Predictive Mean Matching using a factor model, an application to the (Business) Multipurpose Survey

Roberta Varriale, Ugo Guarnera
Italian National Statistical Institute

In 2012 the Italian National Institute of Statistics (Istat) carried out - together with the Business census - a Multipurpose sample Survey (MPS) on enterprises, involving 260,110 enterprises from a target population of about 1.6 million.

MPS survey design is the result of the combination of a census and a sample survey. In particular, it is composed of: census survey on medium and large enterprises (more than 20 employees); sample survey on micro and small firms (3 to 19 employees); sample survey on a subset of micro-enterprises with less than 3 employees (with non-microenterprises characteristics). Enterprises with more than 20 employees are 72,771, representing the 5% of the target population.

The MPS utilizes two models of questionnaires according to the dimension of the enterprises ([6]). Both questionnaires are quite complex and aim at identifying specific *business profiles*.

Out of the 72,771 enterprises with more than 20 employees, the non-respondent units are about 15%. We managed missing values with imputation techniques mainly for two reasons: first, the availability of auxiliary information, in the majority continuous, on the entire dataset and the complexity of the phenomenon under investigation makes weighting methods complex to apply; second, Istat needs to meet specific demands of external users carrying out analysis with standard software.

In order to overcome the difficulties related to the high number of variables to be analyzed, a "natural" imputation method is the Nearest Neighbor Donor (NND) that is matching completely observed units (donors) with incomplete units (recipients), based on some distance function, and transferring values from donors to recipients. We used a Predictive Mean Matching (PMM) approach ([4]), that is a NND imputation technique based on a distance function where the auxiliary variables are weighted through their predictive power with respect to the variables that have to be imputed.

In a multivariate context with continuous target variables, a typical application of the PMM uses a regression model to estimate the relevant predictive means for both complete and incomplete units. Then, a distance function based on these predictive means is used to select donors. In this work, we propose a version of the PMM based on a factor model in order to deal with categorical target variables (like the ones in MPS): this approach allows us to define "similarity" between donors and recipients in terms of the predicted values of a single numeric (latent) variable,

regardless of the nature of the target variables. As a typical application of the PMM with continuous target variables, the selection of an appropriate distance function is based on the preservation of the distributional characteristics of certain "target" variables, usually indicated by subject matter experts. The important advantages of the PMM based on a factor model are that: it can deal with indicators of any scale type and allows us to deal with a big number of target variables. In the work, the proposed PMM technique is compared with common imputation methods, such as a genuine NND based on Euclidean distance. For the comparison, data from the 2012 Istat MPS are used.

The paper is organized as follows. In Sect. 2, the PMM method using a latent variable is illustrated. The evaluation of the method through a simulation study is described in Sect. 4.

2 Predictive Mean Matching with latent variables

The PMM is an NND imputation technique based on the minimum distance donor, where the influence of covariates on the selection of donors depends on the their predictive power with respect to the variables that have to be imputed.

Let $Y_1,\ldots,Y_H$ be the variables of a sample survey to be imputed, and let $X_1,\ldots,X_Q$ be the set of variables available for all units (covariates). Let Y_{obs} be the observed values in Y, Y_{mis} the missing values in Y, and X_{obs} the covariates observed for all units. In a multivariate context, when the variables are continuous and in presence of arbitrary patterns of missing items, a typical application of the PMM is the following ([3]):

1. The parameters of the conditional distribution of Y given X are estimated with standard methods using all the available data (complete and incomplete).

2. Based on the estimates from step 1, for each missing pattern predictive means $Y^* \equiv E(Y_{mis} \mid X, Y_{ob})$ are computed for both incomplete units and complete data.

3. For each recipient u_r, a donor u_d is selected in order to minimize the Mahalanobis distance $D(u_d, u_r) \equiv (y_d^* - y_r^*)^T S^{-1} (y_d^* - y_r^*)$, where y_d^* and y_r^* are the predictive means estimates on donor and recipient, respectively, and S is the relevant residual variance-covariance matrix of the regression model corresponding to the current missing pattern.

4. Each u_r is imputed by transferring the Y values from its closest donor.

For (approximately) continuous variables, the most widely used imputation model is the multivariate normal model. Whereas the most "natural" choice of an imputation model for categorical variables is the log-linear model, there are two main limitations in the use of PMM based on log-linear models: first, it can be applied only when the number of variables used in the imputation model is small (i.e. only when we are able to set up and process the full multi-way cross-tabulation required for the log-linear analysis) ([7]); second, the distance function (as the

Mahalanobis distance for continuous variables) has to take into account both the distance between the expected frequencies of the multi-way cross-tabulation and the variability due to the estimation process. We propose a PMM using a latent variable to overcome these two limitations. Furthermore, the PMM using a latent variable can be used to deal with target variables of various scale types.

The model (also known as Structural Equation Model ([2]) or Multiple Indicators Multiple Causes model) is composed of two parts: a factor model linking the latent factor to the observed indicators and a regression model, linking the covariates to the latent factor. Figure 1 depicts the model. Following the conventions, circles represent latent variables and rectangles observed variables, arrows connecting latent and/or observed variables represent direct effects, which do not need to be linear, and arrows pointing at latent or observed variables represent residuals.

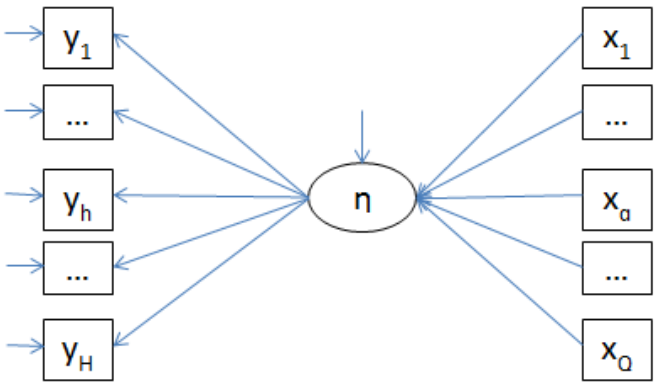

Fig.1 – Factor model with covariates.

Factor Analysis (FA) is a statistical method for describing the associations among sets of observed variables in terms of a small number of underlying continuous latent variables ([1]). We use the term factor model to refer to models for both continuous and categorical indicators. Even if the extension to more latent variables is straightforward, we present the model used for PMM that has only one factor.

Let y_{hi} denote the observed response of individual i ($i = 1, \ldots, N$) on indicator h ($h = 1, \ldots, H$); and h_i the unobserved score of individual i on common factor η, where N and H are the total number of individuals and items. In FA, a series of H regression models are used to define the relationships between the latent variable η_i and the indicators y_{hi}. To accommodate for the various possible scale types of the indicators, we use response models from the generalized linear modeling family, which are specified via a linear predictor v_{hi}, a link function $g(.)$, and an error distribution from the exponential family ([5]).

In a factor analytic model the linear predictor has the following form:

$v_{hi} = \mu_h + \lambda_h \eta_i$

where μ_h is an item intercept and λ_h a factor loading.

The linear predictor is connected to y_{hi} as follows:

$g(E(y_{hi} \mid \eta_i)) = v_{hi}$.

After applying an appropriate transformation $g(.)$, the expected value of y_{hi} conditional on the latent factors equals the linear predictor. The choice of the link function depends on the scale type of the indicators. The definition of the H response models is completed by the specification of the distribution of the indicators' residuals $e_{hi} = y_{hi} - E(y_{hi} \mid \eta_i)$ or, equivalently, of the conditional density of y_{hi} given the latent variables $f(y_{hi} \mid \eta_i)$.

The relationship between the continuous latent factor h and the covariates $X_1, \ldots, X_Q$ is expressed by the regression model:

$\eta_i = \alpha_0 + \alpha_1 X_1 + \ldots + \alpha_Q X_Q + u_i$

where u_i represents the error component of the model. Usual assumption on the regression model are made.

The PMM with latent variables adapts the typical process of the PMM as follows:

1. The parameters of a factor model are estimated using all the available data.

2. Based on the estimates of the regression component of the factor model, for each missing pattern predictive means $\eta^* \equiv E(\eta \mid X)$ are computed for both incomplete units and complete data.

3. For each recipient u_r, a donor u_d is selected in order to minimize the distance between the predictive means η^*.

4. Each u_r is imputed by transferring the Y values from its closest donor.

3 Empirical evaluation

In order to evaluate the PMM using a factor model we used data from the MPS, available at the end of December 2012, with a total number of observations equal to N = 3982.

The target variables are: type and nationality of decision management (Y_1, 4 categories), employees with high skills (Y_2, 2 categories), type of partnership (Y_3, 3 categories) and delocalization of specific production functions (Y_4, 2 categories). As covariates in the models we used the number of employees (X_1, continuous), added value (X_2, continuous), turnover (X_3, continuous), membership in an enterprise group (X_4, 2 categories). The variables Section of economic activity and Export/import activity have been used as stratification variables in the imputation process.

We compared the PMM using a factor model (Factor.Donor) with other imputation methods commonly used in the context of official statistics. Specifically, other two NND imputation methods have been considered, both using an Euclidean distance to match recipient and donor units. In the first one, X.Donor, the matching variables are $X_1,\ldots, X_Q$, while in the second one, Logit.Donor, the matching variables are the probabilities of each category of $Y_1,\ldots,Y_H$, estimated through a multinomial logit model with $X_1,\ldots, X_Q$ as explanatory variables. The estimated probabilities from the multinomial logit model and the factor model have also been used to directly draw realizations of $X_1,\ldots, X_Q$ (methods Logit.Rnd and Factor.Rnd, respectively).

To evaluate the additional variability introduced by the random drawing, we also computed the expected frequencies according to the different models (Logit, Factor). It should be noted that the multinomial logit model and the factor model assume that the Y variables are independent conditionally on the covariates and the latent variable, respectively.

The experiment is based on a Monte Carlo simulation study with 200 replications. At each replication, we first simulate item nonresponse (20% of the total number of observations) on Y variables according to a Missing at Random (MAR) mechanism (Little and Rubin 2002), where the nonresponse probabilities for $Y_1,\ldots, Y_4$ depend on the observed values of the variable X_3: the higher the turnover, the higher the nonresponse probability. Subsequently, we estimate the marginal and joint frequencies corresponding to the dropped units using each method previously described. Finally, we evaluate the different methods averaging the Hellinger distance between the true and estimated frequencies obtained at each iteration. Frequencies are compared separately for each estimation domain defined by the Section of economic activity (X_D).

Table 1. Simulation study on BCS data, Hellinger distances evaluated in the estimation domains defined by the Section of economic activity.

	Logit	Factor	Logit.Rnd	Factor.Rnd	X.Donor	Logit.Donor	Factor.Donor
All	0.2247	0.2024	0.2442	0.2232	0.2016	0.2016	0.1998
Y_1	0.1201	0.0956	0.1246	0.0997	0.0976	0.0970	0.0953
Y_2	0.0587	0.0712	0.0632	0.0744	0.0676	0.0716	0.0680
Y_3	0.0833	0.0636	0.0871	0.0689	0.0773	0.0809	0.0788
Y_4	0.0575	0.0568	0.0615	0.0614	0.0618	0.0644	0.0636

Table 1 shows the results of the MC simulation study. The Hellinger distances for each method are quite similar. The reduction of dimensionality performed by the PMM with a factor model seems to not harm the results of the imputation process compared to those obtained with Donor.Logit and X.Donor. The performance of the NND methods is very similar to that one of the

corresponding methods based on directly drawing from the estimated probability distribution (columns 6-7 vs columns 3-4). The advantage of using a NND is that it allows us to impute all variables of each incomplete record, rather than only the "target" variables. Finally, as expected, the additional variability introduced by the random drawing methods results in a small increase of the Hellinger distance values (columns 3-4 vs columns 1-2). The results show a quite good performance of the PMM with a factor model, thus encouraging further research to carefully exploit all the features of the proposed method.

References

1. Bartholomew, D.J., Knott, M. (1999). *Latent variable models and factor analysis*. Arnold, London
2. Bollen, K.A. (1989). *Structural equations with latent variables*. J.Wiley, New York
3. Di Zio, M., Guarnera, U. (2009). Semiparametric predictive mean matching. *AStA - Advances in Statistical Analysis.* 93, 175-186
4. Little, R.J.A. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*. 6, 287-296
5. Skrondal, A., Rabe-Hesketh, S. (2004). *Generalized latent variables modeling: multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC
6. Lombardi, S., Lorenzini F., Verrecchia F. (2012). Three Pillars for a New Statistical System on Enterprises: Business Register, Thematic Surveys and Business Census 2011. In: Fourth International Conference on Establishment Surveys (ICES-IV), Montreal, Canada, 11-14 June
7. Vermunt, J.K., van Ginkel, J.R., van der Ark, L.A., Sijtsma K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology* 33, 369–297