Predictive Mean Matching using a Factor Model, an application to the Business Multipurpose Survey

Roberta Varriale, Ugo Guarnera
Istat

Nuremberg, 09/09/2013

Istat

Motivation: (Business) Multi Purpose Survey

Sample survey on businesses carried out by Istat together with the Business and Services Census 2011

Features

Aims: identifying specific ***business profiles***
Complex questionnaire
Large number of variables (different nature)

Nonresponse-rate of 72k enterprises (more than 20 employees): 15%
Difficulty to use an explicit parametric method for imputation

Nearest Neighbor Donor (NND): consists in matching completely observed units (*donors*) with incomplete units (*recipients*), based on some distance function, and transferring values from donors to recipients

Predictive Mean Matching (PMM): distance function "weights" the covariates used in NND with its relative predictive power with respect to the "target" variables

Predictive Mean Matching

Unità	Y1	Y.	YP	X1	X.	XQ
1	$Y_{missing}$			$X_{observed}$		
.						
.						
.						
.	$Y_{observed}$			$X_{observed}$		
.						
.						
N						

Unità	Y_1	Y.	YP	Y*1	Y*.	Y*P	X1	X.	XQ
1	$Y_{missing}$			Y^*_{11}	$Y^*_{1.}$	Y^*_{1P}	$X_{observed}$		
.				.	.	.			
.				.	.	.			
.				.	.	.			
.	$Y_{observed}$			.	.	.	$X_{observed}$		
.				.	.	.			
.				.	.	.			
N				Y^*_{N1}	$Y^*_{N.}$	Y^*_{NP}			

Predictive Mean Matching

$Y = Y_1,\ldots,Y_P$ variables of a sample survey to be imputed
$X = X_1,\ldots,X_Q$ variables available for all units (covariates)

Y continuous, a typical application of PMM:

1. the parameters of the regression model of Y on X are estimated with standard methods
2. based on the estimates from step 1, predictive means $Y^* \equiv E(Y \mid X)$ are computed both for the units with missing values (recipients) and units with complete data (donors)
3. for each receiver u_r a donor u_d is selected in order to minimize the Mahalanobis distance defined by:

   $$D(u_d, u_r) \equiv (y_d^* - y_r^*)T\, S^{-1} (y_d^* - y_r^*)$$

   where y_d^* and y_r^* are the predictive means estimates on donor and recipient, respectively, and S is the residual variance-covariance matrix of the regression model
4. each u_r is imputed by transferring the Y values from its closest donor

Predictive Mean Matching

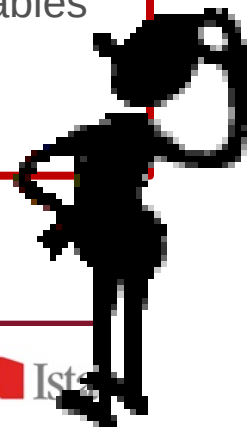
PMM → "weights" the covariates to be used in NND with its relative predictive power with respect to the target variables

Mahalanobis distance (residual covariance matrix from the regression model) → establishes the distance between units weighting the "goodness of prediction" of the single target variables → largest "weights" to the target variables of the predictive means with the smallest prediction error

Research problem:

Which *model* and which *distance function* has to be used when the target variables are not all continuous?

MPS: target variables are all categorical



Predictive Mean Matching

In PMM for categorical variables to be imputed,

- the “natural” imputation model to define an appropriate distance is the log-linear model with covariates *X*
- distance function on the estimated probabilities of the categories of the target variables

2 main limitations:

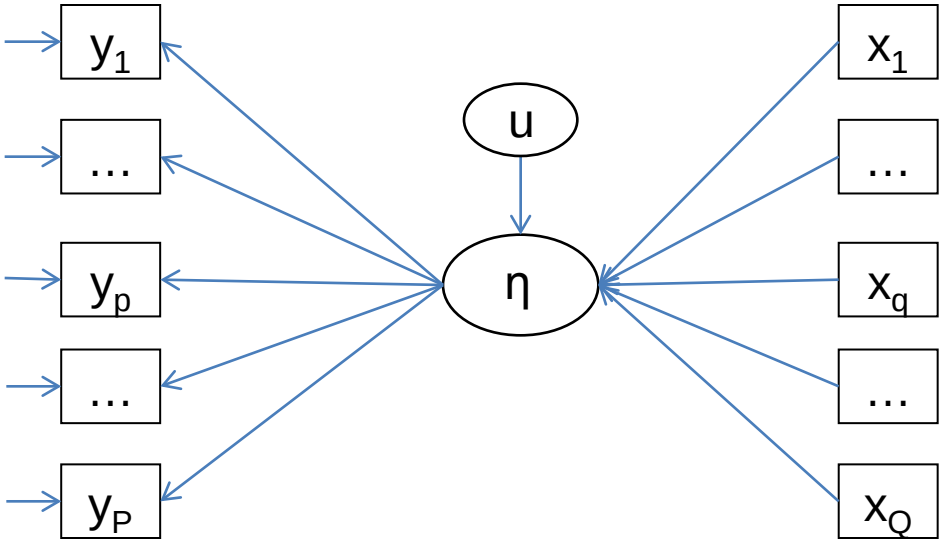
1. the model is very complex when the number of variables used in the imputation model is big (i.e. only when we are able to set up and process the full multi-way cross-tabulation required for the log-linear analysis)
2. the distance function has to take into account both the distance between the expected frequencies of the multi-way cross-tabulation and the variability due to the estimation process

PMM with latent variables (factor model)

Factor model (with covariates)

The model is composed of 2 parts:

1. factor model → links the latent factor to the observed indicators
2. regression model → links the covariates to the latent factor



Structural Equation Model

MIMIC (Multiple Indicators Multiple Causes) model

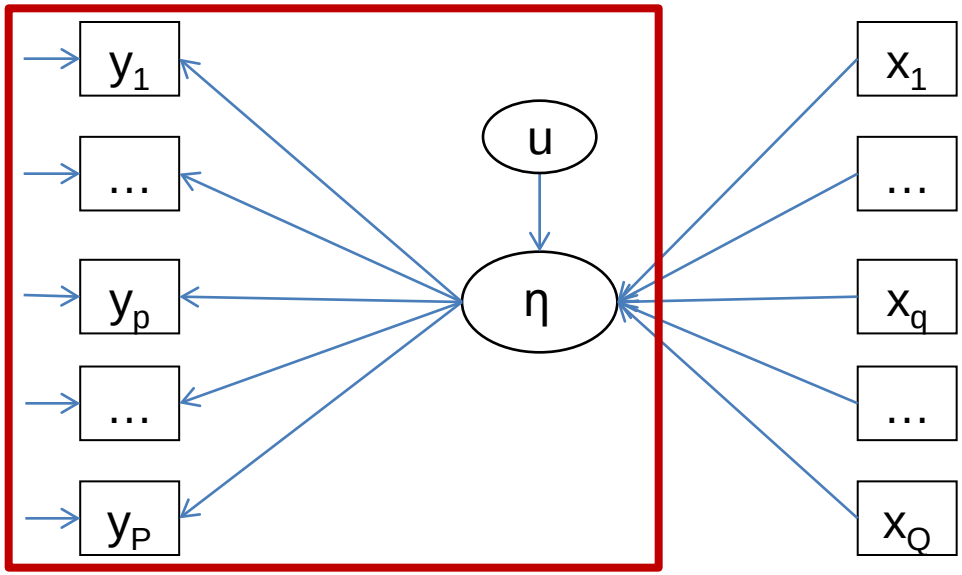
Factor model (with covariates)

The model is composed of 2 parts:

1. factor model → links the latent factor to the observed indicators
2. regression model → links the covariates to the latent factor

$v_p = \mu_p + \lambda_p \eta$

$g_p\,(E(y_p|\eta)) = v_p$

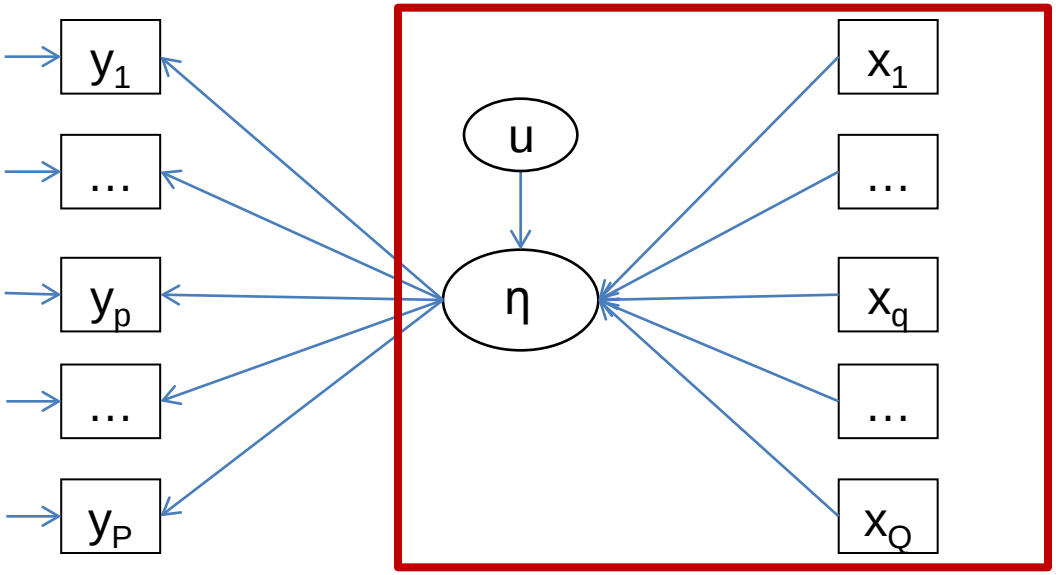


Istat

Factor model (with covariates)

The model is composed of 2 parts:

1. factor model → links the latent factor to the observed indicators
2. regression model → links the covariates to the latent factor



$$\eta = \alpha_0 + \alpha_1 X_1 + \ldots + \alpha_Q X_Q + u$$

Predictive Mean Matching

Unità	Y1	Y.	YP	X1	X.	XQ
1	$Y_{missing}$			$X_{observed}$		
.						
.						
.						
.	$Y_{observed}$			$X_{observed}$		
.						
.						
N						

Unità	Y1	Y.	YP	η	X1	X.	XQ
1	$Y_{missing}$			η^*_1	$X_{observed}$		
.				.			
.				.			
.				.			
.	$Y_{observed}$			.	$X_{observed}$		
.				.			
.				.			
N				η^*_N			

PMM with one latent factor

$Y = Y_1,\ldots,Y_P$ variables of a sample survey to be imputed
$X = X_1,\ldots,X_Q$ variables available for all units (covariates)

Y categorical (or not all continuous), the PMM with a latent factor:

1. the parameters of the factor model with covariates are estimated with standard methods
2. based on the estimates from step 1, predictive means $\eta^* \equiv E(\eta \mid X)$ are computed both for the units with missing values (recipients) and units with complete data (donors)
3. for each receiver u_r a donor u_d is selected in order to minimize the distance between η values
4. each u_r is imputed by transferring the Y values from its closest donor

The "matching variable" of PMM:

(expected value of) the latent factor η

- summarizes data variability (response variables y_p)
- is related to the variables $X_1,..,X_Q$

(Business) Multi Purpose Survey

Target variables:

Y_1: type and nationality of decision management (4 categories)
Y_2: employees with high skills (2 categories)
Y_3: type of partnership (3 categories)
Y_4: delocalization of specific production functions (2 categories)

- 0=other
- 1=physical person/family, Itaian
- 2= other firm, Italian
- 3=abroad

High/not high

- 0=no partnership
- 1= both formal and informal partn.
- 2= only formal partnership

yes/no

Covariate:

X_1: number of employees (continuous)
X_2: added values (continuous)
X_3: turnover (continuous)
X_4: membership in an enterprise group (2 categories)

Stratification variables in the imputation process:

Section of economic activity (ATECO)
Export/import activity

Simulation study

Monte Carlo simulation study

Data MPS, available at 20/12/2012

enterprises: **39826** **(20 or more than 20 employees)**

Software: *R, Latent Gold*

200 iterations

1. **Simulation of the item nonresponse**
   (20% of the total number of observations) on *Y* variables according to a Missing at Random (MAR) mechanism

2. **Estimation of the marginal and joint frequencies of the categories of $Y_1,\ldots,Y_4$ for the dropped units,** using different methods

3. **Evaluation of the different methods by comparing the true and estimated frequencies** obtained at each iteration with the Hellinger distance, and averaging the results over the 200 replications (frequencies are compared separately for each estimation domain defined by the ATECO)

Simulation study

2. **Estimation of the marginal and joint frequencies of the categories of *Y1,…,Y4* for the dropped units,** using different methods

Simulation study

2. **Estimation of the marginal and joint frequencies of the categories of *Y1,…,Y4* for the dropped units,** using different methods

Computation of the expected frequencies according to the estimates obtained by the models:

Logit: multinomial logit model

Factor: factor model with covariates

Simulation study

2. **Estimation of the marginal and joint frequencies of the categories of *Y1,…,Y4* for the dropped units,** using different methods

Computation of the expected frequencies according to the estimates obtained by the models:

Logit: multinomial logit model

Factor: factor model with covariates

Draw realization of $Y_1,\ldots,Y_4$ using the probabilities estimated by the models:

Logit.Rnd: multinomial logit model

Factor.Rnd: factor model with covariates

Simulation study

2. **Estimation of the marginal and joint frequencies of the categories of *Y1,…,Y4* for the dropped units,** using different methods

Computation of the expected frequencies according to the estimates obtained by the models:

Logit: multinomial logit model

Factor: factor model with covariates

Draw realization of $Y_1,\ldots,Y_4$ using the probabilities estimated by the models:

Logit.Rnd: multinomial logit model

Factor.Rnd: factor model with covariates

Imputation method NND (Euclidean distance):

X.Donor, matching variables: $X_1,\ldots,X_4$

Logit.Donor, matching variables: probabilities of each category of $Y_1,\ldots,Y_4$ estimated by a multinomial logit model with covariates $X_1,\ldots,X_4$

Factor.Donor, matching variable: value of the latent factor estimated by a factor model with covariates (**PMM with a latent factor**)

Simulation study - results

	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

- The Hellinger distance computed for each method is similar
- The reduction of dimensionality performed by the PMM with a factor model does not harm the results of the imputation process
- The performance of the NND methods is very similar to that ones of the corresponding methods based on directly drawing from the estimated probability of the categories of *Y* → an advantage of using a NND is that it allows us to impute all variables of each incomplete record, rather than only the target variables
- The additional variability introduced by the random drawing methods result in a small increase of the Hellinger distance values

Simulation study - results

	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

- The Hellinger distance computed for each method is similar
- The reduction of dimensionality performed by the PMM with a factor model does not harm the results of the imputation process
- The performance of the NND methods is very similar to that ones of the corresponding methods based on directly drawing from the estimated probability of the categories of *Y* → an advantage of using a NND is that it allows us to impute all variables of each incomplete record, rather than only the target variables
- The additional variability introduced by the random drawing methods result in a small increase of the Hellinger distance values

Simulation study - results

	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

- The Hellinger distance computed for each method is similar
- The reduction of dimensionality performed by the PMM with a factor model does not harm the results of the imputation process
- The performance of the NND methods is very similar to that ones of the corresponding methods based on directly drawing from the estimated probability of the categories of *Y* → an advantage of using a NND is that it allows us to impute all variables of each incomplete record, rather than only the target variables
- The additional variability introduced by the random drawing methods result in a small increase of the Hellinger distance values

Simulation study - results

	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

- The Hellinger distance computed for each method is similar
- The reduction of dimensionality performed by the PMM with a factor model does not harm the results of the imputation process
- The performance of the NND methods is very similar to that ones of the corresponding methods based on directly drawing from the estimated probability of the categories of *Y* → an advantage of using a NND is that it allows us to impute all variables of each incomplete record, rather than only the target variables
- The additional variability introduced by the random drawing methods result in a small increase of the Hellinger distance values

Future research

The results show a quite good performance of the PMM with a factor model

Further analysis:

- (additional) analysis of the nonresponse process (nonresponse pattern)
- (additional) analysis of the variable “meanings”

} **subject matter**

- Monte Carlo experiments using simulated data

} **methodological matter**

- Suggestions?

Thank you for your attention

varriale@istat.it, guarnera@istat.it

References

Bartholomew, D.J., Knott, M. (1999). Latent variable models and factor analysis. Arnold, London

Bollen, K.A. (1989). Structural equations with latent variables. J.Wiley, New York

Di Zio, M., Guarnera, U. (2009). Semiparametric predictive mean matching. AStA - Advances in Statistical Analysis. 93, 175-186

Little, R.J.A. (1988). Missing-data adjustments in large surveys. Journal of Business & Economic Statistics. 6, 287-296

Skrondal, A., Rabe-Hesketh, S. (2004). Generalized latent variables modeling: multilevel, longitudinal, and structural equation models. Boca Raton, FL: Chapman & Hall/CRC

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

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