Should You Impute That? A brief look into the art of poverty imputation

Paul Corral, Andres Ham, Peter Lanjouw, Leonardo Lucchetti, and Henry Stemmler

December 5, 2024

1 Introduction

Sporadic household survey collection has been a thorn on the side of those monitoring global poverty long before the COVID-19 pandemic. As H.-A. Dang et al. (2017) note, countries where poverty is mostly concentrated are the countries where there are long lapses between household survey collection efforts. For example, household survey data for monitoring welfare has not been made publicly available in India since 2011. In Nigeria, there was a lack of household survey data for poverty monitoring between 2009 and 2018.

Collecting household survey data that can be used for welfare monitoring is not a simple task. As noted by Yoshida et al. (2021) it is a long and costly endeavor. Beyond just the costs incurred for the design of the survey instrument, the collection of household consumption is also a burden on households since consumption diaries must be filled every day for all household members, or a recall survey is necessary to measure household consumption. These modules are often lengthy and depending on which is chosen, can lead to different types of bias (Beegle et al. 2012).

Household survey efforts across the globe during the COVID 19 Pandemic were completely halted. Household surveys are crucial for the monitoring of living conditions and during one of the greatest health and economic shocks the world was left in the dark about the impact on poverty of the pandemic. Given the lack of survey data which collected information necessary for the construction of welfare aggregates needed to monitor global poverty international organizations had to resort to alternate methods for estimating poverty. One of those methods, survey to survey imputation, was heavily used in conjunction with high frequency phone surveys to obtain an estimate of poverty during the pandemic in countries were HFPS were conducted.[1]

Survey-to-survey imputation methods are a common approach to address the lack of a survey-based measurement of poverty. The basics behind the method build upon the poverty mapping approach introduced by Hentschel et al. (1998) and then further refined by Elbers et al. (2003). The method relies on obtaining the joint distribution between expenditure and household and individual characteristics, and applying the estimated parameters of that distribution to those same household and individual characteristics in a different dataset. In the case of Elbers et al. (2003) the different dataset was the census which allowed estimating poverty for small areas. In survey-to-survey imputation the approach

[1] For global monitoring see Mahler et al. (2021) Global projections of poverty rely on GDP projections during the pandemic. The GDP projection for each country is used to shift the welfare distribution of a given country by the GDP per capita growth rate observed between year corresponding to the household survey collection and the desired year in the future.

is implemented from one sample to another sample where, for any number of reasons, welfare is not adequate for poverty monitoring or is entirely absent.

Both, Hentschel et al. (1998) and Elbers et al. (2003) lay out the preconditions required for survey-to-survey imputation to work. Mainly that the characteristics used to predict welfare must be present in both datasets and correspond to a similar period since the variables should have similar distributions. Deviations, from these preconditions can result in severely biased poverty estimates.

This paper explores how ignoring the basic tenets underlying survey-to-survey imputation may lead to considerably biased estimates. Most papers where survey-to-survey imputation is discussed place a considerable focus on the variables selected for the method to work. That is ignored here. The goal here is to illustrate that survey-to-survey imputation over long periods of time is a futile attempt. To illustrate this, the paper relies on simulated data since it is easier to manipulate and determine what may be driving the observed results. Two main findings are presented; i) in the face of very different data, reweighting so that the means of the covariates match those of population will still yield biased estimates, and ii) survey-to-survey imputation will mostly replicate the welfare distribution of the survey used to construct the model and thus is unlikely to yield useful information on inequality.

The results here are of relevance since survey-to-survey imputation has become more prevalent in recent times to overcome data scarcity. One example is the multiple applications in India, where survey data useful for global poverty monitoring have not been released since 2011. India's importance in global poverty measurement cannot be overstated. Its large population implies that even a slight shift in poverty in the country can make or break the World Bank's pledge on ending extreme poverty. Consequently, since 2011 there are at least 3 different studies which attempt to estimate a poverty rate for the country making use of survey-to-survey imputation (see Edochie et al. (2022), Newhouse and Vyas (2019), and Sinha Roy and Van Der Weide (2022)). These papers obtain a welfare model constructed on the 2011 data and apply it to several more recent data which lack a welfare measure to obtain poverty predictions. These survey-to-survey applications may provide a false sense of certainty on poverty where data does not exist.

2 The basics behind survey to survey imputation (S2S)

Survey to survey (S2S) imputation for poverty measurement relies on the assumption that a population's welfare distribution can be captured by a linear model. Hence, the assumed data generating process (DGP) for transformed welfare $\ln y_i$ is:

$$\ln\, y_i = x_i\beta + e_i;\ e_i \sim N(0, \sigma_e^2) \tag{1}$$

There are many consideration one must take into account when doing S2S over time, mostly because when doing S2S there is a considerable share of the variation of welfare that is unexplained, e_i. The distribution of the unexplained/unobserved portion of welfare must be estimated using actual data. Hence, the portion of the welfare distribution that is unobserved will always correspond to the value of $\hat{\sigma}_e^2$ estimated using the survey where welfare is present.

In essence, we are decomposing the variance of $\ln\, y_i$, into an explainable component, σ_{xb}^2, and a random component, σ_e^2. Thus, what the often used R^2 reflects is the share of the variance of the dependent variable that is explained by the model and is equal to:

$$R^2 = \hat{\sigma}^2_{xb}/(\hat{\sigma}^2_{xb} + \hat{\sigma}^2_{e})$$

When imputing across time, we lack information on the true unobservable component, σ^2_e and must rely on the estimated one that comes from a different point in time, $\hat{\sigma}^2_e$. What this means is that when imputing across time, households with the same observable characteristics will have the same probability of being poor in the imputed data as in the observed data. When doing S2S across surveys conducted in the same period and corresponding to the same population, this is a desirable feature, but when imputing across time, this is a stronger assumption which is unlikely to hold. Consequently, over time the model is not only assumed to hold (meaning that the characteristics used to model welfare will remain the same), but that unobservable factors will affect households equally over time.

Because we assume normally distributed data, for any given household, i, the probability of being poor is entirely dependent on its expected welfare, $x_i\hat{\beta}$, and its error, e_i, which is assumed to follow $e_i \sim N(0, \sigma^2_e)$.

$$FGT_{0i} = \Phi\,(\frac{\ln z - x_i\hat{\beta}}{\hat{\sigma}^2_e})$$

where $\ln z$ is the natural log of the poverty line, and Φ is the standard normal distribution. Note that when imputing across time, the error is assumed to have the same distribution as the one estimated from the original survey where the model is fit, $\hat{\sigma}^2_e$. Hence, households with similar $x_i\hat{\beta}$ over time will have the same probability of being poor.

For the population, differences in poverty between the original survey and the imputed survey (*new*) will be entirely dependent on the composition of the population and the distribution of the linear fit, $\hat{\sigma}^2_{xbnew}$, and the imputed mean transformed welfare which will be given by $\bar{X}_{new}\hat{\beta}$, where $\bar{X}_{new}$ is a matrix of characteristic means for the population of the target survey. Consequently, for the population, poverty is given by:

$$FGT_0 = \Phi\,(\frac{\ln z - \bar{X}_{new}\hat{\beta}}{\hat{\sigma}^2_{xbnew} + \hat{\sigma}^2_e})$$

Assuming that the original model (Eq. 1) assumptions hold, then differences in poverty between the original welfare and the imputed welfare will be due to changes in the population's characteristics. Thus, it will still be considerably dependent on the past $\hat{\sigma}^2_e$, but now the $\hat{\sigma}^2_{xbNEW}$ also plays a role.

Differences between $\hat{\sigma}^2_{xbNEW}$ and $\hat{\sigma}^2_{xb}$ may be due to several reasons, for example $\hat{\sigma}^2_{xbNEW}$ could be smaller if poorer households catch up in education to richer households. The difference may also be due to differences in sampling discrepancies; for example, if the survey where we are imputing to has an over sample of richer households, this may be an issue as it could inflate $\bar{X}_{new}\hat{\beta}$ and also yield a different $\hat{\sigma}^2_{xbnew}$. For example, if the target survey yields an upward estimate of TV ownership (used in the model) and TV ownership is related to higher welfare values, then all else equal, this would yield a higher value for $\bar{X}_{NEW}\hat{\beta}$,which will yield a different poverty rate.

Finally, the same issues affecting poverty will affect Gini when imputing across time. Assuming that welfare is lognormally distributed then Gini is equal to (Crow and Shimizu 1987):

$$Gini = 2\Phi(\frac{\sigma}{\sqrt{2}}) - 1$$

where σ is the standard deviation of $\ln y$. Consequently, the imputed Gini across time is also dependent on $\hat{\sigma}_e^2$, which is estimated in an older survey since

$$\sigma^2 = \hat{\sigma}_{xbNEW}^2 + \hat{\sigma}_e^2$$

and consequently, also subject to changes in the sample's distribution of the observed characteristics used in the model.

Because under most good modeling scenarios the model's R^2 ranges from 0.40 to 0.60, the unexplained portion $\hat{\sigma}_e^2$ may be quite considerable. Note that differences between $\hat{\sigma}_e^2$, which could be estimated using data from a different point in time, and actual unobserved distributional changes could be due to observable factors which may now matter for welfare in the imputed years. For example, the introduction of a cash transfer program, which is not captured in the original model. Or due to other types of shock, for example, currency demonetization.

3 Survey to survey imputation a brief overview

Much of the work of survey to survey imputation of poverty began with the work of Elbers et al. (2003) and Hentschel et al. (1998). The authors focus on obtaining poverty estimates for smaller localities in Ecuador by using parameters obtained from a linear model where the dependent variable is household welfare fit on the household survey.[2] The authors rely on covariates which are readily found in the household survey as well as in the population census. The estimated parameters are applied to the country's census data which covers the entire country but lacks a welfare measure which is apt for poverty measurement. With model parameters in hand it is possible to recreate in the census population the welfare distribution observed in the survey. From the imputed welfare distribution it is possible to then obtain estimates of any indicator as if one had the actual welfare distribution available. For an indicator such as headcount poverty, being able to replicate the welfare distribution accurately is essential to ensure unbiased estimates as noted by Corral et al. (2022).

There are basic preconditions required for survey to survey imputation to work. The most basic one is that the covariates used in both datasets must correspond to the same population. This means that the distribution of these covariates must have similar moments, e.g. mean and variance. Differences between covariates across datasets can lead to biased estimates.

The methodology of survey to survey imputation where actual welfare data are unavailable has gained a bit of prominence within the World Bank. A good example is the Survey of Well-being via Instant and Frequent Tracking (SWIFT) program. In essence the program relies on survey to survey imputation where a small and cost effective sample of data are collected in the field and poverty estimates for the collected data are obtained by imputation. SWIFT has also been used to obtain poverty estimates in a survey which has been collected much more recently than the one used for the modeling, which has led to incorrect poverty estimates. Such an example is that of Afghanistan in 2015, where the model was trained on 2011 data and failed to capture an increase in headcount poverty. As an answer to this instance the SWIFT program added variables which change along economic conditions. However, the program notes that without updating the training data where the model parameters are estimated there is always a risk of not properly capturing changes in the face of economic shocks (Yoshida et al. 2021).

[2] The method falls under the academic literature of small area estimation. For more details on small area estimation refer to Rao and Molina (2015).

Because survey to survey imputation may yield biased estimates when the data used to obtain parameter estimates and the data imputed to are very different, multiple recommendations on addressing the differences have been made. For example, Stifel and Christiaensen (2007) note the estimated parameters are assumed stationary over time and that inclusion of key time varying variables such as rainfall and prices are important and may allow to capture changes over time. However, the authors also indicate that predictions too far in the future or in the past should be avoided, mostly due to the restrictiveness of the stationary parameter assumption. In essence, the stationary assumption implies that any changes in poverty captured over time are entirely attributable to changes in covariates used in the model and not due to changes in unobservables or rates of return to the covariates used (Dang, Lanjouw and Serajuddin 2014). Christiaensen et al. (2012) note that in fast growing economies, like India, the stationary assumption may be controversial.

The bias of estimates obtained with data that correspond to very different time periods or data where the covariates are considerably different has mostly been studied using real world data. For example, Dang, Lanjouw and Serajuddin (2014) conduct experiments using the Household Expenditure and Income Survey and the Unemployment and Employment Survey in Jordan; Stifel and Christiaensen (2007) rely on survey data for Kenya to conduct experiments; Dang et al. (2021) applies the method to several countries and note that estimates are well within margins of error.[3]

Christiaensen et al. (2012) undertakes an empirical validation of survey to survey imputation methods over time. The authors perform survey to survey imputation over time in scenarios where there is a comparable expenditure data which provides a “true” estimate of poverty. The authors validate their approach using data for Vietnam and for China using a rural household panel dataset. In Vietnam, the authors obtain a model using the 1992/93 data and predict poverty using the 1997/98 data. They note that the method works relatively well and depending on the covariates used, differences between predicted and observed poverty rates were on average 3.4 percentage points during a period where poverty fell by 23.2 percentage points. For the Chinese regions where the method was tested, the authors also find that the methods work relatively well. However, depending on the model used differences between predicted and observed rates were considerable.

3.1 The Case of India

Applications of survey to survey imputation have also made their way to global poverty monitoring. This is mostly due to India’s lack of recent survey data. The last expenditure survey published by the National Sample Survey agency of India dates back to 2011. Since then, official poverty estimates have been lacking. In 2017, when the next expenditure survey for the country was supposed to be released it was scrapped due to concerns regarding its validity. There were leaks of the report, however, and these suggested that between 2011 and 2017/18 consumption in the country had fallen by 3.7 percent.[4] While rural consumption fell by 8.8 percent, urban areas fared somewhat better and grew by 2 percent over the period. Given the lack of actual data to validate the leaked report there have been multiple attempts to obtain a prediction for headcount poverty in the country (see Edochie et al. (2022), Newhouse and Vyas (2019), and Sinha Roy and Van Der Weide (2022)). All attempts rely on the 2011 survey to obtain parameters which are then applied to more recent data from different sources.

Newhouse and Vyas (2019) obtain model parameters estimated using the 2011 data and apply these to the NSSO expenditure on sevices and durables survey of 2014/15 to obtain a headcount poverty rate

[3] The countries are: Ethiopia, Malawi, Nigeria, Tanzania, and Vietnam

[4] https://www.thehinducentre.com/the-arena/current-issues/article30265409.ece

for 2015.[5] The authors suggest that between 2011 and 2015 headcount poverty rates in the country had fallen by nearly 8 percentage points, with a predicted poverty rate in 2015 of 14.6 percent.[6]

A similar exercise to the one of Newhouse and Vyas (2019) was undertaken to obtain a poverty rate for India for 2017 by Edochie et al. (2022). In that study the authors find that in 2017 the share of Indians who live on less than $1.90 (USD PPP) per person per day is 9.9 percent. The drop in poverty between 2011 and 2017 (22.5% to 9.9%) is in sharp contrast to the leaked results from the 2017 leaked report from India's NSO which suggested poverty had increased in the country. The authors rely on model parameters estimated using the 2011 data and apply these to the 2017/18 Survey on Social Consumption (SCS) on Health. The SCS is a nationally representative survey and thus the authors' concern is mostly centered on imputing across time and recognize that they assume parameters remain constant over time. Consequently, the authors assume that the entire change is explained by the changes in covariates. The authors corroborate their results by using a pass-through approach to obtain a projected poverty rate of 10.4 percent.[7]

To obtain a picture of how poverty in India has evolved since 2011, Sinha Roy and Van Der Weide (2022) follow a similar route to Edochie et al. (2022) and Newhouse and Vyas (2019) by obtaining model parameters estimated using the 2011 data and applying these to the Consumer Pyramids Household Survey (CPHS). The CPHS survey is conducted by a private company and has been cited not representative. Specifically, Somanchi (2021) notes that the CPHS under-samples women and children, over-represents more educated households and underrepresents the poor. Moreover, Somanchi (2021) also notes that sampling issues in the CPHS data may have gotten worse over time. To address the issues with the biased data, Sinha Roy and Van Der Weide (2022) implement a reweighting procedure to yield adjusted survey weights. The adjusted weights are obtained via a minimum-cross entropy procedure which uses the weighted means of the target variables between the CPHS and other nationally representative surveys. The authors find that in 2019 the share of Indians who live on less than $1.90 (USD PPP) per person per day is 12.3 percentage points lower than in 2011 where the poverty rate stood at 22.5 percent.

3.2 The Case of Afghanistan

Afghanistan's poverty estimates for 2023 rely on a survey-to-survey imputation model (Barriga-Cabanillas et al., 2023). This model uses an expenditure model estimated on the third quarter of the 2019-2020 Expenditure and Labor Force Survey (IE-LFS), which measured the national poverty rate at 52.3 percent. The imputed poverty rate for the corresponding period in 2023 stands at 48.3 percent, indicating a 4-percentage-point drop. The authors state that this national trend masks important regional differences: urban poverty rates slightly increased, while a decrease in rural areas offset this rise and contributed to the overall decline.

Nevertheless, the country's economy during that period faced considerable challenges. Between 2019 and 2023, Afghanistan experienced a significant GDP contraction, particularly following the change in administration and the departure of the U.S. Armed Forces. According to the country's Macro-Poverty Outlook for the 2023 annual meetings, this GDP drop coincided with falling food prices, leading to increased labor force participation and a doubling of unemployment. Agricultural households—predominantly in rural areas—were likely the most affected by these lower prices. Yet, Barriga-Cabanillas et al. (2023) suggest that the overall poverty reduction is driven by declines in rural poverty, possibly indicating that falling

[5] The authors offer multiple models, the preferred model includes data from past expenditure surveys.

[6] Using $1.90 USD PPP per capita per day as the threshold.

[7] The authors use a pass-though rate of 0.67 applied to growth in Household Final Consumption Expenditure in national accounts. The pass-through rate times the growth rate are then applied to shift the welfare distribution neutrally to obtain a projected poverty rate.

prices might offset income losses in these areas. However, this interpretation is speculative, as it is not explicitly modeled or thoroughly discussed by the authors.

Barriga-Cabanillas et al. (2023) argue that much of the observed welfare changes stem from dummy variables capturing consumption patterns, such as the presence or absence of specific goods like apples. Their model test results (Table 4) show that imputed estimates using the same data as the one used for the model yielded an estimate that was nearly 1 percentage point off the real value. The model when applied to data for 2021, was already giving estimates that were 1.7 percentage points above actual urban values and 2.1 points above rural values, suggesting that a 4-point margin of error for 2023 is plausible. Given the economic indicators suggesting worsening conditions, the reported 4-percentage-point poverty drop seems questionable. While reduced conflict might partially explain this trend, such factors are not included in the model and remain speculative narratives rather than data-driven conclusions.

3.3 The Case of Zambia

Zambia had a gap in survey data between 2015 and 2022. Nevertheless, the 2022 consumption data is not comparable to the 2015 consumption data. The reason for the lack of comparability is that the food module in 2022 uses a different recall period than the 2015 survey. The 2015 survey relied on a fixed recall period, whereas the 2022 survey only used a fixed recall period when inquiring if the household consumed the good over that period. Nevertheless, respondents in 2022 were allowed to select a different reference period when providing information on quantities and value of their consumption. The difference in reporting periods leads to a lack of comparability in the data, which compromises the construction of a poverty trend (Beegle et al., 2012).

To solve the lack of comparability, the Zambia team relied on survey-to-survey methods to impute the 2015 welfare aggregate on the 2022 data. Through that approach, the team finds that international poverty ($2.15 2017 USD PPP) in the country worsened between 2015 and 2022 by close to 4 percentage points, from 60.8 to 64.4 percent. At the same time the team reports a drop in inequality, measured by the Gini index, from 55.9 to 51.5. Average consumption is also reported to have declined during the period, from $2.97 (2017 PPP) to $2.53 (2017 PPP), a 15 percent drop in consumption. All these indicators, except for inequality, point towards a worsening situation in Zambia.

Despite the negative story from the imputations, the macro story is not as negative. The country's GDP per capita between 2015 and 2022 is relatively the same. After a considerable drop during the COVID-19 pandemic, GDP per capita in constant terms reached its 2015 levels by 2022, and by 2023 it is above its peak in 2018. Hence, the imputed consumption aggregate suggests that there has been an increase in the gap between national accounts data and household survey data which is not explained. One of the suggested reasons for the gap between national accounts and household surveys has been under reporting of incomes in surveys (Ravallion, 2003), although there is evidence that the gap shrinks as countries become richer (Prydz et al., 2022).

The comparable components of expenditure correspond to 33.7 percent of the 2015 survey expenditure, but does not include food, and frequent non-food components. The comparable component consists mainly of health, a sub-set of education, clothing, financial services, durables, and housing. The last item, housing, corresponds to imputed rent. In urban areas, the comparable component suggests consumption has decreased in real terms. In rural areas there is no discernible change. The authors validate their results applying a method from Deaton (2003) that relies on a comparable subset of the welfare aggregate nad is aligned to that based on the imputation model.

The authors also validate their model by imputing on the same data as the one used for the model. Their validations already point toward a slighupward bias in their model for urban areas, same for their Gini predictions, both likely driven by the residuals not meeting the model's assumptions (Table 6). There are concerns with the approach taken in this imputation exercise. First, the models fit have a surprisingly high R^2 value – 0.8 for rural, and 0.91 for urban areas. Such a high R^2 value may be suggestive of overfitting. The rural model includes as a covariate the natural logarithm of comparable reat consumption per addult equivalent, and its square – the coefficients for both are positive.[8] The rural model also includes number of items purchased in logarithms, where presumably $\ln(0)$ is treated as 0, which can lead to biased coefficients particularly when the proportion of 0 is large (Battese, 1997).[9] Finally, the model likely includes many covariates that are potentially highly correlated. For example, number of tubers consumed from own consumption and purchased.

The model for urban areas has an R^2 value of 0.91, one of the higre values obsevred in such an exercise. The model includes the number of inactive household members, which a priori one would expect is negatively related to consumption, but is positive in this case. Moreover, the model suggests that between 2015 and 2022 the number of inactive members increased by nearly 1 person, from 1.6 to 2.4. Beyond the issue of the inactive members, the urban and rural models share many of the same limitations. Mainly, the high $R^2$is suggestive of overfitting that would limit its predictive out-of-sample capacity, particularly when applying the model to data that is 7 years ahead.

The next section discusses two topics regarding survey to survey imputation. The first is related to imputations made on to biased samples and how reweighting as implemented in most studies is not an adequate solution. The second is related to imputations over time assuming constant parameters. The section relies on simulated data to illustrate the potential problems that may arise under cross temporal imputation and imputation to biased samples.

4 Simulations under a controlled scenario

In this section some of the aspects discussed in Section 2 are explored. Mainly the section will rely on simulated data to explore how violations of the underlying tenets of survey to survey imputation may affect imputed poverty estimates.

4.1 Imputing to the Same Population

The section first looks at imputations to the same population - those that correspond to the same period. For example, you have a living standards survey and a contemporaneous Demographic and Health Survey (DHS). In this case, imputing welfare to the DHS can be used to determine the incidence of stunting or some other health indicator across the welfare distribution. In essence, the assumption is that the living standards survey and the DHS are samples of the same population.

The section first details how populations for the simulations shown are created. Since it is possible for a sample to be biased, biased samples are also created. The section first explores how imputation from a model fit to a sample of the population works when imputing to a separate sample of the same population. The section then explores how imputing to a biased sample may affect predictions and how one may overcome biased samples for poverty imputation.

[8] Includes health, a subset of education, clothing, financial services, durables, and housing. It has a correlation of 0.987 with total consumption and corresponds to 33.7 percent of consumption as noted by the authors.

[9] The authors include the nat. log. of hoes owned as well a fishing and hunting gear which likely have multiple 0.

4.2 Creating populations

We create 1,000 populations of 20,000 households where the welfare of the population is generated with the following DGP:

$$\ln y_i = 3 + 0.1x_{1_i} + 0.5x_{2_i} - 0.25x_{3_i} + 0.2x_{4_i} - 0.15x_{5_i} + e_i$$

where $e_i \sim N(0, 0.5^2)$

1. x_i is a discrete variable, simulated as the rounded integer value of the maximum between 1 and a random Poisson variable with mean $\lambda = 4$
2. x_2 is a binary variable, taking value 1 when a random uniform number between 0 and 1 is less than 0.2
3. x_3 is a binary variable, taking value 1 when a random uniform number between 0 and 1 is less than 0.5 as long as $x_2 = 1$, otherwise it is equal to 0
4. $x_4 \sim N(2.5, 2^2)$
5. x_5 is a variable drawn from a Student's t distribution with 5 degrees of freedom and scaled by 0.25

The Gini for this distribution is 0.38, and the covariates explain roughly 47 percent of the variation of $\ln y_i$.

4.2.1 Extracting samples

Under each of the 1,000 populations, the samples are taken in the following manner :

1. **Random sample:** a 20 percent simple random sample (SRS) of the population.
2. **Bottom biased sample:** here, the poorest quintile is purposely undersampled.
   (a) For the bottom 20, take a c percent as a sample. This means that for the 20th centile, we take an SRS sample of 20%, for the 19th centile, we sample 19 percent, for the 18th centile, we sample 18 percent, and so forth.
3. **Top and bottom biased sample:** Create a biased sample, where the poorest quintile and the richest quintile are purposely under sampled.
   (a) For the top 20, sample $100 - c$ percent. This means that for the 80th centile, we take an SRS sample of 20%, for the 81st centile we sample 19 percent, for the 82nd centile, we sample 18 percent, and so forth. Note that we do not sample the top 1 percent.
   (b) For the bottom 20, we sample c percent. This means that for the 20th centile, we take an SRS sample of 20%, for the 19th centile, we sample 19 percent, for the 18th centile, we sample 18 percent, and so forth. Note that we do not sample the bottom 1 percent.
   (c) For all other centiles (21-79), an SRS sample by centile is taken.

4.2.2 Adjusting for potential bias in the sample through reweighting

The purpose of this section is to test the implications of working with a biased sample. Specifically, samples that under-sample segments of the population. Re-weighting is applied by Sinha Roy and Van Der Weide (2022) for the case of India where there were concerns regarding the sampling of the target survey. Although referred to by the authors as a max-entropy approach, the method used to adjust the target sample weights is in reality minimum cross-entropy where the sampling weights in the target survey are adjusted by the smallest extent possible given the constraints (Wittenberg 2010).[10] The constraints, in the case of India's imputation, correspond to the mean of a set of variables that are found in other nationally representative surveys.

Three possible adjustments for the covariate mismatch is proposed:

1. **Covariate match**: The simple random sample's mean of each covariate is used to calibrate the weights of the biased samples. For re-weighting, I use minimum cross-entropy to adjust prior weights by the smallest amount possible so that the constraints (i.e., the means) are satisfied.
2. **Linear fit match**: The linear fit's mean of the simple random sample is used to calibrate the weights of the biased samples. Weights are calibrated using minimum cross-entropy to adjust prior weights by the smallest amount possible so that the constraints (i.e., the means) are satisfied.
3. **Linear fit & variance match**: The linear fit's **mean** and **variance** of the simple random sample are used to calibrate the weights of the biased samples. Weights are calibrated using minimum cross-entropy to adjust prior weights by the smallest amount possible so that the constraints (i.e., the mean and the variance of the linear fit) are satisfied.[11]

4.2.3 Adjusting for potential bias in the sample through standardizing covariates

The basics of the method is simple. In principle, from the data where the model is fit (source data) the mean and variance of each covariate is known. Consequently, covariates in the sample to be imputed (target data) can be adjusted so that the mean and variance of the covariates are aligned to that of the source data. This is what H.-A. H. Dang et al. (2014) recommend for making different survey samples more comparable. However, a key assumption here is that the covariates follow a normal distribution. In tests done by H.-A. H. Dang et al. (2014) with data from Jordan the method is judged to be valid.

To test the validity of covariate standardization, the same samples detailed in section 4.2.1. Note, however, that the covariates here do not follow a normal distribution. Covariate standardization is done by adjusting the covariate from the *new* survey in the follwoing manner:

$$x_{new} = (x_{new} - \hat{\mu}_{x_{new}})\frac{\hat{\sigma}_x}{\hat{\sigma}_{x_{new}}} + \hat{\mu}_x$$

The following adjustments are considered:

1. **Standardize each covariate:** Every covairate in the model is standardize to ensure the mean and variance of each covariate in the biased samples matches those of the simple random sample.

[10] When the original sampling weights of the target survey are ignored, it is assumed that every observation had a similar probability of being selected, then the re-weighting method is max-entropy (see Golan, Judge and Miller (1996) for a more thorough exposition of maximum entropy).

[11] Since the simulated data does not have sampling weights, the priors are equal to 1 and thus minimum cross-entropy is in reality max-entropy in this instance.

2. **Standardize the linear fit:** The linear fit's **mean** and **variance** of the simple random sample are used to standardize the linear fit in the biased samples.

4.2.4 Results

To remove the potential for other sources of bias, instead of refitting a model on the sample, the actual values for the coefficients β and the error distribution σ_e^2 are used. For each sample, under each population, 100 vectors in the sample are created by adding the linear fit to a randomly drawn error term $e_i \sim N(0, 0.5^2)$ to every observation; note that the linear fit is kept fixed. Then, poverty and Gini estimates are obtained for each of the 100 created vectors using the newly created survey weights or the standardized covariates to calculate poverty and Gini for the samples. The average value of the 100 Gini and poverty estimates is used as the final estimate. Bias is calculated for each measure of interest by comparing the estimate with the actual value from the population. Results are presented in Table 1.

Under the simple random sample, bias is nearly 0 for all indicators of interest.

Poverty estimates under the biased samples are downward biased. The bias is more considerable for the top and bottom biased samples. Additionally, the bias for $FGT0$ is lower for low thresholds and increases before falling again for higher thresholds. The direction of the bias would eventually change as the distributions cross paths. Thus, the size and direction of the bias are dependent on the threshold chosen.

The results show that standardizing covariates as suggested by H.-A. H. Dang et al. (2014) is a better approach than reweighting covariates, of course this result may not hold under different covariates. However, even better results are obtained by standardizing the linear fit ($X\beta$). Standardizing $X\beta$ yields phenomenal results under both biased samples tested. Given that the standardizing relies on normally distributed data, perhaps a better option is achieved from reweighting.

Results presented in Table 1 suggest that not all types of reweighting work well. Re-weighting to match means, as done in an application in India by Sinha Roy and Van Der Weide (2022), still yields biased results (see Table 1, rows Covariate mean match (x100)). Since the β and σ_e^2 used are always the true values, one can surmise that the bias from the sampling cannot be overcome by simply reweighting the data. This also holds true for cases where the weights are adjusted so that the linear fit in the biased sample matches the mean of the SRS data. This introduced bias will be in addition to any other potential sources of bias, for example, non-normally distributed residuals or using a σ_e that does not correspond to the same period. The bias is in addition to all the other potential sources of bias of the imputation, such as deviation from the model's assumptions or assuming that the distribution of unobservables is the same as it was in a different period.

Adjusting weights so that the linear fit ($X\beta$) and its variance in the biased sample matches that of the SRS yields excellent results which rival those from standardizing $X\beta$ for bottom biased samples and supersedes them in the top & bottom biased sample. The added advantage with the reweighting is that there are limited assumptions imposed on the distribution of the linear fit.

For Gini, a similar issue is at play. Because biasing a sample by excluding households at the bottom of the distribution will likely lead to lower var$[x\beta]$, which is not addressed by the re-weighting procedure used,[12] keeping everything else equal leads to a downward biased estimate of Gini. This is worse under the top and bottom biased samples because the value of $var[x\beta]$ is considerably smaller than under the

[12] Note that under the minimum cross entropy procedure chosen it would be possible to set the variance as one of the constraints. However, this is not implemented here because in practice $var[x\beta]$ for a given point in time, is not known.

Table 1: Bias (x100) due to re-weighting

		Headcount poverty rate								
	Gini	*5%*	*10%*	*15%*	*20%*	*25%*	*30%*	*35%*	*40%*	*45%*
Simple random sample (x100)	-0.0174	0.0065	0.0134	0.0009	0.0119	0.0098	0.0204	0.0124	0.0206	0.0292
Bottom biased samples										
Standardize all covariates (x100)	-0.331	-0.251	-0.310	-0.306	-0.274	-0.224	-0.162	-0.087	-0.025	0.038
Standardize only $X\beta$ (x100)	0.117	-0.074	-0.061	-0.026	0.011	0.046	0.080	0.116	0.132	0.143
Covariate mean match (x100)	-0.741	-1.208	-1.918	-2.416	-2.772	-3.021	-3.181	-3.257	-3.280	-3.246
Covariate mean and variance match (x100)	-0.741	-1.208	-1.918	-2.416	-2.772	-3.021	-3.181	-3.257	-3.280	-3.246
$X\beta$ match (x100)	-0.923	-0.536	-0.699	-0.732	-0.692	-0.608	-0.492	-0.348	-0.205	-0.056
$X\beta$ and var $[X\beta]$ match (x100)	0.180	-0.106	-0.088	-0.042	0.011	0.062	0.109	0.158	0.186	0.205
Top & bottom biased sample										
Standardize all covariates (x100)	-1.159	-0.423	-0.605	-0.685	-0.705	-0.679	-0.618	-0.522	-0.417	-0.298
Standardize only $X\beta$ (x100)	-0.120	0.047	0.037	0.021	-0.002	-0.025	-0.047	-0.060	-0.082	-0.101
Covariate mean match (x100)	-2.160	-0.767	-1.020	-1.073	-1.009	-0.860	-0.654	-0.396	-0.125	0.162
Covariate mean and variance match (x100)	-2.160	-0.767	-1.020	-1.073	-1.009	-0.860	-0.654	-0.396	-0.125	0.162
$X\beta$ match (x100)	-2.150	-0.900	-1.263	-1.407	-1.421	-1.335	-1.181	-0.963	-0.720	-0.451
$X\beta$ and var $[X\beta]$ match (x100)	-0.083	0.024	0.027	0.027	0.018	0.009	-0.004	-0.008	-0.026	-0.044

bottom biased sample. Consequently, adjusting the weights so that not just the mean of $X\beta$ matches the value observed in the SRS but also var$[x\beta]$ yields much better imputed values for Gini.

4.3 Imputation over time

A key assumption when imputing across time is that the parameters are constant over time. This means that the relationship between the covariates and the dependent variable remain constant and that the distribution of the unobservables remain constant. In essence, the assumption is that any change in the welfare distribution will be entirely driven by changes in the covariates. Nevertheless, under variaous early work of S2S the common recommendation was to rely on covariates that are relatively stable and do not change (). Recommending the use of covariates that show minimal change is contradictory to the previous statement if the goal is to capture changes in poverty. In this section the discussion focuses on how changes in the parameters can affect the poverty predictions. In practice, it is possible that any of the parameters used for the imputation (β, σ) has changed over time, however it is impossible to know if changes compound or cancel eachother.

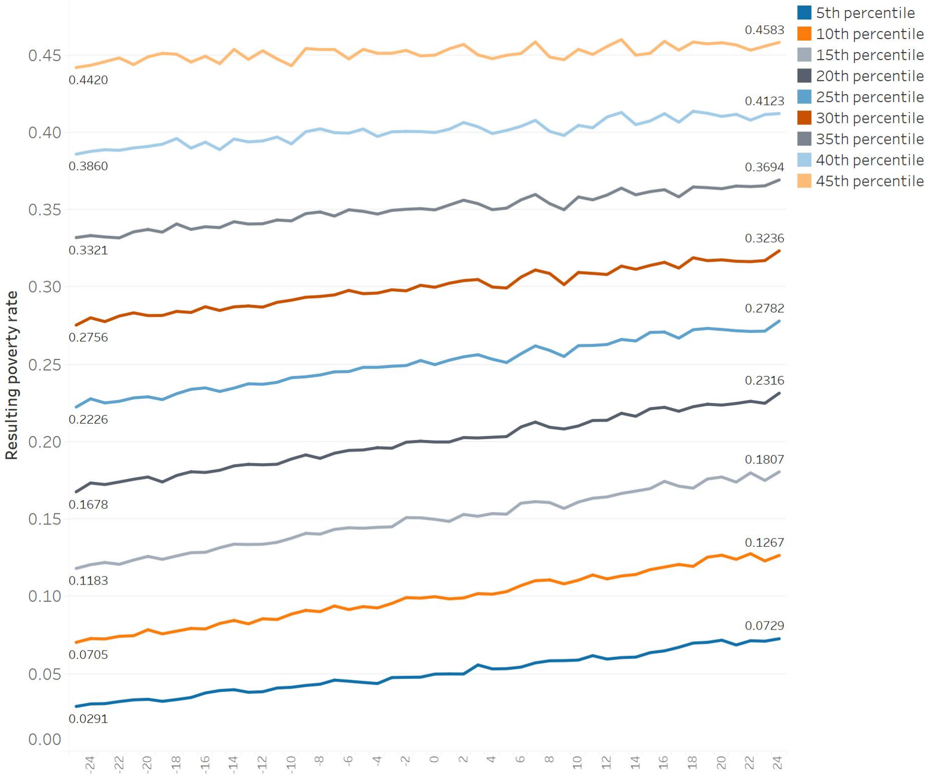
4.3.1 Changes in the Residual's Distribution

First note that it is possible, although unlikely, for the relationship between covariates and the dependent variable to be constant over time and there to be a change in poverty. The change can be driven entirely by changes in the unobservables. Changes in the unobservables also implies a change in inequality. To illustrate this, the same population created in section 4.2 is used. The simulation consists in keeping the linear fit constant and only change the value of σ_e. Results are presented in Figure 1, and illustrate how poverty rates change under different lines determined at the percentile of the original welfare. Hence, if the original poverty rate was 25 percent, and using that same threshold when lowering the value of σ_e by 25 percent, the resulting poverty rate would be of 22.3 percent. At higher threholds the gap between the original poverty rate and the new rates is less noticeable, although still present.

4.4 Changes in the Constant Term

When conducting imputations over time, changes to the constant term are seldomly considered. Failing to capture a change in the constant term can lead to relatively stagnant predictions over time. It will

Figure 1: Change in poverty prediction if σ changes by $x\%$



also lead to predictions that are considerably off. In Figure 2 the true constant term is adjusted by $x\%$ in the data generating process for the target data. Even slight changes lead to considerable differences in poverty. Imagine an unlikely scenario where transformed welfare for everyone has increased by 5%, yet everything else about the welfare distribution remains the same. This would entail a neutral distribution shift to the right, and thus everyone's transformed welfare is improved by 5%. However, relying on a model fit on data before the increase would predict a constant term that is lower than what it really is after the increase in welfare. This would lead to imputations that considerably overestimate poverty for the new period.

5 References

References

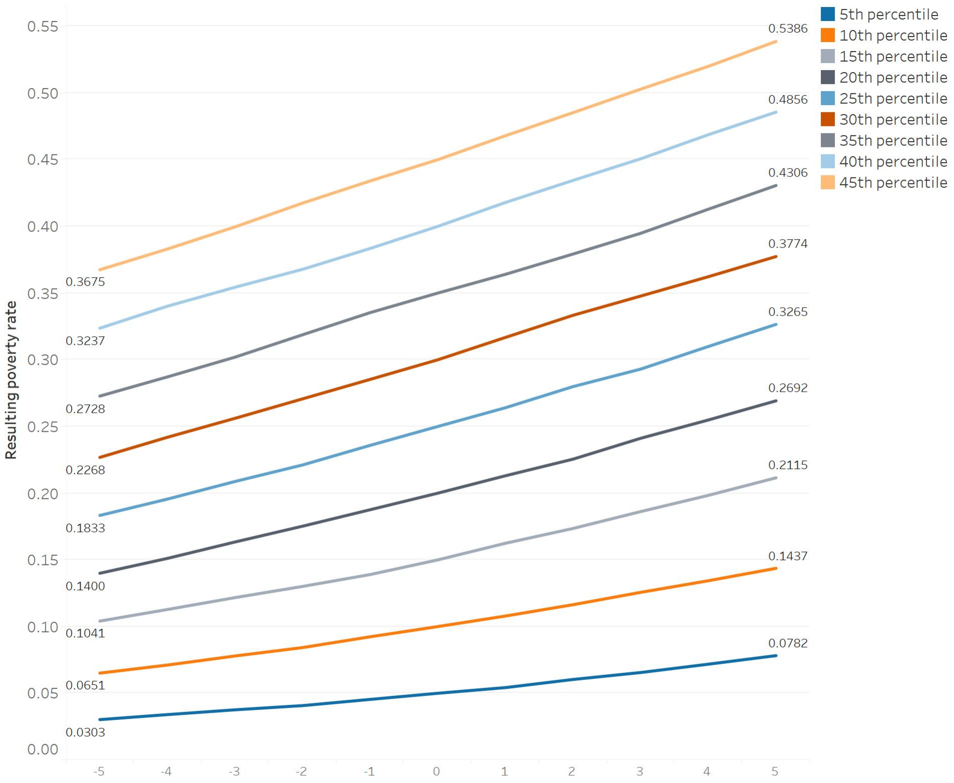
Barriga-Cabanillas, O., Chawla, P., Redaelli, S., & Yoshida, N. (2023). *Updating poverty in afghanistan using the swift-plus methodology* (tech. rep.). The World Bank.

Battese, G. E. (1997). A note on the estimation of cobb-douglas production functions when some explanatory variables have zero values. *Journal of agricultural Economics*, *48*(1-3), 250–252.

Beegle, K., De Weerdt, J., Friedman, J., & Gibson, J. (2012). Methods of household consumption measurement through surveys: Experimental results from tanzania. *Journal of development Economics*, *98*(1), 3–18.

Christiaensen, L., Lanjouw, P., Luoto, J., & Stifel, D. (2012). Small area estimation-based prediction methods to track poverty: Validation and applications. *The Journal of Economic Inequality*, *10*(2), 267–297.

Figure 2: Resulting poverty prediction if constant term changes by $x\%$



Source: Based on simulated data illustrated in Section 4.2. The colors represent percentiles at which poverty thresholds were determined under the source data. Thus, if true welfare shifted neutrally to the right by 5 percent and the imputation model fails to account for that, then poverty would be overestimated.

Crow, E. L., & Shimizu, K. (1987). *Lognormal distributions: Theory and applications*. Marcel Dekker New York.

Dang, H.-A., Jolliffe, D., & Carletto, C. (2017). *Data gaps, data incomparability, and data imputation: A review of poverty measurement methods for data-scarce environments*. The World Bank. https://doi.org/10.1596/1813-9450-8282

Dang, H.-A. H., Kilic, T., Carletto, C., & Abanokova, K. (2021). Poverty imputation in contexts without consumption data: A revisit with further refinements.

Dang, H.-A. H., Lanjouw, P. F., & Serajuddin, U. (2014). *Updating poverty estimates at frequent intervals in the absence of consumption data: Methods and illustration with reference to a middle-income country*. The World Bank. https://doi.org/10.1596/1813-9450-7043

Edochie, I. N., Freije-Rodriguez, S., Lakner, C., Moreno Herrera, L., Newhouse, D. L., Sinha Roy, S., & Yonzan, N. (2022). What do we know about poverty in india in 2017/18?

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

Hentschel, J., Lanjouw, J. O., Lanjouw, P., & Poggi, J. (1998). Combining census and survey data to study spatial dimensions of poverty. *World Bank Policy Research Working Paper*, (1928).

Mahler, D., Yonzan, N., Lakner, C., & Castaneda Aguilar, H., R.A. abd Wu. (2021, June). Updated estimates of the impact of covid-19 on global poverty: Turning the corner on the pandemic in 2021? https://blogs.worldbank.org/opendata/updated-estimates-impact-covid-19-global-poverty-turning-corner-pandemic-2021

Maximum entropy econometrics: Robust estimation with limited data. (1996). Chichester [England] ; New York : Wiley, c1996.

Newhouse, D. L., & Vyas, P. (2019). Estimating poverty in india without expenditure data: A survey-to-survey imputation approach. *World Bank Policy Research Working Paper*, (8878).

Prydz, E. B., Jolliffe, D., & Serajuddin, U. (2022). Disparities in assessments of living standards using national accounts and household surveys. *Review of Income and Wealth*, *68*, S385–S420.

Rao, J., & Molina, I. (2015). *Small area estimation* (2nd). John Wiley & Sons.

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.

Sinha Roy, S., & Van Der Weide, R. (2022). Poverty in india has declined over the last decade but not as much as previously thought.

Somanchi, A. (2021). Missing the poor, big time: A critical assessment of the consumer pyramids household survey.

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

Wittenberg, M. (2010). An introduction to maximum entropy and minimum cross-entropy estimation using stata. *The Stata Journal*, *10*(3), 315–330.

Yoshida, N., Chen, X., Takamatsu, S., Yoshimura, K., Malgioglio, S., & Shivakumaran, S. (2021). The concept and empirical evidence of swift methodology. *unpublished manuscript*.