Multiple Imputation with Remittances Data

Ben Mazzotta

Thursday, October 02, 2014

Overview

IBGC is benchmarking the cost of cash worldwide. In order to do that, we rely on estimates of the rate of international remittances, domestic remittances and cash-out transcations in countries around the world. Missing data is a gigantic problem with remittances and cash transactions.

Multiple imputation, first described in Rubin (1987), improved regression estimates from datasets corrupted by missing observations. Software to accomplish this in R is available in open source, using packages `Amelia` and `Zelig` from the Comprehensive R Archive Network (Honaker, King and Blackwell, 2010; Owen, Imai, King and Lau, 2010; R Core Team, 2014).

Cookbook

1. Impute financial inclusion and prices of cash transactions.
2. Model national cash transaction volume.
3. Model national cash fee, transit and time value from transaction volume.

That (preceding) is the cookbook I followed. It produced these estimtes using Rubin's Rules (1987) for scalar estimates from multiple imputation.

```r
setwd("../data")
require(mi); require(Amelia); require(ggplot2); require(data.table)
load("cashfees_miviaAmelia01.Rdata")
```

```r
summary(feeshat)
```

```
##
## Amelia output with 12 imputed datasets.
## Return code:  1
## Message:  Normal EM convergence.
##
## Chain Lengths:
## --------------
## Imputation 1:  452
## Imputation 2:  265
## Imputation 3:  719
## Imputation 4:  136
## Imputation 5:  190
## Imputation 6:  212
## Imputation 7:  156
## Imputation 8:  122
## Imputation 9:  655
## Imputation 10:  232
## Imputation 11:  470
## Imputation 12:  672
```

```
##
## Rows after Listwise Deletion:  71 
## Rows after Imputation:  144 
## Patterns of missingness in the data:  11 
##
## Fraction Missing for original variables: 
## -----------------------------------------
##
##            Fraction Missing
## iso3c              0.078818
## cashprice          0.571429
## country            0.000000
## year               0.004926
## wb3c               0.078818
## rec_paymt          0.290640
## rec_remit          0.290640
## rec_wage           0.290640
## acc_active         0.290640
## payXrem            0.290640
## iso2c              0.000000
## income             0.078818
## region             0.078818
## lending            0.078818
## gdp                0.078818
## gdpreal            0.078818
## gdpcap             0.078818
## pop                0.078818
## remit_MM           0.201970
## remit_TRX          0.201970
## cashfees           0.650246
##
## Post-imputation transformed variables: 
## -----------------------------------------
##
##                                        Transformations
## cashfees = rec_paymt/rec_remit * remit_TRX * cashprice
```

I was not able to calculate the mean of 12 imputations with Amelia efficiently. Instead it's faster to write to CSV and then summarize by country.

Amelia makes it very easy to give the mean of each variable; but not to collapse imputations to a singe point estimate. That is perhaps the point: if you're modeling something with MI, feed the imputations into your model and then predict the values you want to observe.

So: since we have lots of missing data on cashprices, let's rewrite with a different cookbook.

1. Cash prices as a function of GDPcap, ATMdensity, and active accounts.
2. MI of cash prices, model inputs, and covariates rho, nu and X. == MI and model fit stage
3. Linear combination of rho, nu, X, and cashprice will give you cashfees. == simulate stage

At 3,

$$fees = \frac{\nu}{\rho}\hat{X}_{rem}p_{cash}$$

Problem

The algorithm will not converge with variables that are linear multiples of one another. It is brittle to collinear sparse data.

Hierarchical imputation and regression will yield purely speculative results. The variances of parameter estimates will not be accurately reported once imputations are reduced to estimated values.

More technically: the imputation procedure fits a distribution and not a point estimate. These estimates are great for multivariate regression–specifically, they are efficient and unbiased for applications with OLS and perhaps also generalized linear models. The point estimates so derived, on the other hand, are not reflective of the full distribution. Substituting predicted values in among independent variables for regression analysis is qualitatively different from multiple imputation, in that it overestimates the precision of the estimates.

So: recipe A does not work; but recipe B is fine.

Recipe A (fail)

1. Estimate some transaction price and market volume data from sparse, observed data.
2. Join to geography and economy data.
3. As a final step, impute the missing values across derived variables.

Recipe B (perhaps)

1. Join as much data as is available concerning geography and financial inclusion.
2. Impute missing values.
3. Combine imputed dataset according to a linear model.
4. Report predicted values for countries in the linear model, including those that lack some crucial finanical inclusoin and economy data.

Methodology notes

Multiple imputation (MI) was invented in the 1980s (Rubin 1987), based on work dating to 1976. Most social science data suffers from missing observations. MI is a statistical approach that estimates the *distribution of the missing data* through simulation, rather than approximating the missing data with a point estimate. Important advances in computational approaches were made in the 1990s. By the early 2000s, two main algorithms had been developed for MI inference: IP and EMis. Expectation maximization with importance sampling (EMis) is the main method used by the `{Amelia}` software.[1]

MI has important advantages relative to listwise deletion, related to bias and efficiency. Deleting obserations under typical condition leads to bias in model estimates. It is also inefficient because information in the partial observations is omitted from model fit. The size of the difference has to do with both the ratio of the data that are missing, and also whether data are missing completely at random (MCAR), missing at random (MAR), or not at random (MNAR). The type of missing data problem depends whether the rate of missing data correlates with observed or unobserved data, and to what extent. With data missing completely at random (MCAR), missingness does not systematically bias results. With data systematically or purposively removed from a dataset (MNAR), multiple imputation cannot accurately estimate the original.

The procedure described in Honaker and King (2010) has three steps: impute, fit, and summarize. *Imputation* requires 3-12 iterations to converge using the EMis algorithm. Under Amelia software, the imputation software stores the results of each iterations in a list of simulated data. From a matrix of n=1000 observations

[1] Honaker and King explain EMis is reasonably accurate and fast, as compared to the more computationally intensive and technically demanding imputation-posterior (IP) algorithm.

with 7 imputations, the imputation step would produce a list of 7 observations, each of n=1000. During the *fit* stage, each of these 7 elements would then be fit to a model, such as a conditional mean estimation or a regression analysis. Every model parameter would be estimated seven different times (the number of imputations). Amelia automates the *summary* step for every model parameter and variance.

Standard statistics such as p-values can be reported from these estimates; though that is not necessarily the best approach to summarizing the quality of results (Gelman 2013).

`{Amelia II}` has important advantages relative to other R packages such as `{mi}` and `{mice}`. It takes a structured approach to time series cross-sectional data. Time series data are fit with polynomial smoothing.[2] Cross-sectional identifiers, such as "country", or "individual" are also fitted with fixed effects for imputation, if not in the model fitting step itself. For values with skewed or logistic distributions, logarithmic transformations yield passable approximations for most social science purposes.

No priors are incorporated into this analysis. `{Amelia II}` is capable of implementing priors where data sparsity presents a challenge and analysts have some beliefs about the distribution of missing data. Priors can be based on experience, logical deduction, and observed distributions of similar variables.

MI packages typically fit the missing data to a multivariate normal distribution. Categorical, discrete, ordinal and skewed variables can be accommodated by mapping the distribution in a number of simple ways. `{Amelia II}` makes appropriate assumptions about appropriate distributions for MI based on the structure of data types in R data frames, such as integers, factors, and floats. Discrete distributions are approximated as if they were continuous. Categorical variables are transformed into sets of binary variables and estimated using a logistic function.

□

Bibliography

1. Gelman, A, and H Stern. 2006. "The Difference between 'significant' and 'not Significant' Is Not Itself Statistically Significant." *The American Statistician*. http://www.tandfonline.com/doi/abs/10.1198/000313006X152649.

2. Honaker, J, and G King. 2010. "What to Do about Missing Values in Time-series Cross-section Data." *American Journal of Political Science* 54(2):561–581. http://onlinelibrary.wiley.com/doi/10.1111/j.1540-5907.2010.00447.x/full.

3. James Honaker, Gary King, Matthew Blackwell (2011). Amelia II: A Program for Missing Data. *Journal of Statistical Software*, 45(7): 1–47. http://www.jstatsoft.org/v45/i07/

4. Matt Owen, Kosuke Imai, Gary King and Olivia Lau (2013). Zelig: Everyone's Statistical Software. R package version 4.2-1. http://CRAN.R-project.org/package=Zelig

5. Rubin, DB. 1987. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York.

6. R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

See also

1. King, G, and J Honaker. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." American Political Science Review 95(1):49–69. http://journals.cambridge.org/abstract_S0003055401000235.

Benjamin D. Mazzotta is a postdoc at the Institute for Business in the Global Context ([IBGC](http://fletcher.tufts.edu/ibgc)), The Fletcher School, Tufts University.

[2] Order is limited to $k \leq 3$.