Estimating Poverty for Refugee Populations:

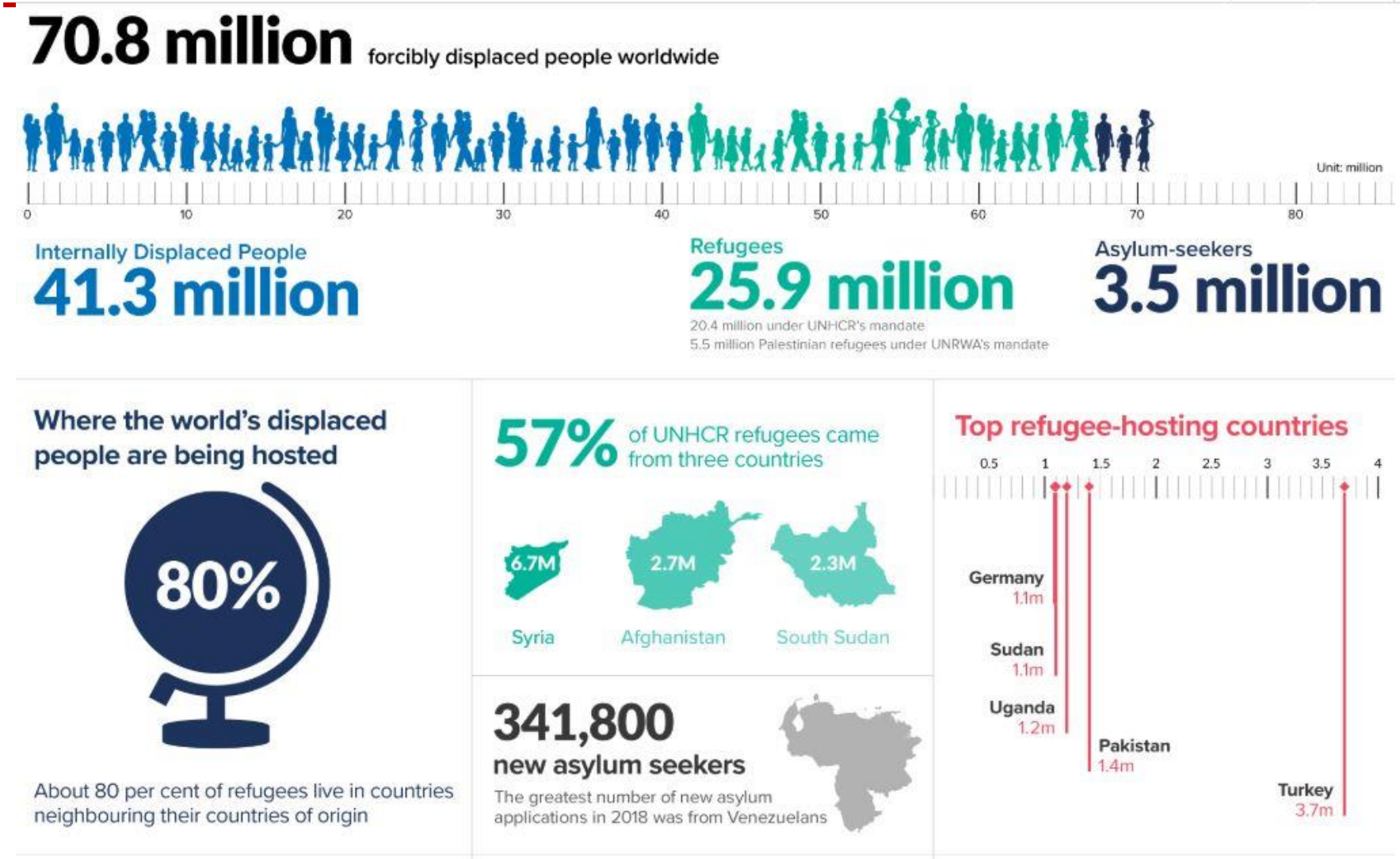
Can Cross-survey Imputation Methods Substitute for Data Scarcity?

Hai-Anh Dang and Paolo Verme
IARIW

November 2019

Washington DC

Motivation


70.8 million forcibly displaced people worldwide
Unit: million
0
10
20
30
40
50
60
70
80
Internally Displaced People
41.3 million
Refugees
25.9 million
20.4 million under UNHCR's mandate
5.5 million Palestinian refugees under UNRWA's mandate
Asylum-seekers
3.5 million
Where the world's displaced people are being hosted
80%
About 80 per cent of refugees live in countries neighbouring their countries of origin
57% of UNHCR refugees came from three countries
6.7M
Syria
2.7M
Afghanistan
2.3M
South Sudan
341,800
new asylum seekers
The greatest number of new asylum applications in 2018 was from Venezuelans
Top refugee-hosting countries
0.5
1
1.5
2
2.5
3
3.5
4
Germany 1.1m
Sudan 1.1m
Uganda 1.2m
Pakistan 1.4m
Turkey 3.7m

Motivation

- Increasing growth of forced displacement worldwide is of concern for
  - Policy makers (poverty reduction policies)
  - International organizations (global poverty count)

- Very few studies offer nationally representative poverty estimates for refugees
  - Little data (refugees and IDPs excluded from national surveys)
  - Scarce and very recent literature (Verme et al. 2016)

Contribution

- Tests poverty imputation methods to a refugee context
- Uses UNHCR administrative registration data to estimate refugee poverty
- If validated, the method proposed can potentially result in significant cost-saving for WB and UNHCR operations by reducing surveys' sample size and frequency
- Provides results that are robust to:
  - varying poverty lines
  - disaggregated population groups
  - stronger parametric modelling assumptions

Analytical Framework (1)

Dang, Lanjouw & Serajuddin (2017)'s imputation model

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

where the error term μ_{ij} is broken down into two components: a cluster random effects (υ_{cj}) and the idiosyncratic error term (ε_j). Suppose that consumption data is available for survey 1 but missing for survey 2. Let z_2 be the poverty line in period 2, and y_2^1 be the imputed consumption for survey 2. Poverty in period 2 is then estimated as $P_2 = P(y_2^1 \leq z_2)$ (2)

Analytical Framework (2)

Assumptions:

1) x_j are comparable for both surveys;

2) changes in x_j between the two periods can capture the change in poverty rate in the next period

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

where β_1' are estimated using Equation (1), and $\tilde{\hat{v}}_{1,s}$ and $\tilde{\hat{\varepsilon}}_{1,s}$ represent the sth random draw from their estimated distributions, for s= 1,…, S.

Data

- Two data sets for Jordan in 2014:
  - UNHCR Home Visits 2014 (survey-type data, 1/3 refugee population)
  - UNHCR ProGres registration data 2014 (census-type data)

- About 40,000 household, only refugees living outside camps. Same consumption aggregate and poverty line as in Verme et al. (2016)

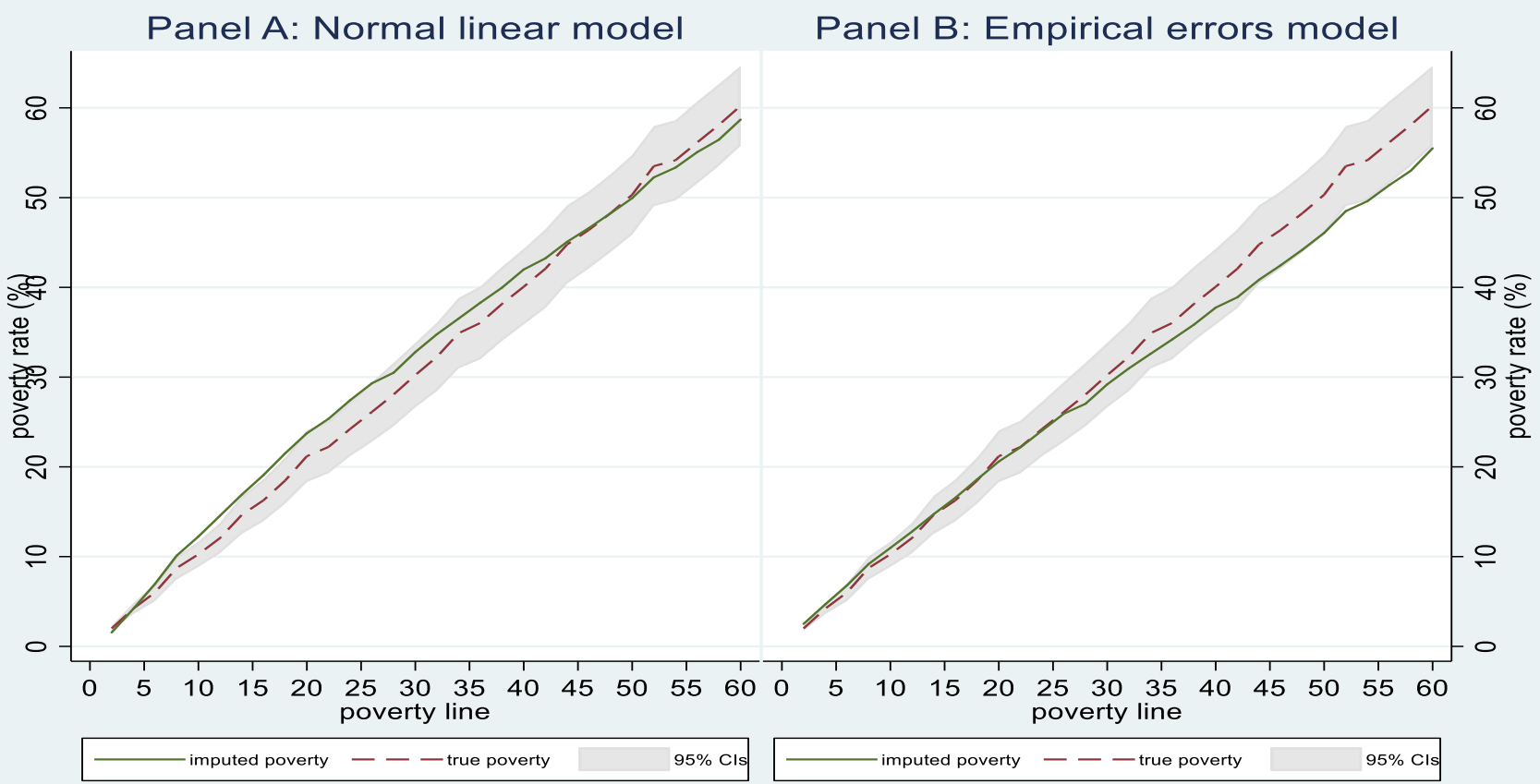
- Two random samples: Sample 1 and Sample 2
  - Impute from Sample 1 into Sample 2, then check estimates against true values in Sample 2
  - Vice versa, impute from Sample 2 into Sample 1, then check estimates against true values in Sample 1

Main Results – DLS Model

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)	

Robustness – Poverty Lines

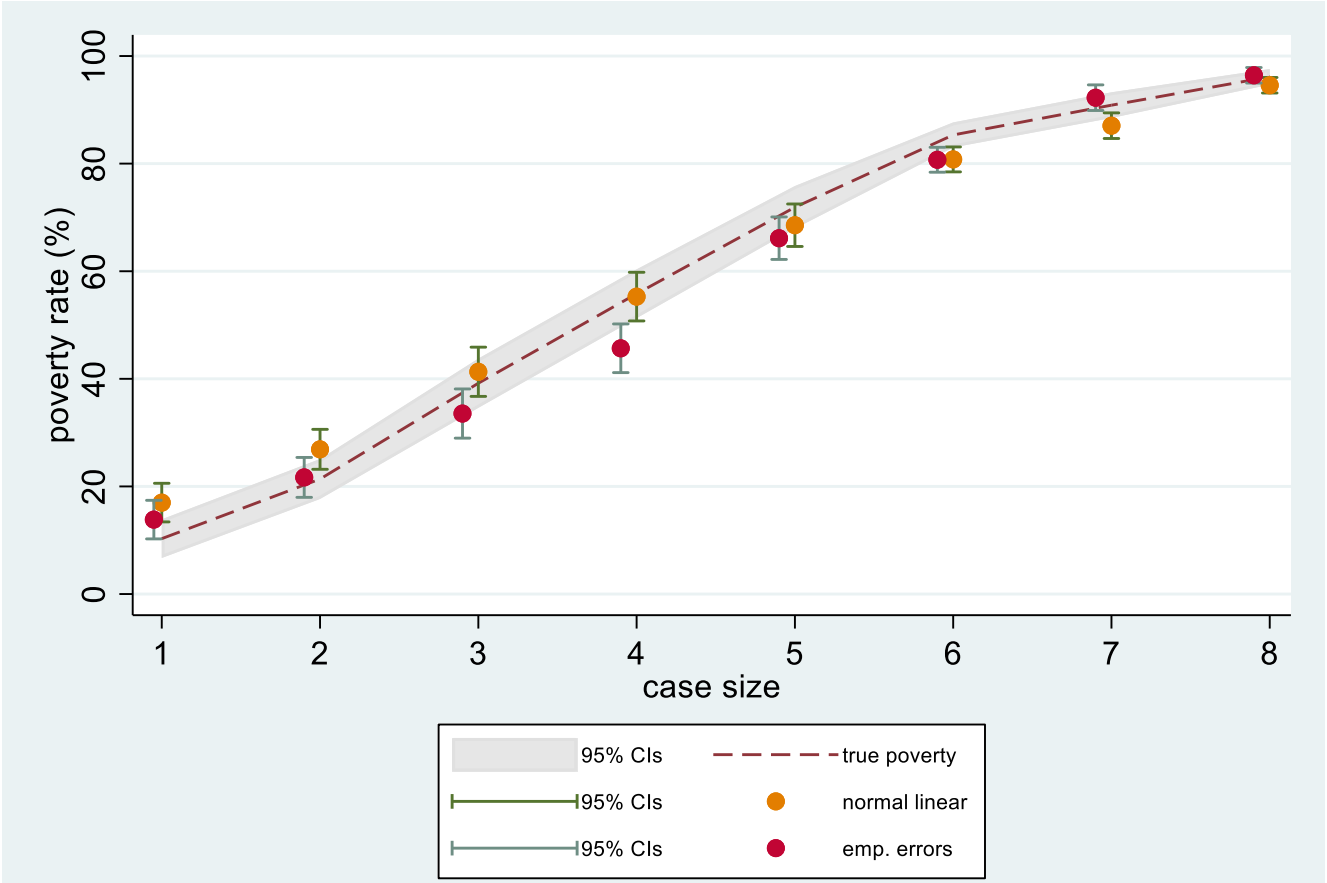
Figure 1. Predicted Poverty Rates for Different Poverty Lines



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

Robustness – Household size

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



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

Comparison – OLS-PMT Model

Method	Sample 1			Sample 2		
	Spec. 1	**Spec. 2**	**Spec. 3**	**Spec. 1**	**Spec. 2**	**Spec. 3**
Proxy means test	53.7	52	49.7	55.3	53.7	53.5
	(1.8)	(2)	(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)	

Comparison – Categorical Model

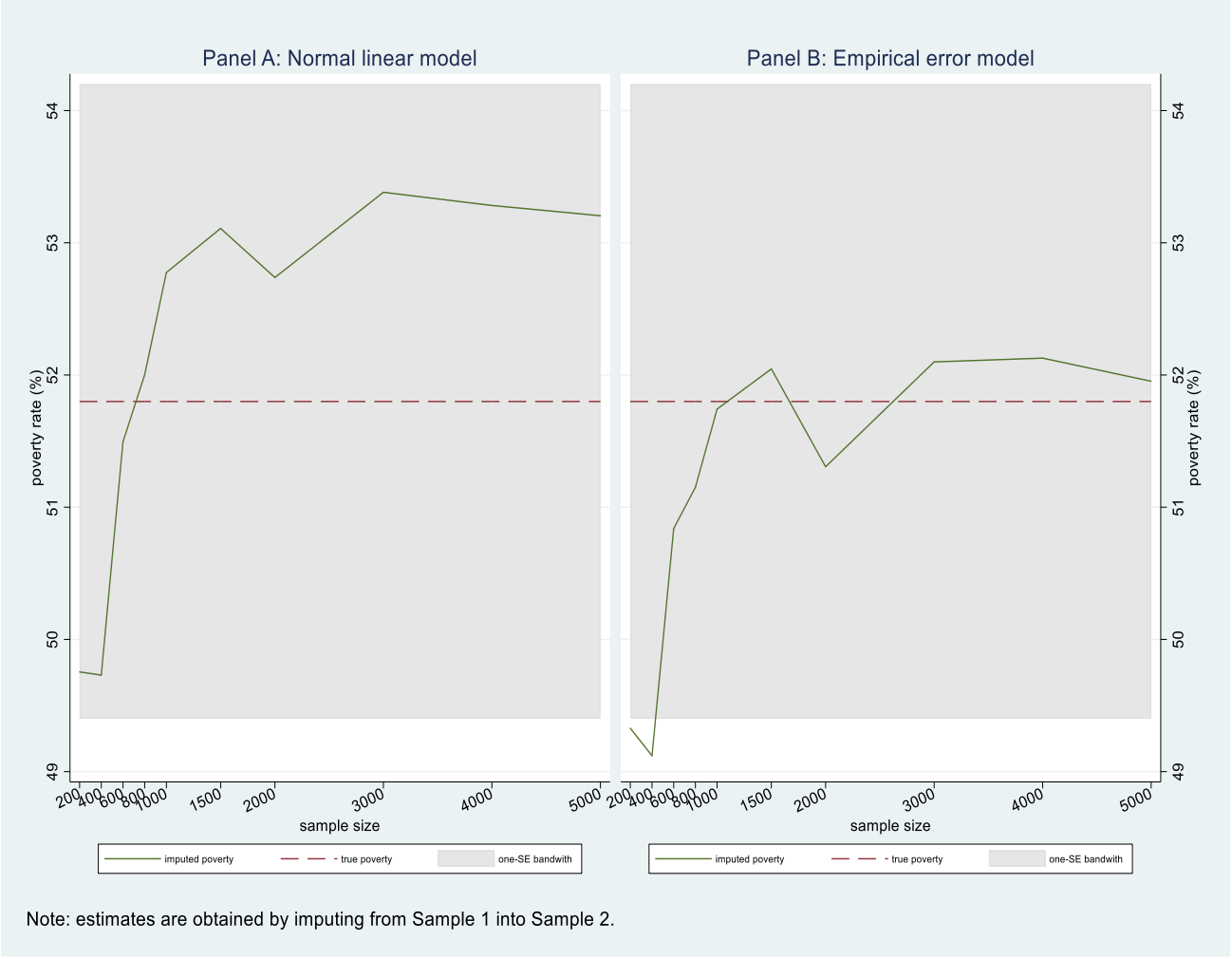
Method	Sample 1			Sample 2		
	Spec. 1	**Spec. 2**	**Spec. 3**	**Spec. 1**	**Spec. 2**	**Spec. 3**
Probit model	55.7	54.2	54.8	54.1	52.8	53.4
	(1.2)	(1.5)	(1.7)	(1.4)	(1.6)	(1.9)
Control variables						
Demographics & employment	Y	Y	Y	Y	Y	Y
Household assets & house characteristics	N	Y	Y	N	Y	Y
Shock-coping 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)	

Comparison – Asset index

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	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	13.2	12.6	22.6
Richest quintile	20	24.6	26.2	19.2	23.4	25.3
Correlation with household consumption	0.2	0.22	0.22	0.21	0.22	0.23
N	19,028	19,028	18,602	19,028	19,028	18,620

Each cell in the first five rows shows the share of the population that would be correctly captured for each quintile if the asset index would be used in place of consumption per capita.

Sample size



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

Sample Size

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

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

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

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

Sample size

ε	γ		
	0.99	0.95	0.9
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

VI. Conclusion

- First application of poverty imputation methods for refugees
  - encouraging performance
  - better performance than simple OLS, probit model and asset indexes
  - robust to different validation checks

- Promising for data-scarce environments
  - cost-effective
  - rapid assessment

Table 2.1. Estimation Model, 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.