Multiple Imputation in the Complex National Nursery Survey Data by Fully Conditional Specification

Wan Xu
Graduate Research Assistant
Food and Resource Economics Department
University of Florida
Gainesville, FL 32611
Email: wanxu@ufl.edu

Hayk Khachatryan
Assistant Professor
Food and Resource Economics Department
Mid-Florida Research & Education Center
University of Florida
2725 Binion Road, Apopka, FL 32703
Email: hayk@ufl.edu

Selected Poster Prepared for Presentation at the Agricultural & Applied Economics Association's 2014 AAEA Annual Meeting, Minneapolis, MN, July 27-29, 2014.

Multiple Imputation in the Complex National Nursery Survey Data by Fully Conditional Specification

Wan Xu and Hayk Khachatryan

Food and Resource Economics Department, University of Florida, Gainesville, FL 32611

Introduction

Missing data problems are always prevalent and inevitable in survey data such as national nursery survey when we investigate the primary factors influencing sales revenue. It often bias the statistical results and result in invalid inferences due to significant missing information in the observations. Instead of imputing each missing value with a single known, multiple imputation (MI) is a useful and popular method in handling missing data by filling-in a set of simulating values to account for uncertainty in the missing data [1] . MI has three steps including filling, analyzing and pooling. Different methods are used for different patterns of missing data [1]. Although the binary variable could be imputed by the Markov Chain Monte Carlo (MCMC) method with rounding approximation, some literatures have stated that it violated the normality assumption and such rounding method can even cause bias in the estimates [2].

Objectives

Firstly, a more flexible and semi-parametric imputation approach-fully conditional specification (FCS) method [3] is applied to estimate the sales revenue in the U.S. national nursery industry, which assumes a joint distribution existed for all variables to impute the missing data for both continuous and discrete variables (i.e. binary, nominal categorical, and ordered categorical variables) in the complex national nursery survey data. Secondly, by comparing with the MCMC method based on rounding approximation, we will show that the FCS method performs better in terms of overall performance and efficiency measure.

References

1. Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
2. Horton, N. J., S. R. Lipsitz, and M. Parzen. 2003. “A potential for bias when rounding in multiple imputation.” *American Statistician* 57: 229-232.
3. van Buuren, S. 2007. “Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification.” *Statistical Methods in Medical Research* 16: 219–242.
4. Heitjan, F. and R. J. A. Little. 1991. “Multiple Imputation for the Fatal Accident Reporting System.” *Applied Statistics* 40:13–29.

Methods

FCS Predictive Mean Matching Method **(*Continues Variable*) [4]**

- Fit a linear model for each continuous missing variable (x_i) given other variables as covariates, and obtain $\hat{\beta}$ and corresponding covariance matrix $\widehat{V}_i = \hat{\sigma}_i^2 S_i = \hat{\sigma}_i^2 (X'X)^{-1}$
- Simulate new parameters $\hat{\beta}^*$ and $\hat{\sigma}_i^{2^*}$ from the posterior distribution of the parameters $\hat{\beta}$ and $\widehat{V}_i$: $\hat{\sigma}_i^{2^*} = \hat{\sigma}_i^2(n_i$-k-1) /m, where n_i is the number of observed subjects for x_i, and m is a Chi-squared random variable with d. f. of n_i-k-1. $\hat{\beta}^* = \hat{\beta} + \hat{\sigma}_i^* U_i' Z$, where U_i is the upper triangular matrix in the Cholesky decomposition ($S_i = U_i' U_i$), and Z is the vector of (k+1) i.i.d. normal variables.
- Compute the predicted value for the continuous missing variable by: $x_i^* = \hat{\beta}_0^* + \hat{\beta}_1^* x_1 + \hat{\beta}_2^* x_2 + \cdots + \hat{\beta}_{i-1}^* x_{i-1} + \hat{\beta}_{i+1}^* x_{i+1} + \cdots + \hat{\beta}_k^* x_k$
- Generate a set of *d* observed subjects whose predicted values are nearly matching to x_i^*, and then fill-in the missing variables by random draw from d observed values.

FCS Logistic Regression Method **(*Discrete Variable*) [1]**

- Fit a logistic regression model for each binary missing variable given other variables as covariates, obtain $\hat{\beta}$ and corresponding covariance matrix $\widehat{V}_i$.
- Simulate new parameters $\hat{\beta}^*$ from the posterior distribution of the parameters $\hat{\beta}$ and $\widehat{V}_i, \hat{\beta}^* = \hat{\beta} + U_i' Z$
- Calculate the expected probability of missing values: $p_i = \frac{e^{\mu}}{1+e^{\mu}}$,
- Simulate μ from Uniform (0,1) distribution and set p_i as the cutoff
- Ordered logistic regression can be extended to impute the ordinal categorical missing variables.

Conclusion

We applied a semi-parametric FCS multiple imputation method to address for missing data problems in the national nursery survey, and analyzed the sales revenue in the U.S. national nursery industry In comparison of the MCMC method with the strict normality assumption. We showed that the FCS method is more robust and superior than the MCMC method. However, the further performance of the FCS method should still be thoroughly investigated by simulations. Since the FCS method is more flexible, different conditional distributions can be tailored for different types of covariates with missing information. An extension of exploring the performance of the FCS method under different conditional distributions would be useful and valuable.

Results

Table 1: Regression Result and Variance Information for FCS MI

Parameter	Regression			Variance		
Dep. var: log (sales)	Estimate	Std Error	Pr > |t|	Between	Within	Total
Intercept	12.882	0.110	<.0001	1.8E-04	1.2E-02	1.2E-02
Opreate_Other	0.670	0.199	0.001	1.9E-03	3.7E-02	4.0E-02
Forward Contracting	-0.005	0.323	0.989	3.2E-04	1.0E-01	1.0E-01
Firm Age	0.005	0.002	0.004	3.7E-08	2.9E-06	3.0E-06
Computer Tech. Usage	0.617	0.078	<.0001	6.8E-05	6.1E-03	6.1E-03
Employee	0.005	0.001	<.0001	2.0E-08	2.7E-07	2.9E-07
Trade Show	0.017	0.008	0.033	1.2E-05	4.8E-05	6.1E-05
Product Uniqueness	-0.044	0.079	0.579	1.2E-05	6.3E-03	6.3E-03
Region_Southeast	0.018	0.068	0.788	5.8E-06	4.7E-03	4.7E-03
Region_Northeast	-0.182	0.077	0.018	4.8E-05	5.9E-03	5.9E-03
Region_Pacific	0.271	0.078	0.001	4.2E-05	6.1E-03	6.1E-03
Region_Midwest	-0.035	0.086	0.684	3.2E-05	7.4E-03	7.4E-03
IPM Practice	0.027	0.009	0.004	1.0E-06	8.6E-05	8.7E-05

Table 2: Comparison of MI Efficiency: FCS vs MCMC

	FCS			MCMC		
Parameter	r	λ	RE	r	λ	RE
Intercept	0.017	0.017	0.998	43.026	0.981	0.911
Opreate_Other	0.056	0.054	0.995	2.868	0.768	0.929
Forward Contracting	0.003	0.003	1.000	0.023	0.023	0.998
Firm Age	0.014	0.014	0.999	0.423	0.311	0.970
Computer Tech. Usage	0.012	0.012	0.999	3.764	0.814	0.925
Employee	0.081	0.076	0.992	1.622	0.647	0.939
Trade Show	0.275	0.224	0.978	0.561	0.377	0.964
Product Uniqueness	0.002	0.002	1.000	0.426	0.312	0.970
Region_Southeast	0.001	0.001	1.000	0.164	0.144	0.986
Region_Northeast	0.009	0.009	0.999	1.312	0.596	0.944
Region_Pacific	0.008	0.007	0.999	2.927	0.772	0.928
Region_Midwest	0.005	0.005	1.000	0.310	0.246	0.976
IPM Practice	0.013	0.013	0.999	0.885	0.494	0.953

Note: r: Relative Increase in Variance, λ: Fraction of Missing Information, RE: Relative Efficiency

Contact Information

Wan Xu, Email: wanxu@ufl.edu