2019

IARIW-World Bank

Special IARIW-World Bank Conference "New Approaches to Defining and Measuring Poverty in a Growing World" Washington, DC, November 7-8, 2019

Estimating Poverty for Refugee Populations: Can Cross-survey Imputation Methods Substitute for Data Scarcity?

Hai-Anh H. Dang

Paolo Verme

Paper Prepared for the IARIW-World Bank Conference

Washington, DC, November 7-8, 2019

Session 2B: Data Methods for Improved Poverty Measurement

Time: 14:15 – 16:15, September 7

Estimating Poverty for Refugee Populations: Can Cross-survey Imputation Methods Substitute for Data Scarcity?

Hai-Anh H. Dang and Paolo Verme*

September 2019

Abstract

The increasing growth of forced displacement worldwide has led to a stronger interest by policy makers and humanitarian and development organizations in measuring poverty among refugee populations. We offer the first application of cross-survey imputation methods in a refugee context where household consumption data are not readily available. We exploit a rich database consisting of administrative and survey data for the Syrian refugees living in Jordan to offer various validation tests for the accuracy of imputation-based poverty estimates. These estimates are found to either perform better or have smaller standard errors than those based on asset indexes or proxy means testing and are robust to varying poverty lines. Furthermore, we find that accurate poverty estimation requires surprisingly small samples and relatively few variables that are already available in the United Nations High Commissioner for Refugees' global registration system. If these encouraging results are replicated in other refugee contexts, this would open prospects for cost-effective and rapid measurement of poverty among refugees worldwide.

JEL: C15, I32, O15

Keywords: poverty imputation, Syrian refugees, household survey, missing data, Jordan

* Dang (hdang@worldbank.org; corresponding author) is an economist in the Analytics and Tools Unit, Development Data Group, World Bank and is also affiliated with Indiana University and Vietnam's Academy of Social Sciences; Verme (pverme@worldbank.org) is a lead economist and manager of the Research Program on Forced Displacement at the World Bank. We would like to thank Theresa Beltramo, Jose Cuesta, Peter Lanjouw, David Newhouse, Franco Peracchi, Matthew Wai-Poi, Tara Wishvanath, and participants at the WB-UNHCR training course on poverty imputation for helpful comments and discussions on earlier versions. This work is part of the program "Building the Evidence on Protracted Forced Displacement: A Multi-Stakeholder Partnership". The program is funded by UK aid from the United Kingdom's Department for International Development (DFID), it is managed by the World Bank Group (WBG) and was established in partnership with the United Nations High Commissioner for Refugees (UNHCR). The scope of the program is to expand the global knowledge on forced displacement by funding quality research and disseminating results for the use of practitioners and policy makers. We further thank DFID for additional funding support through its Knowledge for Change (KCP) program. This work does not necessarily reflect the views of DFID, the WBG or UNHCR.

I. Introduction

The sharp growth in the global count of forcibly displaced people during the past decade has created new challenges for host governments and aid organizations that will require a new approach to the measurement of poverty.[1] Host governments are keen to know the number and status of refugees living in their countries as they struggle to maintain internal order while assisting the new comers. Humanitarian organizations charged with managing displacement crises are confronted with increasing financial needs and, when these needs are not met by donors, with budget cuts and a shift from universal to means-tested targeting. The increasingly protracted nature of displacement also challenges development organizations to design sustainable poverty reduction programs for displaced people and host communities. For all these actors, measuring poverty among displaced populations has become a key ingredient for any effective economic policies. It also becomes increasingly clear that achieving the SDG (Sustainable Development Goals) number one goal of poverty reduction will not be possible if the forcibly displaced are excluded from the count.

This is not an easy task. Measuring poverty among refugees is more complex than for regular populations because refugees are mobile. They also live in areas often difficult to reach due to environmental or security barriers. Indeed, the global count of the poor excludes, for the most part, displaced populations because these populations are not usually captured by censuses and, as a consequence, are largely excluded from consumption surveys, the main instruments used to measure poverty. The various challenges related to micro survey data collection such as survey administration, sampling, and questionnaire design or funding are exasperated for displaced populations and will require years of efforts to meet the poverty measurement standards that we

[1] The UNHCR estimates that the number of forcibly displaced people at the end of 2018 was 71.4 m, the largest number since the beginning of records in 1951.

are now accustomed to see in (most) low-income countries. Organizations such as the United Nations Commissioner for Refugees (UNHCR) and the World Bank are now fully committed to bridging this data gap but past experiences with measuring poverty in low-income countries suggest that this is going to be a long-term process.[2] In the meantime, the development of various methodologies designed to estimate poverty in contexts where income or consumption data are not available can provide a useful alternative to producing reliable poverty figures for displaced communities.

In this paper, we propose to employ recent advances in cross-survey imputation techniques (Dang, Lanjouw, and Serajuddin, 2017) to estimate poverty for refugees, making use of data typically available to the UNHCR. All individuals seeking protection, assistance and refugee status are expected to register with the host government or the UNHCR and, for this purpose, the UNHCR maintains a profile Global Registration System (proGres). This system contains biometric and socio-economic information on asylum seekers and refugees and serves the purpose of identifying the persons most in need and determine the type of protection and assistance required. ProGres does not offer information on income, consumption or expenditure but contains a rich list of variables that are potentially closely associated with these monetary indicators. In addition, the UNHCR and partner organizations routinely collect information on household consumption by means of sample surveys designed to address specific issues such as measuring food security or determining various types of vulnerabilities.

We combine the UNHCR home visits surveys and the proGres registration database to estimate poverty for all the Syrian refugees in Jordan that are registered in the latter database and live

[2] Over the past five years, these two organizations have sharply increased their cooperation and they recently announced the establishment of a joint data center with the objective of addressing this data challenge.

outside camps.[3] We also discuss further method extensions that are relevant to refugees, such as the minimum required survey sample size, and comparison with alternative techniques commonly used in practice including proxy means test and targeting. To our knowledge, this is the first experiment of its kind. Poverty studies that made use of cross-survey imputation methods have now become more frequent (see, e.g., Dang, Jolliffe, and Carletto, 2019 for a recent review), but none of these works has focused on refugee populations.

Indeed, hardly any studies exist that measure poverty among refugees, perhaps because of data scarcity and because the economics profession has paid very little attention to displacement issues until very recently. For example, the protection mandate given to humanitarian organizations has prevented these organizations from sharing their data for public research purposes. This is now changing and the proGres data set used in this paper was the first provided by the UNHCR to an external organization for the purpose of poverty research.[4] We attempt to bridge this gap with the analysis in this paper.

We find that imputation-based poverty estimates are generally statistically not different from the true poverty rates, and this result is robust to various validation test, including varying poverty lines and disaggregated population groups. These estimates are found to perform better or have smaller standard errors than other poverty measures based on asset indexes or proxy means testing. Moreover, our imputation models are rather parsimonious and use variables that are already available from the UNHCR's proGres database, which are consistent with the findings in recent studies for imputation-based poverty estimates for regular populations. We provide both

[3] Refugees living inside camps are assisted with shelter, cash transfers, education, health and other services. This group was excluded because it is problematic to construct a relevant consumption aggregate. More than 80% of refugees in Jordan lived outside camps at the time of data collection in 2014.

[4] These data were provided by the UNHCR to the World Bank in the context of a joint poverty assessment of Syrian refugees in Jordan and Lebanon conducted between 2014 and 2015 (Verme et al, 2016).

theoretical and empirical evidence that a smaller-sample survey may be fielded for refugees, and data from this survey can be combined with those from the census-type registration system to provide cost-effective and updated estimates of poverty.

The paper consists of five sections. Section II provides the basic theory and analytical framework. Section III provides the country background, a description of the data and the empirical results including robustness tests. Section IV discuss further extensions in other contexts and Section V concludes.

II. Analytical Framework

Where consumption data are either incomparable across two survey rounds or missing in one survey round but not the other, but other characteristics (x_j) that can help predict consumption data are available in both survey rounds, we can apply survey-to-survey imputation methods. In particular, Dang *et al.* (2017) propose an imputation framework that builds on earlier studies (Elbers, Lanjouw, and Lanjouw, 2003; Tarozzi, 2007).[5] We briefly describe this imputation method before discussing its extensions to the refugee context.

Let x_j be a vector of characteristics representing the main observable factors that determine a household's consumption, where j indicates the survey type. More generally, j can indicate either another round of the same household expenditure survey, or a different survey (census), for j= 1,

[5] Elbers *et al.* (2003) provide a method that imputes household consumption from a survey into a population census to measure poverty, which is commonly known as "poverty mapping". Adapting this approach for survey-to-survey imputation, Christiaensen *et al.* (2012) impute poverty estimates using data from several countries, including China, Kenya, the Russian Federation, and Vietnam; other studies analyze data from Uganda (Mathiassen, 2013). Compared to previous studies, Dang *et al.*'s (2017) method provides a more explicit theoretical modeling framework, with new features such as model selection and standardization of surveys of different designs (e.g., for imputing from a household survey into a labor force survey). This technique has recently been applied to data from several African countries (Beegle *et al.*, 2016), India (Dang and Lanjouw, 2018), Tunisia (Cuesta and Ibarra, 2017), and Vietnam (Dang *et al.*, 2019).

2.[6] Subject to data availability, x_j can include household variables such as the household head's age, sex, education, ethnicity, religion, language (i.e., which can represent household tastes), occupation, and household assets or incomes. Occupation-related characteristics can generally include whether the household head works, the share of household members that work, the type of work that household members participate in, as well as context-specific variables such as the share of female household members that participate in the labor force, or some variables at the region level. Other community or regional variables can also be added since these can help control for different labor market conditions.

The following linear model is typically employed in empirical studies to project household consumption on household and other characteristics (x_j)

$$y_j = \beta_j' x_j + \upsilon_{cj} + \varepsilon_j \qquad (1)$$

where υ_{cj} is a cluster random effects, ε_j is the idiosyncratic error term, y_j is household consumption typically modelled in log form. Note that we suppress the subscript that indexes households to make the notation less cluttered.[7] For convenience, we also refer to the survey that we are interested in imputing poverty estimates for as the target survey, and the survey that we can estimate Equation (1) on as the base survey. The former survey is usually more recent (or offers more disaggregated information, as in the case of a census) and has no consumption data, while the latter is usually older and has consumption data.

[6] More generally, j can indicate any type of relevant surveys that collect household data sufficiently relevant for imputation purposes such as labor force surveys or demographic and health surveys.

[7] Conditional on household characteristics, the cluster random effects and the error terms are usually assumed uncorrelated with each other and to follow a normal distribution such that $\upsilon_{cj}|x_j \sim N(0, \sigma_{\upsilon_j}^2)$ and $\varepsilon_j|x_j \sim N(0, \sigma_{\varepsilon_j}^2)$. While the normal distribution assumption results in the standard linear random effects model that is more convenient for mathematical manipulations and computation, it is not necessary for this type of model. As can be seen later, we can remove this assumption and use the empirical distribution of the error terms instead, albeit at the cost of somewhat more computing time.

Assume that the explanatory variables x_j are comparable for both surveys (Assumption 1), and that the changes in x_j between the two periods can capture the change in poverty rate in the next period (Assumption 2). Dang *et al.* (2017) define the imputed consumption y_2^1 as

$$y_2^1 = \beta_1' x_2 + \upsilon_1 + \varepsilon_1 \quad (2)$$

and estimate it as

$$\hat{y}_{2,s}^1 = \hat{\beta}_1' x_2 + \tilde{\hat{\upsilon}}_{1,s} + \tilde{\hat{\varepsilon}}_{1,s} \quad (3)$$

where the parameters β_1' are estimated, and $\tilde{\hat{\upsilon}}_{1,s}$ and $\tilde{\hat{\varepsilon}}_{1,s}$ represent the s^{th} random draw from their estimated distributions using Equation (1), for $s = 1, \ldots, S$. Using the same notation as in Equation (3), the poverty rate P_2 in survey (or period) 2 and its variance can then be estimated as

$$\text{i) } \hat{P}_2 = \frac{1}{S}\sum_{s=1}^{S} P(\hat{y}_{2,s}^1 \leq z_1) \quad (4)$$

$$\text{ii) } V(\hat{P}_2) = \frac{1}{S}\sum_{s=1}^{S} V(\hat{P}_{2,s}|x_2) + V(\frac{1}{S}\sum_{s=1}^{S} \hat{P}_{2,s}|x_2) \quad (5)$$

It is important to check on both Assumption 1 and Assumption 2 before efforts are made to obtain imputation-based poverty estimates for refugees. In particular, Assumption 1 is testable; we can implement a t-test for equality of the means for the same variables in the two surveys. More sophisticated tests for comparing the two distributions (e.g., the non-parametric Kolmogorov-Smirnov) may be used if necessary. Assumption 2, on the other hand, is not testable. But if data are available for (at least) two previous survey rounds, we can obtain some indirect supportive evidence for Assumption 2 using a decomposition test.

Yet, it should be noted that Assumption 1 may not hold in the context for refugees if any survey round on refugees is non-random and is not representative of the whole refugee population. For example, administrative agencies (or NGOs) may oftentimes conduct surveys just on a couple refugee camps to provide a rapid assessment of refugees' welfare situation. Assumption 2 may be violated if there are policy changes that result in structural change in the relationship between

household consumption and refugee characteristics in Equation (1). Such changes can occur if, say, refugees are allowed to work in certain professions which are different from their previous professions, or they can receive their work permits more easily (or at lower costs).[8]

This survey-to-survey imputation method can be employed to impute poverty either contemporaneously (i.e., from a smaller survey into a census) or track poverty trends over time. In this paper, we focus on the first objective (and so do not necessarily need to use Assumption 2). The UNHCR typically maintains a census-type administrative database on refugees which is updated at regular intervals, usually every six months, or when refugees make contact with the UNHCR. At the same time, more detailed data on refugees are also collected in smaller household consumption surveys. These two sources of data can be combined to provide poverty estimates for refugees in an efficient and cost-effective manner. We build the model specifications in Equation (1) on those used in Verme *et al.* (2016), and add the district random effects. The list of explanatory variables are described in detail in Section III.2. We discuss these data sources for the Syrian refugees in Jordan in the next section.

III. Application to Syrian Refugees in Jordan

III.1. Country Background and Data

The Syrian refugee crisis is one of the largest refugee crises ever recorded in history if we consider the numbers of displaced people relatively to the country of origin and the countries of

[8] Prohibitive survey costs, particularly in conflict and violence situations, may pose another challenge to the implementation of a fully representative survey; this in turn can violate Assumption 1. Another potential concern with Assumption 2 is that refugees derive their main source of income from transfers and subsidies, rather than from work as with most of the regular population. Still, if their consumption is well captured by Equation (1) (as indicated by a high R^2 value), we can still track the change in poverty for refugees using the changes in the relevant variables. For example, the amount of cash transfer to a refugee household typically depends on the number of people in the household, thus a refugee's household size can determine their total income (or consumption). Put differently, the changes in the size of the refugee's household should be strongly correlated with the changes of their income.

destination. The crisis started in the spring of 2011 following clashes between protestors and Government forces in several major cities and quickly descended into a complex civil war. By 2014, 6.7 m people had been displaced internally in the country, about 1.5 m people fled the country with their own means and an additional 3.7 m people were hosted as refugees mostly in neighboring countries. As a result, about half of the Syrian population was considered displaced in 2014. For some countries, Syrian refugees also represented a major population shock. In 2014, Syrian refugees accounted for about 20% of the population of Lebanon and about 10% of the population in Jordan. The incidence of such immigration for these countries is among the highest ever recorded in history (Verme and Schuettler, 2019).

The UNHCR has the mandate to protect and assist refugees in host countries and its role in the aftermath of a crisis is to find shelter, provide food and cash assistance and assist with basic services such as health and education. In order to provide these services, the UNHCR employs a system of mandatory registration for all refugees or asylum seekers requiring assistance that implies the collection of personal biometric and socio-economic information. This proGres registration system is the most comprehensive database on refugees in any country where the UNHCR manages the registration of refugees.[9] This is the case of Jordan, the country we consider in this paper.

In addition to the registration system, the UNHCR conducts sample surveys and home visits for a variety of purposes such as protection of different categories of vulnerable populations or assistance of targeted programs such as the cash or food assistance program. In the case of Jordan and the Syrian crisis, the UNHCR and the World Food Program (WFP) have been conducting a

[9] In some countries such as Turkey, the host government or other agencies manage the registration process.

variety of surveys as well as extensive home visits that allowed researchers to analyze refugee conditions as it had never been done before.

The paper uses two data sets: the Jordan proGres registration system (PG for short) as of December 2014 and the Jordan Home Visits survey, round II data (HV for short) collected between November 2013 and September 2014. Both data sets were provided by the UNHCR in the context of the joint World Bank-UNHCR study on the welfare of Syrian refugees (Verme *et al.*, 2016). These comprehensive data sets have the distinct advantage that they can be linked by a common identification number. We can therefore trace the same individuals and households across the two sources of data.

The proGres registration system is what we consider the "census" of refugees. This data set has no information on consumption but contains socio-economic characteristics for all registered individuals and households. Variables available in the PG data include, among others, date of birth, place of birth, gender, date and reasons of flight, arrival date in Jordan, registration date, ethnicity, religion, education, professional skills, and occupations in the countries of origin and asylum.

The HV data have been collected in successive rounds since 2013 for the purpose of targeting refugees with cash assistance programs and they contain information on income and expenditure as well as a large set of individual and household socio-economic characteristics. Although this is not a sample survey, for the purpose of this study we will consider this data set as our hypothetical sample survey. The HV data we use cover about one third of all registered persons in Jordan in 2014 and are therefore a sub-sample of the PG data.[10]

[10] Verme *et al.* (2016) used a t-test for partly overlapping groups to test for differences in covariates between the HV and PG data and found only 5 out of 22 covariates to be statistically similar. This is because the PG data rely on shorter and quicker questionnaires and are somewhat more outdated compared with the HV data. However, we offer analysis that uses these covariates from each data set alone in the empirical analysis in Section III.

As unit of observation, we use what the UNHCR refers to as the "case". A case is a group of individuals who register at the UNHCR together with a principal applicant (PA) who takes responsibility for the group. This group may be a family, a household or an extended household. For simplicity and practical purposes, we will consider a case and the PA as a household and its head respectively. The poverty line used is 50 JD/month/person, which is what the UNHCR used in 2014 to select beneficiaries of the cash assistance program. In 2014, this poverty line was higher than the international poverty line and lower than the poverty line used for the Jordanian population. In our case, this poverty line is more relevant than either the national or international poverty line as it corresponds to what the UNHCR—the UN agency specialized on refugees—considers a sufficient amount to meet basic needs.

As for the welfare aggregate, we use a combination of two retrospective questions on expenditure consisting of 16 consumption items and based on a one month recall period. With an experiment on regular populations in Tanzania, Beegle *et al.* (2012) found that a long recall period and a small number of consumption items lead to underreporting. While there is no prior evidence to suggest that these same findings similarly apply to refugee populations, our estimated mean consumption may be lower than the true figure. Yet, this bias is likely smaller for refugees than for regular populations because refugees consume a much more restricted number of items due to their restricted conditions and resources (including limited labor market opportunities and mobility). Verme *et al.* (2016) provides a full discussion of these issues together with details related to the poverty line and the consumption aggregate as well as the statistical tests used to validate the aggregate.

III.2. Estimation Results

For the purpose of this paper, the HV data are considered the "survey" data containing information on consumption and the PG registration data are our "census" data containing predictors of consumption but no consumption data. The primary objective of the exercise is, therefore, to test how accurate are poverty figures estimated using a model built with the HV data using the PG data only.

As a first step, we generated two samples by extracting 50% of observations from the whole HV sample randomly (Sample 1) and using the remaining observations as second sample (Sample 2). We then impute from Sample 1 to Sample 2 to obtain the imputation-based poverty rate in Sample 2, and we compare this imputed poverty rate with the "true" poverty rate that can be directly calculated from Sample 2 for robustness checks. We also implement this imputation process the other way around by imputing from Sample 2 to Sample 1 and then compare with the true poverty rate in Sample 1.

We also consider three model specifications based on different sets of regressors for further comparison. Specification 1 employs the variables that are only available in the PG dataset (HV-specific variables), which include household size and the PA's demographic and employment characteristics (age, gender, the highest education achievement, occupation group, marital status, religion, and the governorate or city of original residence in Syria). Specification 1 also includes variables related to the PA's immigration status such as the type of border crossing point and the legal status of entry. It is the main model specification. Specification 2 adds to Specification 1 several variables that are only available in the HV data and that are related to household assets, utilities, and the physical characteristics of the house. These variables include the status of the kitchen, electricity, ventilation system, house size, whether the house is made of concrete, and the availability of tap water and piped sewerage system. Specification 3 further adds to Specification

2 HV-specific variables related to the household's shock-coping strategies (i.e., whether receiving humanitarian assistance, help from the host family, or from the host community), whether the household has a valid certificate of asylum, and whether the household receives UNHCR financial assistance.

We are particularly interested in examining whether adding HV-specific variables to the main specification in Specification 1 can improve the accuracy of the estimates. If we find that some key predictors of household expenditure—that are not available in the PG data—can improve the accuracy of the poverty predictions significantly, this provides a strong argument for collecting this information upfront when refugees are first registered. Vice-versa, if poverty estimates imputed with the PG data are not statistically different from those produced with HV data, this would suggest that existing PG variables are already suitable to produce reliable poverty estimates.

We also use two alternative models to estimate regression errors: one where we assume a standard normal distribution for the error term, and another where we remove this assumption and use the (non-parametric) empirical distribution of the error terms instead. If the error term is not distributed normally, our poverty estimates would be biased, and a non-parametric model based on the empirical distribution would likely perform better.

Table 1 present the summary results and Table 2.1 in Annex provides the full regression results. Table 1 shows that all the estimates using the normal linear regression model fall within the 95 percent confidence interval (CI) of the true poverty rate, for both Sample 1 and Sample 2. In other words, these estimates are not statistically significantly different from the true poverty rates reported at the bottom of the table.[11] Estimates using Specification 2 with more variables on household assets and house characteristics are somewhat better and closer to the true poverty rate

[11] The standard errors around the true poverty estimates are larger than those for the imputation-based estimates since the latter are model-based. See Dang *et al.* (2019) for more discussion.

than those using Specification 1 for both samples. For example, the poverty estimate using Specification 1 (Table 1, first column) is 52.6 percent, which is 0.8 percentage points larger than the true poverty estimate of 51.8 percent. The poverty estimate using Specification 2 (Table 1, second column) is 51.3 percent, which is 0.5 percentage points less than the true poverty estimate. The improvement is even more noticeable for Sample 2, where the difference from the true poverty estimate for Specification 1 is 2.2 percent, twice as large as that of Specification 2 (0.8 percent).[12] Yet, since the standard error around the true poverty rate is 2.3 percent for Sample 1 and 2.6 percent for Sample 2, all these differences are in fact still within one standard error of the true poverty estimates. As such, statistically speaking, the differences between Specification 1 and Specification 2 and the true poverty rates are negligible.

Notably, Specification 2 performs slightly better than Specification 3, which has more control variables. While this result may appear counter-intuitive, one possible reason is that Specification 3 may overfit the data and thus does not offer more accuracy than Specification 2. This concurs with evidence from other studies for India and Jordan, where adding too many variables to the imputation model is not found to improve estimates (Dang *et al.*, 2017; Dang and Lanjouw, 2018).[13] In any case, this difference in performance is not statistically significant, as similarly seen with Specification 1.

The alternative imputation model based on the empirical distribution of the error terms (Table 1, row 2) performs worse than those based on the normal linear regression, although both methods provide estimates within the 95 percent CI of the true poverty rates. Finally, since the HV dataset is originally a non-random subsample of the PG database, we also re-run Table 1 using only the

[12] Imputation models that include household assets are usually found to perform better than those that do not. See, e.g., Christiaensen *et al.* (2012) and Dang *et al.* (2019).

[13] A recent study also suggests that for misspecified regressions, adding more variables may result in larger inconsistency (De Luca, Magnus, and Peracchi, 2018).

variables that are available in the HV dataset. Results are shown in Table 2.2 in Appendix 2. Poverty estimates are underestimated as compared to Table 1 but they are qualitatively similar when compared to the true poverty values.

In summary, the set of variables available in the PG registration data seems sufficiently powerful to predict the true poverty rate with a 95% accuracy level. This is very encouraging considering that these variables were not selected for this purpose when the registration system was designed.

III.3. Robustness Checks and Extensions

This section provides some simple robustness tests for the results presented on the Jordan case in Table 1. We test robustness to different sample sizes, changes in the poverty line, more disaggregated population groups, and alternative estimation methods.

Sensitivity to the poverty line

One important question relates to the performance of the model specifications when the poverty line and the poverty level change. With the poverty rate close to 50%, we have half of the sample below and half above the poverty line. But estimating poverty accurately when the poverty rate is around 5-10 percent may be more difficult. In Figure 1, we used variations of the poverty line ranging from 0 to 60 percent of the population (i.e., 0 to 60^{th} percentile of the consumption distribution) to reproduce poverty estimates using imputations from Sample 1 to Sample 2 and the two error models described. Results show that with a low poverty line and a low poverty rate the empirical errors model is more accurate in estimating true poverty than the normal linear model, whereas this is reversed when the poverty line and the poverty rate are high. Both methods result in predictions that are within the 95% CI of the true values but these two methods clearly differ in

accuracy as the poverty line and the poverty rate change. Estimation results are similar if we impute from Sample 2 to Sample 1 (Figure 2.1). A possible explanation is that, as the number of poor households (sample size) increases, the distribution of the error term approaches a normal distribution. Therefore, as a rule of thumb, we should expect the normal linear model to perform well with larger samples.

Disaggregated population groups

The next question is whether results are sensitive to changes in the specified population groups. We know from our regressions that the most important predictor of poverty is case size (see also Verme *et al.*, 2016). If the prediction capacity of the model specification is sensitive to changes in household characteristics, changing case size would likely have the most impact. In Figure 2 we impute from Sample 1 to Sample 2 and re-estimate poverty using the two error estimation models and plot the estimated poverty rates against case size. Both methods provide similar results and both sets of results are within 95% CI of the true values. In this case, we do not observe any sharp difference between the two error estimation models. As before, we repeat the exercise imputing from Sample 2 to Sample 1 (Figure 2.2) and find that results are virtually unchanged. As such, the performance of the two error models is related to the case size rather than population groups. Moreover, given the association between case size and poverty, we should also exclude that the poverty level is associated with a better or worse performance of one of the error estimation models.

Models with a stronger parametric assumption

One alternative approach to the present poverty estimation models is to run a probit or logit model on poverty status rather than a linear model on expenditure. In this case, the population is first divided into poor and non-poor groups using the poverty line and this variable is then used as dependent variable in a logit or probit model to predict poverty. The difference with a probit (or logit) model is that we need to make a stronger parametric modelling assumption on the dependent variable, which can results in more accurate estimation results if this assumption is correct. But the disadvantage with such models is that estimation results may be worse if the modelling assumption is violated. Furthermore, the conversion of the continuous expenditure variable into a binary variable indicating poverty status can result in loss of information and generally less efficient estimation (Ravallion, 1996). Indeed, Table 2.3 in Appendix 2 shows that results are less accurate than those provided with the consumption models and, for this reason, we do not consider this model further.[14]

IV. Methodological Challenges in Other Contexts

The data on Syrian refugees in Jordan that we analyze are of relatively high quality in the context of refugee populations. In this section, we discuss methodological challenges in other contexts where data quality may not be as good.

IV.1. Small Survey Sample Sizes

[14] A promising direction to improve these estimates is by shifting the probability threshold with the objective of minimizing the leakage rate while maximizing the coverage rate. This methodology builds on a popular method (ROC curves) used by epidemiologists to estimate false negatives and false positives in clinical trials, it has been applied before in the context of targeting humanitarian programs and it was tested with the same data we use in this paper (see, e.g., Verme and Gigliarano, 2019).

One practically relevant question is how large the imputation sample should be to obtain accurate poverty estimates?[15] On the one hand, a large sample size can provide estimates with more accuracy and generally better statistical properties than a small sample size; but on the other hand, it is also more expensive and demands more logistical and technical resources to implement. A balance thus should be reached between these tradeoffs. In most conflict situations, however, the logistical and technical constraints may pose especially severe challenges for data collection efforts.

Park and Dudycha (1974) offer some theoretical guidance on selecting the appropriate sample size for obtaining regression-based prediction estimates. In particular, we want to find the sample size *n* such that

$$\Pr[(\rho^2 - \rho_c^2) \leq \varepsilon] = \gamma \qquad (7)$$

where ρ^2 is the maximum (or true) multiple correlation coefficient (R^2) possible for Equation (1) in the population, and ρ_c^2 is the correlation between the predicted value using Equation (1) and the original *y* variable. ρ_c^2 is usually referred to as the squared cross-validity correlation coefficient.[16] A good sample size would ensure that the probability of obtaining an estimate within an acceptable error interval (ε) around ρ^2 has reasonably good power (γ). In other words, after we specify some (acceptable) values for ε and γ, the sample size *n* that satisfies Equation (7) can be derived as follows

$$n = \left[\delta^2 \frac{1-\rho^2}{\rho^2}\right] + p + 2 \qquad (8)$$

[15] Note that this challenge of finding an appropriate sample size is in the context of predicted values based on regression models, which is different from calculating the sample sizes for other purposes such as hypothesis testing. For the latter, see, e.g., Cohen (1998) for a textbook treatment.

[16] The intuition is that, since the best job that we can do with prediction is to reproduce the original *y* variable, the correlation between the original *y* variable and its predicted value should always be less than or equal to the true correlation in the population.

where δ^2 is the noncentrality parameter for the noncentral Student's *t* distribution with *p-1* degrees of freedom associated with Equation (7), and p is the number of predictors (i.e., explanatory variables) in the estimation model. We provide a more detailed description of Park and Dudycha (1974)'s derivations in Appendix 1.

We apply Equations (7) and (8) above and calculate the sample sizes where ε ranges from 0.01 to 0.05, and γ ranges from 0.90 to 0.99.[17] These ranges should cover most of the cases of interest, with a smaller value for ε and a larger value for γ requiring a larger sample size. In particular, the smallest sample size given these values would be where ε and γ are respectively 0.05 and 0.90, or the probability that ρ_c^2 falls within a bandwidth of 0.05 around the true value of ρ^2 is 0.90. Increasing this probability to, say, 0.95 and tightening ε to 0.02 would require a larger sample size. We also assume that ρ^2 is 0.38 and the number of predictors p is 11, which are the parameters obtained under Specification 1 for both samples in Table 1. Estimates provided in Table 2 suggest that the minimum sample size is 196 observations (where ε and γ are respectively 0.05 and 0.90), and a reasonably good sample size is 989 observations (where ε and γ are respectively 0.01 and 0.90). Table 2 also indicates that the largest sample size required to increase γ to its maximal value of 0.99 and reduce ε to its minimal value of 0.01 is 1,437 observations.

Yet, while Park and Dudycha's formulae provide useful theoretical guidance on the appropriate sample size, these formulae were originally developed for the simple OLS model. As such, their model does not explicitly take into account our cluster random effects. Thus it remains an empirical question whether these formulae can apply to our context.

We address this question and show estimation results in Figure 3. Estimations in this figure are restricted to Sample 2 from which 10 sub-samples of different sizes—including 200, 400, 600,

[17] Pituch and Stevens (2016) consider 0.05 (or smaller) and 0.90 (or larger) are respectively good values for ε and γ.

800, 1000, 1500, 2000, 3000, 4000, and 5000 observations—have been extracted randomly. The first five samples represent situations ranging from the theoretical minimum sample size (200) to less than the theoretically ideal sample (1,000), and the last first five samples represent situations ranging from the theoretically ideal sample (1,500) to a common and reasonably good sample size in practice (5,000). Specification 1 is then re-run on each sub-sample, the underlying regression results are provided in Appendix 2, Table 2.4.

Results show that almost all the poverty estimates fall within one standard error of the true poverty rate, and that there appears no strong relationship between the number of observations and the accuracy of results.[18] Yet, plotting all estimation results with the linear and empirical models in Figure 3 yields two additional observations. The first is that estimates fluctuate less around a sample of 1,000 observations with both estimation methods, and the second is that the normal linear model tends to overestimate the true value more than the empirical errors model.[19] We can also observe from Table 2.4 that the estimated R^2 of the model specifications tends to decline and also stabilize as the number of observations increases, which is consistent with the well-known statistical result that these estimates for R^2 in smaller samples may be larger than their population counterparts (see, e.g., Pituch and Stevens, 2016). In essence, good estimates can also be obtained with very small samples but samples of medium size, around 1,000 observations in our case, seem to be the best strategy to obtain consistent and more stable estimates while containing survey costs. This sample size is also consistent with the theoretical results offered in Park and Dudycha (1974).

These results have practical relevance. The HV data used in this study were collected with field visits that covered about 5,000 households per month, or 60,000 households per year. We have

[18] All estimates fall within the 95 percent CI of the true poverty rate but are not shown for lack of space.
[19] Note that we are only considering a single summary statistics for the whole population (the poverty rate). If we were to estimate disaggregated statistics by geographical areas or population groups for example, sample sizes would have to be reconsidered.

shown that covering about one-sixtieth of this number, or 1,000 household per year, would be sufficient to provide reliable poverty statistics.[20]

IV.2. Related Measures of Poverty

Could we produce similar poverty estimates by using alternative estimation methods such as asset (wealth) indexes and proxy-means tests? We examine each of these two alternatives, together with the related exercise of targeting, in this section. This is a particular important question for the UNHCR which uses assets indexes to measure well-being in place of consumption in many places where consumption is not available. Other development organizations such as the WFP also often employs asset indexes to target food assistance programs for refugees; one such recent application was for the Malian refugees in Niger (Beltramo et al., 2019).

Asset index

Again, suppressing the subscript that indexes households to make the notation less cluttered, we consider a variant of Equation (1) where the left-hand side variable, household consumption y_j, is now missing. But we have data on household assets a_j, which is a subset of x_j. Still, we want to generate a wealth index w_j which offers the best combination of (the elements of the different) household assets a_j. This can then be expressed as follows

$$\alpha' a_j = w_j \qquad (9)$$

[20] This result should not be interpreted as suggesting that 1,000 observations are sufficient for a multi-purpose survey. In our case, we estimate this number to be sufficient to estimate one statistic (the poverty rate) whereas most surveys have typically multiple objectives and require the correct estimation of multiple statistics. The latter are the reasons behind common tasks associated with designing a survey such as power calculations, stratification, and clustering of the sample.

where α are the (vector of) weights we place on the a_j to generate the wealth index w_j. A common way to derive α is through Principal Component Analysis (PCA), another way is just to sum up all the assets available in a_j.

We briefly describe here a couple reasons that asset indexes are likely to result in biased estimates of poverty. First, the wealth index w_j does not include the non-asset component, which is equivalent to the well-known issue of omitted variable bias. Second, β_1 and α are generally different from each other, since the estimator for α maximizes the variance in a_j, while the estimator for β maximizes the variance in y_j.[21] Finally, in the refugee context, the temporary nature of displacement likely affects refugees' behaviors in terms of accumulating and using assets. For example, refugees may choose not to invest as much in high-quality durables as a regular household does. This practical aspect may further make assets (alone) an even less reliable data source with poverty estimation for refugees.

Table 3 provides an illustrative example where we generate the wealth (assets) index using both the simple counting method (Table 3, Specification 1) and the PCA method (Table 3, Specifications 2 and 3) on the two samples. Each cell in the first five rows shows the proportion of each quintile of the consumption distribution that is correctly captured by each quintile of the wealth index. In other words, the five quintiles provide five different slices of the consumption distribution. The list of assets for Specification 1 and Specification 2 include the status of the kitchen, electricity, ventilation system, whether the house is made of concrete, and the availability of tap water and piped sewerage system. Specification 3 adds to Specification 1 the house size and the condition of household furniture.

[21] See Rencher (2002, pp. 389) for a graphical illustration of the general difference between principal component analysis and OLS methods, and Dang *et al.* (2019) and Dang (forthcoming) for further discussion on asset indexes.

Consistent with our earlier discussion, the quintiles based on the wealth index can only capture between 13 percent and 33 percent of the corresponding quintile based on the consumption distribution. For example, the poorest wealth index quintile in Specification 3 can correctly capture only 32 percent (35 percent) of the poorest consumption quintile in Sample 1 (Sample 2). The correlation between asset indexes and household consumption is slightly higher for the PCA wealth index than for the simple aggregation method (e.g., this correlation is 0.20 for Specification 1 but 0.22 for Specifications 2 and 3 using Sample 1). In fact, these correlation coefficients between the wealth indexes and consumption are in fact even weaker than those observed in Filmer and Scott (2012) for 11 countries around the world (which range from 0.39 to 0.72). This provide supportive evidence for our earlier discussion that asset indexes may perform even worse as measures of household consumption and poverty in the case of refugees.

Proxy means test

Most of the estimates based on proxy means testing are usually estimated as

$$y_j^p = \beta_j^{p\prime} x_{j,p} \qquad (10)$$

where the vector of coefficients β_j^p is often obtained from the regression using another survey (see, e.g., Coady *et al.*, 2014; Ravallion, 2016; Brown, Ravallion, and van de Walle, 2018). As such, proxy mean tests are rather similar to the poverty imputation model expressed in Equation (1) in terms of the deterministic part ($\beta_j^{p\prime} x_{j,p}$). Yet, one key difference between the two methods is that, the error terms $v_{cj} + \varepsilon_j$ in Equation (1) are often omitted in Equation (10). Consequently, the mean and the variance of the predicted consumption based on proxy means testing would likely provide biased estimates of those of household consumption. But when $x_{j,p}$ is identical to x_j—or

when the error terms $(\upsilon_{cj} + \varepsilon_j)$ are negligible—there is no bias in the estimated mean consumption, but there is still bias in the estimated variance.[22]

Table 4 provides poverty estimates using the proxy means test method as in Equation (10). A couple remarks are in order for the results. First, estimates are within the 95 percent CI of the true poverty rate for both samples, which suggests that the error terms $\upsilon_{cj} + \varepsilon_j$ in Equation (1) are negligible. Indeed, the share of the variance of district random effect υ_{cj} out of the total variance of the error term is just 0.01 for Specification 1 (i.e., the estimate for ρ_υ^2 in Appendix 2, Table 2.1).[23] But on the other hand, consistent with our theoretical discussion above, the standard errors for the poverty estimates in Table 4 range from 1.8 to 2.5 percent, which are about twice those based on the poverty imputation methods shown in Table 3.

Targeting

Another useful extension of poverty imputation methods is targeting, whereby we can examine the percentage of the poor population that are correctly identified (i.e., coverage rate) versus the percentage of the population identified as poor who is not poor (i.e., leakage rate).

Estimates based on the empirical errors model, shown in Table 5, suggest that Specification 1 can provide a reasonable coverage rate of 70 percent but a relatively high leakage rate of 43 percent. As we add more control variables to this specification, these rates unsurprisingly improve. In particular, the coverage rate increases by 4 percent, while the leakage rate decreases by 6 percentage points when we switch from Specification 1 to the richer Specification 3. These rates compare favorably with recent estimates of the coverage rate and leakage rate of 64 percent and

[22] Dang *et al.* (2019) offer more detailed discussion and more formal proofs of these results.
[23] Notably, ρ_υ^2 is also estimated to be 0 for Models 2 and 3.

31 percent, using proxy-means test for a similar poverty rate of 40 percent for nine African countries (Brown *et al.*, 2018).

V. Conclusion

We provide a first application of survey imputation methods to obtain poverty estimates for the Syrian refugees living in Jordan. Imputation-based poverty estimates are generally statistically not different from the true poverty rates, and this result is robust to various validation tests. These estimates are found to perform better or have smaller standard errors than other poverty measures based on asset indexes or proxy means testing. Furthermore, our imputation models are rather parsimonious and use variables that are already available from the UNHCR's global registration system. These encouraging results are consistent with the findings in recent studies for imputation-based poverty estimates for regular populations.

Estimation results also point to further research on an alternative and promising method of obtaining poverty estimates for refugees where it is expensive or logistically challenging to implement a large-scale survey. We provide both theoretical and empirical evidence for Jordan that a smaller-sample survey may be fielded for refugees, and data from this survey can be combined with those from the census-type registration system to provide cost-effective and updated estimates of poverty. While these results are encouraging, they are not definitive and need to be replicated in other contexts, possibly using surveys that have a more detailed consumption module. If further validated in other contexts, these findings can potentially lead to significant reductions in data collection costs in the context of refugee operations.

References

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

Beegle, Kathleen, De Weerdt, Joachim, Friedman, Jed and John Gibson. (2012) *Journal of Development Economics*, 98(1):13-18.

Beltramo, Theresa, Christina Wieser, Chiara Gigliariano and Robert Heyn. (2019). "Identifying Poor Refugees for Targeting of Food and Multi-Sectoral Cash Transfers in Niger". *mimeo*.

Brown, Caitlin, Martin Ravallion, and Dominique van de Walle. (2018). "A poor means test? Econometric targeting in Africa." *Journal of Development Economics*, 134:109-124.

Christiaensen, Luc, Peter Lanjouw, Jill Luoto, and David Stifel. (2012). "Small Area Estimation-based Prediction Models to Track Poverty: Validation and Applications." *Journal of Economic Inequality,* 10(2): 267-297.

Coady, David, Margaret Grosh, and John Hoddinott. (2014). "Targeting Outcomes Redux". *World Bank Research Observer*, 19:61–85.

Cohen, Jacob. (1988). *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Erlbaum: Hillsdale, NJ.

Cuesta, Jose, and Gabriel Lara Ibarra. (2017). "Comparing Cross-Survey Micro Imputation and Macro Projection Techniques: Poverty in Post Revolution Tunisia." *Journal of Income Distribution,* 25(1): 1-30.

Dang, Hai-Anh. (forthcoming). "To Impute or Not to Impute, and How? A Review of Poverty Estimation Methods in the Absence of Consumption Data". *Development Policy Review*.

Dang, Hai-Anh and Peter Lanjouw. (2018). "Poverty and Vulnerability Dynamics for India during 2004-2012: Insights from Longitudinal Analysis Using Synthetic Panel Data". *Economic Development and Cultural Change,* 67(1): 131-170.

Dang, Hai-Anh, Peter Lanjouw, Umar Serajuddin. (2017). "Updating Poverty Estimates at Frequent Intervals in the Absence of Consumption Data: Methods and Illustration with Reference to a Middle-Income Country." *Oxford Economic Papers,* 69(4): 939-962.

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

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

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

Filmer, Deon and Kinnon Scott. (2012). “Assessing Asset Indices.” *Demography*, 49 (1): 359–92.

Mathiassen, Astrid. (2013). “Testing Prediction Performance of Poverty Models: Empirical Evidence from Uganda”. *Review of Income and Wealth* 59, no. 1:91–112.

Park, Colin N. and Arthur L. Dudycha. (1974). "A cross-validation approach to sample size determination for regression models." *Journal of the American Statistical Association*, 69(345): 214-218.

Pituch, Keenan A. and James P. Stevens. (2016). *Applied Multivariate Statistics for the Social Sciences: Analyses with SAS and IBM’s SPSS*. Routledge: New York.

Ravallion, Martin. (1996). "Issues in Measuring and Modelling Poverty." *Economic Journal*, 106(438): 1328-1343.

---. (2016). *The Economics of Poverty: History, Measurement, and Policy*. New York: Oxford University Press.

Rencher, Alvin C. (2002). *Methods of Multivariate Analysis*. USA: John Wiley & Sons.

Tarozzi, Alessandro. (2007). “Calculating Comparable Statistics from Incomparable Surveys, With an Application to Poverty in India”. *Journal of Business and Economic Statistics* 25, no. 3:314-336.

Verme, Paolo, and Chiara Gigliarano. (2019). "Optimal targeting under budget constraints in a humanitarian context." *World Development*, 119: 224-233.

Verme, Paolo, Chiara Gigliarano, Christina Wieser, Kerren Hedlund, Marc Petzoldt, and Marco Santacroce. (2016). *The welfare of Syrian refugees: evidence from Jordan and Lebanon*. World Bank: Washington, DC.

Verme, Paolo and Kirsten Schuettler. (2019) The Impact of Forced Displacement on Host Communities: A Review of the Empirical Literature in Economics, *Household in Conflict Network Working Paper* No. 302

Table 1. Predicted Poverty Rates for Syrian Refugees Based on Imputation, ProGres and HV Data 2014 (percentage)

Method	Sample 1			Sample 2		
	Spec. 1	**Spec. 2**	**Spec. 3**	**Spec. 1**	**Spec. 2**	**Spec. 3**
1) Normal linear regression model	52.6	51.3	50.5	53.7	52.3	52.9
	(1.2)	(1.4)	(1.7)	(1.2)	(1.4)	(1.7)
2) Empirical errors model	48.5	48.5	48.7	48.6	48.5	48.9
	(1.2)	(1.5)	(1.8)	(1.4)	(1.5)	(1.9)
Control variables						
Demographics & employment	Y	Y	Y	Y	Y	Y
Household assets & house characteristics	N	Y	Y	N	Y	Y
Shock-coping strategies & receiving UNHCR assistance	N	N	Y	N	N	Y
Overall R2	0.38	0.43	0.48	0.38	0.43	0.48
N	19028	19028	19028	19028	19028	19028
True poverty rate		51.8			51.5	
		(2.3)			(2.6)	

Note: The full regression results are provided in Table 2.1, Appendix 2. Specification 1 employs variables from the ProGres database only, and Specifications 2 and 3 employs variables from both the ProGres and HV databases. The estimation sample is generated by splitting the data into two random samples named Sample 1 and Sample 2. The imputed poverty rate for Sample 1 and Sample 2 are shown in the first and second three columns respectively. The true poverty rate for each sample is shown at the bottom of the table. Robust standard errors in parentheses are clustered at the district level. We use 1,000 simulations for each model run.

Table 2. Theoretical Sample Size as a Function of the Population Parameters

ε	γ		
	0.99	0.95	0.90
0.01	1437	1133	989
0.02	718	566	494
0.03	478	376	328
0.04	358	282	246
0.05	285	225	196

Note: Estimates are based on the formulae provided in Park and Dudycha (1974). We use the given parameters, the ρ^2 value of 0.38 and the number of predictors of 11 under Specification 1 from Table 1.

Table 3. Population Distribution by Asset Indexes vs. Consumption

Per capita consumption	2012			2014		
	Spec. 1	**Spec. 2**	**Spec. 3**	**Spec. 1**	**Spec. 2**	**Spec. 3**
Poorest quintile	33.1	32.3	32.1	34.6	34.0	34.6
Quintile 2	25.3	25.8	20.9	23.7	23.9	19.4
Quintile 3	29.1	24.6	20.8	29.7	25.7	21.6
Quintile 4	12.9	12.8	23.0	13.2	12.6	22.6
Richest quintile	20.0	24.6	26.2	19.2	23.4	25.3
Correlation with household consumption	0.20	0.22	0.22	0.21	0.22	0.23
N	19,028	19,028	18,602	19,028	19,028	18,620

Note: Each cell in the first five rows shows the percentage of the population that would be correctly captured for each consumption quintile if asset index was used. Specification 1 provides a simple count of the number of assets a household possesses, while Specification s 2 and 3 construct the asset index using principal component method. The list of assets for Specification 1 and Specification 2 include the status of the kitchen, electricity, ventilation system, whether the house is made of concrete, and the availability of tap water and piped sewerage system. Specification 3 adds to Specification 1 the house size and the condition of household furniture.

Table 4. Predicted Poverty Rates for Syrian Refugees Based on Proxy Means Test, Home Visit Data 2014 (percentage)

Method	Sample 1			Sample 2		
	Spec. 1	**Spec. 2**	**Spec. 3**	**Spec. 1**	**Spec. 2**	**Spec. 3**
Proxy means test	53.7	52.0	49.7	55.3	53.7	53.5
	(1.8)	(2.0)	(2.3)	(1.8)	(2.2)	(2.5)
Control variables						
Demographics & employment	Y	Y	Y	Y	Y	Y
Household assets & house characteristics	N	Y	Y	N	Y	Y
Shock-coping strategies & receiving UNHCR assistance	N	N	Y	N	N	Y
R2	0.38	0.43	0.48	0.38	0.43	0.48
N	19028	19028	19028	19028	19028	19028
True poverty rate		51.8			51.5	
		(2.3)			(2.6)	

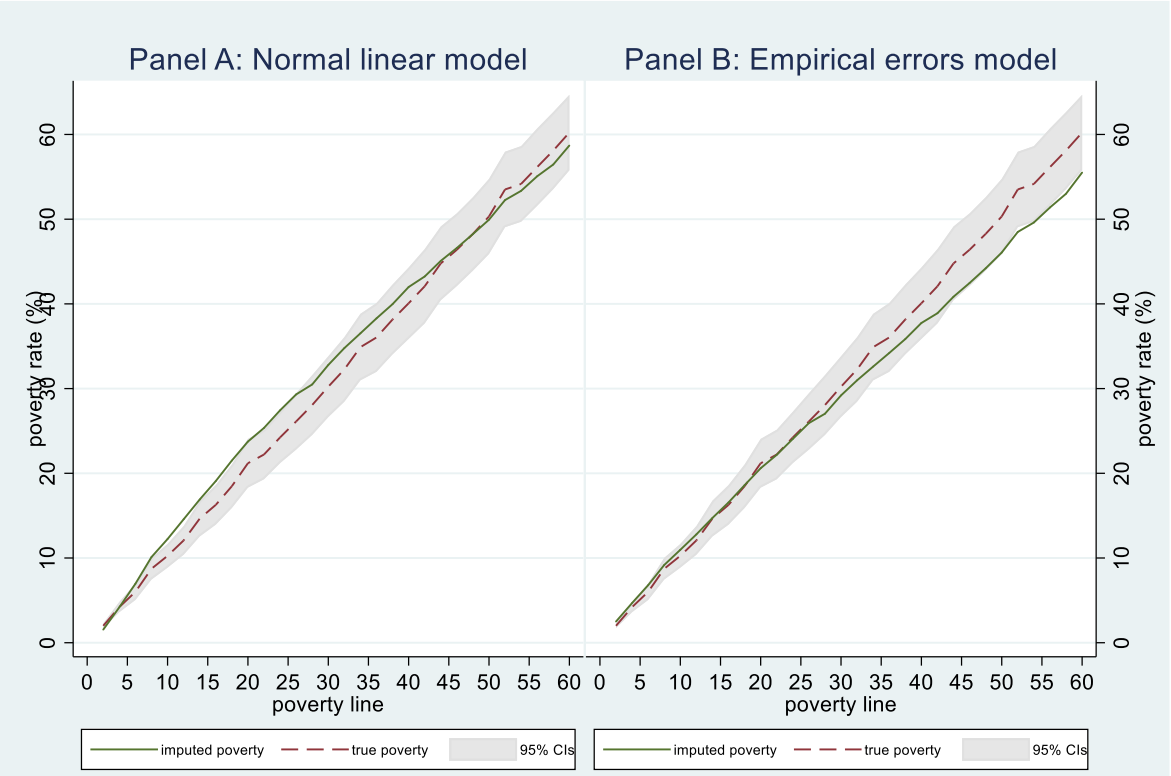
Note: The full regression results are provided in Table 2.1, Appendix 2. The estimation sample is generated by splitting the data into two random samples named sample 1 and sample 1. We then impute from Sample 1 to Sample 2 to obtain the imputed poverty rate in Sample 2, which are shown in the first and third rows. The true poverty rate for each sample is shown at the bottom of the table. Robust standard errors in parentheses are clustered at the district level. We use 1,000 simulations for each model run.

Table 5. Coverage and Leakage Rates Based on Imputation, ProGres and Home Visit Data (percentage)

	Spec. 1	**Spec. 2**	**Spec. 3**
Coverage rate	70.0	71.4	73.8
Leakage rate	42.5	39.7	36.5
Control variables			
Demographics & employment	Y	Y	Y
Household assets & house characteristics	N	Y	Y
Shock-coping strategies & receiving UNHCR assistance	N	N	Y
R2	0.44	0.48	0.54
N	18992	18992	18992

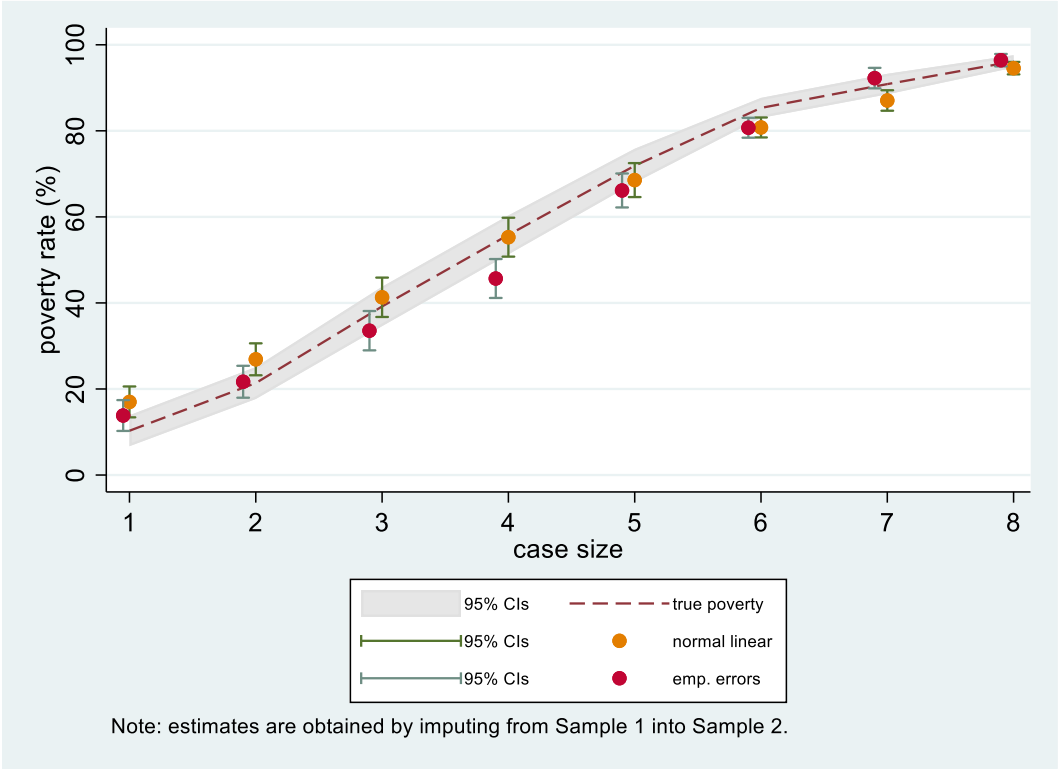
Note: The full regression results are provided in Table 2.1, Appendix 2. Specification 1 employs variables from the ProGres database only, and Specifications 2 and 3 employs variables from both the ProGres and HV databases. The estimation sample is generated by splitting the data into two random samples named Sample 1 and Sample 2. The imputed poverty rates are shown Sample 2, and the true poverty rate is shown at the bottom of the table. Robust standard errors in parentheses are clustered at the district level. We use 1,000 simulations for each model run.

Figure 1. Predicted Poverty Rates for Different Poverty Lines



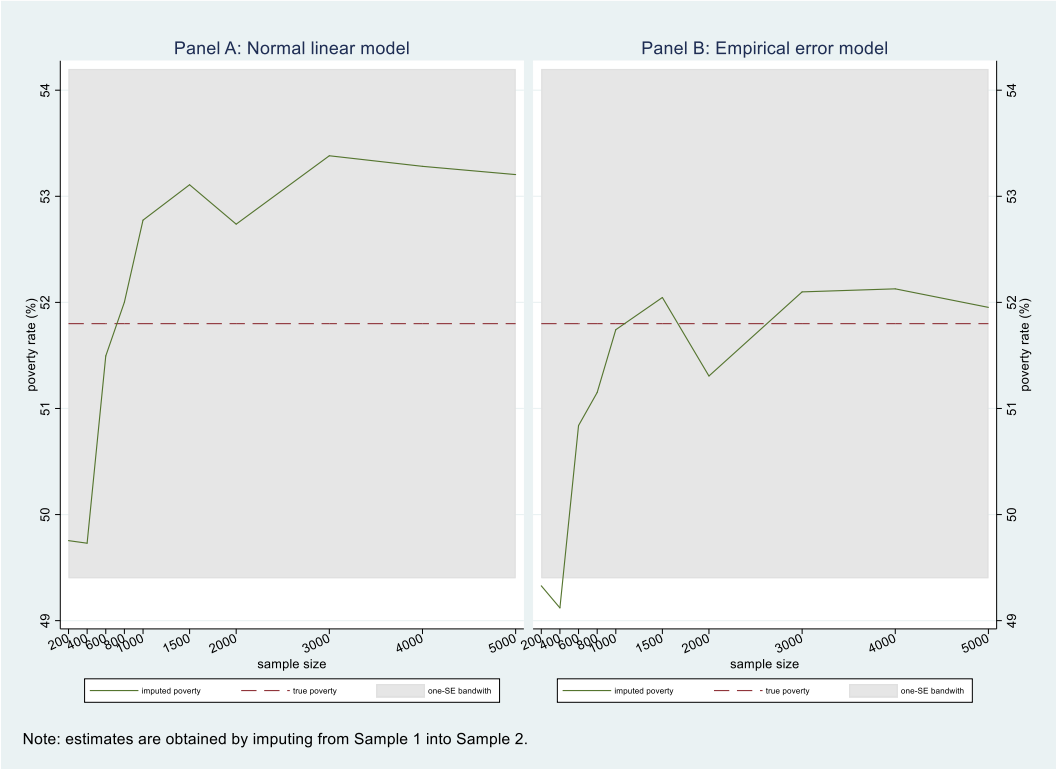
Note: estimates are obtained by imputing from Sample 1 into Sample 2.

Figure 2. Predicted Poverty Rates for Different Population Sub-groups



Note: estimates are obtained by imputing from Sample 1 into Sample 2.

Figure 3. Predicted Poverty Rates for Different Sample Sizes



Note: estimates are obtained by imputing from Sample 1 into Sample 2.

Appendix 1. Description of Park and Dudycha (1974)'s Derivations

We provide a more detailed description of Park and Dudycha (1974)'s derivations for their formulae in this appendix. In particular, we want to find the sample size *n* such that

$$P(\rho^2 - \rho_c^2) \leq \varepsilon = \gamma \quad (1.1)$$

where ρ^2 is the maximum (or true) multiple correlation possible for Equation (1) in the population, and ρ_c^2 is the correlation between the predicted value using Equation (1) and the original *y* variable. ρ_c^2 is usually referred to as the squared cross-validity correlation coefficient. A good sample size would ensure that the probability of obtaining an estimate within an acceptable degree of loss of precision (ε) around ρ^2 has reasonably good power (γ).

Park and Dudycha (1974) also show that the following relationship holds for ρ_c^2 and ρ^2

$$\rho_c^2 = \frac{\rho^2}{1+\frac{p-1}{F_{1,(p-1),\delta}}} \quad (1.2)$$

where $F_{1,(p-1),\delta}$ has a noncentral F distribution with the noncentrality parameter δ.

From Equation (1.2), we have for any positive ε

$$P(\rho^2 - \rho_c^2) \leq \varepsilon = P\left\{-(p-1)^{\frac{1}{2}}\left[\left(\frac{\rho^2}{\varepsilon}\right)-1\right]^{\frac{1}{2}} \leq t_{(p-1),\delta} \leq (p-1)^{\frac{1}{2}}\left[\left(\frac{\rho^2}{\varepsilon}\right)-1\right]^{\frac{1}{2}}\right\} \quad (1.3)$$

In other words, after we specify some (acceptable) values for ε and γ, we can obtain the value of the noncentrality parameter δ^2 for the noncentral Student's *t* distribution with *p-1* degrees of freedom that satisfies Equation (1.3).

Finally, given this value for δ^2, we can derive the sample size *n* that satisfies Equation (1.1) as follows

$$n = \left[\delta^2 \frac{1-\rho^2}{\rho^2}\right] + p + 2 \quad (1.4)$$

Appendix 2. Additional Tables and Figures

Table 2.1. Estimation Specification, Using Sample 1

	Model 1	Model 2	Model 3
csize_pg	-0.229***	-0.227***	-0.223***
	(0.00)	(0.00)	(0.00)
edu_highest	0.072***	0.054***	0.043***
	(0.00)	(0.00)	(0.00)
empl_occ_grp	0.006	-0.004	-0.004
	(0.00)	(0.00)	(0.00)
dem_age	0.002***	0.001***	0.001***
	(0.00)	(0.00)	(0.00)
dem_marriage	0.013**	0.026***	0.032***
	(0.01)	(0.01)	(0.01)
dem_pafemale	-0.060***	-0.074***	-0.050***
	(0.01)	(0.01)	(0.01)
dem_religion	-0.008***	-0.009***	-0.009***
	(0.00)	(0.00)	(0.00)
dem_origin_admlevel1	-0.003***	-0.003**	-0.002*
	(0.00)	(0.00)	(0.00)
arr_crosspoint_grp	-0.017***	-0.022***	-0.009**
	(0.00)	(0.00)	(0.00)
arr_legal	0.139***	0.122***	0.132***
	(0.01)	(0.01)	(0.01)
house_kitchen		0.038***	0.100***
		(0.01)	(0.01)
house_electricity		0.045***	0.042***
		(0.01)	(0.01)
house_ventilation		0.050***	0.040***
		(0.01)	(0.01)
house_rent_owned		0.580***	0.613***
		(0.02)	(0.02)
concrete_house		0.035	0.077***
		(0.02)	(0.02)
house_areapp		0.001***	0.001***
		(0.00)	(0.00)
wash_piped		0.022*	0.025**
		(0.01)	(0.01)
nfi_1_dummy			-0.070***
			(0.01)
pov_cop_aid			-0.177***
			(0.01)
pov_cop_share			-0.278***
			(0.01)
pov_cop_comm			-0.083***
			(0.01)
prot_cert_valid			0.109***
			(0.01)
pov_inc_unhcr			-0.410***
			(0.02)
_cons	4.808***	4.058***	3.897***
	(0.14)	(0.13)	(0.13)
sigma_e	0.69	0.67	0.64
sigma_u	0.06	0.04	0.00
rho	0.01	0.00	0.00
r2_o	0.38	0.43	0.48
N	19028	19028	19028

Note: The dependent variable is log of per capita household expenditure, net of UNHCR cash assistance.

Table 2.2. Predicted Poverty Rates for Syrian Refugees Based on Imputation, Home Visit Data (percentage)

Method	Sample 1			Sample 2		
	Spec. 1	**Spec. 2**	**Spec. 3**	**Spec. 1**	**Spec. 2**	**Spec. 3**
1) Normal linear regression model	50.3	50.0	49.7	50.8	50.7	50.7
	(1.1)	(1.3)	(1.6)	(1.1)	(1.3)	(1.7)
2) Empirical errors model	48.0	47.9	48.0	48.5	48.5	49.0
	(1.1)	(1.4)	(1.7)	(1.1)	(1.4)	(1.8)
Control variables						
Demographics & employment	Y	Y	Y	Y	Y	Y
Household assets & house characteristics	N	Y	Y	N	Y	Y
Shock-coping strategies & receiving UNHCR assistance	N	N	Y	N	N	Y
R2	0.40	0.45	0.50	0.40	0.45	0.50
N	19048	19048	19048	19072	19072	19072
True poverty rate		51.3			51.9	
		(2.4)			(2.5)	

Note: The full regression results are provided in Table 2.1, Appendix 2. All specifications employ variables from the HV database only The estimation sample is generated by splitting the data into two random samples named Sample 1 and Sample 2. The imputed poverty rate for Sample 1 and Sample 2 are shown in the first and second three columns respectively. The true poverty rate for each sample is shown at the bottom of the table. Robust standard errors in parentheses are clustered at the district level. We use 1,000 simulations for each model run.

Table 2.3. Predicted Poverty Rates for Syrian Refugees Based on Imputation with Probit Model, ProGres and HV Data (percentage)

Method	Sample 1			Sample 2		
	Spec. 1	**Spec. 2**	**Spec. 3**	**Spec. 1**	**Spec. 2**	**Spec. 3**
Probit model	55.7	54.2	54.8	54.1	52.8	53.4
	(1.2)	(1.5)	(1.7)	(1.4)	(1.6)	(1.9)
Control variables						
Demographics & employment	Y	Y	Y	Y	Y	Y
Household assets & house characteristics	N	Y	Y	N	Y	Y
Shock-coping strategies & receiving UNHCR assistance	N	N	Y	N	N	Y
Chi2	10.5	14.8	20.5	10.2	22.4	28.3
N	19028	19028	19028	19028	19028	19028
True poverty rate		51.8			51.5	
		(2.3)			(2.6)	

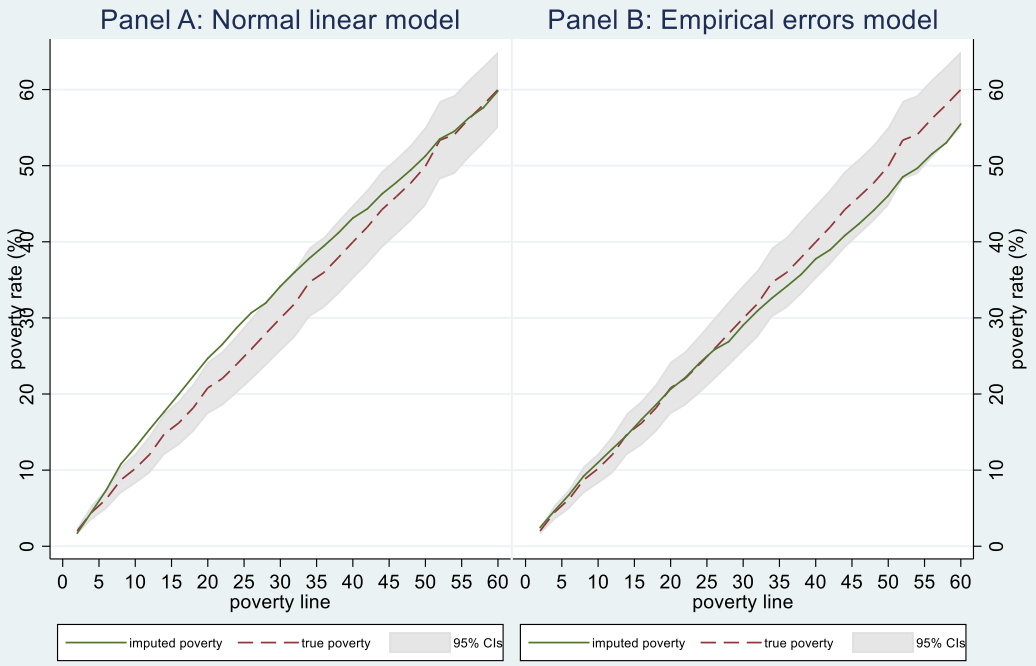
Note: The full regression results are provided in Table 2.1, Appendix 2. The estimation sample is generated by splitting the data into two random samples named sample 1 and sample 1. We then impute from Sample 1 to Sample 2 to obtain the imputed poverty rate in Sample 2, which are shown in the first and third rows. The true poverty rate for each sample is shown at the bottom of the table. Robust standard errors in parentheses are clustered at the district level. We use 1,000 simulations for each model run.

Table 2.4. Estimation Results for Subsamples of Different Sizes

	Subsample 1	Subsample 2	Subsample 3	Subsample 4	Subsample 5	Subsample 6	Subsample 7	Subsample 8	Subsample 9
	1000	1500	2000	2500	3000	3500	4000	4500	5000
csize_pg	-0.235***	-0.230***	-0.229***	-0.228***	-0.228***	-0.231***	-0.232***	-0.231***	-0.230***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)
edu_highest	0.115***	0.081***	0.083***	0.078***	0.078***	0.069***	0.068***	0.068***	0.061***
	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
empl_occ_grp	-0.013	-0.009	-0.013	-0.010	-0.010	0.002	0.000	-0.005	-0.007
	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
dem_age	0.002	0.003*	0.002	0.002*	0.002*	0.002**	0.002***	0.003***	0.003***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
dem_marriage	-0.006	0.009	0.005	0.009	0.009	0.021	0.018	0.022*	0.021*
	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
dem_pafemale	0.028	-0.022	-0.038	-0.057*	-0.057*	-0.076***	-0.074***	-0.066**	-0.080***
	(0.05)	(0.05)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)
dem_religion	-0.021**	-0.017**	-0.018**	-0.013*	-0.013*	-0.009	-0.012*	-0.013**	-0.011*
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
dem_origin_admlevel1	-0.003	-0.004	-0.003	-0.004	-0.004	-0.004	-0.005*	-0.004*	-0.003
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
arr_crosspoint_grp	0.000	0.001	-0.003	-0.003	-0.003	-0.010	-0.007	-0.012	-0.011
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
arr_legal	0.138***	0.131***	0.158***	0.182***	0.182***	0.146***	0.182***	0.190***	0.140***
	(0.05)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)
_cons	5.320***	5.136***	5.224***	5.008***	5.008***	4.786***	4.956***	4.986***	4.873***
	(0.42)	(0.39)	(0.38)	(0.34)	(0.34)	(0.30)	(0.29)	(0.27)	(0.25)
sigma_e	0.68	0.69	0.70	0.69	0.69	0.69	0.69	0.69	0.69
sigma_u	0.06	0.11	0.08	0.00	0.00	0.13	0.00	0.00	0.11
rho	0.01	0.02	0.01	0.00	0.00	0.03	0.00	0.00	0.02
r2_o	0.41	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38
N	1000	1500	2000	2500	2500	3500	4000	4500	5000

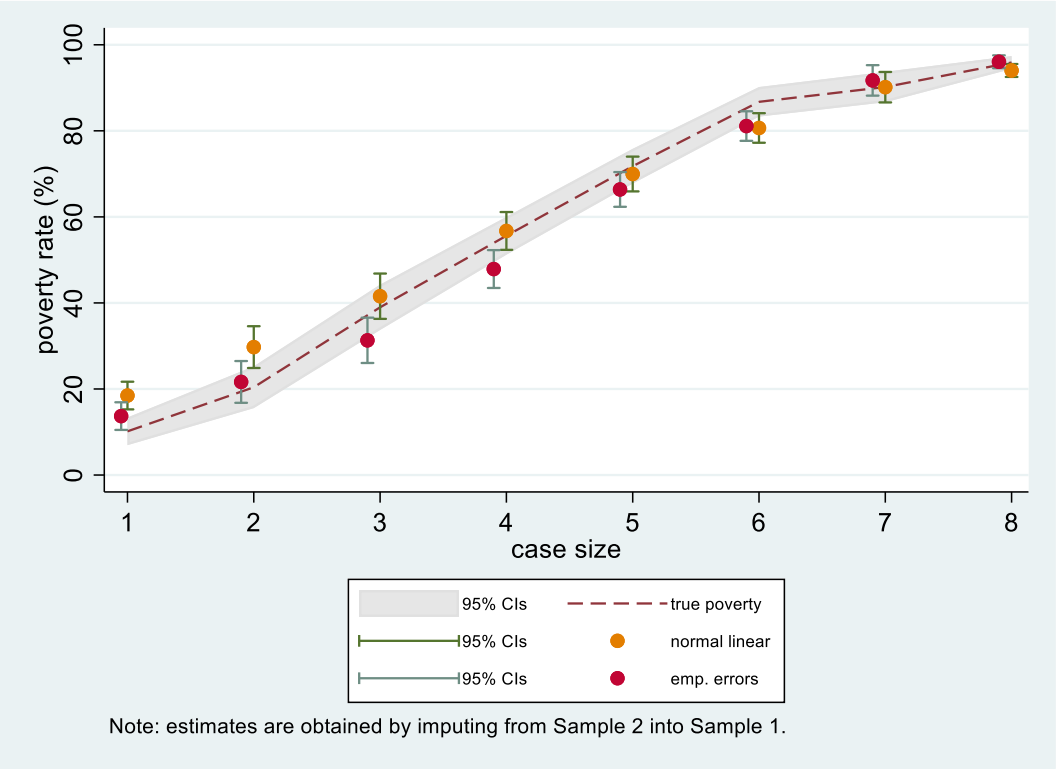
Note: The dependent variable is log of per capita household expenditure, net of UNHCR cash assistance.

Figure 2.1. Predicted Poverty Rates for Different Poverty Lines



Note: estimates are obtained by imputing from Sample 2 into Sample 1.

Figure 2.2. Predicted Poverty Rates for Different Population Sub-groups



Note: estimates are obtained by imputing from Sample 2 into Sample 1.