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.

Methodology notes

Multiple imputation (MI) was invented in the 1980s (Rubin 1987). Most social science data suffers from missing observations. MI is a statistical approach that discovers the distribution of the missing data through simulation, rather than approximating the missing data with a single 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. The benefits of MI are even greater when data are missing only partially at random. With data missing completely at random (MCAR), missingness does not systematically bias results. With data systematically and strategically removed from a dataset (missing not at random, or MNAR), multiple imputation cannot recreate the original dataset purposively destroyed.

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 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]

[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.

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

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.

□

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.