15-24	0.16	0.17	0.17	0.17
Share 25-34	0.15	0.15	0.15	0.16
Share 35-49	0.18	0.19	0.19	0.20
Share 50-64	0.10	0.11	0.11	0.11
Religion and social group				
Hindu	0.84	0.84	0.83	0.83
Share Hindu in district	0.82	0.82	0.82	0.81
Share sched caste in district	0.71	0.73	0.74	0.76
Household type				
Self-employed	0.56	0.52	0.55	0.58
Share self-employed in district	0.54	0.51	0.53	0.55
Share casual labor in district	0.03	0.03	0.03	0.04
Share reg wage worker in dist.	0.07	0.07	0.07	0.08
Principal industry				
Agriculture	0.65	0.62	0.58	0.62
Industry	0.15	0.18	0.20	0.19
Share industry in district	0.17	0.20	0.22	0.21
HH expenditure				
Avg misc services in district	1587	1536	1828	2602
Misc services	1464	1361	1619	2256
Avg rec services in district	168	147	187	326
Recreational services	141	116	152	276
Avg transport in district*	530	585	742	959
Transport*	410	479	611	811
District rainfall shock				
July-September	-0.22	-0.10	0.42	0.02
July-September (squared)	0.27	0.21	0.38	0.22
October-December	-0.30	0.26	-0.62	-0.07
October-December (squared)	0.23	0.43	0.51	0.26
January-March	0.30	-0.30	-0.22	0.63
January-March (squared)	0.43	0.22	0.29	1.15
April-June	-0.21	-0.14	-0.15	0.34
April-June (squared)	0.35	0.43	0.17	0.34

Does not include expenditure on fuel or bus.

Table 5: Predicted Poverty Rates ($1.90 per Day) from Different Models

	Model 1	Model 2	Model 3	Model 4
National	12.7	18.8	17.4	15.4
Urban	10.4	13.1	15.4	10.0
Rural	13.8	21.6	18.4	18.0

Sources: India National Sample Survey Office (NSSO) Surveys.

Table 6: Comparison of Actual Poverty in 2011-12 with Forward Prediction of Models

	Actual	Model 1	Model 2	Model 3	Model 4	Line-Up
National	21.1	21.1	24.3	19.7	25.1	22.9
Urban	13.4	16.5	18.7	15.9	23.9	14.6
Rural	24.8	23.2	27.0	21.6	25.7	27.0

Table 7: Comparison of Actual Poverty in 2004-05 with Reverse Projection of Models

	Actual	Model 1	Model 2	Model 3	Model 4	Line-Up
National	37.5	40.7	48.9	64.1	28.5	47.3
Urban	25.4	30.4	37.0	48.4	19.5	32.0
Rural	43.4	45.8	54.7	71.7	32.9	54.8

Sources: India National Sample Survey Office (NSSO) Surveys.

Model 1: Final Model

Model 2: District dummies*Time Trend

Model 3: Expenditures at the Extensive Margin

Model 4: Constant Coefficient Model

Table 8: Preferred Model (Model 1) vs. Typical Line-Up Method Predictions for 2014-15

	Model	95% C.I		Line-up
National	12.7	11.2	14.2	12.1
Urban	10.4	9.1	11.7	7.7
Rural	13.8	12.3	15.4	14.2

Source: India National Sample Surveys (NSSO) Surveys.

Table 9: Understanding the Drivers of Changes in Log Welfare

	Urban	**Rural**
HH Size (Dist)	1.9	1.4
HH Size	0.8	0.5
Age Category (Dist)	5.1	0.4
Age Category	0.5	2.2
Hindu (Dist)	0.3	0.1
Hindu	-0.1	-0.0
Low Caste (Dist)	2.5	1.6
Low Caste	-0.3	-0.3
HH Type (Dist)	1.0	0.4
HH Type	-2.4	-0.7
Occupation (Dist)	0.1	-2.5
Occupation	0.5	-0.2
Sector (Dist)	-0.2	-0.1
Sector	-0.0	-0.2
Expd on Services (Dist)	2.6	8.1
Expd on Services	-0.3	3.1
Expd on Recreation (Dist)	1.2	0.8
Expd on Recreation	0.8	-0.3
Expd on Transportation (Dist)	0.0	1.3
Expd on Transportation	-0.2	-0.1
Rainfall	-3.4	-0.1
Difference	8.8	15.4
Mean Log Welfare (2011-12)*100	473.3	438.9
Mean Pred Log Welfare (2014-15)*100	482.1	454.3

Sources: India National Sample Survey Office (NSSO) Surveys.

Table 10: Elasticity of Poverty to Growth by Model

	Elasticity	**Semi-Elasticity**
2004-05 to 2009-10	-0.6	-21.4
2009-10 to 2011-12	-3.0	-85.4
Model 1	-2.8	-50.6
District dummies*Time Trend	-0.7	-14.3
Expd at the Extensive Margin	-1.1	-22.7
Constant Coefficient Model	-1.8	-34.6
Typical line-up method	-3.4	-59.4

Sources: India National Sample Survey Office (NSSO) Surveys.

Figure 2: Changes in Poverty Across States: 2011 to 2014

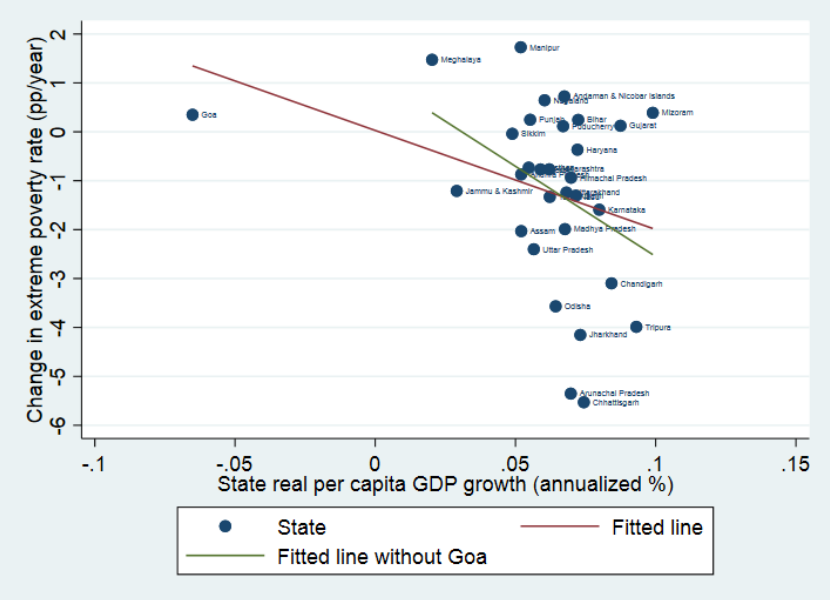


Figure 3: Changes in Poverty Across States: 2004 to 2011

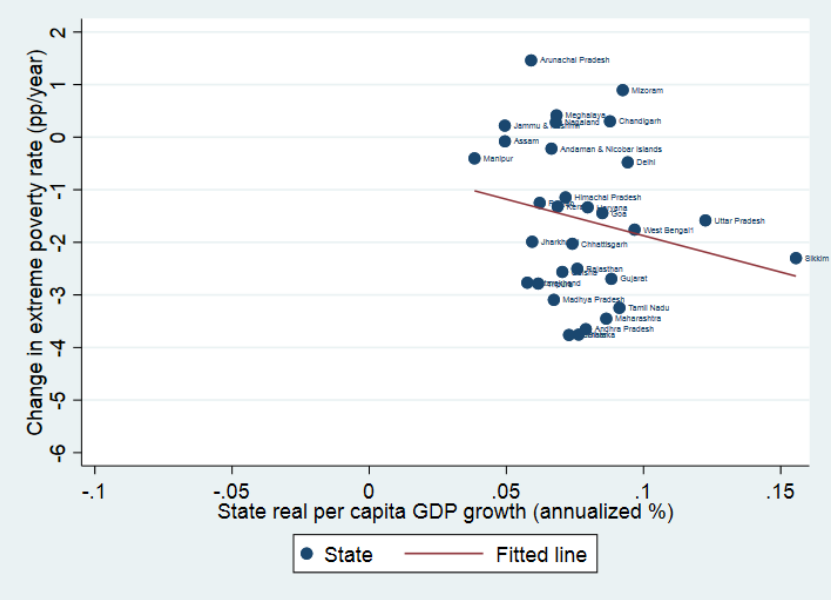


Table 11: Actual and Predicted Poverty Rates at $3.20 Per Day

	2004-05	2009-10	2011-12	2014-15*
National	74.6	69.7	60.4	44.2
Urban	58.4	51.5	43.3	33.4
Rural	82.1	78.1	68.3	49.4

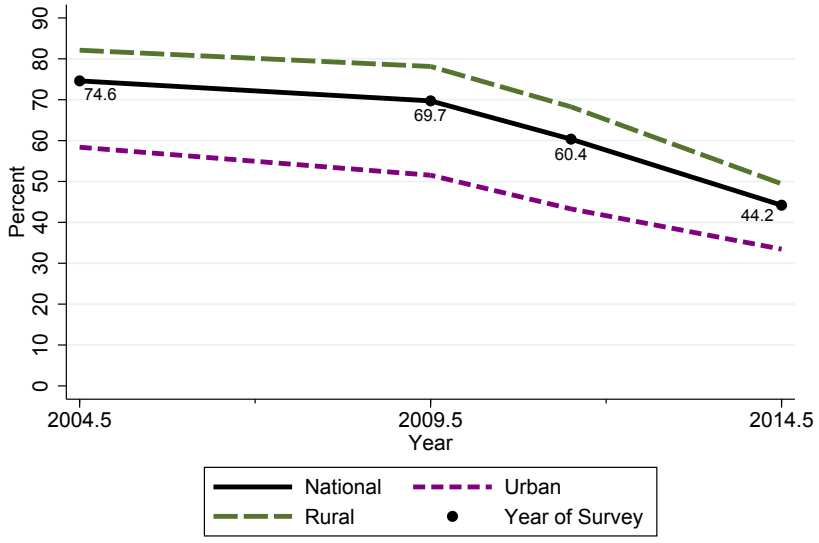
Sources: India National Sample Survey Office (NSSO) Surveys.
*Preferred Model predictions.

Table 12: Actual and Predicted Poverty Rates at $5.50 Per Day

	2004-05	2009-10	2011-12	2014-15*
National	95.7	94.3	89.7	77.3
Urban	95.4	92.1	84.5	65.1
Rural	95.8	95.4	92.1	83.2

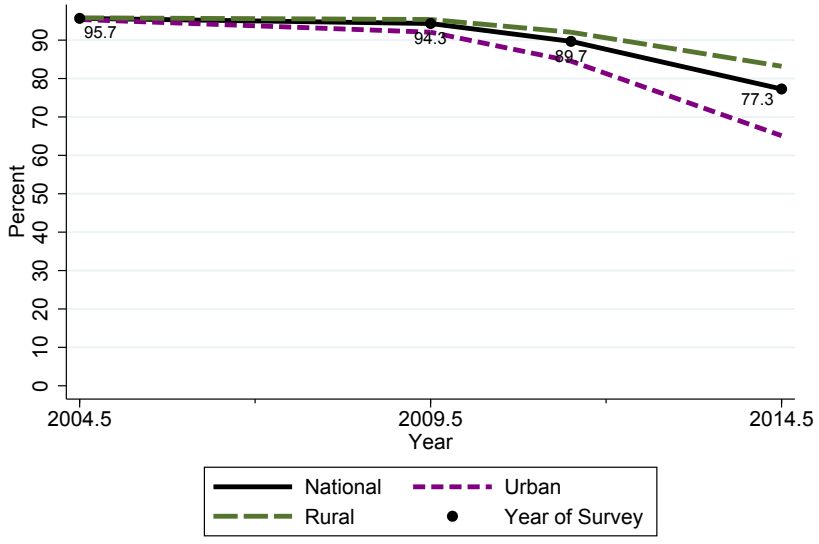
Sources: India National Sample Survey Office (NSSO) Surveys.
*Preferred model predictions.

Figure 4: Poverty Rates: $3.20 Per Day



Sources: India National Sample Survey (NSSO) Surveys.

Figure 5: Poverty Rates: $5.50 Per Day



Sources: India National Sample Survey (NSSO) Surveys.

VII Appendix

Table 13: Descriptive Statistics

	Mean	Min	Max	SD	N
HH size					
1 or 2 (Dist)	0.06	0.0	0.8	0.04	410,761
3 (Dist)	0.10	0.0	0.4	0.05	410,761
4 (Dist)	0.20	0.0	0.8	0.09	410,761
5 (Dist)	0.20	0.0	0.6	0.07	410,761
6+ (Dist)	0.44	0.0	0.9	0.17	410,761
1 or 2	0.06	0.0	1.0	0.24	410,761
3	0.10	0.0	1.0	0.29	410,761
4	0.20	0.0	1.0	0.40	410,761
5	0.20	0.0	1.0	0.40	410,761
6+	0.44	0.0	1.0	0.50	410,761
Prop HH					
0-15 yrs (Dist)	0.33	0.1	0.9	0.07	410,761
16-24 yrs (Dist)	0.17	0.0	0.4	0.03	410,761
25-34 yrs (Dist)	0.16	0.0	0.5	0.03	410,761
35-49 yrs (Dist)	0.19	0.0	0.5	0.04	410,761
50-64 yrs (Dist)	0.11	0.0	0.2	0.03	410,761
65+ yrs (Dist)	0.05	0.0	0.2	0.02	410,761
0-15 yrs	0.33	0.0	1.0	0.22	410,758
16-24 yrs	0.17	0.0	1.0	0.20	410,758
25-34 yrs	0.16	0.0	1.0	0.18	410,758
35-49 yrs	0.19	0.0	1.0	0.19	410,758
50-64 yrs	0.11	0.0	1.0	0.17	410,758
-					
Hindu (Dist)	0.05	0.0	1.0	0.12	410,758
Hindu	0.82	0.0	1.0	0.18	410,761
Low Caste (Dist)	0.82	0.0	1.0	0.39	410,761
Lowcaste	0.72	0.0	1.0	0.20	410,761
Lowcaste	0.72	0.0	1.0	0.45	410,761

Descriptive Statistics (contd)

	Mean	Min	Max	SD	N
HH type					
Self Employed (Dist)	0.51	0.0	1.0	0.15	410,761
Self Employed	0.51	0.0	1.0	0.50	410,761
Casual Urban (Dist)	0.04	0.0	0.5	0.04	410,761
Casual Urban	0.04	0.0	1.0	0.19	410,761
Regular Wage Worker (Dist)	0.11	0.0	1.0	0.13	410,761
Regular Wage Worker	0.11	0.0	1.0	0.31	410,761
-					
Middle Skill Occupation(Dist)	0.11	0.0	0.7	0.07	410,761
Middle Skill Occupation	0.11	0.0	1.0	0.31	410,761
High Skill Occupation (Dist)	0.13	0.0	1.0	0.10	410,761
High Skill Occupation	0.13	0.0	1.0	0.34	410,761
Princip Ind					
Agri (Dist)	0.46	0.0	1.0	0.20	410,761
Agri	0.46	0.0	1.0	0.50	410,761
Industry (Dist)	0.22	0.0	0.7	0.11	410,761
Industry	0.22	0.0	1.0	0.41	410,761
-					
HH services expenses (Dist)	2192.54	0.0	15555.8	1,500.28	410,761
HH services expenses	2192.70	0.0	359904.6	4,706.94	410,348
Rec services expenses (Dist)	251.86	0.0	4264.4	324.15	410,761
Rec services expenses	251.83	0.0	376508.5	1,477.84	410,348
Transp services expenses (Dist)	4328.45	0.0	40845.0	3,495.47	410,761
Transp services expenses	4329.40	0.0	2600945.8	11,789.84	410,348
Rainfall Q3	0.05	-1.3	1.9	0.53	406,829
Rainfall Q4	-0.17	-1.5	2.3	0.57	406,829
Rainfall Q1	0.08	-1.1	2.3	0.72	406,829
Rainfall Q2	-0.02	-1.3	2.1	0.56	406,829
Rainfall $Q1^2$	0.52	0.0	5.2	0.83	406,829
Rainfall $Q2^2$	0.32	0.0	4.6	0.43	406,829
Rainfall $Q3^2$	0.28	0.0	3.5	0.37	406,829
Rainfall $Q4^2$	0.36	0.0	5.3	0.50	406,829

Table 14: Model 1

	Urban	**Rural**
HH size 1 or 2	0.58***	0.33***
	(0.01)	(0.01)
HH size 1 or 2*t	0.00	0.01***
	(0.00)	(0.00)
HH size 3	0.43***	0.27***
	(0.01)	(0.01)
HH size 3*t	0.00	0.01***
	(0.00)	(0.00)
HH size 4	0.33***	0.21***
	(0.01)	(0.00)
HH size 4*t		0.00***
		(0.00)
HH size 5	0.19***	0.14***
	(0.01)	(0.00)
Prop HH 0-15 yrs	-0.34***	-0.40***
	(0.01)	(0.01)
Prop HH 16-24 yrs	-0.11***	-0.07***
	(0.01)	(0.01)
Prop HH 25-34 yrs		0.06***
		(0.01)
Prop HH 25-34 yrs*t		0.00*
		(0.00)
Prop HH 35-49	0.15***	0.16***
	(0.01)	(0.01)
Prop HH 50-64 yrs	0.02	
	(0.01)	
R^2	0.54	0.40
N	128260	197310
F-Stat	3504	2775

Table 14: Model 1 (contd.)

	Urban	**Rural**
Hindu	0.06***	0.02***
	(0.00)	(0.00)
Low caste	-0.13***	-0.12***
	(0.00)	(0.00)
HH type: Self Employed	-0.11***	0.15***
	(0.01)	(0.00)
HH type: Self Employed*t	-0.02***	-0.00***
	(0.00)	(0.00)
HH type: Casual Laborer	-0.35***	
	(0.01)	
HH type: Regular Wage Worker*t	-0.01***	
	(0.00)	
Princip Ind: Agri		-0.09***
		(0.00)
Princip Ind: Industry	-0.01	-0.07***
	(0.00)	(0.00)
High Skill Occupation	0.25***	0.18***
	(0.01)	(0.01)
High Skill Occupation*t		-0.01***
		(0.00)
Middle Skill Occupation	0.06***	0.01**
	(0.01)	(0.01)
Middle Skill Occupation*t	0.01***	0.01***
	(0.00)	(0.00)
R^2	0.54	0.40
N	128260	197310
F-Stat	3504	2775

Table 14: Model 1 (contd.)

	Urban	**Rural**
Recreation services expenses	0.00***	0.00***
	(0.00)	(0.00)
Recreation services expenses*t		-0.00***
		(0.00)
Household services expenses	0.00***	0.00***
	(0.00)	(0.00)
Household services expenses*t	-0.00***	
	(0.00)	
HH type: Self Employed (Dist)		0.18***
		(0.02)
HH type: Casual Laborer (Dist)*t		0.06***
		(0.01)
HH type: Regular Wage Worker (Dist)	0.30***	0.13**
	(0.05)	(0.05)
HH type: Regular Wage Worker (Dist)*t	-0.01	
	(0.01)	
Hindu (Dist)	-0.04	-0.08***
	(0.03)	(0.02)
Princip Ind: Industry (Dist)	0.23***	
	(0.04)	
High Skill Occupation (Dist)	-0.03	0.25***
	(0.04)	(0.06)
Recreation services expenses (Dist)	0.00	0.00***
	(0.00)	(0.00)
Household services expenses (Dist)	0.00***	0.00***
	(0.00)	(0.00)
R^2	0.54	0.40
N	128260	197310
F-Stat	3504	2775

Table 14: Model 1 (contd.)

	Urban	**Rural**
HH size 4 (Dist)	0.10**	0.01
	(0.05)	(0.03)
Prop HH 25-34 yrs (Dist)	-0.27*	
	(0.14)	
Prop HH 35-49 (Dist)		-0.14*
		(0.07)
Rainfall Q1	0.00	0.00
	(0.00)	(0.00)
Rainfall $Q1^2$		0.01
		(0.00)
Rainfall $Q3^2$	0.04***	0.04***
	(0.01)	(0.00)
Rainfall Q4	-0.01***	
	(0.00)	
Rainfall $Q4^2$		-0.01**
		(0.00)
R^2	0.54	0.40
N	128260	197310
F-Stat	3504	2775

References

Belloni, A., and V. Chernozhukov (2013): "Least squares after model selection in high-dimensional sparse models," *Bernoulli*, 19(2), 521–547.

Blinder, A. S. (1973): "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources*, pp. 436–455.

Christiaensen, L., and D. Stifel (2007): "Tracking poverty over time in the absence of comparable consumption data," *World Bank Economic Review*, 21(2), 317–341.

Dang, H.-A., P. Lanjouw, and U. Serajuddin (2017): "Updating Poverty Estimates in the Absence of Regular and Comparable Consumption Data: Methods and Illustration with Reference to a Middle-Income Country," *Oxford Economic Papers*, 69(4), 939–962.

Deaton, A., and J. Dreze (2002): "Poverty and Inequality in India: A Re-Examination," *Economic and Political Weekly*, 37(36), 3729–3748.

Deaton, A., and V. Kozel (2005): "Data and Dogma: The Great Indian Poverty Debate," *World Bank Research Observer*, 20(2), 177–199.

Douidich, M., A. Ezzrari, R. Van der Weide, and P. Verme (2016): "Estimating Quarterly Poverty Rates Using Labor Force Surveys: A primer," *World Bank Economic Review*, 30(3), 475–500.

Edward, P., and A. Sumner (2013): "The Future of Global Poverty in a Multi-Speed World: New Estimates of Scale, Location and Cost," *Working Paper, International Policy Centre for Inclusive Growth, No. 111, International Policy Centre for Inclusive Growth (IPC-IG), Brasilia*, pp. 1–65.

Elbers, C., J. O. Lanjouw, and P. Lanjouw (2003): "Micro-Level Estimation of Poverty and Inequality," *Econometrica*, 71(1), 355–364.

Elbers, C., P. Lanjouw, and P. G. Leite (2008): "Brazil within Brazil: testing the poverty map methodology in Minas Gerais," *Policy Research Working Paper, World Bank*, February(4513), 1–41.

Funk, C., A. Verdin, J. Michaelsen, P. Peterson, D. Pedreros, and G. Husak (2015): "A global satellite assisted precipitation climatology," *Earth System Science Data Discussions*, 8(1), 401–425.

Hillebrand, E. (2008): "Poverty, Growth, and Inequality Over the next 50 Years," *Expert Meeting on how to feed the World in 2050*, pp. 1–23.

Jolliffe, I. (2014): *Principal Component Analysis*. Wiley & Sons.

Kijima, Y., and P. F. Lanjouw (2003): "Poverty in India During the 1990s A Regional Perspective," *Policy Research Working Paper, World Bank*, (October).

Korinek, A., J. A. Mistiaen, and M. Ravallion (2006): "Survey nonresponse and the distribution of income," *Journal of Economic Inequality*, 4(1), 33–55.

Newhouse, D., S. Shivakumaran, S. Takamatsu, and N. Yoshida (2014): "How Survey-to-Survey Imputation Can Fail," *Policy Research Working Paper, World Bank*, July(6961), 1–33.

Nguyen, M. C., P. Corral, J. P. Azevedo, and Q. Zhao (2017): "Small Area Estimation: An Extended ELL Approach," *mimeo*.

Nguyen, M. C., and R. van der Weide (2018): "Estimating poverty without a new income survey and without assuming a time-invariant model," *mimeo*.

Oaxaca, R. (1973): "Male-Female Wage Differentials in Urban Labor Markets," *International Economic Review*, pp. 693–709.

Ravallion, M. (2003): "Measuring Aggregate Welfare in Developing Countries : How Well Do National Accounts and Surveys Agree," *Review of Economics and Statistics*, 85(3), 645–652.

Rubin, D. B. (2004): *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, 58(1), 267–288.