**GOAL:**

The goal of this paper is to impute poverty rates of St. Lucia for years when expenditure/ consumption data is not available with the help of the consumption model built upon the household expenditure data.

**METHODs:**

The survey-to-survey imputation of poverty rates essentially is a problem of missing values. There are several ways we can achieve our goal. One is the single imputation, where we replace missing values with predicted scores from a regression equation.

Another one is the stochastic imputation, where a residual term is randomly drawn from a normal distribution with mean zero and variance equal to the residual variance from the regression model, is added to the predicted scores from the regression imputation [1].

The third method is the Multivariate Imputation and we apply this method to our St. Lucia data in this paper.

*“Missing data analyses are difficult because there is no inherently correct methodological procedure. In many (if not most) situations, blindly applying maximum likelihood estimation or multiple imputation will likely lead to a more accurate set of estimates than using one of the [previously mentioned] missing data handling techniques” (p.344, Applied Missing Data Analysis, 2010).*

Multiple imputation is essentially an iterative form of stochastic imputation. However, instead of filling in a single value, the distribution of the observed data is used to estimate multiple values that reflect the uncertainty around the true value. These values are then used in the analysis of interest, such as in a OLS model, and the results combined. Each imputed value includes a random component whose magnitude reflects the extent to which other variables in the imputation model cannot predict its true values (Johnson and Young, 2011; White et al, 2010). Thus, building into the imputed values a level of uncertainty around the “truthfulness” of the imputed values [1].

The regular linear regression imputation is included as a comparison.

We implement the MI method in STATA using the MI package in the following way:

*// prediction with MI*

*mi set mlong*

*mi register imputed ln\_pcexpae\_def11 $head\_char $hh\_char*

*mi impute mvn ln\_pcexpae\_def11 $head\_char $hh\_char = sex, add(10) rseed(1234) force*

*mi estimate, saving(miest, replace): reg ln\_pcexpae\_def11 sex $head\_char $hh\_char*

*mi predict ln\_pcexpae\_def11\_hat using miest*

*gen pcexpae\_def11\_hat=exp(ln\_pcexpae\_def11\_hat)*

*gen poor\_hat=1 if pcexpae\_def11\_hat <= povline\_def11*

*replace poor\_hat=0 if pcexpae\_def11\_hat > povline\_def11*

*gen poverty=1 if pcexpae\_def11 <= povline\_def11*

*replace poverty=0 if pcexpae\_def11 > povline\_def11*

*// prediction with simple regression*

*regress ln\_pcexpae\_def11 sex $head\_char $hh\_char*

*predict ln\_pcexpae\_def11\_hat0*

*gen pcexpae\_def11\_hat0=exp(ln\_pcexpae\_def11\_hat0)*

*gen poor\_hat0=1 if pcexpae\_def11\_hat0 <= povline\_def11*

*replace poor\_hat0=0 if pcexpae\_def11\_hat0 > povline\_def11*

*bys survey: sum poverty poor\_hat poor\_hat0 [aw=nweight] if pcexpae\_def11\_hat!=. & year==2016*

*bys year: sum poverty poor\_hat poor\_hat0 [aw=nweight] if pcexpae\_def11\_hat!=. & survey=="LFS"*

**KEY ASSUMPTIONS:**

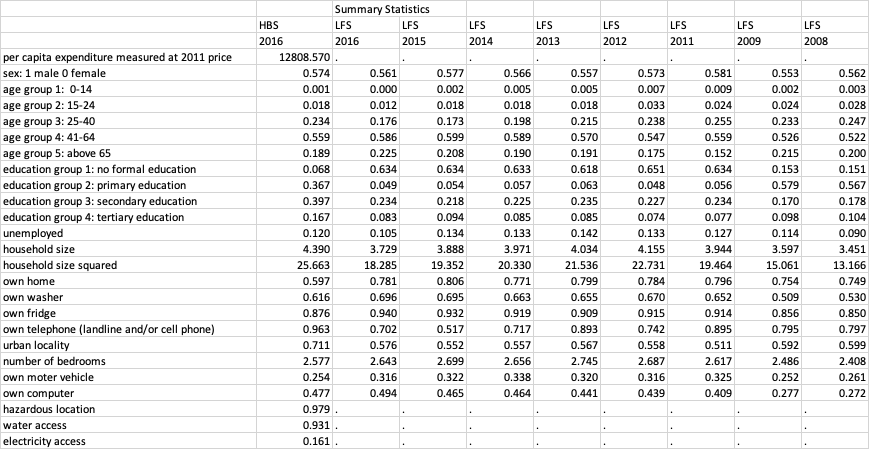
1. the questions in the two surveys are consistent. The common variables in the two surveys must measure the same version of a concept.
2. The model must be time-invariant so that the estimated parameters from one survey one years can be useful in imputing consumptions in other data in other year.
3. Common variables of the two surveys can explain a large share of the variation in household expenditures and poverty.
4. Sampling design should be comparable across the surveys

This survey-to-survey imputation relies on these key assumptions above, violations of them could result in a biased imputation of the poverty rates.

**DATA:**

Household Budget Survey (HBS) 2016 and Labor Force Survey (LFS) 2008-2016 (missing 2010). Table 1 below is the summary statistics for variables common across survey.

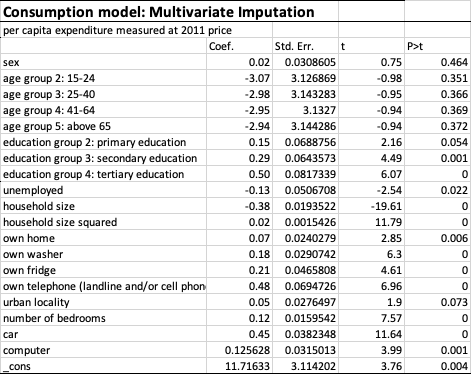
Table 1: summary statistics



Most variables are comparable across survey and year.

**RESULTS:**

Table 2:



The resulting imputed poverty rates are in Table 3. Notice that the imputed poverty rate (24.8%) using MI method from LFS 2016 is very close to the direct calculation (26.6%) from HBS 2016 (1.8% higher than MI result), which is also close to the poverty rated imputed using the HBS 2016 (24.5 %). The closeness between the imputed poverty rates from HBS 2016 and LFS 2016 gives us confidence on our imputation model. The third column reports the imputation results from a simple linear regression model which underperforms.

To conclude, by applying MI method to HBS 2016 and LFS 2008-2016 (2010 missing), we successfully imputed poverty rates for St. Lucia for the years lack of consumption data.

Table 3: poverty rates from direct calculation, MI method, and simple regression

|  |  |  |  |
| --- | --- | --- | --- |
|  | Direct calculation | MI | simple regression |
| HBS 2016 | 0.266 | 0.245 | 0.243 |
| LFS 2016 |  | 0.248 | 0.217 |
| LFS 2015 |  | 0.390 | 0.187 |
| LFS 2014 |  | 0.387 | 0.205 |
| LFS 2013 |  | 0.285 | 0.255 |
| LFS 2012 |  | 0.401 | 0.241 |
| LFS 2011 |  | 0.299 | 0.282 |
| LFS 2009 |  | 0.314 | 0.276 |
| LFS 2008 |  | 0.286 | 0.277 |

Source: Calculated by the author.

For the detailed calculation, see the attached excel sheet.

**Reference:**

[1] <https://stats.idre.ucla.edu/stata/seminars/mi_in_stata_pt1_new/>

[2] <https://www.stata.com/manuals/mimiimputemvn.pdf#mimiimputemvn>

[3] <https://academic.oup.com/wber/article/30/3/475/2449745>

[4] <https://pdfs.semanticscholar.org/678f/e7e47cb3067712ea380a6d0fcd382031c283.pdf>

[5] <http://documents.worldbank.org/curated/en/763111468164962978/How-survey-to-survey-imputation-can-fail>