CONTRIBUTI DI RICERCA CRENOS

# ASSESSING THE EFFECTIVENESS OF A STOCHASTIC REGRESSION IMPUTATION METHOD FOR ORDERED CATEGORICAL DATA

**Isabella Sulis**

**Mariano Porcu**

WORKING PAPERS

CUEC

2008/04

**CENTRO RICERCHE ECONOMICHE NORD SUD**
**(CRENOS)**
**UNIVERSITÀ DI CAGLIARI**
**UNIVERSITÀ DI SASSARI**

Il CRENoS è un centro di ricerca istituito nel 1993 che fa capo alle Università di Cagliari e Sassari ed è attualmente diretto da Raffaele Paci. Il CRENoS si propone di contribuire a migliorare le conoscenze sul divario economico tra aree integrate e di fornire utili indicazioni di intervento. Particolare attenzione è dedicata al ruolo svolto dalle istituzioni, dal progresso tecnologico e dalla diffusione dell'innovazione nel processo di convergenza o divergenza tra aree economiche. Il CRENoS si propone inoltre di studiare la compatibilità fra tali processi e la salvaguardia delle risorse ambientali, sia globali sia locali.
Per svolgere la sua attività di ricerca, il CRENoS collabora con centri di ricerca e università nazionali ed internazionali; è attivo nell'organizzare conferenze ad alto contenuto scientifico, seminari e altre attività di natura formativa; tiene aggiornate una serie di banche dati e ha una sua collana di pubblicazioni.

**www.crenos.it**
**info@crenos.it**

CRENOS – CAGLIARI
VIA SAN GIORGIO 12, I-09100 CAGLIARI, ITALIA
TEL. +39-070-6756406; FAX +39-070- 6756402

CRENOS - SASSARI
VIA TORRE TONDA 34, I-07100 SASSARI, ITALIA
TEL. +39-079-2017301; FAX +39-079-2017312

Titolo: ASSESSING THE EFFECTIVENESS OF A STOCHASTIC REGRESSION IMPUTATION METHOD FOR ORDERED CATEGORICAL DATA

ISBN: 978-88-8467-445-6

Prima Edizione: Marzo 2008

Via Is Mirrionis, 1
09123 Cagliari
Tel./Fax 070291201
www.cuec.it

# Assessing the effectiveness of a stochastic regression imputation method for ordered categorical data[†]

Isabella Sulis & Mariano Porcu[‡]
Università di Cagliari
& CRENoS

January 2008

## Abstract

The main aim of this paper is to describe a workable method based on stochastic regression and multiple imputation analysis (MISR) to recover for missingness in surveys where multi-item Likert-type scale are used to measure a latent attribute (namely, the *quality of university teaching*). A simulation analysis has been carried out and results have been compared in terms of bias and efficiency with other missing data handling methods, specifically: Complete Cases Analysis (CCA) and Multiple Imputation by Chained Equations (MICE). The authors provide also functions (implemented in R language) to apply the procedure to a matrix of ordered categorical items. Functions described allow: (i) to simulate missing data *at random* and *completely at random*; (ii) to replicate the simulation study presented in this work in order to assess the accuracy in distribution and in estimation of a multiple imputation procedure.

**Keywords:** multiple imputation analysis, validation process, MAR, MCAR, MICE.
**JEL Classification:** C15

[†] This research has been partially founded by the Fondazione Banco di Sardegna.

[‡] Address: Università degli Studi di Cagliari, Dipartimento di Ricerche Economiche e Sociali, Viale S. Ignazio 78, 09123 Cagliari. Email: isulis@unica.it; mrporcu@unica.it.

# 1 Introduction

When a data matrix displays empty observations, the solution of limiting the analysis to units not affected by missingness is the default procedure applied by many statistical softwares. This approach is known as the *Complete Case Analysis* (CCA). Unfortunately, this way to cope with missing information could bias the final results depending on the mechanism which has generated missing observations. Schafer (1997), recommends not to ignore a fraction of missing information higher than 5%. Furthermore, dealing with real data, a CCA could produce a selection bias; e.g., in an applied context such as the evaluation of quality of university teaching CCA could lead to bound the analysis on a specific group of students that are not representative of the overall population, specifically those who provide more attention in answering questionnaires. An alternative way to deal with the problem of missingness is to perform an *Available Case Analysis* (ACA) using the information provided by partially observed units for each of the variables. Nevertheless, this technique is not recommended in a regression analysis context where it may produce bias and less efficient estimates (Haitovsky, 1968). Considering the drawbacks of CCA and ACA another approach to deal with missingness is to recover empty observations with plausible values generated on the basis of some reasonable criteria. This is the frame of the *Multiple Imputation Analysis* (MIA) techniques (Rubin, 1987) which consist of imputing *M* plausible values for each missing value. In this way *M* complete data sets are generated and separately considered. Results obtained in each data set are next summarized in a single inferential statement using results provided by Rubin (1987).

This work proposes a multiple imputation analysis (MIA) based on *stochastic regression* (MISR) (Little and Rubin, 2002) to cope with missing values in surveys where variables measured on Likert-type scale (with the same number of response categories) define the same underlying attribute(Sulis and Porcu, 2007). The MISR approach replaces missing values with random draws from a distribution whose parameters have been estimated by parametric regressions. The method uses both the information provided by the observed values of the variables affected by missingness and the multivariate structure of the data in order to recover partially observed units. The application here implemented, simultaneously tackles with data-sets affected by different rates of missingness and by two different missing data generating mechanisms: *Missing Completely at Random* (MCAR) and *Missing at Random* (MAR) (Rubin, 1976). The procedure has been validated according to the criteria of *Accuracy in Distribution* (AD) and *Accuracy in Estimation* (AE) (Chambers, 2001;

Table 1: Example of data matrix affected by missingness

| unit | items | | | | | | |
|---|---|---|---|---|---|---|---|
| ♯ | $I_0$ | $I_1$ | $I_2$ | $I_3$ | $I_4$ | $I_5$ | $I_6$ |
| 1 | · | 1 | 2 | 2 | 1 | 1 | · |
| 2 | 4 | 4 | 4 | · | 3 | 4 | 4 |
| 3 | 3 | 3 | · | 4 | 2 | · | 3 |
| 4 | 3 | 3 | 3 | · | 3 | 3 | 3 |
| 5 | · | 3 | · | 4 | 2 | · | 3 |

Borgoni and Berrington, 2004). All simulations have been carried out by imputing the data-sets using multiple imputation by stochastic regression (MISR) and multiple imputation by chain equations (MICE). Results arisen from both missing data multiple imputation methods have been compared with the CCA. Functions to replicate the simulation study are provided in the Appendix.

# 2 An imputation procedure to recover for missingness

This MISR procedure works in two steps. Let's define a data matrix with $n$ units and $p$ items. For the sake of simplicity, the method is described by supposing that responses are recorded on a 4 (K) category Likert-type scale (we suppose that items are questions on some teaching attributes): 1 =*Definitely No – DN*, 2 =*More No than Yes – MN*, 3 =*More Yes than No – MY*, 4 =*Definitely Yes – DY*. Table 1 (for $i = 5$ obs. and $p = 7$ items) shows the first five units.

## 2.1 Step 1

The procedure starts by building up for each unit $i$ the distribution of the relative frequencies of ratings in each of the $K$ response categories, as Table 2 shows. From Table 1 arises that the rate of response for unit ♯1 is:

$$DN = \frac{3}{5} = 0.60; \quad MN = \frac{2}{5} = 0.40; \quad MY = DY = \frac{0}{5} = 0.00.$$

Unobserved items for unit $i$ are replaced by drawing values from a *Multinomial* distribution with parameters set equal to the relative frequencies of ratings observed for each category (see Table 3).

Table 2: Response pattern for each unit

| unit | counts: 4 categories | | | |
|---|---|---|---|---|
| ♯ | DN | MN | MY | DY |
| 1 | 3 | 2 | 0 | 0 |
| 2 | 0 | 0 | 1 | 5 |
| 3 | 0 | 1 | 3 | 1 |
| 4 | 0 | 0 | 6 | 0 |
| 5 | 0 | 1 | 2 | 1 |

Table 3: Parameters for the Multinomial random draws

| unit | vector of parameters of the Multinomial distribution | | | |
|---|---|---|---|---|
| | $\pi_{i1}$ | $\pi_{i2}$ | $\pi_{i3}$ | $\pi_{i4}$ |
| 1 | 0.60 | 0.40 | 0.00 | 0.00 |
| 2 | 0.00 | 0.00 | 0.17 | 0.83 |
| 3 | 0.00 | 0.20 | 0.60 | 0.20 |
| 4 | 0.00 | 0.00 | 1.00 | 0.00 |
| 5 | 0.00 | 0.25 | 0.50 | 0.25 |

For unit $\sharp 1$ unobserved items $I_0$ and $I_6$ (see Table 1) are replaced by generating $M \times 2$ values from a *Multinomial* $(0.60, 0.40, 0.00, 0.00)$. These values fill in the two unobserved records $I_0$ and $I_6$ in the $M$ data-sets. Similarly, the unobserved values in items $I_0$, $I_2$ and $I_5$ for unit 5 are replaced by drawing $M \times 3$ values from a *Multinomial* with parameters $(0.00, 0.25, 0.50, 0.25)$. Table 4 shows the first $M$ random draws generated for unit $\sharp 1$

Table 4: Values for the imputation of missing items $I_0$ and $I_6$ (unit $\sharp 1$)

| Missing Items | ***M*** randomly generated data sets | | | | | | |
|---|---|---|---|---|---|---|---|
| unit $\sharp 1$ | 1 | 2 | 3 | 4 | 5 | ... | $M$ |
| $I_0$ | **1** | **1** | **1** | **2** | **2** | ... | **1** |
| $I_6$ | **1** | **1** | **1** | **1** | **1** | ... | **1** |

The first step uses just the information provide from unit response pattern in order to generate plausible imputed values for each record affected by missingness.

## 2.2 Step 2

Lets consider one of the $M$ imputed data-set obtained as described in Step 2.1. In this second step a stochastic regression approach is used: $p$ regression equations are specified (one for each of the items in the data-set) where each of the $p$ item is considered as a response variable whose values depend upon the set of $(p-1)$ remaining predictors, and $(p-1)$ times as predictor (e.g. the value of $I_0$ is assumed to depend on $I_1 - I_6$, the value of $I_1$ is assumed to depend on the predictors $I_0, I_2 - I_6$ and so forth).

By adopting a *proportional odds* logistic regression model (Agresti, 2002) to predict the probability to answer a category lower rather than greater than $k$

$$\text{logit}[P(Y \leq k|\boldsymbol{x})] = \alpha_k + \boldsymbol{\beta}'\boldsymbol{x}, \tag{1}$$

the probability to provide a response in each category is expressed as

$$\pi_k = \Big[\frac{\exp(\alpha_k + \boldsymbol{\beta}'\boldsymbol{x})}{1 + \exp(\alpha_k + \boldsymbol{\beta}'\boldsymbol{x})} - \frac{\exp(\alpha_{k-1} + \boldsymbol{\beta}'\boldsymbol{x})}{1 + \exp(\alpha_{k-1} + \boldsymbol{\beta}'\boldsymbol{x})}\Big]. \tag{2}$$

The $\hat{\alpha}_k$s and $\hat{\boldsymbol{\beta}}_k$s are estimated using the complete data-set generated in Step 2.1. Next, for each unobserved unit, 1 random draw is generated from a *Multinomial* distribution with vector of parameters $[\hat{\pi}_1(\boldsymbol{x}), \ldots, \hat{\pi}_K(\boldsymbol{x})]$ estimated using equations (1) and (2). The procedure is iterated in each of the $M$ data-sets. Conditional regression methods allow enormous flexibility for predicting missing values (Raghunathan, 2004). They consider both the information provided by the observed values of the variables affected by missingness and the multivariate structure of the data.

## 3 An application of MISR to a survey on university course quality

MISR has been tested on data provided by the survey on university course quality carried out at the University of Cagliari. Specifically the data-set concerns questionnaires gathered at the first level degree scheme at the Faculty of Engineering in 2004-05 academic year. The study aims to assess the extent to which the imputation procedure fulfills the two criteria of AD and AE. To validate the method with respect to a benchmark data-set, a complete data set has been built up discarding all the records with missing observations. The final CD contains 1725 records concerning 24 courses and 10 items: 7 items are related to student's evaluation of lecturer ($L_1 - L_7$), one is related to students' overall satisfaction towards the course ($S$), one is related to student's attendance at classes ($A$), one to student's interest toward the topic ($I$). Courses evaluated by less than 50 students have been not considered in the analysis. The final data set contains 1725 records on 24 courses and it will be called in the next sections *complete data-set* (CD). The imputation procedure has been applied to seven items related to student's evaluation of lecturer ($L_1 - L_7$) and to the item concerning students' overall satisfaction ($S$) (Table 5); all are measured on a four-category Likert scale: *Definitely No*, *More No than Yes*, *More Yes than No*, *Definitely Yes*.

Five data sets with an increasing rate (5%, 10%, 15% ,20%, 25%) of missing units have been generated deleting observations from the CD. Records have been set missing according to two different mechanisms: MCAR and MAR. The two mechanisms have be simulated using function miss.CAR (see § A.3) and miss.AR (see § A.4). The former sets an observation missing independently from any response scheme (see §A.4). The latter fixes the probability to set an observation missing on the bases of some significant units' covariates. In the study two students' covariates, both measured on a four category scale, have been selected as predictors of the probability of non response: *Students' attendance at classes* (1 =*Always*; 4 =*Very rarely*) and *Students' interest toward the topic* (1 =*Definitely No*; 4 =*Definitely Yes*).

Table 5: Item considered for the application

| Item | Contents |
|---|---|
| $L_1$ | Lecturer ability on motivating students |
| $L_2$ | Lecturer highlights topics |
| $L_3$ | Lecturer answers questions during the class |
| $L_4$ | Lecturer clarifies goals of the course |
| $L_5$ | Lecturer clearly explains topics |
| $L_6$ | Lecturer suggests how to study |
| $L_7$ | Lecturer gives classes on schedule |
| $S$ | Global satisfaction |
| $A$ | Student's attendance at classes |
| $I$ | Student's interest toward the topics |

In the CD data-set, the cross-classification of units according to the two covariates provides 16 clusters of students; each of them has a different probability ($\pi$) to skip an item:

$$\pi_i = \frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)}{1+\exp(\boldsymbol{\beta}'\boldsymbol{x}_i)}. \tag{3}$$

The $\boldsymbol{\beta}$ vector has been defined attaching the lowest probability $\pi_i$ to skip an item to students who say to be *Definitely Yes* interested on the topic and who *Always* attend classes, instead the highest $\pi_i$ is attached to students who are *Definitely No* interested and who have *Rarely* attended classes. A unit $i$ in the matrix has been set missing if the result of a random draw from a Bernoulli($\pi_i$) was 1 (see §A.4). Five data-sets have been generated according to the MCAR criteria and five according to the MAR criteria. The 10 data-sets have been imputed by MISR and by MICE. The latter procedure has been implemented by Van Buuren & Oudshoorn in the mice package for the R environment (Van Buuren and Oudshoorn, 2000). MICE generates multiple imputations for incomplete multivariate data by Gibbs Sampling (Van Buuren and Oudshoorn, 2004; Schafer, 1997). The algorithm imputes an incomplete column by generating appropriate imputation values given other columns in the data matrix. In this application the predictors are the set of the remaining columns in the data. The imputation function specified is polyreg, which is the default method for polytomous variables.

## 3.1 AD

The AD has been assessed by comparing the agreement between the marginal distribution of each item in the CD with the marginal distributions of the same item in each of the $M = 100$ randomly imputed data sets (Sulis, 2007). The *Dissimilarity index* $z'$ (Leti, 1983) for ordinal variables has been used to measure the discrepancy. For each item, 100 comparisons have been made. Function multipledissimilarity.index (see § A.6) calculates the average values of the index taken over the 100 data-sets. Results are depicted in Table 6.

For each of the five rates of missingness, the *dissimilarity index* exhibits better performances when the MCAR assumption holds. Another measure proposed to evaluate the discrepancy between distributions is the *Chi-square* statistic. The Chi-square test highlights those items for which the lack of agreement between the benchmark and the imputed distributions may not be considered random. Results obtained by using MISR shows that none of the 100 distribution imputed for each item departs significantly from the reference. Function multiplechisq.test (see §A.5) returns for each item the average value of the Chi-square statistic taken over the M imputed data-sets (the greatest value observed is $\bar{\chi}^2 = .099$ for $L_7$). Under both MCAR and MAR none of the average values signal a significant departure from the benchmark distribution. The values of multiplechisq.test highlights a good performance of both the imputation procedures in terms of AD. The overall degree of agrement is high also when the rate of missingness in the data matrix is equal to 25% (the highest value assumed by the index is 0.02). However, even though for any rate of missingness MICE seems to perform slightly better than MISR, differences in absolute terms may be considered no relevant.

## 3.2 AE

The AE has been assessed by comparing the parameters of a *random intercept* logit model estimated using the CD with the one obtained as a synthesis of the corresponding estimates observed in the 100 randomly imputed data sets. The logit model with random intercept has been estimated using the glmmML function implemented in the R package glmmML (Broström, 2007). Function multiglmmML (see § A.7) summarizes in a single inferential statement results observed in each of the $M$ randomly imputed data sets using the formula provided by Rubin (1987). By indicating with $\hat{\theta}_m$ an estimate for a parameter $\theta$ in data set $m$, the final estimate for $\theta$ is the mean of $\hat{\theta}_m$ taken over $M$ data sets:

Table 6: Accuracy in distribution: average values taken over $M$ data sets

| | Dissimilarity Index | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| % miss | $L_1$ | $L_2$ | $L_3$ | $L_4$ | $L_5$ | $L_6$ | $L_7$ | $S$ |
| MCAR | | | | | | | | |
| MISR | | | | | | | | |
| 5% | 0.0015 | 0.0019 | 0.0014 | 0.0027 | 0.0016 | 0.0025 | 0.0023 | 0.0013 |
| 10% | 0.0038 | 0.0035 | 0.0036 | 0.0024 | 0.0029 | 0.0041 | 0.0055 | 0.0025 |
| 15% | 0.0081 | 0.0057 | 0.0054 | 0.0026 | 0.0032 | 0.0042 | 0.0078 | 0.0026 |
| 20% | 0.0111 | 0.0046 | 0.0075 | 0.0041 | 0.0030 | 0.0080 | 0.0115 | 0.0038 |
| 25% | 0.0157 | 0.0077 | 0.0091 | 0.0067 | 0.0038 | 0.0152 | 0.0199 | 0.0057 |
| MICE | | | | | | | | |
| 5% | 0.0018 | 0.0016 | 0.0013 | 0.0023 | 0.0016 | 0.0031 | 0.0016 | 0.0015 |
| 10% | 0.0028 | 0.0025 | 0.0022 | 0.0021 | 0.0032 | 0.0049 | 0.0019 | 0.0029 |
| 15% | 0.0048 | 0.0039 | 0.0024 | 0.0028 | 0.0041 | 0.0057 | 0.0022 | 0.0029 |
| 20% | 0.0053 | 0.0035 | 0.0024 | 0.0031 | 0.0031 | 0.0044 | 0.0033 | 0.0039 |
| 25% | 0.0067 | 0.0053 | 0.0028 | 0.0037 | 0.0028 | 0.0049 | 0.0038 | 0.0071 |
| MAR | | | | | | | | |
| MISR | | | | | | | | |
| 5% | 0.0018 | 0.0019 | 0.0016 | 0.0023 | 0.0013 | 0.0029 | 0.0020 | 0.0016 |
| 10% | 0.0049 | 0.0031 | 0.0027 | 0.0024 | 0.0021 | 0.0038 | 0.0039 | 0.0041 |
| 15% | 0.0120 | 0.0065 | 0.0052 | 0.0042 | 0.0037 | 0.0049 | 0.0072 | 0.0062 |
| 20% | 0.0119 | 0.0044 | 0.0064 | 0.0056 | 0.0038 | 0.0102 | 0.0110 | 0.0068 |
| 25% | 0.0128 | 0.0108 | 0.0106 | 0.0079 | 0.0039 | 0.0143 | 0.0217 | 0.0067 |
| MICE | | | | | | | | |
| 5% | 0.0015 | 0.0022 | 0.0012 | 0.0019 | 0.0014 | 0.0034 | 0.0017 | 0.0015 |
| 10% | 0.0033 | 0.0027 | 0.0019 | 0.0023 | 0.0024 | 0.0031 | 0.0019 | 0.0044 |
| 15% | 0.0084 | 0.0040 | 0.0025 | 0.0037 | 0.0042 | 0.0040 | 0.0034 | 0.0056 |
| 20% | 0.0081 | 0.0030 | 0.0025 | 0.0031 | 0.0044 | 0.0033 | 0.0044 | 0.0077 |
| 25% | 0.0075 | 0.0042 | 0.0040 | 0.0029 | 0.0033 | 0.0038 | 0.0045 | 0.0057 |

$$\bar{\theta} = M^{-1} \sum_{m=1}^{M} \hat{\theta}_m. \tag{4}$$

For the $V_m$ associated variances, the overall variance of $\theta$ is a combination of the *Within* imputation variance and the *Between* imputation variance:

$$T = \text{Within} + (1 + M^{-1})\text{Between}; \tag{5}$$

$$\text{Within} = M^{-1} \sum_{m=1}^{M} \hat{V}_m; \tag{6}$$

$$\text{Between} = (M-1)^{-1} \sum_{m=1}^{M} (\hat{\theta}_m - \bar{\theta})^2. \tag{7}$$

The model fitted specifies the probability to be or not to be globally satisfied (item *S*) as a function of items $L_1 - L_7$. Both response and predictor variables have been previously dichotomized. The model is defined as

$$\text{logit}[Pr(Y_{ig} = 1 | u_g)] = \alpha + \sum_{j=1}^{p} \beta_j x_{ij} + u_g \tag{8}$$

where $i = 1, \ldots, n_g$ are students' evaluations for the $g_{th}$ course and the random intercept $u_g \sim N(0, \sigma^2)$. Results depicted in Table 7 and 9 show that MISR method produces satisfactory estimates of coefficients regression parameters under both MCAR and MAR assumptions for rates of missing records in the matrix not over 10%.

The advantage of adopting MICE in respect of MISR seems to be higher under the MCAR than MAR. Moreover, the convenience increases as the rate of missingness in the data set becomes severe. From the simulation study arises that both multiple imputation methods do not perform well in estimating the intercept parameter ($\hat{\alpha}$) and the parameter ($\hat{\beta}_{L_7}$). Nevertheless, the latter is better estimated by MICE. The estimates of $\hat{\alpha}$ become strongly unreliable when the rate of missing units increases. Both procedures show a better performance in respect of the CCA results; as depicted in Table 7 and 9, the latter leads to bias and inefficient estimates of many parameters in data-sets strongly affected by missingness. In the MAR datasets, the two multiple imputation procedure provides results quite similar for rates of missingness under 15%. Tables 8 and 10 show , as could be expected, that both multiple imputation procedure, MICE and MISR, produce accurate estimates of

Table 7: MCAR: Coefficient parameters for the random intercept logit model

| % miss | $\hat{\alpha}$ | $\hat{\beta}_{L_1}$ | $\hat{\beta}_{L_2}$ | $\hat{\beta}_{L_3}$ | $\hat{\beta}_{L_4}$ | $\hat{\beta}_{L_5}$ | $\hat{\beta}_{L_6}$ | $\hat{\beta}_{L_7}$ | $\hat{\sigma}$ |
|---|---|---|---|---|---|---|---|---|---|
| | | | | MCAR | | | | | |
| | | | | MISR | | | | | |
| 5% | −4.048 | 1.484 | 0.599 | .770 | .507 | 1.702 | .734 | 1.713 | 0.559 |
| 10% | −3.942 | 1.399 | 0.630 | .839 | .592 | 1.683 | .707 | 1.503 | 0.571 |
| 15% | −3.469 | 1.369 | 0.632 | .727 | .465 | 1.649 | .667 | 1.219 | 0.553 |
| 20% | −3.155 | 1.274 | 0.590 | .712 | .555 | 1.492 | .742 | 1.002 | 0.552 |
| 25% | −2.961 | 1.156 | 0.679 | .599 | .453 | 1.404 | .899 | 0.994 | 0.532 |
| | | | | MICE | | | | | |
| 5% | −4.131 | 1.490 | .587 | .745 | .527 | 1.739 | .709 | 1.806 | .561 |
| 10% | −4.294 | 1.409 | .622 | .857 | .605 | 1.785 | .673 | 1.788 | .566 |
| 15% | −3.918 | 1.434 | .579 | .750 | .482 | 1.781 | .611 | 1.613 | .554 |
| 20% | −3.633 | 1.404 | .488 | .841 | .489 | 1.694 | .644 | 1.382 | .565 |
| 25% | −3.627 | 1.282 | .637 | .765 | .265 | 1.684 | .852 | 1.555 | .524 |
| | | | | CCA | | | | | |
| 5% | −4.009 | 1.606 | .660 | .510 | 639 | 1.672 | .610 | 1.710 | .428 |
| 10% | −4.370 | 1.630 | .710 | .516 | 671 | 1.905 | .612 | 1.775 | .359 |
| 15% | −3.294 | 1.756 | .733 | .141 | 360 | 1.946 | .636 | 1.140 | .441 |
| 20% | −3.071 | 1.254 | .678 | .085 | 549 | 2.008 | .977 | .887 | .000 |
| 25% | −3.492 | 1.078 | .954 | −.009 | 333 | 1.609 | 1.017 | 1.492 | .000 |
| | | | | CD | | | | | |
| - | −4.254 | 1.494 | 0.666 | 0.658 | 0.575 | 1.742 | 0.767 | 1.871 | 0.599 |

Table 8: MCAR: SE of the coefficient parameters for the random intercept logit model

| % miss | $\hat{\alpha}$ | $\hat{\beta}_{L_1}$ | $\hat{\beta}_{L_2}$ | $\hat{\beta}_{L_3}$ | $\hat{\beta}_{L_4}$ | $\hat{\beta}_{L_5}$ | $\hat{\beta}_{L_6}$ | $\hat{\beta}_{L_7}$ | $\hat{\sigma}$ |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | MCAR | | | | |
| | | | | | MISR | | | | |
| 5% | .415 | .184 | .203 | .235 | .202 | .182 | .189 | .345 | .133 |
| 10% | .423 | .192 | .210 | .243 | .209 | .191 | .196 | .346 | .138 |
| 15% | .398 | .193 | .225 | .252 | .221 | .193 | .204 | .340 | .137 |
| 20% | .391 | .198 | .240 | .270 | .220 | .195 | .205 | .330 | .140 |
| 25% | .394 | .207 | .214 | .261 | .228 | .202 | .201 | .336 | .140 |
| | | | | | MICE | | | | |
| 5% | .418 | .183 | .205 | .239 | .203 | .184 | .189 | .344 | .134 |
| 10% | .447 | .196 | .218 | .248 | .217 | .196 | .203 | .375 | .141 |
| 15% | .444 | .204 | .240 | .257 | .229 | .205 | .211 | .373 | .142 |
| 20% | .449 | .218 | .251 | .288 | .242 | .211 | .214 | .390 | .139 |
| 25% | .463 | .221 | .243 | .286 | .259 | .214 | .239 | .397 | .138 |
| | | | | | CCA | | | | |
| 5% | .438 | .212 | .235 | .270 | .230 | .210 | .215 | .367 | .142 |
| 10% | .571 | .274 | .317 | .357 | .301 | .261 | .278 | .484 | .206 |
| 15% | .616 | .334 | .383 | .442 | .377 | .324 | .341 | .580 | .263 |
| 20% | .728 | .429 | .521 | .554 | .474 | .426 | .445 | .693 | .344 |
| 25% | 1.003 | .542 | .626 | .681 | .621 | .565 | .523 | .978 | .358 |
| | | | | | CD | | | | |
| - | .417 | .177 | .197 | .230 | .195 | .176 | .182 | .347 | .138 |

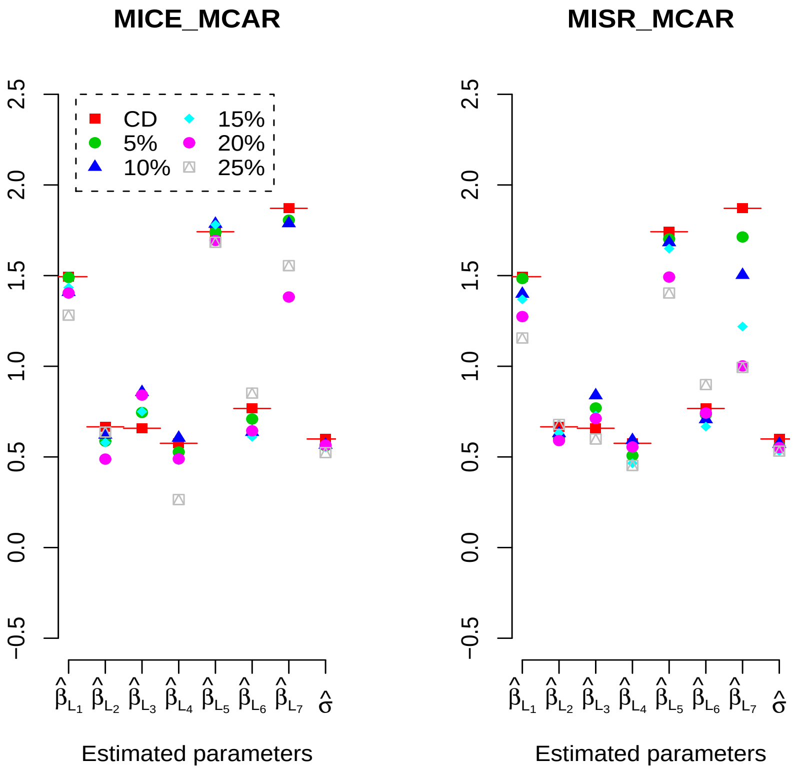
Table 9: MAR: Coefficient parameters for the random intercept logit model

| % miss | $\hat{\alpha}$ | $\hat{\beta}_{L_1}$ | $\hat{\beta}_{L_2}$ | $\hat{\beta}_{L_3}$ | $\hat{\beta}_{L_4}$ | $\hat{\beta}_{L_5}$ | $\hat{\beta}_{L_6}$ | $\hat{\beta}_{L_7}$ | $\hat{\sigma}$ |
|---|---|---|---|---|---|---|---|---|---|
| | | | | MAR | | | | | |
| | | | | MISR | | | | | |
| 5% | −4.109 | 1.481 | .572 | .774 | .566 | 1.752 | .715 | 1.733 | .558 |
| 10% | −3.726 | 1.436 | .733 | .725 | .489 | 1.683 | .759 | 1.341 | .521 |
| 15% | −3.421 | 1.342 | .727 | .697 | .433 | 1.537 | .689 | 1.264 | .539 |
| 20% | −3.256 | 1.310 | .724 | .679 | .462 | 1.482 | .778 | 1.111 | .536 |
| 25% | −3.086 | 1.189 | .792 | .726 | .408 | 1.479 | .769 | 0.967 | .533 |
| | | | | MICE | | | | | |
| 5% | −4.175 | 1.502 | .535 | .755 | 0.584 | 1.783 | .705 | 1.809 | .554 |
| 10% | −3.853 | 1.486 | .660 | .681 | 0.535 | 1.785 | .687 | 1.486 | .538 |
| 15% | −4.067 | 1.399 | .610 | .762 | 0.399 | 1.735 | .549 | 1.897 | .519 |
| 20% | −3.894 | 1.408 | .546 | .721 | 0.416 | 1.743 | .719 | 1.761 | .546 |
| 25% | −3.597 | 1.312 | .838 | .701 | 0.290 | 1.712 | .749 | 1.373 | .544 |
| | | | | CCA | | | | | |
| 5% | −3.977 | 1.578 | .570 | .518 | .496 | 1.731 | .703 | 1.817 | .438 |
| 10% | −3.678 | 1.632 | .813 | .422 | .561 | 1.668 | .556 | 1.382 | .290 |
| 15% | −3.673 | 1.894 | 1.080 | .118 | .393 | 2.013 | .415 | 1.182 | .001 |
| 20% | −4.187 | 1.696 | 1.248 | −.059 | .723 | 2.409 | .635 | 1.093 | .000 |
| 25% | −4.803 | 1.759 | 1.533 | −.015 | .355 | 2.490 | .745 | 1.690 | .001 |
| | | | | CD | | | | | |
| - | −4.254 | 1.494 | 0.666 | 0.658 | 0.575 | 1.742 | 0.767 | 1.871 | 0.599 |

Table 10: SE of coefficient parameters for the random intercept logit model

| % miss | $\hat{\alpha}$ | $\hat{\beta}_{L_1}$ | $\hat{\beta}_{L_2}$ | $\hat{\beta}_{L_3}$ | $\hat{\beta}_{L_4}$ | $\hat{\beta}_{L_5}$ | $\hat{\beta}_{L_6}$ | $\hat{\beta}_{L_7}$ | $\hat{\sigma}$ |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | MAR | | | | |
| | | | | | MISR | | | | |
| 5% | .415 | .183 | .206 | .235 | .200 | .183 | .187 | .347 | .137 |
| 10% | .421 | .189 | .212 | .249 | .217 | .197 | .196 | .364 | .136 |
| 15% | .415 | .192 | .225 | .253 | .218 | .207 | .207 | .356 | .144 |
| 20% | .405 | .201 | .230 | .257 | .232 | .200 | .206 | .353 | .144 |
| 25% | .392 | .207 | .227 | .258 | .231 | .199 | .212 | .343 | .146 |
| | | | | | MICE | | | | |
| 5% | .419 | .182 | .205 | .239 | .202 | .180 | .189 | .349 | .135 |
| 10% | .429 | .191 | .215 | .249 | .224 | .191 | .198 | .373 | .136 |
| 15% | .467 | .200 | .243 | .273 | .239 | .203 | .209 | .412 | .139 |
| 20% | .465 | .209 | .251 | .289 | .256 | .211 | .227 | .400 | .143 |
| 25% | .470 | .205 | .281 | .274 | .262 | .223 | .236 | .423 | .148 |
| | | | | | CCA | | | | |
| 5% | .438 | .203 | .231 | .268 | .226 | .204 | .212 | .375 | .144 |
| 10% | .495 | .241 | .276 | .320 | .269 | .240 | .248 | .436 | .200 |
| 15% | .612 | .306 | .381 | .415 | .362 | .302 | .321 | .523 | .170 |
| 20% | .731 | .351 | .467 | .477 | .421 | .347 | .359 | .586 | .181 |
| 25% | .925 | .435 | .585 | .583 | .529 | .438 | .443 | .721 | .303 |
| | | | | | CD | | | | |
| - | .417 | .177 | .197 | .230 | .195 | .176 | .182 | .347 | .138 |

Figure 1: Comparisons between MISR and MICE estimates under MCAR assumption

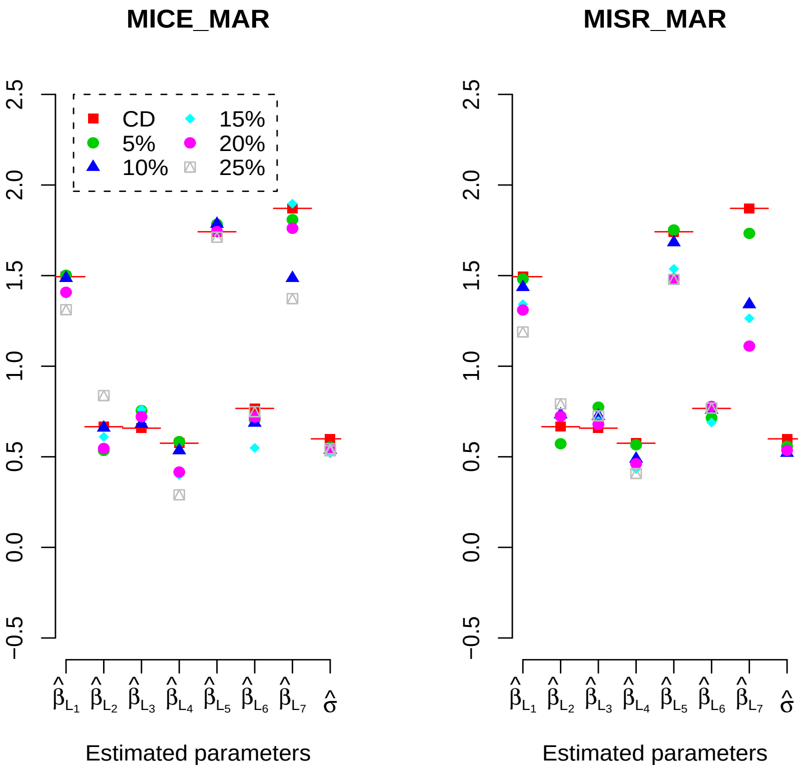


standard errors. In the *Mixed effect model* framework the greatest advantage of the MICE and MISR approaches is the accuracy in the estimation of the random term (see last column of Tables 7 and 9). Figures 1 and 2 better highlight the accuracy in estimation of both multiple imputation methods and make easier the comparisons between them under the two missing data generating mechanisms.

## 4 Some final remarks

In this article a multiple imputation approach based on stochastic regression models has been described, implemented and evaluated in respect of the widely validated MICE approach. The proposed MISR procedure is an *ad hoc* method to recover for missingness in data where items measured on Likert-type scale define the same

Figure 2: Comparisons between MISR and MICE estimates under MAR assumption



latent trait. The examination of subject pattern of responses provides information on the way students score categories and help us to learn if subject tends to use high, low or middle categories. This motivates the first step of the procedure. MISR seems to produce unbias and efficient estimates of many coefficient parameters in *Mixed effect models* framework. Estimates provided by MISR under MAR assumption do not seems to show a remarkable departure from the one obtained using MICE library, at least when the rate of missingness in the data matrix does not become severe.

# A Some functions implemented in R-language

## A.1 Multiple imputation procedure for a single item affected by missingness: the function imputrm

**Description:**

A multiple imputation procedure to impute unobserved units in a single categorical item measured on a $K$ category ordinal scale.

**Use:**

```
imputrm(B,m)
```

**Arguments:**

**B**: A data matrix composed by $p$ categorical variables all measured on a $K$ category ordinal scale. The variable which has to be imputed is in the first column of the $\boldsymbol{B}$ data matrix.

**m**: The number of $M$ randomly imputed variables.

**Function:**

```
imputmr<-function(B, m){
library(nnet)
library(MASS)
y<-B[,1]
cat<- length(table(B[,1]))
n<-nrow(B)
ca<-c(1:cat)
distfreq<-matrix(NA, n, cat)
   for(i in 1:n)  {
   for(j in 1:cat){
            distfreq[i,j]<-length(B[i,][B[i,]==ca[j]&
            !is.na(B[i,])])
                 }
                 }
j<-ncol(B)
nrisp<-apply(distfreq,1,sum)
```

```
freq<-distfreq/nrisp
ceck<-is.na(freq)*5
impm<-array(NA, c(n,cat,j,m))
   for(i in 1:n)  {
            impm[i,,,]<-ifelse( array(rep(ceck[i,],j*m),
            c(cat,j,m))==array(rep(5,j*m),c(cat,j,m)),
            array(rep(freq[i,],j*m),c(cat,j,m)),
            array( rmultinom(j*m,1, c(freq[i,])),c(cat,j,m)))
                  }
   for(k in 1:cat){
impm[,k,,]<-ifelse(impm[,k,,]==1, k, impm[,k, ,])
                  }
imp<-array(NA,c(n,j,m))
   for(t in 1:j)  {
   for(s in 1:m)  {
imp[,t,s]<-apply(impm[,,t,s],1, sum, na.rm=TRUE)
                   }
                   }
mat<-array(NA, c(n,j,m))
    for(t in 1:j) {
    for(s in 1:m) {
        mat[,t,s]<-ifelse(is.na(B[,t])==TRUE & imp[,t,s]!=0 ,
        imp[,t,s],B[,t])
                  }
                  }
p<-cat
pred<-array(NA, c(n,p,m))
mat1<-mat
   for(s in 1:m){
     mat1[,,s]<-apply(mat1[,,s],2,factor)
                }
  for(s in 1:m)  {
      reg<- polr(factor(mat1[,1,s]) ~., data=mat1[,-c(1),s],
      na.action=’na.exclude’)
      x<-mat1[,-c(1),s]
      hat<-predict(reg, newdata=x, type="prob")
      pro<-as.matrix(hat)
      p<-ncol(pro)
      ceck<-is.na(pro)*1
         for(i in 1:n) {
            pred[i, ,s]<-ifelse(ceck[i,]==rep(1,p), pro[i,],
            rmultinom(1,1, c(pro[i,])) )
                       }
                       }
        for(k in 1:p){
         pred[,k,]<-ifelse(pred[,k,]==1, k, pred[,k,])
                       }
         imp<-matrix(NA,n,m)
          for(t in 1:m) {
          imp[,t]<-apply(pred[,,t],1, sum, na.rm=TRUE)
                          }
          for(t in 1:m) {
          imp[,t]<-ifelse(is.na(y)==TRUE  ,imp[,t], y)
                          }
          return(imp)
          }
```

## A.2 Multiple imputation procedure for a set of items affected by missingness: the function imputmult

**Description:**

Multiple imputation procedure to impute missing observations in a set of categorical items all measured on a $K$ category ordinal scale.

**Use:**

```
imputmult(B,m)
```

**Arguments:**

**B**: A data matrix of $p$ categorical variables all measured on a $K$ category ordinal scale. The procedure starts imputing firstly the first column of the $\boldsymbol{B}$ data matrix, using the set of the $p-1$ categorial items as a predictors, next it carries on imputing the second column using the remaining items as predictors and so forth.

**m**: The number of $M$ randomly imputed data sets.

**Function**

```
imputmult<-function(B,m){
n<-nrow(B)
j<-ncol(B)
prova<-array(NA,c(n,m,j))
for(v in 1:j) {
prova[,,v]<-imputmr(cbind(B[,v],B[,-c(v)]),m)
              }
return(prova)
              }
```

## A.3 Function to simulate MCAR observation in the data matrix: the function miss.CAR

**Description:**

Function to simulate a given rate of completely at random missing values in each of the *p* items of the data matrix *B*

**Use:**

```
miss.CAR(B,pi, numbers)
```

**Arguments:**

**B**: A data matrix of *p* categorical variables all measured on a *K* category ordinal scale.

**pi**: The rate of observations simulated missing in each item.

**numbers**: Seed of the random numbers generator.

**Function:**

```
miss.CAR<-function(B, pi, numbers){
n<-nrow(B)
c<-ncol(B)
```

```
  set.seed(numbers)
  Binom<-matrix(NA,n,c)
  for(j in 1:c){
  Binom[,j]<-rbinom(n,1,pi)
               }
item2<-matrix(NA,n,c)
 for(j in 1:c) {
 item2[,j]<-ifelse(Binom[,j]==1,NA, item[,j])
               }
  return(item2)
               }
```

## A.4 Function to simulate MAR observation in the data matrix: the function miss.AR

**Description:**

Function to simulate a given rate of missing values at random in each of the *p* items of the data matrix *B*

**Use:**

```
miss.AR(B, X, numbers, b)
```

**Arguments:**

**B**: A data matrix of *p* categorical variables all measured on a *K* category ordinal scale.

**X**: A data matrix where the first column is a vector of ones and the remaining *J* columns are the predictors of the probability of non response. For each predictor measured on a *K* categories scale are introduced $(K-1)$ dummy variables.

**numbers**: Seed of the random numbers generator.

**b**: The vector of coefficient parameters corresponding to *X*.

**Function:**

```
miss.AR<-function(B, X, numbers, b){
n<-nrow(B)
c<-ncol(B)
set.seed(numbers)
pi=exp(X%*%b)/(1+exp( X%*%b))
Binom<-matrix(NA,n,c)
for(j in 1:c){
Binom[,j]<-rbinom(n,1,pi)
          }
item2<-matrix(NA,n,c)
for(j in 1:c) {item2[,j]<-ifelse(Binom[,j]==1, NA, item[,j])}
return(item2)
          }
```

## A.5 Function to assess the accuracy in distribution 1: the function multiplechisq.test

**Description:**

The function calculates for each of the *M* randomly imputed data sets the Chi-squared test between the marginal distribution of the randomly imputed variables and the true distribution in the CD data set. The statistic is calculated for all items involved in the imputation procedure. For each item the function returns the average value of the Chi-square Test taken over the *M* results.

**Use:**

```
multiplechisq.test<-function(imp,item,K)
```

**Arguments:**

**imp**: An array of dimensions $n \times j \times M$ which contains the M imputed data sets

**item**: The Complete data set (CD).

**K**: Number of categories of the item in the data matrix .

**Function:**

```
multiplechisq.test<-function(imp,item,K){
f.count<-function(var){
freq<-table(var) /sum(table(var))
}
        m<-dim(imp)[2]
        j<-dim(imp)[3]
    distr<-apply(imp, c(2,3), f.count)
    orig<-apply(item,2,f.count)
    cumor<-apply(orig,2, cumsum)
    cumdist<-array(NA,c(m,K,j))
    for(t in 1:m){
    cumdist[t,,]<- apply(distr[,t,],2,cumsum)
    }
    onesm<-rep(1,m)
    cumorm<-outer(onesm,cumor)
    distrm<-array(NA,c(m,K,j))
       for(t in 1:m) {
       distrm[t,,]<-distr[,t, ]
       }
       orm<-outer(onesm,orig)
       chisq<-(distrm-orm)^2/orm
       r.chisq<-matrix(NA,m,j)
    for(g in 1:j) {
    r.chisq[,g]<-apply(chisq[,,g],1,sum)
    }
    list(chisq=apply(r.chisq,2,mean))
    }
```

## A.6 Function to assess the accuracy in distribution 2: the function multipledissimilarity.index

**Description:**

The function calculates in each of the *M* randomly imputed data sets the Dissimilarity Index for ordinal categorical variables between the marginal distribution of the randomly imputed variables and the true distribution in the

CD data set. The statistic is calculated for all items involved in the imputation procedure. For each item the function returns the average value of the Dissimilarity Index taken over the *M* results.

**Use:**

```
multidissimilarity.test<-function(imp,item,K)
```

**Arguments:**

**imp**: An array of dimensions $n \times j \times M$ which contains the M imputed data sets

**item**: The Complete data set (CD).

**K**: Number of categories of the item in the data matrix .

**Function:**

```
multipledissimilarity.index<-function(imp,item,K){
f.count<-function(var){
freq<-table(var) /sum(table(var))
                      }
         m<-dim(imp)[2]
         j<-dim(imp)[3]
     distr<-apply(imp, c(2,3), f.count)
     orig<-apply(item,2,f.count)
     cumor<-apply(orig,2, cumsum)
     cumdist<-array(NA,c(m,K,j))
    for(t in 1:m){
          cumdist[t,,]<- apply(distr[,t,],2,cumsum)
                  }
     onesm<-rep(1,m)
     cumorm<-outer(onesm,cumor)
     absdiff<-abs(cumdist-cumorm)
     dissimilarity<-matrix(NA,m,j)
        for(g in 1:j){
        for(t in 1:m){
    dissimilarity[t,g]<-(sum(absdiff[t,,g])/(K-1))
       }
     }
  list(dissimilarity.mean=apply(dissimilarity,2,mean),
      dissimilarity.m=dissimilarity)
       }
```

## A.7 Function to fit a logit model with random intercept in each of the $M$ multiple imputed data-sets: the function multiglmmML

**Description:**

The function multiglmmML fits GLMs with random intercept by Maximum Likelihood and numerical integration (see function glmmML implemented in the package glmmML) in each of the $M$ randomly imputed data sets and summarizes multiple results in a single inferential statement. The function *multiglmmML* returns the following estimates:

- mean.beta: the average value of coefficient parameters taken over the $M$ imputed data-sets;
- W.beta: the within data-sets variance of the coefficient parameters;
- B.beta: the between data-sets variance of the coefficient parameters;
- sigma.coeff: the average value of the random parameter taken over the $M$ imputed data-sets;
- W.sigma : the within data-sets variance of the random parameter
- B.beta: the between data-sets variance of the random parameter

**Use:**

```
multiglmmML<- function(Y, X, I, npar)
```

**Arguments:**

**Y**: A matrix of dimensions $n \times M$: the dependent variable in each of the $M$ randomly imputed data-sets.

**X**: An array $n \times J \times M$: the matrix of $J$ predictors in each of the $M$ randomly imputed data-sets.

**I**: A matrix of dimensions $n \times M$: the group variable in each of the $M$ randomly imputed data-sets.

**npar**: Number of coefficients

**Function:**

```
 multiglmmML<- function(Y, X, I, npar){
               library(glmmML)
               m<-dim(X)[2]
               coeff<-matrix(NA, m,npar)
               varianza<-matrix(NA,m,npar)
               random<-rep(NA,m)
               var.random<-rep(NA,m)
              for(t in 1:m){
              mod1<-glmmML(Y[,t]~X[,t,], cluster=I[,t],
                    family=binomial)
## See the help of function glmmML  to add/change some options to mod1
                   coeff[t,]<-mod1$coeff
                varianza[t,]<-(mod1$coef.sd)^2
                   random[t]<-mod1$sigma
               var.random[t]<-(mod1$sigma.sd)^2
               }
   list(mean.beta=apply(coeff,2,mean), W.beta=apply(varianza,2,mean),
   B.beta=apply(coeff,2,var),   sigma.coeff=mean(random),
   W.sigma=mean(var.random), B.sigma=var(random))
   }
```

# References

Agresti A. (2002) *Categorical Data Analysis*, Wiley-Intersciencie, Hoboken, New Jersey.

Borgoni R. and Berrington A. (2004) A tree based procedure for multivariate imputation, in: *Atti XLII Convegno della Società Italiana di Statistica. Bari, 9-11 Giugno 2004*, Società Italiana di Statistica.

Broström G. (2007) *glmmML: Generalized linear models with clustering*, R-package.

Chambers R. (2001) Evaluation criteria for statistical editing and imputation, *National Statistics Methodological Series*, 28, 1–41.

Haitovsky Y. (1968) Missing data in regression analysis, *Journal of the Royal Statistical Society, B*, 30, 67–82.

Leti G. (1983) *Statistica Descrittiva*, il Mulino, Bologna.

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. (1976) Inference and missing data, *Biometrika*, 63, 581–592.

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

Schafer J. (1997) *Analysis of Incomplete Multivariate Data*, Chapman and Hall.

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

Sulis I. and Porcu M. (2007) A multiple imputation approach in a survey on university teaching evaluation, in: *Classification and Data Analysis 2007*, Meeting of the Classification and Data Analysis Group of the Italian Statistical Society, Macerata 11-12 September 2007.

Van Buuren S. and Oudshoorn C. (2000) Multivariate imputation by chained equations: Mice v1.0 user's manual, Technical Report Report PG/VGZ/00.038, Prevention and Health, Leiden.

Van Buuren S. and Oudshoorn C. (2004) *MICE: Multivariate Imputation by Chained Equations*, R-package.

## Ultimi Contributi di Ricerca CRENoS

*I Paper sono disponibili in:* http://www.crenos.it

**08/03** *Manuele Bicego, Enrico Grosso, Edoardo Otranto,* “Recognizing and Forecasting the Sign of Financial Local Trends Using Hidden Markov Models”

**08/02** *Juan de Dios Tena, Edoardo Otranto,* “A Realistic Model for Official Interest Rates Movements and their Consequences”

**08/01** *Edoardo Otranto,* “Clustering Heteroskedastic Time Series by Model-Based Procedures”

**07/16** *Sergio Lodde,* “Specialization and Concentration of the Manufacturing Industry in the Italian Local Labor Systems”

**07/15** *Giovanni Sulis,* “Gender Wage Differentials in Italy: a Structural Estimation Approach”

**07/14** *Fabrizio Adriani, Luca G. Deidda, Silvia Sonderegger,* “Over-Signaling Vs Underpricing: the Role of Financial Intermediaries In Initial Public Offerings”

**07/13** *Giovanni Sulis,* “What Can Monopsony Explain of the Gender Wage Differential In Italy?”

**07/12** *Gerardo Marletto,* “Crossing the Alps: Three Transport Policy Options”

**07/11** *Sergio Lodde* “Human Capital And Productivity Growth in the Italian Regional Economies: a Sectoral Analysis”

**07/10** *Axel Gautier, Dimitri Paolini,* “Delegation, Externalities and Organizational Design”

**07/09** *Rinaldo Brau, Antonello E. Scorcu, Laura Vici,* “Assessing visitor satisfaction with tourism rejuvenation policies: the case of Rimini, Italy”

**07/08** *Dimitri Paolini,* “Search and the firm’s choice of the optimal labor contract”

**07/07** *Giacomo Carboni,* “Shape of U.S. business cycle and long-run effects of recessions”

**07/06** *Gregory Colcos, Massimo Del Gatto, Giordano Mion and Gianmarco I.P. Ottaviano,* “Productivity and firm selection: intra-vs international trade”

**07/05** *Silvia Balia,* “Reporting expected longevity and smoking: evidence from the share”

**07/04** *Raffaele Paci, Stefano Usai,* “Knowledge flows across European regions”

**07/03** *Massimo Del Gatto, Giordano Mion and Gianmarco I.P. Ottaviano,* “Trade Integration, firm selection and the costs of non-europe”

**07/02** *Vittorio Pelligra,* “Intentions, trust and frames: a note on sociality and the theory of games”

**07/01** *Oliviero A. Carboni and Giuseppe Medda,* “Government size and the composition of public spending in a neoclassical growth model”

**06/17** *Susana Iranzo, Fabiano Schivardi and Elisa Tosetti,* “Skill dispersion and firm productivity: an analysis with employer-employee matched data”

**06/16** *Luigi Guiso and Fabiano Schivardi,* “What determines entrepreneurial clusters?”

**06/15** *Vittorio Pelligra,* “Trust Responsiveness: on the dynamics of fiduciary interactions”

**06/14** *Rinaldo Brau, Alessandro Lanza and Francesco Pigliaru,* “How fast are small tourist countries growing? The 1980-2003 evidence”

Finito di stampare nel mese di Marzo 2008
Presso Editoria&Stampa
Zona Industriale Predda Niedda str. n. 10
07100 Sassari

www.crenos.it