Multiple Imputation of Missing Values in Economic Surveys: Comparison of Competing Algorithms

Masayoshi Takahashi*
National Statistics Center, Tokyo, Japan mtakahashi@nstac.go.jp

Takayuki Ito
National Statistics Center, Tokyo, Japan titou@nstac.go.jp

Due to the missing values in a dataset, not only available data size shrinks and efficiency decreases, but also bias is likely to exist if there is a systematic difference between respondents and non-respondents. Therefore, we almost always need to deal with missing values in one way or another, and multiple imputation has been proposed as a method to handle missing data. While the theoretical concept of multiple imputation is simple and has been around for decades, the implementation is difficult and contentious because making a random draw from the posterior distribution is a complicated matter. As a result, there are many competing computational algorithms in software. The original version of multiple imputation proposed by Donald B. Rubin is based on the well-known Bayesian computational algorithm, called Markov chain Monte Carlo (MCMC), also known as joint modeling, such as R Package Norm and SAS PROC MI. Recently, two alternative algorithms have been proposed. One is Fully Conditional Specification (also known as chained equations), such as R Package MICE (Multivariate Imputation by Chained Equations), SPSS Missing Values, and SOLAS. Another emerging algorithm is the Expectation-Maximization with Bootstrapping (EMB), such as R Package Amelia II. Thus, we have one theory of multiple imputation, but many ways to conduct it. To this date, however, it is unknown which of the three algorithms outperforms the others under what circumstances. In this paper, we describe the mechanisms of these multiple imputation algorithms and compare their performance in a variety of situations to determine which algorithm is best suited to the imputation of missing values in official economic statistics. Each of the algorithms will be judged in many dimensions, such as accuracy in comparison with the true values, computational efficiency, and so on. A real application on Turnover in the EDINET (Electronic Disclosure for Investors' NETwork) data will be used to illustrate the arguments.

Key Words: Official economic statistics, Markov chain Monte Carlo (MCMC), multivariate imputation by chained equations (MICE), expectation-maximization with bootstrapping (EMB)