Multiple Imputation on EFH

EFH Team

Outline

I. Multiple Imputation (MI)
   1. What is it? How do we do it? How do we use it?
   2. Methods available

II. MI on the Survey of Household Finances (EFH)
   1. A crash introduction to EFH
   2. Practical adjustments
   3. Choosing a method

III. Results and Discussion
   1. Empirical results
   2. MI on surveys (Spain's experience)

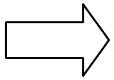
MI: What is it?

- In empirical applications researchers must work with incomplete datasets.
- A “solution” to these is Multiple Imputation (MI)
- MI implies changing our incomplete dataset with a set of complete ones.

Original

m=0

obs	*X*	*Y*	*Z*
1			
2			NA
3			
4	NA		
	NA		
		NA	
		NA	
	NA		NA
n			



Imputaciones

m=1

obs	*X*	*Y*	*Z*
1			
2			
3			
4			
n			

m=2

obs	*X*	*Y*	*Z*
1			
2			
3			
4			
n			

m=3

obs	*X*	*Y*	*Z*
1			
2			
3			
4			
n			

MI: How do we do it?

- The methods used here for doing MI relies on the assumption Missing At Random (MAR):
  - "…the probability of missing data on Y is unrelated to the value of Y, after controlling for other variables in the analysis." (Allison, 2002)
- In contrast to Missing Completely At Random (MCAR) in which discarding the missing observations produces no changes to the analysis, while under MAR dropping missing observations could give biased results.
- However, Missing Not At Random (MNAR) in which the probability of missing is related to the value, implies that a model for missing pattern should be considered for that analysis. This is the case of the Heckman's selection model.
- Validity of MAR, MCAR or MNAR is beyond the scope of this analysis.

MI: How do we do it?

- To get a more sensitive idea of MI, consider as an example the Hot-Deck procedure
  - As in the diagram we have 3 variables (X, Y, and Z) for which some of the individuals (or firms, etc.) do not have fully-observed records.
  - Consider firm k, for which we do not have an X value.
  - For other firms we do observe X values so we draw one those and we impute that selected value to firm k.
  - Repeating the m-times, we have m simulated-values for missing the X value of firm k.
- More general methods such as Multivariate Normal relies on a joint distribution which is estimated by observed data. Imputed values are obtained from simulations of the parameters.

MI: How do we use it?

- As stated above, the application of MI will lead to have m complete-datasets.
- Inference should be done using Rubin's steps
  - Compute parameters of interest in each complete-dataset.
  - Consider the average effect and its significant under an adjusted standard error. The latter is a weight sum of the Within and Between imputation-variances.
  - Within is obtained as the average variances along imputed datasets meanwhile Between is the variance of estimated parameters across imputed datasets.
  - The asymptotic distribution of those parameter is t-Student, in which the degrees of freedom depend on the number of imputed datasets, and the ratio within-between variances.

MI: Methods available

- Hot-Deck
  - Replace missing observations with an observed one taken randomly within a specific group: males with college education.
  - Individual within each group must be as homogenous as possible, also each group should have enough data
  - Informal conversations: SE's biased toward zero
- Univariate
  - Regress variable with missing observations on exogenous variables with no missing.
  - Draw posterior distribution of estimators (beta & sigma) and "predict" missing values.

MI: Methods available

- Chained Equations
  - Based on Univariate method, but with the possibility of having missing values in exogenous variables.
  - Reversing equations, missing values of “exogenous” variables could be computed.
  - It is difficult to get the joint distribution.
- Normal
  - Assume Multivariate Normal.
  - Estimate parameters using EM algorithm (initial value)
  - Draw imputations using Data Augmentation procedure.
  - Theory relies on the convergence of EM.

EFH: A crash introduction

- Chilean households surveys: CASEN, and EPS.
  - CASEN was created to measure poverty.
  - EPS was created to evaluate the pension system.
- At the Central Bank we have been using these surveys to analyze financial fragility of households.
- However, CASEN and EPS were not created for this purpose. We need new sample designs.
- In 2007, we started a new survey designed for our purposes (household balance-sheet). We followed mostly the experience of Spain.
- Household Survey of Finance (aka EFH).

EFH: A crash introduction

- EFH has different levels of information
  - Personal information of each member of the household. For example: age, year of education, labor income.
  - Aggregate information of the household. For example: value of assets (e.g: cars, house, financial instruments), debts (mortgage, consumer loans, educational loans, etc.)
- Some variables of interest are irrelevant for a set of households in the sample.
  - Some households have only credit from retail-firms. It is likely that they do not have access to banking credit.
  - Few households have personal savings invested in financial instruments such as stocks, or bonds.

EFH: Practical adjustments

- Using conditional methods, we could attach the constraints to the imputation procedure.
  - We are able to impute labor income for each member of the household, considering only individual level variables.
  - At the household level, we could impute “banks loans” in a sub-sample of households that declared to have that kind of debt. We use as exogenous variables age, years of education, and gender of interviewee.
  - We impute “debt in retail companies” with a different sub-sample but with the same exogenous variables.
  - But we cannot impute “debt in retail companies” with “banks loans” because sub-samples may be different.

EFH: Practical adjustments

- Dividing households intro groups
  - Suppose that a household without a house, we could pretend that the value of its house is “zero”. However, that will affect the correlation between value of the house and total amount of debts.
  - Following Alfaro and Fuenzalida (2009) we extend the analysis of creating groups based on dummy variables.
  - Those dummies indicate if a particular household reports financial assets, banking debt, real estate property and/or educational expenditure.
  - Using that information six groups are created. The following tables show the differences between groups.

EFH: Practical adjustments

Variable	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Total
Financial Assets	YES	YES	YES	YES			
Real Estate Property		YES	YES		YES	YES	
Banking Debt[1]			YES	YES	YES	YES	
Educational Expenditure[2]						YES	
Observations							
Fully-observed	755	1,072	66	234	124	224	2475
With missing values	218	725	225	177	73	128	1546
Total	973	1,797	291	411	197	352	4021

Source: Authors' calculation based on EFH 2007.

[1] Banking debt includes credit card, credit lines, and consumer loans.

[2] Educational expenditure includes only tuition.

EFH: Descriptive Statistics (m=0)

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Total
Income	737.113	1470.566	2966.450	1262.097	1360.589	1754.282	1399.483
	(40.200)	(74.346)	(251.045)	(73.300)	(121.484)	(114.781)	(42.350)
Debt	817.717	3790.253	20731.190	4533.358	9439.887	16276.730	5373.159
	(144.160)	(285.284)	(1615.745)	(405.820)	(1208.236)	(1179.599)	(225.699)
Cons Debt	785.167	626.267	6927.814	4789.503	5205.933	6934.836	2197.240
	(140.899)	(89.336)	(807.665)	(414.642)	(654.368)	(581.711)	(110.316)
Assets	7865.025	57063.080	121545.200	22365.740	72822.870	69133.500	47298.710
	(1617.420)	(2589.219)	(7885.144)	(4594.891)	(9058.544)	(4686.999)	(1609.238)

Source: Authors' calculation based on EFH 2007.

Note: Unweighted Mean and SE (in parenthesis)

EFH: Practical adjustments

- Keeping the observed interval
  - Some methods of imputation rely on the simulation of the parameters. Thus the support of imputed values depends on the distributional assumptions.
  - For example, for a model with X and Y variables, the bivariate normal distribution assumption implies that the support of both X and Y is the real line.
  - In practical applications variables are bounded. For that we consider the logistic transformation in order to preserve the observed interval.
  - Let M and m be the maximum and minimum of variable X, then we consider Y=logit[X/(M-m)] as the new scaled variable. Clearly the support of Y is the real line.

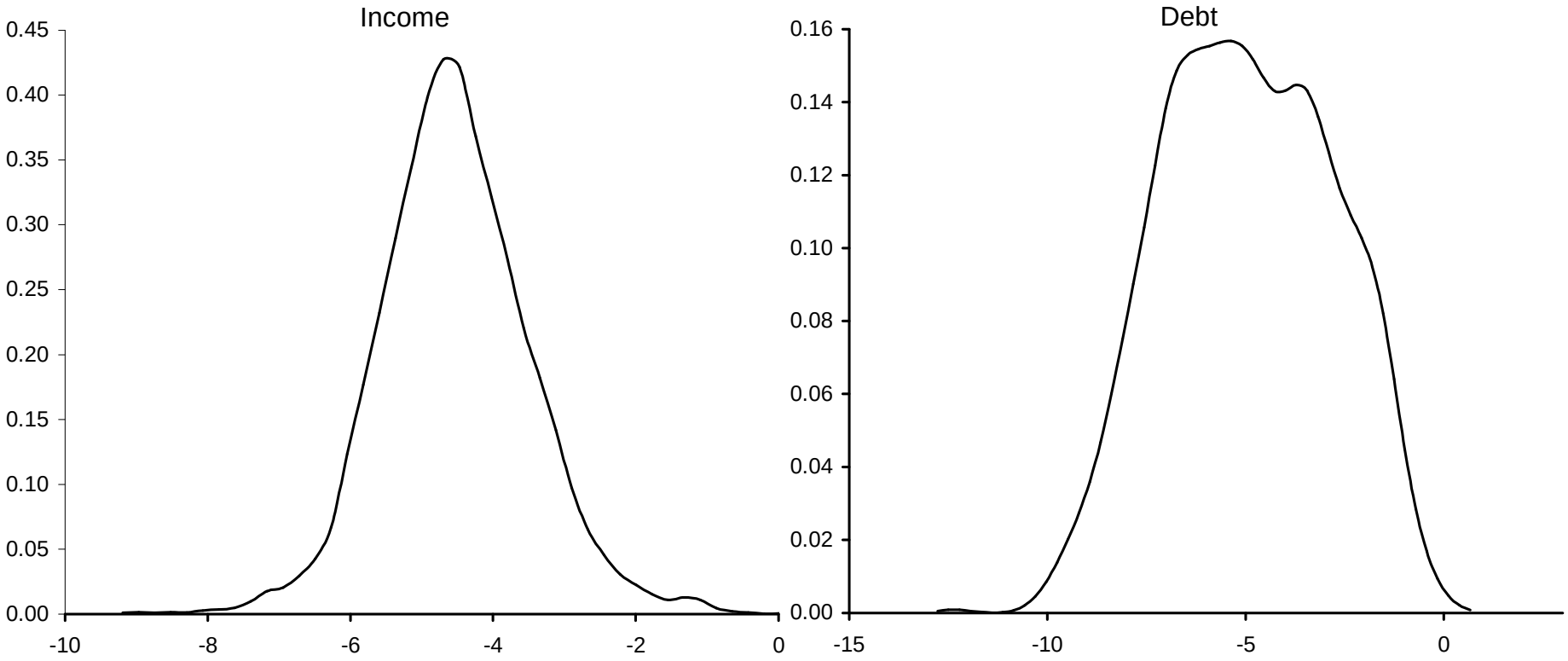
EFH: Choosing a method

- Desired properties
  - Proper: Rubin (1987) defines the proper methods as those that consider the fact that imputed values cannot reduce sample variance.
  - Replicable: anyone should able to redo it.
  - Efficient: it should consider all available information.
- Practical points
  - Hot Deck is a proper method, but in practices reduces the sample variance due to the finite-sample problem.
  - Practical replication means implemented in a well-known package. Hopefully with corporate support.
  - Finally, distributional assumptions behind the method should match observed data.

EFH: Choosing a method

- The Multivariate Normal Approach (MNA)
  - MNA is a well-known process for MI (Schafer, 1997).
  - MNA is an efficient method given that is based on EM. Data Augmentation (DA) step uses simulated parameters to get imputed values.
  - MNA has a strong assumption: joint distribution of the variables is multivariate normal.
  - This means that the support of each variable is the real line. This fact is "solved" with logistic transformation.
  - Joint distribution is complicated but we could rely on Quasi-Maximum Likelihood (QML) to consider general standardized distributions
  - Anyway, logistic variables look "close" to normal.

EFH: Choosing a method



OUT: Empirical Results

- Overview
  - Each group gets convergence on the EM procedure after a small number of iterations.
  - Given non-monotone patterns in the data, sweep procedure was not available.
  - No failure in DA process for 30 imputed datasets
  - Most of the variables end up with similar statistics as the observed one. This is explained by the small fraction of missing observations.
  - There is a small increment in SE for aggregate variables. Looking at each variable the increment is more significant.

OUT: Descriptive Statistics (m=5)

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Total
Income	738.646	1515.510	3715.573	1271.046	1394.849	1772.441	1478.336
	(40.258)	(75.730)	(338.650)	(73.834)	(123.958)	(115.772)	(46.527)
Debt	769.082	3573.028	19625.420	4633.153	8420.255	15635.240	5508.784
	(132.602)	(260.698)	(1502.710)	(394.967)	(997.976)	(1061.287)	(218.860)
Cons Debt	738.375	614.582	6700.097	4721.598	4545.621	6876.384	2245.492
	(129.415)	(83.688)	(717.897)	(387.108)	(559.029)	(637.629)	(108.782)
Assets	8277.452	64820.520	131214.800	22694.350	73448.350	73967.630	52972.930
	(1567.251)	(2903.901)	(8519.436)	(4408.042)	(8962.189)	(5683.443)	(1770.543)

Source: Authors' calculation based on EFH 2007.

Note: Unweighted Mean and SE (in parenthesis) using 5 imputed datasets

OUT: Spain versus Chile

	Spain	Chile
Procedure		
Method	Chained equations and Hot-Deck	Multivariate normal
Convergence criteria	Euclidean norm of Median and IQR	EM
Variables		
Type	Continuous, binary and multinomial	Continuous only
Transformation	Logarithm	Logistic
Restrictions	Bounded	Scaled
Others		
Model	Sets of covariates	Groups
Software	SAS	Stata