The evaluation of university teaching. An imputation procedure to recover for missingness [1]

La valutazione della qualità dei corsi universitari. Una procedura di imputazione dei dati mancanti

Isabella Sulis[2] Mariano Porcu
Dipartimento di Ricerche Economiche e Sociali
Università di Cagliari, e-mail: isulis@unipa.it; mrporcu@unica.it

Riassunto: Nel lavoro viene descritto un metodo di imputazione per la ricostruzione di unità parzialmente osservate nell'ambito di indagini sulla valutazione della didattica universitaria in cui vengono impiegate delle scale multi-item categoriali. La procedura di imputazione è basata sulla generazione casuale per ogni valore mancante di m valori plausibili condizionati rispetto ai valori osservati negli altri item (Rubin, 1987; Little and Rubin, 2002). Una simulazione ha permesso la validazione della procedura in termini di accuratezza nel riprodurre sia le distribuzioni bivariate degli item sia alcune stime di parametri di interesse.

Keywords: Multiple imputation, missing values, stochastic regression

1. Introduction

Missing data is a common problem in surveys on university teaching evaluation. The proposed method, based on stochastic regression analysis and multiple imputation, replaces missing data with random draws from the predictive distribution of each unobserved datum conditional upon a selected set of predictors (Little and Rubin, 2002). A simulation study has been carried out in order to evaluate the procedure.

2. A multiple imputation approach to recover for missingness

The method here proposed is an iterative procedure in three steps, which simultaneously generates plausible values for a set of I items ($i = 1, \ldots, I$) measured on a four category Likert scale. **(1)** In the first step a system of regression models is set up. Each equation predicts the unobserved value for missing unit X_{pi} ($p = 1, \ldots, n$) in item i $\{X_{i.obs}, X_{i.miss}\}$ from the distribution of the missing values conditional upon a set of J predictors. Values for $X_{i.miss}$ are thus generated from several conditional distributions according to the pattern of missing observations among the predictors. For each missing unit m plausible values ($\tilde{X}_{pi}$) are drawn from a *Multinomial*($\pi_{pi.1}, \ldots, \pi_{pi.K}$). The vector of probabilities from which the random draws are made is estimated using a *multinomial logit regression model*: $\hat{\pi}_{pi.k} = (\exp(\alpha_k + \boldsymbol{\beta}'_J \boldsymbol{x}_p)/(+\sum_{k=1}^{K-1} \exp(\alpha_k + \boldsymbol{\beta}'_J \boldsymbol{x}_p))$. **(2)** The

[1] This work is being carried on in a National Research Project – PRIN2005: 2005139210_001.
[2] Address of correspondence: V.le S.Ignazio da Laconi 78, 09123 Cagliari.

Table 1: *Bivariate distributions that statistically differ from the reference*

	I_1	I_2	I_3	I_4	I_5	I_6
I_1						
I_2	0					
I_3	0	0.04				
I_4	0	0.04	0			
I_5	0.18	0	0	0		
I_6	0	0	0	0.02	0.02	
$p-$value=0.05						

Table 2: *Estimation accuracy: Ratio between parameters (MIA/ reference)*

item	*20% missing values in each item*				*10% of missing values in each item*			
I_1	1.00	0.99	1.01	0.99	0.99	1.00	1.00	1.00
I_2	0.98	0.98	1.01	1.00	1.02	0.97	1.00	1.00
I_3	0.95	0.97	1.00	1.01	1.02	0.97	1.00	1.00
I_4	1.00	1.01	1.00	1.00	0.99	1.03	0.99	1.00
I_5	0.97	0.99	1.01	1.00	1.01	0.98	1.01	0.99
I_6	1.04	1.02	0.99	1.00	1.01	1.01	0.99	1.00

system of regression models is iterated S times until the whole data matrix $\bar{\boldsymbol{X}}\{\tilde{\boldsymbol{X}}, \boldsymbol{X_{obs}}\}$ is fully observed. At any iteration the response variable is not updated $X_i\{X_{obs}, X_{miss}\}$ while predictors are updated with values from previous iteration. **(3)** In this step all the *n* units $\bar{\boldsymbol{X}}\{\tilde{\boldsymbol{X}}, \boldsymbol{X_{obs}}\}$ are involved in the estimation process. The probability vector from which the m values are drawn is estimated moving from the distribution of $\bar{X}_i$ conditional upon the set of predictors $\bar{\boldsymbol{X}}_{n\times J}$ (Raghunathan, 2004). The proposed procedure has been applied in order to recover partially observed units in a survey on the evaluation of university courses carried out in the 2004-2005 academic year (Sulis, 2007). The imputation procedure has been jointly applied to a set of 6 categorical ordinal items and for each missing unit 50 random draws have been made. In order to validate the method in terms of distributional and estimation accuracy an increasing percentage of units have been set missing at random in each of the 6 items and then imputed. The *Chi-squared* test has been used in order to assess the discrepancy between the bivariate distributions - which are obtained crossing an item with all the others - in the 'reference data-set' and in 50 randomly imputed data-sets. Table 1 shows the relative frequencies of the bivariate distributions that statistically differ from the reference distribution (the rate of missingness is set equal to 20% in each item). The estimation accuracy has been evaluated by comparing the relative frequencies in each of the four categories in the *reference* data set and the ones estimated by adopting a *Multiple Imputation Analysis (MIA)* (Rubin, 1987). Simulation results for a rate of missing values in each item equal to 20% and 10% are depicted in Table 2.

References

Little R.J.A. and Rubin D.B. (2002) *Statistical Analysis with Missing Data, 2nd edition*, New York: John Wiley.

Raghunathan T.E. (2004) What do we do with missing data? Some options for analysis of incomplete data, *Annual Review Public Health*, 25, 99–117.

Rubin D.B. (1987) *Multiple Imputation For Nonresponse in Surveys*, John Wiley.

Sulis I. (2007) *Measuring students' assessments of 'university course quality' using mixed-effects models*, Ph.D. thesis, Università di Palermo.