Lavoro e disoccupazione: questioni di misura e di analisi

Progetto di ricerca cofinanziato dal Ministero per l'Università
e la Ricerca Scientifica e Tecnologica - Assegnazione: 1998
Coordinatore: Ugo Trivellato

A Multiple Imputation Method for School to Work Panel Data

Antonella D'Agostino, Giulio Ghellini, Laura Neri

Dip. di Metodi Quantitativi, Univ. di Siena

Working Paper n. 16 Ottobre 1999

Unità locali del progetto:

Dip. di Economia Politica, Univ. Di Modena	(coord. Michele Lalla)
Dip. di Economia "S. Cognetti De Martiis", Univ. di Torino	(coord. Bruno Contini)
Dip. Di Statistica, Univ "Ca' Foscari" di Venezia	(coord. Tommaso Di Fonzo)
Dip. di Metodi Quantitativi, Univ. di Siena	(coord. Achille Lemmi)
Dip. di Scienze Statistiche, Univ. di Padova	(coord. Ugo Trivellato)

Dip. di Scienze Statistiche
via S. Francesco 33, 35121 Padova

1 Premiss[1]

In panel survey non-response problem seems too be more complex to analyse on one side, but in some way easier to treat on the other.

In fact panel non response, in particular attrition and wave non response, is a very critical aspect of panel data, causing both bias and loss of efficiency of the estimate. The most relevant task for the researcher is first of all to analyse and understand the nature of non response process (the process term is used because during panel life the non response behavior can be seen - and statistically treated, as a social process that causes or not the choice of non responding), specially its ignorable/non-ignorable nature (for more details on it see among others Little and Rubin, 1987; Rubin 1987; O'Muirgheartaigh, 1996; Shafer 1997). If the non-response process seems to be ignorable weighting or imputation has a positive effect in reducing the estimates bias, even if with relevant effect on their variability. Instead, if the process is non-ignorable, in other words the non response mechanism depends on not-observable variables, correct inference can be made only with models for non response, usually difficult to specify and frequently affected of misspecification problems (see Little and Rubin 1989).

On the other side, these relevant disadvantages can be counterbalanced by some remarkable analytical advantages coming from the repeated observation on the same unit (Kalton 1986; Duncan and Kalton, 1987; O'Muircheartaigh, 1989; Lepkowski, 1989). The most important ones are: i) availability of data on individual characteristics and on target variable of the survey in previous/subsequent waves respect the non response one; ii) data on non response behavior during panel life, usually with relevant inertia effect; iii) possibility of relationship between variables estimates across time.

Furthermore, panel data allow a better non response treatment of item non response then cross section one, because the repeated observations make it usually possible to know the value of a missing value at one wave in the previous/subsequent ones.

Within this framework the effort of this research work has been focused on a seven waves panel data on youth transition, coming from a cohort study called LEVA, with relevant problems on wave non-response, (see Bernardi Ghellini, Penello, 1997 for more details). Of course, the final aim is the

[1]Thought it is the results of the common efforts of all authors, A. D'Agostino has written sections 2 and 5, L. Neri sections 3, 4, 6, and G. Ghellini section 1.

reconstruction of the complete data set across the seven waves, even if till now the paper considers just the first three waves and one variable, the youth situation on studying and working life.

The paper is organised in the following way. Section 2 is devoted to a brief presentation of a model for complete binary panel data, the time sequence logit model, used in the final application. In section 3, after the definition of the assumption made on missing data mechanism, it is outlined in details the multiple imputation method applied in the following simulation (section 4) and in the application on real data (section 5). Some final remarks end the paper.

2 A complete data model for binary panel data

One of the most important steps of inference processes is the specification of statistical model on survey data. It is usually rather difficult to select a model that is either easy to treat from a mathematical/computational point of view and adequate to represent the observed reality.

This task becomes harder for panel data, because in this case the treatment regards observations on the same units across time; as widely known such situation determines correlation across time and consequent problems in parameters estimation. Let suppose, for example, to put our interest on a regression model for a binary variable y - observed across survey waves - on a vectors of covariates $\mathbf{x}$; e.g. $\{f(y|\mathbf{x};\boldsymbol{\theta}_1), \boldsymbol{\theta}_1 \in \Theta_1\}$ where $\boldsymbol{\theta}_1$ represents the unknown parameters vectors.

Let N $(i = 1...N)$ the sample size, T $(t = 1...T)$ the survey waves and $\mathbf{y}_i^T = (y_i^1, y_i^2, ..., y_i^T)$, $\mathbf{X}_i^T = (\mathbf{x}_i^1, \mathbf{x}_i^2, ..., \mathbf{x}_i^T)$, and $\mathbf{r}_i^T = (r_i^1, r_i^2, ..., r_i^T)$ identically independently distributed (i.i.d.) variables vectors across reference population units, where r_i^t is an indicator that assumes value 1 if y_i^t is observed and value 0 otherwise. The independent variables vectors $\mathbf{X}_i^T$ are assumed always observed for each i. Temporally such hypothesis is extended to the response variable $\mathbf{y}_i^T$. Therefore $r_i^t = 1$ for each individual and for each wave.

Supposing data to be i.i.d. not only across the N units but across the T times too, the density function for $(\mathbf{y}|\mathbf{X};\boldsymbol{\theta}_1)$, where $\mathbf{y}$ is a matrix $N \times T$, $\mathbf{X}$ is a matrix composed of $(N \times T)$ rows and K columns - where K is the number of covariates -, and $\boldsymbol{\theta}_1$ is a matrix $T \times K$ of unknown parameters,

can be factorized as:

$$f(\mathbf{y}|\mathbf{X};\boldsymbol{\theta}_1) = \prod_{i=1}^{N} f\left(\mathbf{y}_i^T \ |\mathbf{X}_i^T;\boldsymbol{\theta}_1\right) = \prod_{i=1}^{N}\prod_{t=1}^{T} f\left(y_i^t, |\mathbf{x}_i^t;\boldsymbol{\theta}_1\right). \tag{1}$$

However the hypothesis of independence across the T observations on the y, x for the same statistical sample unit is very strong and it is not close to the reality. It seems more realistic to consider the sequential nature of the process generating the units characteristics, so it is acceptable that the dependent variable at time t is correlated to its lagged values. For this reason, the econometric literature suggests to model the variable for the unit i at time t in function of its legged values and the main goal becomes to make inference on dynamic models like $f\left(y_i^t, |\mathbf{x}_i^t, \mathbf{y}_i^{t-1};\boldsymbol{\theta}_1\right)$. The following sequential factorization:

$$f(\mathbf{y}|\mathbf{X};\boldsymbol{\theta}_1) = \prod_{i=1}^{N} f\left(\mathbf{y}_i^T \ |\mathbf{X}_i^T;\boldsymbol{\theta}_1\right) = \prod_{i=1}^{N}\prod_{t=1}^{T} f\left(y_i^t, |\mathbf{x}_i^t, \mathbf{y}_i^{t-1};\boldsymbol{\theta}_1\right), \tag{2}$$

where factors have a common shape, does not need the independence hypothesis among the independent variable at time t and its lagged values.

When the variable y is dichotomous one and it assumes value 1 if the event occurs and 0 otherwise, one of the possible parametric specification from the factors in (2) is a logit model. Therefore, each factors in (2) can be written as

$$\Pr(y_i^t = 1|\mathbf{x}_i^t, \mathbf{y}_i^{t-1}, \boldsymbol{\theta}_1) = (1 + \exp(-\mathbf{x}_i^t\boldsymbol{\beta} - \mathbf{y}_i^{t-1}\boldsymbol{\gamma}))^{-1}, \tag{3}$$

where $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}, \boldsymbol{\gamma})$, $t = 1...T$.

The equation in (3) represents T separated logit equations called in literature as sequential logit models (Clogg, Eliason, Grego, 1990)[2].

From (2) each regression model in (3) can be estimate separately and the presence of lagged variables in each equations allows to study how choices across time are influenced by the choices made in the previous times.

[2] To simplify the model specification we consider only the case where $\mathbf{x}_i^t$ is a vector of time invariant characteristics, because $\mathbf{x}_i^t = \mathbf{x}_i$ for each t .

3 The missing data treatment

3.1 Assumption on the missing data mechanism

In the previous paragraph the assumption that data are completely observed, e.g. $r_i^t = 1$ is been made. In presence of missing data, $\boldsymbol{\theta}_2$ represents the parameters vector for the missingness mechanism and $\boldsymbol{\theta}_1$ the one for the observed data model. In this case (2) can be written as:

$$f(\mathbf{y}, \mathbf{r}|\mathbf{X}; \boldsymbol{\eta}) = \prod_{i=1}^{N} f\left(\mathbf{y}_i^T, \mathbf{r}_i^T \ |\mathbf{x}_i^T; \boldsymbol{\eta}\right) = \prod_{i=1}^{N}\prod_{t=1}^{T} f\left(y_i^t, r_i^t|\mathbf{x}_i^t, \mathbf{y}_i^{t-1}, \mathbf{r}_i^{t-1}; \boldsymbol{\eta}\right), \quad (4)$$

where $\boldsymbol{\eta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$.

In relation to the missing data mechanism, different assumptions can be made. In this case the MAR (Missing at Random) selection mechanism and its uninfluenced for the inference on $f\left(y_i^t, |\mathbf{x}_i^t, \mathbf{y}_i^{t-1}; \boldsymbol{\theta}_1\right)$ -e.g. the selection mechanism is ignorable for likelihood-based inference- are assumed (Rubin, 1976).

3.2 The imputation method

In this work a particular multivariate missing data problem is considered; in fact missing values on a single variable y_i^t registered in different waves are treated.

The imputation strategy adopted comes from a Bayesian perspective (Gelman *et al.* 1995), where, without loss of generality, an uninformative prior distribution on the data model parameters $\boldsymbol{\theta}_1$ is assumed.

Let be the observed part of y^t with y_{obs}^t and the missing part with y_{mis}^t, so that $y^t = (y_{obs}^t, y_{mis}^t)$. In same way let define $\mathbf{x}_{obs}^t$ the part of $\mathbf{x}^t$ related to y_{obs}^t and $\mathbf{x}_{mis}^t$ the other part of $\mathbf{x}^t$.

Having to handle missing panel data on y^t $(t = 1...T)$ it is necessary to perform a multivariate imputation strategy. In this case, the first step is the choice of the starting variables to impute; usually the variable with the least missing values is chosen. In the case of panel data it seems reasonable to start from the missing data of the first wave.

Anyway, fixing a starting variable for the imputation is a subjective choice and the imputation values can be conditioned by it; for this reason it is

necessary to perform subsequent cycles to generate imputed values as drawn from the following factorization:

$$\prod_{t=1}^{T} \left[y^t | \mathbf{x}^t, y^l, l \neq t, l = 1...T \right]. \tag{5}$$

Given that the independent variable with missing data y is a dichotomous variable, a logistic regression model can be specified. The imputation for missing values in y^t is created using the following steps:

1. the first step involves the imputation of y^t $(t = 1)$, in this case we fit a logistic model relating y^t to $\mathbf{x}^t$, based on the units on whom y^t is observed.
$$\Pr(y_i^t = 1 | \mathbf{x}_i^t) = (1 + \exp(-\mathbf{x}_i^t \boldsymbol{\beta}))^{-1}. \tag{6}$$

2. generate a vector z of random normal deviates of dimension rows(B) and finally define $\beta^* = B + \Lambda \times z$, where B is the maximum likelihood estimate of β, Σ its covariance matrix, and Λ is the Cholesky decomposition of Σ (that is, $\Lambda\Lambda' = \Sigma$);

3. Define $\Pi^* = \{1 + \exp(-\mathbf{x}_{mis}^t \beta^*)\}^{-1}$and generate an Uniform (0-1) random realization u_i for each unobserved units. Then impute 1 if the value u_i is less or equal to the corresponding value Π_i^* and 0 otherwise.

   For the imputation of the subsequent y^t $(t = 2...T)$ $\mathbf{x}^t$ is redefined, adding y with imputed values as a column of the previous $\mathbf{x}^t$.

Once all y^t $(t = 1...T)$ are completed the subsequent cycles are performed drawing values from (5).

It is necessary to note that the method used simply draws from an approximate rather than an exact posterior distribution. The basic result of large-sample Bayesian inference is: more and more the sample size is large, the posterior distribution of the parameters vectors approaches a multivariate normal distribution (Gelman *et al.* 1995). The quality of the approximation depends on the size of the sample and unless the sample is unusually small or the rate of missing information is high, the effects of using an approximate rather than an exact posterior are probably very minor.

All mentioned steps can be repeated to create multiple imputations, using two or more (say m) values drawn from the predictive distribution of each missing values. Then complete-data analysis is repeated m times, one for

each imputation. The main reason for this choice is related to the possibility of taking into account the variability of imputed values (Rubin, 1978, 1987). In effect, methods that supply a single imputation are usually deficient to compute appropriate sampling errors of estimates from filled-in data. A single value cannot reflect imputation errors, so standard errors from the filled-in data are too optimistic.

4 Simulation Study

In order to evaluate the performance of the imputation method a simulation study has been conducted. The basic strategy is summarised in the following steps:

- generate a complete data set fixing model parameters for the data generator process;
- delete some values using a ignorable missing data mechanism;
- use the method described in section (3) to impute the missing data;
- estimate data generator process parameters on the completed data set.

The differences in the parameter estimates are then analysed across several independent replications of the basic strategy just outlined.

We fixed the sample size at 2000 and the complete data set with three variables (y^1, y^2, x), where x is a dummy variable taken from the existing data set and y^t $(t = 1, 2)$ are a dichotomous variables generated respectively from:

$$\Pr(y_i^1 = 1|x_i) = (1 + \exp(-\alpha_1 - x_i\beta_1))^{-1}, \tag{7}$$

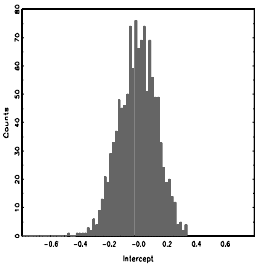
$$\Pr(y_i^2 = 1|x_i, y_i^1) = (1 + \exp(-\alpha_2 - x_i\beta_2 - \gamma y_i^1))^{-1}, \tag{8}$$

where $\alpha_1 = 1.2$, $\beta_1 = -3$, $\alpha_2 = 2$, $\beta_2 = -2.8$, $\gamma = 0.5$.

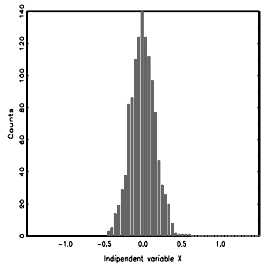
An ignorable missing data mechanism generated by (7) and (8) is assumed; the parameters values are $\alpha_1 = 1$, $\beta_1 = 2$, $\alpha_2 = 1.5$, $\beta_2 = -1$, $\gamma = 0.8$.

Table 1: True values and estimated ones: results of t-ratio test

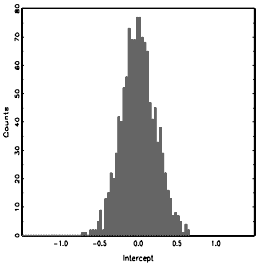
Model on $\mathbf{y}^1$	True values	Estimate	T-test	p-value
Intercept	1.2	1.193	-1.619	0.106
Indip. variable x	-3.0	-3.000	-0.031	0.970
Model on $\mathbf{y}^2$				
Intercept	2.0	2.002	0.331	0.740
Indip. variable x	-2.8	-2.803	-0.512	0.608
Lag variab. $\mathbf{y}^1$	0.5	0.510	1.460	0.143



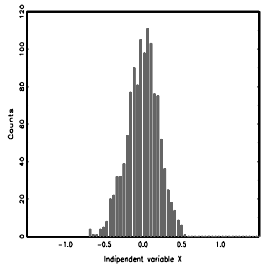
1.a



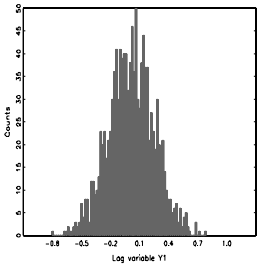
1.b



2.a



2.b



2.c

Table 2: Differences between true parameters values and estimated ones: 1.a Intercept (model on y^1); 1.b Indipendent variable x (model on y^1); 2.a Intercept (model on y^2); 2.b Indipendent variable x (model on y^2); 2.c Lag variable y^1 (model on y^2)

Model on y^1 Model on y^2

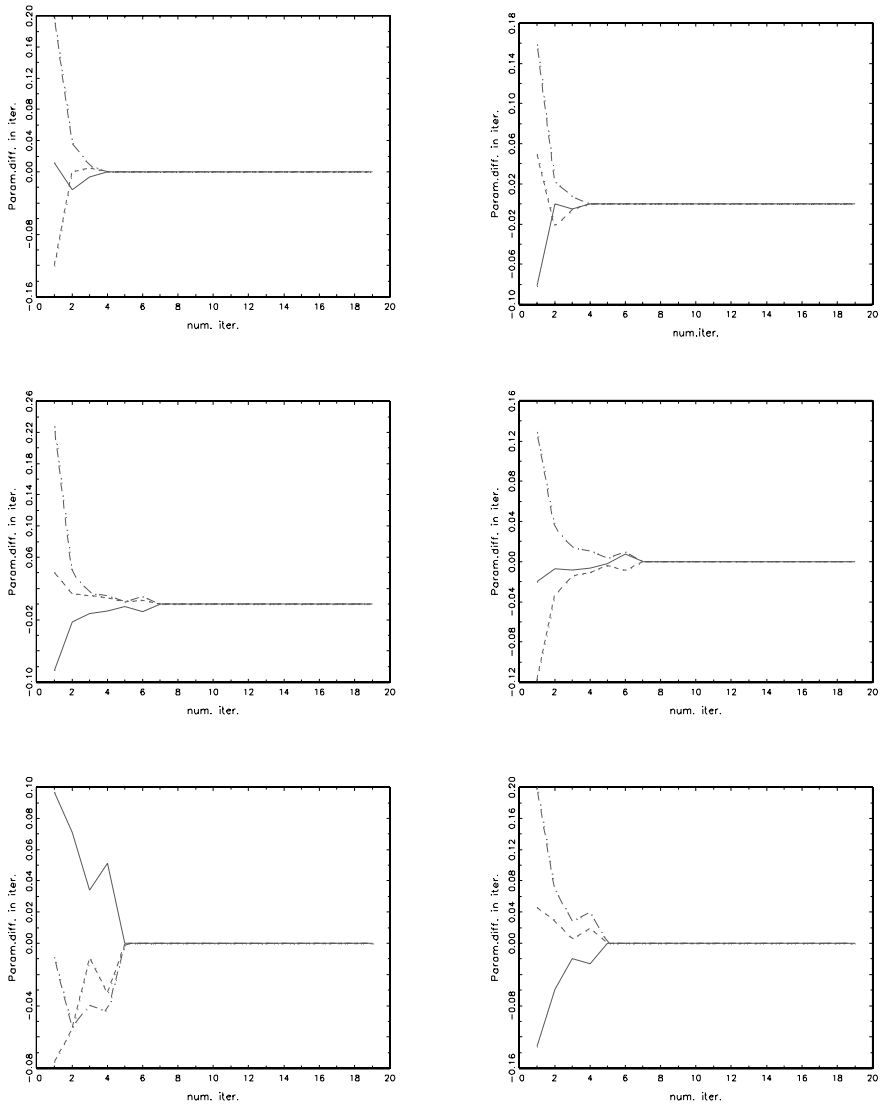


Table 3: Convergences on parameters in different replications

This mechanism creates an amount of missing data equal to 14.8% for y^1 and 21.8% for y^2.

Different trials with different number of iteration for a single replication has been realized. In each trial the number of iterations for reaching the convergence has been less then 20. For this reason the maximum number of iterations for each replication has been fixed to 20. The results have been observed across 1200 replications. In table 3 are reported the differences of the parameters estimates across the iterations for both models on $\mathbf{y}^1$and $\mathbf{y}^2$. In each case the convergence is obtained in less then 10 iterations.

To compare the estimates obtained after the imputation with the real ones a $t - ratio$ test has been computed; the results are shown in Table 1, with the following null hypothesis:

$$H_0 : \Delta = parameter\ estimate - true\ value = 0 \tag{9}$$

The statistic Δ has an asymptotic t distribution with $rep - 1$ $g.l.$, where rep is the number of replications. Table 1 shows the null hypothesis is accepted for each parameter of the model.

Table 2 presents the distributions of the differences Δ for each parameter that are concentrated around zero, according to the results in Table 1.

5 An application on school to work panel survey

The aim of this section is to illustrate the use of the imputation method, described above, in the context of a panel Italian survey, named LEVA, conducted by the Statistical Office of the Regione Lombardia. The LEVA survey is designed to follow some cohorts of pupils in order to study their transitions from school to working life over a period of six/seven years after the compulsory education.

The information on the schooling/working condition is collected every year by a mail questionnaire, except the first and last interview that is planned as a face-to-face interview. In particular, the first interview is a class interview; for this reason it is possible to have many information on household and individual variables for the whole sample.

It is well known that generally mail questionnaires present high rates of nonresponse, so to perform a valid inferential analysis on these data it is

necessary to apply a strategy for compensating missing data. The imputation method outlined before has been chosen. In fact, wave nonresponse on a single variable across time can be considered as an item nonresponse in the corresponding longitudinal record (Lepkowski, 1987).

The cohort of 1986 has been considered; in particular the attention has been focused on the first three waves based on mail questionnaires.The pattern of observed wave nonresponse is respectively about 35%, 25.3%, 35% of the total sample ($N = 2422$).

5.1 Model specification for the observed data

The purpose of this section is the specification of a model to explain schooling/working life choices. The general respondent i at time $t = 1...T$ stays at school ($y_i^t = 0$) or leaves it ($y_i^t = 1$).We consider only respondents at the three waves considered ($N = 1303$).

Some preliminary analysis (Ghellini *et al.*, 1993) suggest to consider the following factors as relevant determinants of the students' choices -all variables have been dichotomized to 0/1-: gender, SEX (1 male); regularity in the compulsory schooling path, REGRIT (1 regular); parents educational level TISTUE (1 low); mother employment status, COPROM (1 employed); father sector activity, POPAD (1 blue-collar); presence of younger brothers/sisters, FA7 (1 yes); presence of both parents, FA2 (1 yes); area school level, AREA (1 high). All the overmentioned variables are considered time-invariant[3], and fixed at time $t = 0$.

The specified model is a time sequence logit (see section 2). At time $t = 1$ we estimate the effect of covariates on the choice to stay at or leave the school. At time $t = 2$ we insert in the model the lagged variable $y^1$and at time $t = 3$ we put in the model both lagged variables y^1 and y^2. Table 4 shows parameters estimates and relative standard errors for the time sequence logit models estimated on the units always observed.

Some interesting remarks about the persistence of the covariates effects can be made. At time $t = 1$ the most part of individuals/household factors are significant, their effects become less strong when the lagged variables are held into account. As expected, it seems the effect of first lagged value has the strongest influence on the present decision to stay at or drop out the

[3]We know that the hypothesis of time invariance for some of these variables can be strong. Anyway, this assumption is constrained by the panel information we have.

Table 4: Maximum Likelihood Estimates of Time-Sequence Logit Models on the observed data (standard errors)

	logit(y^1)	logit(y^2)	logit(y^3)
Intercept	-0.7749 (0.392)**	-1.7525 (0.444)***	-1.9196 (0.536)***
γ (lagged effects)			
y^1	.	4.0800 (0.333)***	1.6110 (0.551)***
y^2	.	.	4.8580 (0.386)***
β (covariates)			
SEX	0.2886 (0.189)	0.2846 (0.185)	0.1602 (0.203)
REGRIT	-2.3037 (0.244)***	-1.7409 (0.314)***	-0.9948 (0.434)**
POPROP	0.5162 (0.205)**	0.7674 (0.200)***	0.5966 (0.216)***
COPROM	-0.3933 (0.216)*	-0.2911 (0.204)	-0.1482 (0.217)
TISTUE	1.1629 (0.259)***	1.3615 (0.246)***	0.8745 (0.239)***
AREA	-0.8071 (0.193)**	-0.2592 (0.190)	-0.2752 (0.210)
FA2	1.0785 (0.388)**	1.1156 (0.415)*	0.2364 (0.554)
FA7	0.0270 (0.192)	0.2142 (0.186)	0.0763 (0.204)

school. It is also to underline that regularity in the compulsory school, father sector activity and parents educational level have a persistent effect across time.

5.2 Imputation and Results

The results shown above are based on the subsample of respondents in three waves. For the reasons explained at the beginning of the section 5, it is reasonable that the results can be biased by the high presence of wave non-response.

Some preliminary analysis conducted with logit models has shown that regularity at the compulsory school and parents educational level influence persistently the nonresponse. In particular the subsample of respondents seems to be composed of the individuals regular at the compulsory school with parents having an high educational level. Then using the methodology presented in section 3.2, the imputation model consider as covariates the same introduced in Table 4.

The imputation strategy adopted on real data can be outlined as follow:

1. a preliminary run was performed to guess starting values for y^1, in this

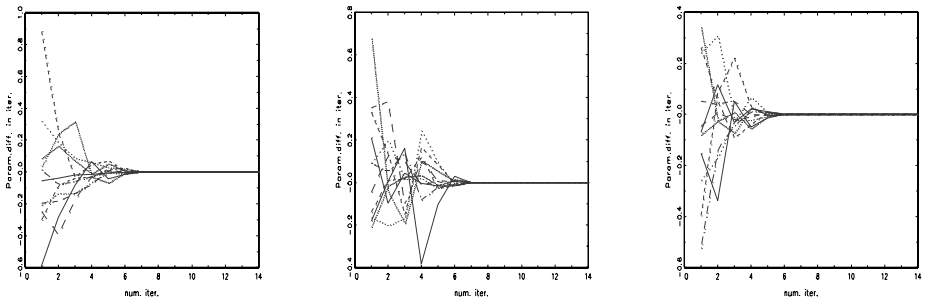


Table 5: Convergence on parameters

step, guess values for missing data on y^1 are obtained performing steps described in section 3.2;

2. a similar logistic regression model was used for guessing values on y^2, excepting that the covariates matrix at this steps was defined as $(\mathbf{x}^t, y^{1*})$, where y^{1*} includes the observed and imputed values obtained in the previous step;

3. again a similar logistic regression model was used for guessing values on y^3, now the covariate matrix was defined as $(\mathbf{x}^t, y^{1*}, y^{2*})$, where y^{2*} includes the observed and imputed values too;

4. the data matrix has been completed and next imputations, generated across the iterations drawn from (5).

As in the simulation study, we choose to monitor the convergence through the behavior of successive values of the parameters of the applied model, instead that monitoring successive values of y^t_{mis}. For each model differences between parameters estimates across iterations has been plotted. Table 5 shows the convergence for the whole set of parameters is reached within 20 iterations.

Multiple imputations ($m = 5$) has been created repeating the steps of the imputation strategy; each imputation is independent respect to each other, because it starts from different random-number generator seed.

A preliminary analysis on the observed and completed data after the five imputation (1), (2), (3), (4), (5) considering the school to work paths has been conducted.

Table 6: Schooling/working paths across waves

Paths	**Observed %**	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
000	69.3	61.80	60.0	61.10	58.50	59.0
001	7.80	6.27	6.20	6.70	7.30	7.80
010	0.38	0.60	0.28	0.24	0.37	0.37
011	10.3	13.50	11.60	11.30	15.10	11.00
100	0.46	0.53	0.66	0.70	0.33	0.45
101	0.46	0.33	0.28	0.20	0.24	0.61
110	0.23	0.37	0.16	0.16	0.12	0.12
111	10.8	16.40	20.64	19.30	17.70	20.30

Reminding that 0 means stay at school at time t and 1 means to leave it, the estimates of the paths after the imputations (see Table 6), seems to be reasonable. In fact the unusual paths 010, 100, 101, 110 present either on the observed and on the completed data low percent frequency. On the contrary, the comparisons for the more usual patterns 000 -indicating people staying at school across waves– and 111 -relative to definitive drops-out from school- indicate an overestimation in the observed data for pattern 000 and an underestimation for 111.

These results reflect the fact people always staying at school (000) has an higher response rate than people leaving school. So inference based on the observed data is biased towards a longer school life.

As shown in Table 6, the imputation strategy adopted seems to work on the right direction to adjust the nonresponse selection bias in the sample.

The next step of the analysis concerns the estimation of a time sequence logit model on the completed data, focusing the attention on the comparison in terms of efficiency the inference deducted on the observed and on the completed data. To perform this analysis we estimated the time sequence logit model for each completed data set and we combined inference results according to the formulation introduced in Rubin (1987) and Schafer (1997).

The result of the analysis are summarized in Table 7. For each coefficient are displayed the combined point estimate $\overline{Q}$ and their standard errors computed as the root square of the estimate of the total variance T. The hypothesis of $Q = 0$ against the two side alternative (where Q represent each parameter of the logistic regression models) is based on the t-statistic $\frac{\overline{Q}}{\sqrt{T}}$, where the degrees of freedom ν of the Student's t-approximation are

Table 7: Multiple imputation inference for time sequence logit regression: model **logit**(y^3)

	logit(y^1)	logit(y^2)	logit(y^3)
Intercept	-0.8404 (0.394)*	-1.4688 (0.363)***	-2.067 (0.476)***
γ (lagged effects)			
y^1	-	4.3310 (0.268)***	1.4457 (0.637)**
y^2	-	-	5.2980 (0.457)***
β (covariates)			
SEX	0.1842 (0.176)	0.2897 (0.170)	0.3377 (0.164)
REGRIT	-2.1941 (0.288)***	-1.7680 (0.270)***	-0.8004 (0.517)**
POPROP	0.3857 (0.154)**	0.7741 (0.216)***	0.4726 (0.173)***
COPROM	-0.0835 (0.178)	-0.3845 (0.214)*	-0.0127 (0.238)
TISTUE	1.3579 (0.266)***	1.2344 (0.279)***	0.7524 (0.279)**
AREA	-0.5269 (0.170)***	-0.3328 (0.148)**	-0.3508 (0.189)*
FA2	0.5421 (0.291)*	1.1242 (0.336)***	0.2201 (0.485)
FA7	0.1334 (0.246)	0.3133 (0.202)	-0.2072 (0.176)

calculated as $\nu = (m-1)\left[1 + \frac{\overline{U}}{(1+m^{-1})B}\right]$, where $\overline{U}$ is the within-imputation variance, B the between-imputation variance and m the number of imputations (see Schafer 1997 for details).

Comparing the results obtained in the observed data (Table 4) and the ones obtained after the imputation (Table 7), we note that in both cases the effect of covariates become smaller across time, since lagged values become the most important effects. In addition all the parameters estimates hold the same signs but we observe that some effect not significant before become significant and some other become more significant on the model estimated on the completed data. In particular, the effect of the variable AREA becomes significant for y^2 and y^3, meaning that people living in a zone with a high school level involves a minor probability to leave the school in each time; the effect to have a mother employed (COPROM) becomes significant also for y^2, that means a persistent effect of a specific family background on the choice to leave the school.

The effect of the variable FA2 is stronger and becomes more significant, that is: the absence of both parents makes the probability to leave the school stronger both for the initial state of the process and for one year after.

6 Final Remarks

The present work is an attempt to use a multivariate imputation strategy for binary panel data. The applied method seems to adjust the nonresponse bias in the right direction, as shown in section 5; for this reason, the assumptions made on the missing data mechanism and on the prior distribution seems reasonable.

Some interesting remarks can be made regard the possibility of considering more complex data set with missing data. In effect in real data some covariates x^t cannot be observed for each t; in this situation the imputation alghoritm has to be modified to take into account missing data on x^t.

We note that our method made the assumption of observations independence. In LEVA survey this hypothesis could be dropped considering correlation of elementary units within school classes using multilevel models.

From an applied point of view, we have neglected the presence of some unobservable factors that can influence the choices during the three years of analysis. It would be interesting to consider in the future work also this aspect, specifying sequential logit models with unobservable components.

References

Bernardi L., Ghellini G., Penello C., (1996), A Panel Survey Design for the Study of School Careers and Work Paths, *Proceedings of the forth International Social Science Methodology Conference*, Colchester, Univ. of Essex, 1-4-July 1996, Mimeo.

Clogg C.C., Ealison S.R., Grego J.M., (1990), Models for the Analysis of change in discrete variables, in A. Von Eye (ed.), *Statistical Methods in Longitudinal Research*, vol. II, Academic Press.

Duncan G.J., Kalton G., (1987), Issue of Design and Analysis of Survey Across Time, *International Statistical Review*, 55, 97-117.

Gelman A., Carlin J.B., Stern H.S., Rubin D.B. (1995), *Bayesian Data Analysis*, Chapman & Hall.

Ghellini G., Lauro S, Lemmi A., Regoli A., (1993), The Transition from School to Working Life: a Longitudinal Approach, *Statistica anno LIII*, n. 3.

KALTON G., KASPRZYK (1986), The Treatment of Missing Survey Data, *Survey Methodology*, **12**, 1–16.

KALTON G., (1986), Handling Wave Nonresponse in Panel Surveys, *Journal of Official Statistics*, vol.2, n°3,303-314.

LEPKOWSKI J., (1989), Treatment of Wave Nonresponse in Panel Surveys, *Panel Surveys,* Kasprzyk D., Duncan G.J., Kalton G., Singh M.P., Wiley & Sons.

LI K.H., RAGHUNATHAN T.E., RUBIN D.B. (1991), Large-Samples Significance Levels from Multiply Imputed Data Using Moment-Based Statistics and an F Reference Distribution, *Journal of the American Statistical Association*, **86**, 1065-1073.

LITTLE R.J.A. (1988), Missing Data Adjustment in Large Surveys, *Journal of Business and Economics Statistics*, **6**, 287-301.

LITTLE R.J.A. A., RUBIN D.B. (1987), *Statistical Analysis with Missing Data*, J. Wiley, New York.

LITTLE R.J.A. A., RUBIN D.B. (1989), The Analysis of Social Science Data with Missing Values, *Sociological Methods and Research,* **18**, 292-326.

O MUIRCHEARTAIGH (1989), Measurement Errors in Panel Surveys: Implications for Survey Design and for Survey Analysis, *Atti della XXXVIII Riunione Scientifica della Società Italiana di Statistica,* vol.1, 208-218.

RAGHUNATHAN T.E., LEPKOWSKI J., VAN VOEWYK J., SOLENBERGER P. (1997), A Multivariate Technique for Imputing Missing Values Using a Sequence of Regression Models, Technical Report, Survey Methodology Program, Survey Research Center, ISR, University of Michigan.

RUBIN D.B. (1976), Inference with Missing Data, *Biometrika*, **63**, 581-592.

RUBIN D.B. (1978), Multiple Imputation in Sample Surveys -A Phenomenological Bayesian Approach to Nonresponse, *Proceeding of the Survey Research Method Section, American Statistical Association,* 20-34.

RUBIN D.B. (1987), *Multiple Imputation for Nonresponse in Survey*, J. Wiley, New York

SCHAFER J. L. (1997), *Analysis of Incomplete Multivariate Data*, Chapman & Hall, London.

TANNER M.A., WONG W.H. (1987), The Calculation of Posterior Distribution by Data Augmentation, *Journal of the American Statistical Association*, **82**, 528-550.

Un metodo di imputazione multipla per dati panel sulla transizione scuola lavoro

Riassunto

I dati panel, pur presentando maggiori problemi di qualità dei dati legati alla presenza di non risposte, nelle loro varie eccezioni, offrono anche interessanti opportunità di studiare in modo più approfondito, rispetto ai dati trasversali, i processi che generano le non risposte.

In questo lavoro, sulla base di un data set panel sulle transizioni giovanili al lavoro, viene approfondita in particolare la non risposta longitudinale, e viene proposta la specificazione e la stima di un modello a scelta discreta di tipo sequenziale. Tramite tale modello si procede ad una imputazione multipla di una variabile (la condizione scolastico lavorativa) in tre occasioni di indagine adiacenti.

Keywords: panel non response, multiple imputation, sequential logit models

Working Papers già pubblicati

1. E. Battistin, A. Gavosto e E. Rettore, *Why do subsidized firms survive longer? An evaluation of a program promoting youth entrepreneurship in Italy*, Agosto 1998.
2. N. Rosati, E. Rettore e G. Masarotto, *A lower bound on asymptotic variance of repeated cross-sections estimators in fixed-effects models*, Agosto 1998.
3. U. Trivellato, *Il monitoraggio della povertà e della sua dinamica: questioni di misura e evidenze empiriche*, Settembre 1998.
4. F. Bassi, *Un modello per la stima di flussi nel mercato del lavoro affetti da errori di classificazione in rilevazioni retrospettive*, Ottobre 1998.
5. Ginzburg, M. Scaltriti, G. Solinas e R. Zoboli, *Un nuovo autunno caldo nel Mezzogiorno? Note in margine al dibattito sui differenziali salariali territoriali*, Ottobre 1998.
6. M. Forni e S. Paba, *Industrial districts, social environment and local growth. Evidence from Italy*, Novembre 1998.
7. B. Contini, *Wage structures in Europe and in the USA: are they rigid, are they flexible?*, Gennaio 1999.
8. B. Contini, L. Pacelli e C. Villosio, *Short employment spell in Italy, Germany and Great Britain: testing the "Port-of-entry" hypothesis*, Gennaio 1999
9. B. Contini, M. Filippi, L. Pacelli e C. Villosio, *Working careers of skilled vs. unskilled workers*, Gennaio 1999
10. F. Bassi, M. Gambuzza e M. Rasera, *Il sistema informatizzato NETLABOR. Caratteristiche di una nuova fonte sul mercato del lavoro*, Maggio 1999.
11. M. Lalla e F. Pattarin, *Alcuni modelli per l'analisi delle durate complete e incomplete della disoccupazione: il caso Emilia Romagna*, Maggio 1999.
12. A. Paggiaro, *Un modello di mistura per l'analisi della disoccupazione di lunga durata*, Maggio 1999.
13. T. Di Fonzo e P. Gennari, *Le serie storiche delle forze di lavoro per il periodo 1984.1-92.3: prospettive e problemi di ricostruzione*, Giugno 1999.
14. S. Campostrini, A. Giraldo, N. Parise e U. Trivellato, *La misura della partecipazione al lavoro in Italia: presupposti e problemi metodologici di un approccio "time use"*, Ottobre 1999.
15. A. Paggiaro e N. Torelli, *Una procedura per l'abbinamento di record nella rilevazione trimestrale delle forze di lavoro*, Ottobre 1999.
16. A. D'Agostino, G. Ghellini e L. Neri, *A Multiple Imputation Method for School to Work Panel Data*, Ottobre 1999.

Le richieste di working papers vanno indirizzate alla Sig.ra Patrizia Piacentini, Dipartimento di Scienze Statistiche, via C. Battisti 241-243, 35121 Padova, oppure all'indirizzo e-mail patrizia.piacentini@unipd.it.