Policy Research Working Paper 9076

Estimating Poverty for Refugee Populations

Can Cross-Survey Imputation Methods Substitute for Data Scarcity?

Hai-Anh H. Dang
Paolo Verme

WORLD BANK GROUP
Development Economics
Development Data Group
&
Fragility, Conflict and Violence Global Theme
December 2019

Policy Research Working Paper 9076

Abstract

The increasing growth of forced displacement worldwide has led to the stronger interest of various stakeholders in measuring poverty among refugee populations. However, refugee data remain scarce, particularly in relation to the measurement of income, consumption, or expenditure. This paper offers a first attempt to measure poverty among refugees using cross-survey imputations and administrative and survey data collected by the United Nations High Commissioner for Refugees in Jordan. Employing a small number of predictors currently available in the United Nations High Commissioner for Refugees registration system, the proposed methodology offers out-of-sample predicted poverty rates. These estimates are not statistically different from the actual poverty rates. The estimates are robust to different poverty lines, they are more accurate than those based on asset indexes or proxy means tests, and they perform well according to targeting indicators. They can also be obtained with relatively small samples. Despite these preliminary encouraging results, it is essential to replicate this experiment across countries using different data sets and welfare aggregates before validating the proposed method.

This paper is a product of the Development Data Group, Development Economics and the Fragility, Conflict and Violence Global Theme. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at hdang@worldbank.org.

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

Hai-Anh H. Dang and Paolo Verme*

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, IZA, and Vietnam's Academy of Social Sciences. Paolo 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, Talip Kilic, Peter Lanjouw, Christoph Lakner, David Newhouse, Franco Peracchi, Shinya Takamatsu, Matthew Wai-Poi, Tara Wishvanath, and participants at the IARIW-WB conference on poverty measurement and 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 newcomers. 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 of any effective economic policy. It also becomes increasingly clear that achieving the SDGs (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 that are 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 million, 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 High 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. For example, the UNHCR has attempted to collect consumption data for the Syrian refugees in Jordan using large-scale surveys that interview as many as 5,000 households per month (or 60,000 households per year). In other refugee contexts, where resources and logistical challenges exist, such large-scale surveys may not be feasible or sustainable.[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.

This paper contributes to the poverty measurement literature by applying recent advances in cross-survey imputations to measure poverty among Syrian refugees living in Jordan. 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 determining the type of protection and assistance they require. 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 occasionally collect information on household well-being by means of sample surveys designed to address specific issues, such as measuring food security or determining various types of vulnerabilities. These surveys may contain information on income, consumption

[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.

or expenditure but they are typically administered only in selected areas or sub-samples of the refugee population.

In this paper, we combine proGres administrative data with survey data collected by the UNHCR in Jordan in 2014 and use a recently developed model for cross-survey imputations to estimate poverty among Syrian refugees. We also provide a sensitivity analysis to different poverty lines, estimate minimum sample sizes required for accurate estimation, compare the model proposed with alternative welfare measurement approaches, and test the performance of the model using targeting indicators. To our knowledge, this is the first experiment of its kind. Poverty studies that make 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.

We find that the imputation-based poverty estimates provided by the paper are not statistically different from the non-predicted consumption-based poverty rates (henceforth, the 'true' poverty rate), and that this result is robust to various validation tests, including alternative 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 in the UNHCR's proGres database, which is consistent with the findings in recent studies for imputation-based poverty estimates for regular populations. We also provide both theoretical and empirical evidence that relatively small survey samples can be combined with those from the census-type registration system to provide cost-effective and updated estimates of poverty.

While our estimation results are encouraging, they may not apply to other country contexts, sources of data or welfare measures. Further application of the proposed methodology to other

countries and data sets is essential before this methodology can be fully validated and used in operations.

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, we apply Dang, Lanjouw, and Serajuddin's (2017) imputation framework, which builds on earlier survey-to-census imputation studies (Elbers, Lanjouw, and Lanjouw, 2003; Tarozzi, 2007).[3] 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,

[3] 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.[4] 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, and y_j is household consumption typically modeled in log form. Note that we suppress the subscript that indexes households to make the notation less cluttered.[5] 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.

[4] 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.

[5] 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 $\upsilon_{cj}|x_j \sim N(0, \sigma^2_{\upsilon_j})$ and $\varepsilon_j|x_j \sim N(0, \sigma^2_{\varepsilon_j})$. 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), 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 Assumption 1 before running the models. In our specific case, this assumption is satisfied by the very nature of the data we use, since we restrict our experiment to households that are present in both data sets by matching individuals and households with personal identifiers so that both data sets contain the same individuals. In other words, for our purposes of testing the method, we pretend to have full coverage of the population with both data sets and then split the sample artificially to simulate a cross-survey imputation exercise.[6] Naturally, this is an ideal data scenario that is not easily found elsewhere, which allows us to provide a rigorous test of the cross-survey imputation model proposed.

III. Application to Syrian Refugees in Jordan

III.1. Country Background and Data

[6] For imputation on two surveys that are implemented in two different periods, Dang et al. (2017) make an additional assumption that the changes in x_j between the two periods can capture the change in poverty rate in the next period (Assumption 2). This assumption is not relevant to our case, since we use administrative and survey data that were collected by the UNHCR in the same year.

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 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 million people had been displaced internally in the country, about 1.5 million people fled the country with their own means, and an additional 3.7 million 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.[7] 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

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

and the Syrian crisis, the UNHCR and the World Food Program (WFP) have been conducting a variety of surveys as well as extensive home visits that allowed researchers to analyze refugee conditions as 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. Our experiment is restricted to households present in both data sets, a total of approximately 40,000 households.

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 the same aggregate used by Verme *et al.* (2016) and Verme and Gigliarano (2019).[8]

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 the estimated poverty figures are using the PG data alone (as both the base and the target survey).

As a first step, we generated two samples by extracting 50% of observations from the 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

[8] A full explanation of the consumption aggregate is provided by Verme *et al.* (2016).

other way around by imputing from Sample 2 to Sample 1 and then compare with the true poverty rate in Sample 1.

We 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 data set (PG-specific variables), which include case (household) size and the PA's demographic and employment characteristics (age, gender, different levels of education achievement, occupation group, marital status, religion, and the governorate or city of original residence in the Syrian Arab Republic).[9] 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 quality status of the kitchen, electricity access, and the ventilation system, the living area of the house (as measured by the number of square meters per person), 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

[9] We consider the following five levels of education achievement: 1) below six years of schooling, 2) 6-8 years of schooling, 3) 9-11 years of schooling, 4) 12-14 years of schooling, and 5) university education or higher.

this information upfront when refugees are first registered. Vice-versa, if poverty estimates imputed with the PG data are not statistically different from the true rates (i.e., those produced directly from the 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 term 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 Appendix 2 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. Estimates using Specification 2 with more variables on household assets and house characteristics are somewhat better and closer to the true poverty rate 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 1.1 percentage points larger than the true poverty estimate of 51.5 percent. The poverty estimate using Specification 3 (Table 1, third column) is 52.3 percent, which is 0.8 percentage points less than the true poverty estimate. This is likely because imputation models that include household assets are usually found to perform better than those that do not (Christiaensen *et al.*, 2012; Dang *et al.*, 2019).[10]

[10] On the other hand, adding more control variables does not necessarily lead to a better model fit. While this result may appear counter-intuitive, one possible reason is that doing so may overfit the data and thus does not offer more accuracy, which is shown with empirical evidence from India and Jordan (Dang *et al.*, 2017; Dang and Lanjouw, 2018). A recent theoretical study also suggests that for misspecified regressions, adding more variables may result in

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 the three specifications and the true poverty rates for both samples are negligible.

The alternative imputation model based on the empirical distribution of the error terms (Table 1, row 2) performs even better 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 data set is originally a non-random subsample of the PG database, we also re-run Table 1 using only variables that are available in the HV data set. The estimation results, shown in Table 2.2 in Appendix 2, are very similar to those in Table 1.

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 the poverty line, more disaggregated population groups, and alternative estimation methods. In the next section, we also consider the question of sample size.

Sensitivity to the poverty line

larger inconsistency (De Luca, Magnus, and Peracchi, 2018). Also note that 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.

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 60th percentile of the consumption distribution) to reproduce poverty estimates using imputations from Sample 1 to Sample 2 and the two models described. The 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 the 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. We impute from Sample 1 to Sample 2 and re-estimate poverty for each of the case sizes. To ensure that the estimation sample size is reasonable, we combine all the cases with eight or more

individuals into a single group (which makes up roughly 6 percent of the estimation sample). We employ the two error estimation models and plot the estimated poverty rates against case size in Figure 2.

Both methods provide similar results and both sets of results are within the 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 the 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, both estimation models seem to perform reasonably well.

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 the 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 modeling assumption on the dependent variable, which can result 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 modeling 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 while the estimates using the probit and logit models are still within the 95% CI of the true rates, they are somewhat less accurate than those obtained using the empirical errors model in Table 1.

For example, the estimated poverty rate using Specification 1 and Sample 2 for the logit model is 53.1%, which is 1 percentage point larger than the corresponding figure of 51.8% for the empirical errors model (compared with the true poverty rate of 51.6%).

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

One practically relevant question is how large the imputation sample should be to obtain accurate poverty estimates.[11] 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 should be reached between these trade-offs. 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

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

[11] 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.

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.[12] 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

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

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's (1974) 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.[13] 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 0.45 and the number of predictors *p* is 27, which are the parameters obtained under Specification 1 for Sample 2 in Table 1. The estimates provided in Table 2 suggest

[12] 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.

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

that the minimum sample size is 389 observations (where ε and γ are respectively 0.05 and 0.90), and a reasonably good sample size is 1,068 observations (where ε and γ are respectively 0.02 and 0.95). 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 2,509 observations.

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. The estimates in this figure are restricted to Sample 2 from which 10 sub-samples of different sizes—including 200, 400, 600, 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.

The 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 the results.[14] Yet, plotting all the 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

[14] All estimates fall within the 95 percent CI of the true poverty rate but are not shown for lack of space.

normal linear model tends to overestimate the true value more than the empirical errors model.[15] 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 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 offer reasonably 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 shown that covering about one-sixtieth of this number, or 1,000 households per year, may be sufficient to provide reliable poverty statistics.[16]

IV.2. Related Measures of Poverty

How does our proposed poverty imputation method compare with alternative estimation methods such as asset (wealth) indexes and proxy-means tests? We examine in this section each of these two alternatives, together with the related exercise of targeting. This is a particularly important question for the UNHCR, which uses asset indexes to measure well-being in place of consumption in many places where consumption is not available. Other development organizations

[15] 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.

[16] 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.

such as the WFP also often employ 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

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. Suppressing the household index to make the notation less cluttered, this can be expressed as follows

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

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 of reasons that make asset indexes more likely to result in biased estimates of poverty. First, the wealth index w_j does not include the non-asset components, 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.[17] Finally, in a refugee context, the temporary nature of displacement likely affects refugees' behaviors in terms of accumulation and use of assets. For example, refugees may choose not to invest as much in high-quality durables as a

[17] 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.

regular household does. This practical aspect may further make assets (alone) an even less reliable data source for poverty estimation in a refugee context.

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.

Consistent with our earlier discussion, the quintiles based on the wealth index can only capture between 12 and 35 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 (34 percent) of the poorest consumption quintile in Sample 1 (Sample 2). The correlation between asset indexes and household consumption is not very strong, ranging between 0.21 and 0.23.[18] These are half as strong as a correlation of roughly 0.44 and 0.48 (respectively for Specification 1 and Specification 3 in Table 1) between the original household consumption and the predicted consumption obtained from our method. This provides supportive evidence for our earlier discussion that asset indexes may not be good predictors of household welfare and poverty, particularly in a refugee context.

[18] These correlation coefficients between the wealth indexes and consumption are weaker than those observed in Filmer and Scott (2012) for 11 other countries around the world (which range from 0.39 to 0.72 for these countries).

Proxy means test

Most of the estimates based on proxy means testing start from a general equation that can be described as follows:

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

where the vector of coefficients β_j^p is 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 $\upsilon_{cj} + \varepsilon_j$ 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 household consumption. Even 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.[19]

Table 4 provides poverty estimates using the proxy means test method as in Equation (10). A couple of remarks are in order to illustrate the results. First, the estimates fall outside 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 not negligible. On the other hand, consistent with our theoretical discussion above, the standard errors for the poverty estimates in Table 4 range from 2.5 to 2.9 percent, which are roughly 10 to 25 percent larger than those based on the poverty imputation methods shown in Table 1.

[19] Dang *et al.* (2019) offer more detailed discussion and more formal proofs of these results.

Targeting ratios

The importance of modeling the error terms can also be appreciated when we estimate the targeting ratios following a poverty prediction exercise such as the percentage of the poor population that are correctly identified (i.e., coverage rate) and the percentage of the population identified as poor who are not poor (i.e., leakage rate). Note that just as with the poverty rate, we need to do multiple simulations to estimate these targeting rates. In particular, the formulae for the coverage rate and the leakage rate are as follows:

$$coverage = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{N}\sum_{i=1}^{N} I(\hat{y}^{1}_{2\mathrm{i,s}} \leq z_1 \mid y^{1}_{2\mathrm{i,s}} \leq z_1) \tag{11}$$

$$leakage = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{N}\sum_{i=1}^{N} I(\hat{y}^{1}_{2\mathrm{i,s}} \leq z_1 \mid y^{1}_{2\mathrm{i,s}} > z_1) \tag{12}$$

where $I(.)$ is the indicator function, "|" inside the parentheses is the conditional operator, and the subscript *i* indicates households.

Estimates based on the empirical errors model, shown in Table 5, suggest that Specification 1 can provide a reasonable coverage rate of 70 percent, and a leakage rate of roughly 32 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 3 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 31 percent, using the 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. Our results show that imputation-based poverty estimates are statistically not different from the non-predicted consumption-based 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, and our imputation models are rather parsimonious and use variables that are already available in 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.

The estimation results also point to the need for 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 relatively small surveys 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 should 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.

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	52.5	52.3	53.1	53.0	53.0
	(2.0)	(2.0)	(2.1)	(2.0)	(1.9)	(2.0)
2) Empirical errors model	51.3	51.3	51.5	51.8	51.8	52.2
	(2.2)	(2.2)	(2.3)	(2.2)	(2.1)	(2.1)
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	0.43	0.48	0.53	0.45	0.50	0.54
N	19001	19001	19001	18999	18999	18999
True poverty rate		51.5			51.6	
		(2.5)			(2.4)	

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

Epsilon	Gamma		
	0.99	0.95	0.90
0.01	2509	2137	1954
0.02	1253	1068	976
0.03	835	711	650
0.04	625	533	487
0.05	500	426	389

Note: Estimates are based on the formulae provided in Park and Dudycha (1974). We use the given parameters, the R2 value of 0.45 and the number of predictors of 27 under Specification 1 from Table 1.

Table 3. Population Distribution by Asset Indexes vs. Consumption

Per capita consumption	2012			2014		
	Model 1	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**
Poorest quintile	32.9	32.8	32.4	34.7	33.7	34.0
Quintile 2	27.2	27.0	22.6	26.1	26.9	21.7
Quintile 3	26.5	22.9	19.0	28.1	23.8	21.6
Quintile 4	12.6	12.4	22.2	13.5	12.7	22.1
Richest quintile	19.9	23.6	25.8	19.4	24.2	26.1
Correlation with household consumption	0.21	0.22	0.22	0.21	0.22	0.23
N	19001	19001	18558	18999	18999	18610

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. Model 1 provides a simple count of the number of assets a household possesses, while Models 2 and 3 construct the asset index using principal component method. The list of assets for Model 1 and Model 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. Model 3 adds to Model 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	59.5	59.0	57.1	60.5	59.6	58.1
	(2.7)	(2.8)	(2.9)	(2.7)	(2.5)	(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	0.43	0.48	0.53	0.45	0.50	0.54
N	19001	19001	19001	18999	18999	18999
True poverty rate		51.5			51.6	
		(2.5)			(2.4)	

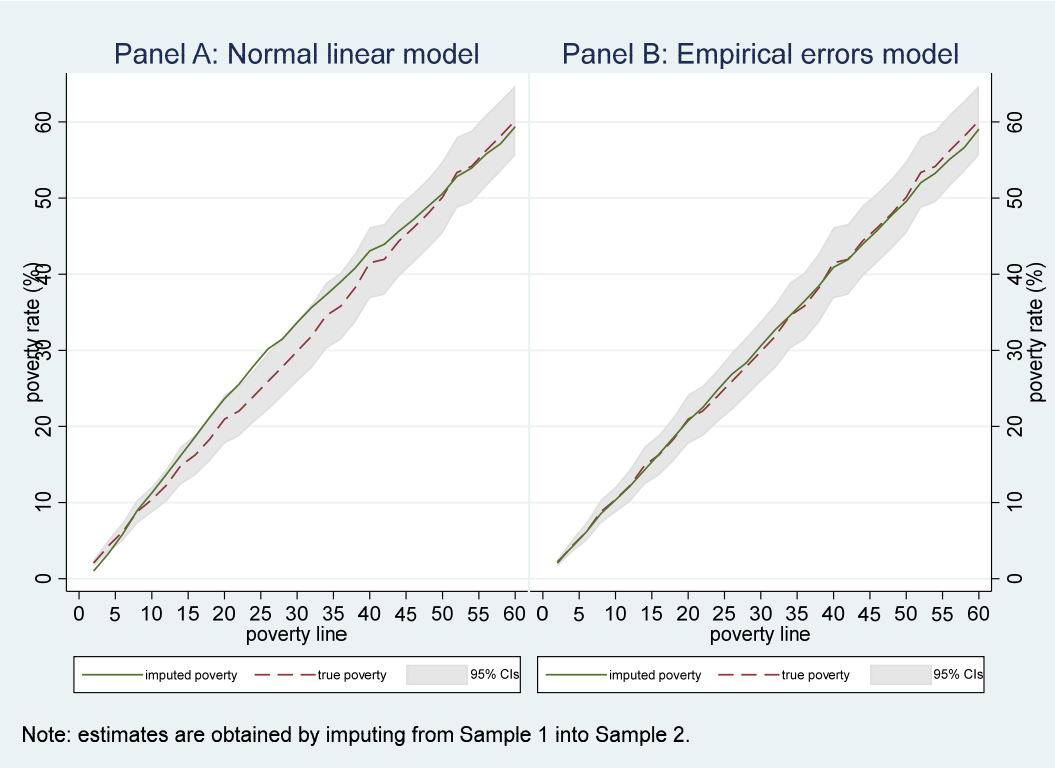
Note: The full regression results are provided in Table 2.1, Appendix 2. The estimation sample is generated by splitting the data into two random samples named sample 1 and sample 1. We then impute from Sample 1 to Sample 2 and vice versa to obtain the imputed poverty rate for each sample. The true poverty 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)

	Model 1	**Model 2**	**Model 3**
Coverage rate	70.0	71.3	73.5
Leakage rate	32.4	30.9	29.4
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

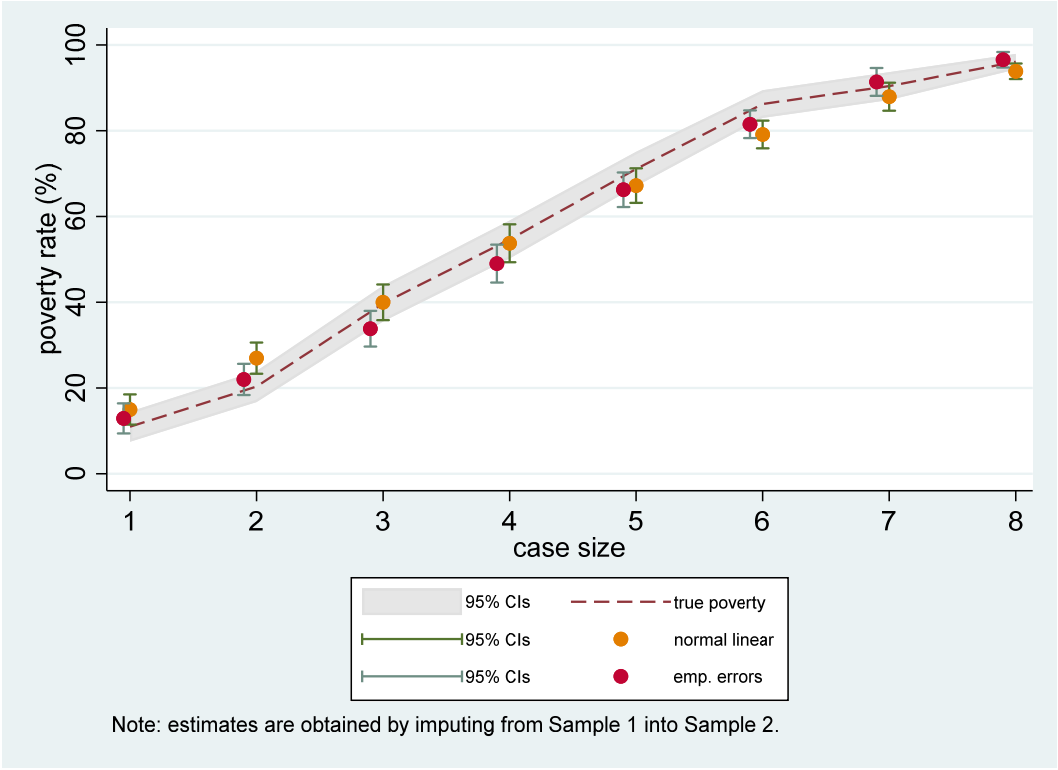
Note: The full regression results are provided in Table 2.1, Appendix 2. Model 1 employs variables from the ProGres database only, and Models 2 and 3 employs variables from both the ProGres and HV databases, using the empirical errors model. The estimation sample is generated by splitting the data into two random samples named Sample 1 and Sample 2. The imputed targeting rates are obtained using the empirical errors on Sample 2. 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



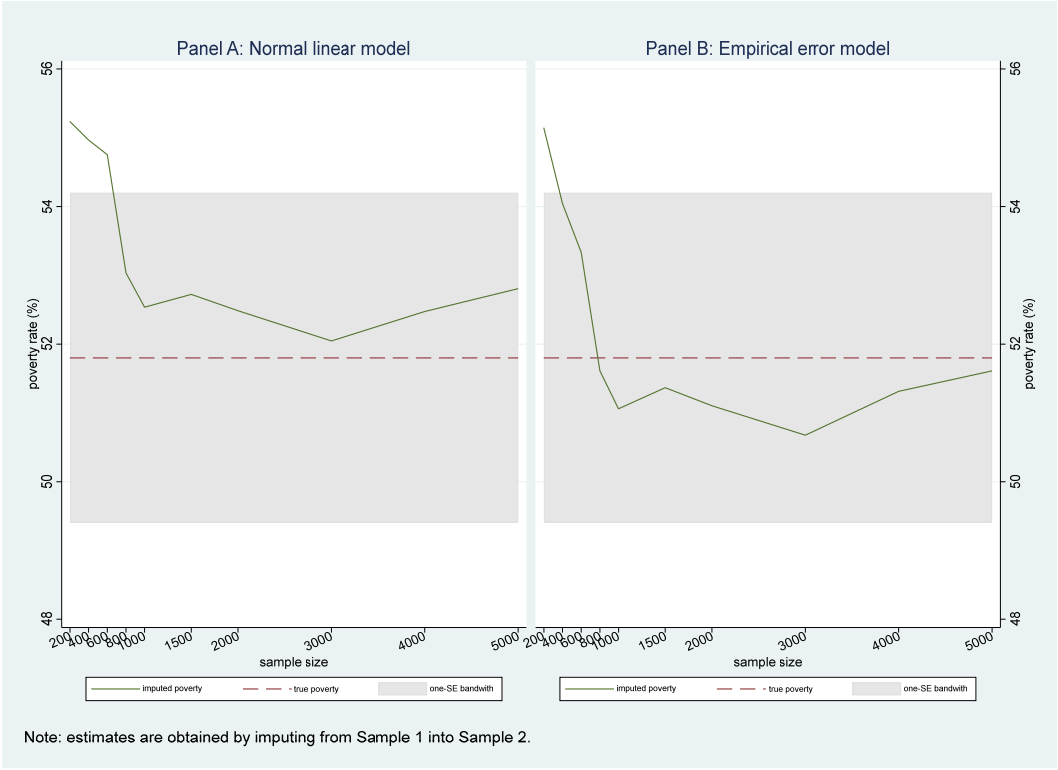
Note: estimates are obtained by imputing from Sample 1 into Sample 2.

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's (1974) Derivations

We provide a more detailed description of Park and Dudycha's (1974) 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
Case size equals 2	-0.543***	-0.551***	-0.547***
	(0.02)	(0.02)	(0.02)
Case size equals 3	-0.927***	-0.938***	-0.934***
	(0.02)	(0.02)	(0.02)
Case size equals 4	-1.162***	-1.166***	-1.164***
	(0.02)	(0.02)	(0.02)
Case size equals 5	-1.320***	-1.330***	-1.320***
	(0.02)	(0.02)	(0.02)
Case size equals 6	-1.531***	-1.538***	-1.522***
	(0.02)	(0.02)	(0.02)
Case size equals 7	-1.602***	-1.607***	-1.586***
	(0.02)	(0.02)	(0.02)
Case size equals 8	-1.665***	-1.664***	-1.656***
	(0.03)	(0.03)	(0.03)
Case size equals 9	-1.724***	-1.706***	-1.682***
	(0.04)	(0.04)	(0.04)
Case size equals 10 or more	-1.797***	-1.742***	-1.723***
	(0.05)	(0.05)	(0.04)
PA completed 6-8 years of schooling	0.075***	0.033**	0.026**
	(0.01)	(0.01)	(0.01)
PA completed 9-11 years of schooling	0.116***	0.066***	0.050***
	(0.02)	(0.02)	(0.02)
PA completed 12-14 years of schooling	0.147***	0.094***	0.071***
	(0.02)	(0.02)	(0.02)
PA had university or higher	0.269***	0.202***	0.169***
	(0.03)	(0.02)	(0.02)
PA is employed in low-skilled occupations	-0.001	0.033	0.044**
	(0.02)	(0.02)	(0.02)
PA is employed in skilled occupations	0.003	0.011	0.016
	(0.02)	(0.02)	(0.02)
PA is employed in high-skilled occupations	0.038*	0.028	0.033*
	(0.02)	(0.02)	(0.02)
PA is employed in professional occupations	0.079***	0.070***	0.072***
	(0.02)	(0.02)	(0.02)
PA's age	0.002***	0.001***	0.001***
	(0.00)	(0.00)	(0.00)
PA is divorced or separated	-0.146***	-0.100***	-0.081**
	(0.04)	(0.04)	(0.03)
PA is widowed	-0.090***	-0.059***	-0.050**
	(0.02)	(0.02)	(0.02)
PA is single	-0.116***	-0.095***	-0.049***
	(0.02)	(0.02)	(0.02)
PA is female	-0.059***	-0.061***	-0.041***
	(0.02)	(0.02)	(0.01)
Border crossing point is Ruwaished- Hadallat	0.020	0.092***	0.071***
	(0.02)	(0.02)	(0.02)
Border crossing point is Tal Shihab	-0.076***	-0.071***	-0.038*
	(0.02)	(0.02)	(0.02)
Border crossing point is Nasib	-0.092***	-0.084***	-0.036**
	(0.02)	(0.02)	(0.02)
Other border crossing points or no data	-0.070***	-0.055***	-0.012
	(0.02)	(0.02)	(0.02)
Arrival is formal	0.111***	0.080***	0.088***
	(0.01)	(0.01)	(0.01)
House: quality of the kitchen		0.049***	0.115***
		(0.01)	(0.01)
House: quality of electricity access		0.036***	0.029***
		(0.01)	(0.01)
House: quality of ventillation system		0.059***	0.049***
		(0.01)	(0.01)
House: rent or owned		0.594***	0.636***
		(0.02)	(0.02)
House: made of concrete		0.069***	0.105***
		(0.02)	(0.02)
House: square meters per person		0.001***	0.001***
		(0.00)	(0.00)
House: having piped water and piped sewarage		0.040***	0.038***
		(0.01)	(0.01)
Receving any type of NFIs			-0.042***
			(0.01)
Poverty coping strategy: humanitarian assistance			-0.154***
			(0.01)
Poverty coping strategy: sharing costs with the host family			-0.313***
			(0.01)
Poverty coping strategy: receiving support from the host community			-0.052***
			(0.01)
Having a valid protection certificate (that gives access to UNHCR services)			0.079***
			(0.01)
Receving UNHCR's monthly financial assistance			-0.378***
			(0.02)
Constant	4.861***	3.956***	3.827***
	(0.04)	(0.04)	(0.04)
sigma_e	0.66	0.63	0.60
sigma_υ	0.00	0.00	0.00
rho	0.00	0.00	0.00
Overall R2	0.45	0.50	0.54
N	19001	19001	19001

Note: The dependent variable is log of per capita household expenditure, net of UNHCR cash assistance. All regressions control for dummy variables indicating the original regions in Syria and the current governorate of residence in Jordan. "PA" stands for the principal case applicant (head of household). The reference categories for the PA's employment categories, marital status, and border crossing point are respectively no employment, being engaged or married, and arrival by air.

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

Method	Sample 1			Sample 2		
	Model 1	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**
1) Normal linear regression model	52.4	52.4	52.2	52.9	52.8	52.8
	(2.0)	(2.0)	(2.1)	(2.0)	(1.9)	(2.0)
2) Empirical errors model	51.2	51.3	51.5	51.7	51.7	52.1
	(2.2)	(2.2)	(2.3)	(2.2)	(2.1)	(2.1)
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	0.45	0.50	0.55	0.47	0.52	0.56
N	19001	19001	19001	18999	18999	18999
True poverty rate		51.5			51.6	
		(2.5)			(2.4)	

Note: The full regression results are provided in Table 2.1, Appendix 2. All models 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		
	Model 1	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**
Probit model	53.5	52.7	53.0	53.2	52.3	52.5
	(2.1)	(2.1)	(2.2)	(1.9)	(2.0)	(2.1)
Logit model	53.4	52.7	52.9	53.1	52.2	52.4
	(2.1)	(2.1)	(2.2)	(1.9)	(2.0)	(2.1)
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
N	19001	19001	19001	18999	18999	18999
True poverty rate		51.5			51.6	
		(2.5)			(2.4)	

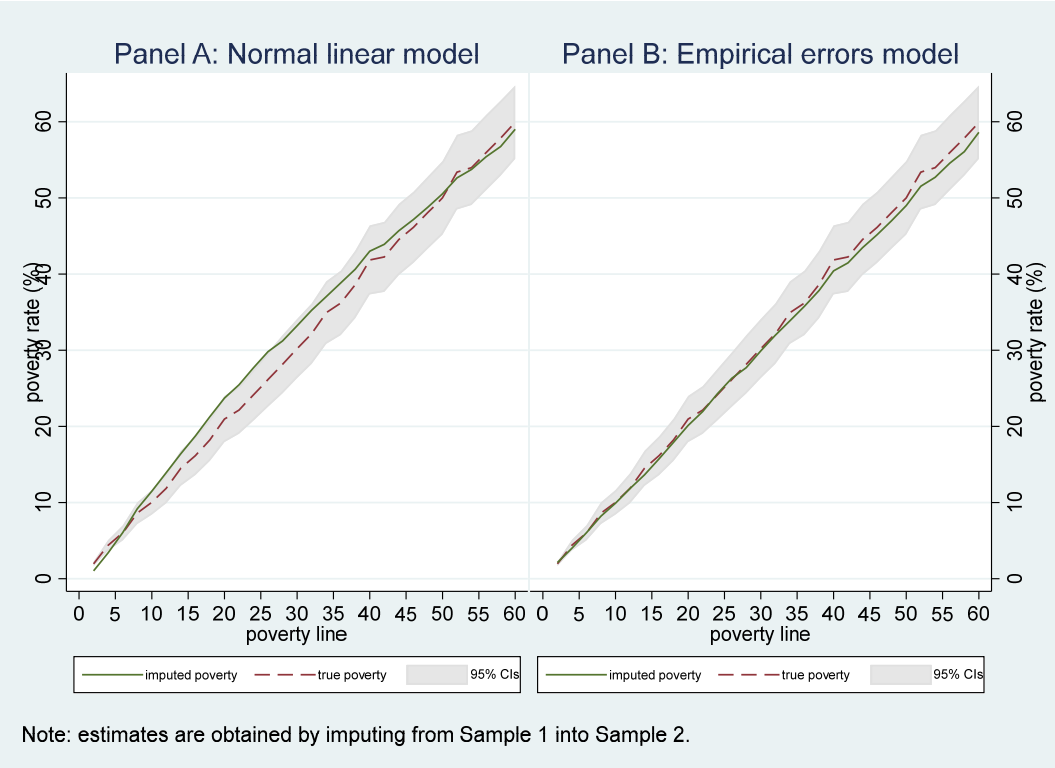
Note: The underlying regression results are obtained from the probit or logit model. Model 1 employs variables from the ProGres database only, and Models 2 and 3 employs variables from both the ProGres and HV databases. The estimation sample is generated by splitting the data into two random samples 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 2.4. 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	Subsample 10
	200	400	600	800	1000	1500	2000	3000	4000	5000
Case size equals 2	-0.394	-0.466***	-0.443***	-0.485***	-0.523***	-0.450***	-0.457***	-0.482***	-0.487***	-0.468***
	(0.25)	(0.16)	(0.12)	(0.10)	(0.08)	(0.07)	(0.06)	(0.05)	(0.04)	(0.04)
Case size equals 3	-0.960***	-0.986***	-0.824***	-0.861***	-0.898***	-0.894***	-0.891***	-0.937***	-0.953***	-0.927***
	(0.24)	(0.15)	(0.11)	(0.10)	(0.08)	(0.06)	(0.06)	(0.05)	(0.04)	(0.04)
Case size equals 4	-0.967***	-0.981***	-0.898***	-1.008***	-1.082***	-1.041***	-1.044***	-1.078***	-1.108***	-1.099***
	(0.24)	(0.14)	(0.11)	(0.09)	(0.08)	(0.07)	(0.06)	(0.05)	(0.04)	(0.04)
Case size equals 5	-1.394***	-1.243***	-1.180***	-1.204***	-1.244***	-1.175***	-1.190***	-1.242***	-1.292***	-1.282***
	(0.25)	(0.15)	(0.12)	(0.10)	(0.08)	(0.07)	(0.06)	(0.05)	(0.04)	(0.04)
Case size equals 6	-1.311***	-1.469***	-1.408***	-1.447***	-1.533***	-1.460***	-1.431***	-1.499***	-1.518***	-1.508***
	(0.22)	(0.15)	(0.12)	(0.10)	(0.09)	(0.07)	(0.06)	(0.05)	(0.04)	(0.04)
Case size equals 7	-1.472***	-1.516***	-1.427***	-1.461***	-1.526***	-1.519***	-1.478***	-1.588***	-1.581***	-1.560***
	(0.31)	(0.17)	(0.13)	(0.11)	(0.10)	(0.08)	(0.07)	(0.06)	(0.05)	(0.05)
Case size equals 8	-1.811***	-1.664***	-1.608***	-1.672***	-1.755***	-1.725***	-1.653***	-1.678***	-1.681***	-1.692***
	(0.32)	(0.21)	(0.16)	(0.14)	(0.13)	(0.10)	(0.09)	(0.07)	(0.06)	(0.06)
Case size equals 9	-1.780***	-1.473***	-1.523***	-1.583***	-1.754***	-1.843***	-1.673***	-1.732***	-1.724***	-1.699***
	(0.44)	(0.26)	(0.23)	(0.21)	(0.19)	(0.16)	(0.13)	(0.10)	(0.09)	(0.08)
Case size equals 10 or more	-1.279	-2.135***	-1.872***	-1.920***	-1.936***	-1.834***	-1.670***	-1.745***	-1.678***	-1.642***
	(0.81)	(0.43)	(0.28)	(0.24)	(0.23)	(0.18)	(0.16)	(0.12)	(0.10)	(0.10)
PA completed 6-8 years of schooling	0.032	0.131	0.069	0.026	0.043	0.009	0.043	0.038	0.067**	0.035
	(0.17)	(0.11)	(0.09)	(0.07)	(0.06)	(0.05)	(0.04)	(0.04)	(0.03)	(0.03)
PA completed 9-11 years of schooling	0.218	0.148	0.083	0.059	0.073	0.049	0.109**	0.101**	0.127***	0.095***
	(0.19)	(0.12)	(0.10)	(0.08)	(0.07)	(0.06)	(0.05)	(0.04)	(0.04)	(0.03)
PA completed 12-14 years of schooling	0.377	0.149	0.085	0.008	0.027	0.043	0.097	0.070	0.122***	0.113***
	(0.25)	(0.16)	(0.12)	(0.10)	(0.09)	(0.07)	(0.06)	(0.05)	(0.04)	(0.04)
PA had university or higher	0.021	0.404*	0.292*	0.298**	0.244**	0.220**	0.245***	0.251***	0.287***	0.245***
	(0.36)	(0.24)	(0.17)	(0.14)	(0.12)	(0.09)	(0.08)	(0.06)	(0.06)	(0.05)
PA is employed in low-skilled occupations	0.085	0.020	0.110	0.005	0.084	0.135*	0.061	-0.001	-0.016	-0.023
	(0.32)	(0.19)	(0.14)	(0.11)	(0.10)	(0.08)	(0.07)	(0.06)	(0.05)	(0.04)
PA is employed in skilled occupations	-0.219	-0.042	0.079	0.006	0.057	0.102	0.067	0.019	0.007	0.017
	(0.30)	(0.17)	(0.12)	(0.10)	(0.09)	(0.07)	(0.06)	(0.05)	(0.04)	(0.04)
PA is employed in high-skilled occupations	-0.349	0.014	0.174	0.133	0.151	0.176**	0.142**	0.076	0.064	0.059
	(0.33)	(0.19)	(0.13)	(0.11)	(0.09)	(0.07)	(0.06)	(0.05)	(0.04)	(0.04)
PA is employed in professional occupations	-0.280	0.076	0.116	0.116	0.140	0.248***	0.205***	0.163***	0.122**	0.121***
	(0.35)	(0.19)	(0.14)	(0.11)	(0.10)	(0.08)	(0.07)	(0.05)	(0.05)	(0.04)
PA's age	-0.002	-0.000	0.002	0.003	0.003	0.001	0.001	0.001	0.001	0.001
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
PA is divorced or separated	-1.034**	-0.330	-0.200	-0.269	-0.302*	-0.272**	-0.239**	-0.185**	-0.134*	-0.141*
	(0.41)	(0.28)	(0.23)	(0.21)	(0.17)	(0.13)	(0.11)	(0.09)	(0.08)	(0.07)
PA is widowed	-0.321	0.097	0.052	0.138	0.046	0.019	0.028	0.019	0.014	-0.018
	(0.30)	(0.18)	(0.15)	(0.12)	(0.10)	(0.08)	(0.07)	(0.05)	(0.05)	(0.04)
PA is single	-0.104	0.019	-0.058	-0.127	-0.133	-0.102	-0.098*	-0.162***	-0.177***	-0.158***
	(0.25)	(0.14)	(0.11)	(0.09)	(0.08)	(0.07)	(0.06)	(0.05)	(0.04)	(0.04)
PA is female	0.161	-0.136	-0.198**	-0.211**	-0.153**	-0.131**	-0.116**	-0.130***	-0.120***	-0.113***
	(0.23)	(0.15)	(0.10)	(0.09)	(0.08)	(0.06)	(0.05)	(0.04)	(0.03)	(0.03)
Border crossing point is Ruwaished- Hadallat	0.208	0.088	0.026	-0.110	-0.001	0.068	0.062	0.079	0.085	0.080*
	(0.29)	(0.19)	(0.15)	(0.13)	(0.11)	(0.09)	(0.08)	(0.06)	(0.05)	(0.05)
Border crossing point is Tal Shihab	-0.429	-0.171	-0.092	-0.113	-0.102	-0.052	-0.033	-0.031	-0.032	0.001
	(0.28)	(0.18)	(0.14)	(0.12)	(0.10)	(0.08)	(0.07)	(0.06)	(0.05)	(0.04)
Border crossing point is Nasib	-0.326	-0.142	-0.081	-0.119	-0.117	-0.066	-0.092*	-0.122***	-0.116***	-0.094***
	(0.20)	(0.13)	(0.10)	(0.08)	(0.07)	(0.06)	(0.05)	(0.04)	(0.04)	(0.03)
Other border crossing points or no data	-0.304	-0.108	0.010	-0.106	-0.080	-0.007	-0.019	-0.058	-0.072*	-0.063*
	(0.22)	(0.14)	(0.11)	(0.09)	(0.08)	(0.07)	(0.06)	(0.05)	(0.04)	(0.04)
Arrival is formal	0.253	0.196*	0.228***	0.153**	0.149**	0.163***	0.181***	0.156***	0.152***	0.157***
	(0.16)	(0.10)	(0.08)	(0.07)	(0.06)	(0.05)	(0.04)	(0.03)	(0.03)	(0.03)
Constant	4.970***	4.770***	4.579***	4.816***	4.850***	4.806***	4.797***	4.950***	4.942***	4.898***
	(0.50)	(0.30)	(0.22)	(0.19)	(0.17)	(0.14)	(0.12)	(0.10)	(0.09)	(0.08)
sigma_e	0.72	0.71	0.69	0.68	0.67	0.65	0.66	0.65	0.66	0.66
sigma_υ	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
rho	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Overall R2	0.54	0.49	0.46	0.47	0.48	0.48	0.46	0.47	0.46	0.45
N	200	400	600	800	1000	1500	2000	3000	4000	5000

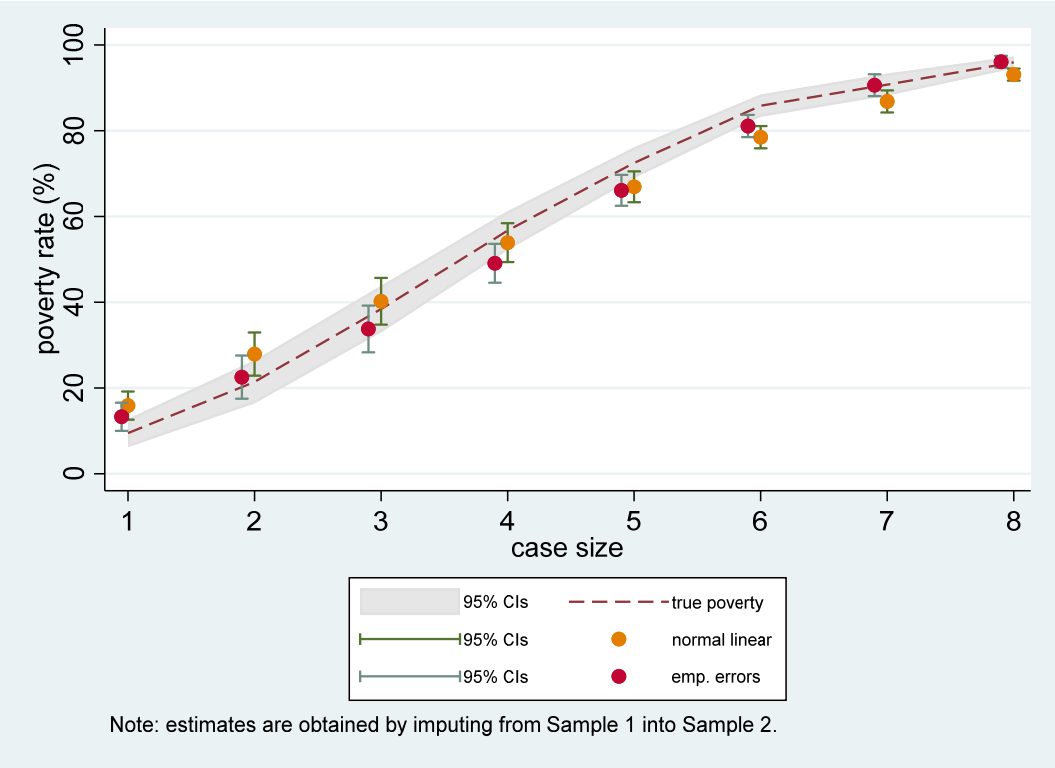
Note: The dependent variable is log of per capita household expenditure, net of UNHCR cash assistance. All regressions control for dummy variables indicating the original regions in Syria and the current governorate of residence in Jordan. "PA" stands for the principal case applicant (head of household). The reference categories for the PA's employment categories and marital status are respectively no employment and being engaged or married.

Figure 2.1. Predicted Poverty Rates for Different Poverty Lines



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

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



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