

A Comparison Between Simple and Multiple Imputation in Applying Logistic Regression Models

*Un Confronto fra Imputazione Semplice e Multipla nell'Impiego di
Modelli di Regressione Logistica*

Lorena C.M. Viviano, Gianfranco Lovison

Dipartimento di Scienze Statistiche e Matematiche "S.Vianelli"

Università degli Studi di Palermo - Viale delle Scienze 90128 Palermo

email: lorena.viviano@tin.it, lovison@unipa.it

Riassunto: Il lavoro si pone come obiettivo il confronto tra due tecniche di imputazione per dati mancanti sulle sole variabili esplicative. In particolare, facendo riferimento ad uno studio sul ruolo dei possibili fattori di rischio e della familiarità del Morbo di Alzheimer, vengono impiegate l'imputazione semplice e multipla per ottenere un insieme di dati completo che consenta di ricavare le stime di massima verosimiglianza di un modello di regressione logistica. Viene poi effettuato uno studio di sensitività, simulando insiemi di dati con varie frazioni di dati mancanti su due variabili esplicative, allo scopo di osservare come cambia la significatività dei parametri dei fattori di rischio.

Keywords: Simple-Multiple Imputation, Missing Data, Sensitivity Analysis.

1. Introduction

The presence of missing data is a common problem in applied statistics, since collecting real data can involve several difficulties: in recruiting units, in complying with economical or ethical constraints, and so on. The basic terminology and methodology used for missing data have been established in the seminal work of Little and Rubin (1987), but still recently there are many papers which propose methods to overcome such problem. The reason is easy to understand, since only recently the developments in hardware and software have made possible it to implement complex algorithms to manage datasets that include several variables with missing values.

In this work two methods of imputation for missing values on covariates are compared: simple and multiple imputation. Simple imputation replaces missing data by plausible estimates and then applies standard complete-data methods to the filled-in data; it underestimates variance and is consistent under MCAR (missing completely at random) missingness, but may be biased if MCAR does not hold. Multiple imputation is a recent method where a single missing value is replaced by a vector of plausible data and then results are pooled. In our study, in employing simple imputation we assume that the vector of missing data for each variable has the same distribution as the observed data, so that missing units are reallocated in the same proportion as observed units.

Multiple imputation is performed by the library `mice` (Multivariate Imputation by Chained Equations) of S-plus. This library is based on a Monte Carlo method by which a single missing value is replaced by $m > 1$ (usually 3-10) values (Rubin, 1996; Rubin and Schafer, 1998). Rubin (Rubin, 1996) computes the estimate of efficiency as $(1 + (\gamma/m))^{-1}$ where γ is the rate of missing information. For example, with $\gamma = 30\%$ and $m = 5$ efficiency is about 90%. The library is based on Gibbs sampling that works

as follows. Data are collected on k random variables $(X_1, X_2, \dots, X_j, \dots, X_k)$ and each of them is made up of observed and missing values $X_j = (X_{j\text{obs}}, X_{j\text{mis}})$. We have to iterate the generic step: “for X_j , draw X_j^{t+1} from $f(X_j | X_1^{t+1}, X_2^{t+1}, \dots, X_{j-1}^{t+1}, \dots, X_k^t)$ ”. The algorithm imputes an incomplete column (the target column) by generating appropriate imputation values given other columns. Each incomplete column acts in turn as a target column, and has its own specific set of predictors. The default predictor set consists of all other columns in the data set. For predictors that are themselves incomplete, the most recently generated imputations are used to complete the predictors prior to imputation of the target column (Van Buuren and Oudshoorn, 2000). Together with point estimates, *mice* gives standard error and confidence intervals (LCI-UCI), that represent uncertainty over data.

In Section 2 we illustrate some results obtained by simple and multiple imputation in a real study on the Alzheimer disease. In Section 3 we briefly describe a sensitivity analysis concerning the proportion of missing data. Finally, in Section 4 we discuss some concluding remarks.

2. Simple and multiple imputation for covariates in an Alzheimer study

This Section describes an Alzheimer disease case-control study carried out at Palermo General Hospital (Viviano, 2002). We thank Rosa Maria Ruggieri and the medical staff for providing the data. There are 109 cases (67 women, 61.47% and 42 men, 38.53%) and 80 controls (48 women, 60% and 32 men, 40%). Patients come mostly from the Center-West of Sicily and are at different stages of disease. Controls were chosen among healthy spouses, patients hospitalised for physical trauma, elderly persons in convents or hospices. The study considers some possible risk factors: sex, years of school, alipoprotein E genotype (ApoE), smoking, moderate consumption of alcohol, familiarity. In particular, ApoE is a lipoprotein which transfers lipids in the blood to make them available for cells; it is made up of any combinations of two alleles ($\epsilon_2, \epsilon_3, \epsilon_4$). Familiarity refers to the existence of a similar inclination to become diseased for subjects belonging to the same family. It can be due to the joint effect of various factors, like genetic predisposition, familiar interaction, shared experiences, environment and lifestyles, and so on. Variables coded as binary are: sex (0 = male; 1 = female), school (years of formal education: 0 = ≤ 8 ; 1 = > 8), smoking (0 = yes; 1 = no); moderate consumption of alcohol (0 = yes; 1 = no). The only categorical variable is ApoE, whose categories are the possible combinations of ϵ alleles (1 = 2/2; 2 = 3/2; 3 = 3/3; 4 = 4/2; 5 = 4/3; 6 = 4/4). Familiarity is treated by considering an average family made up of six members: father (F), mother (M), first and second brother (B1, B2), first and second sister (S1, S2); for each, the presence of Alzheimer disease is coded as a binary variable (0 = no; 1 = yes). Percentages of missing data on covariates for cases and controls respectively are 0.92-10.0 (years of school), 7.34-10.0 (smoking and alcohol consumption), 3.67-33.75 (ApoE). Levels of factors are coded as dummy variables through corner point parameterization.

A logistic regression model has been employed to analyze this dataset, considering the probability of becoming an Alzheimer patient as a function of the risk factors. Table 1(A-B) reports the results only of complete additive models, since the interactions have been excluded as not significant by preliminary tests.

Table 1: Results for complete datasets after imputation (simple and multiple)

A	Estimate	Std.Err.	p-value	B	Estimate	Std.Err.	p-value	LCI	UCI
(Inter)	-11.981	46.330	0.796	(Inter)	0.726	0.367	0.050	0.001	1.451
sex1	-0.467	0.406	0.250	sex1	-0.090	0.087	0.301	-0.262	0.082
scol1	-0.216	0.389	0.579	scol1	0.021	0.083	0.799	-0.145	0.188
gen2	10.914	46.332	0.814	gen2	0.459	0.393	0.245	-0.318	1.236
gen3	11.110	46.327	0.810	gen3	0.571	0.359	0.113	-0.137	1.280
gen4	19.909	59.546	0.738	gen4	0.920	0.443	0.039	0.045	1.795
gen5	12.291	46.329	0.791	gen5	0.848	0.362	0.020	0.134	1.562
gen6	20.222	49.622	0.684	gen6	0.939	0.381	0.015	0.187	1.691
smk1	0.063	0.425	0.882	smk11	0.052	0.103	0.619	-0.157	0.261
alc1	1.196	0.390	0.002	alc1	0.246	0.088	0.008	0.069	0.424
F1	-1.344	0.974	0.167	F1	-0.206	0.155	0.187	-0.512	0.1001
M1	0.646	0.443	0.145	M1	0.107	0.087	0.219	-0.064	0.279
B11	0.594	0.904	0.511	B11	0.075	0.157	0.635	-0.236	0.386
B21	7.477	26.759	0.780	B21	0.153	0.266	0.565	-0.371	0.677
S11	1.770	0.744	0.017	S11	0.208	0.117	0.077	-0.023	0.440
S21	8.289	36.949	0.822	S21	0.377	0.305	0.219	-0.226	0.980

Table 1(A) includes results for the complete dataset obtained by simple imputation as described in Section 1. The only estimates with a p-value that denotes significance correspond to “health status of the first sister” and “alcohol habits”: having a diseased sister and not making moderate use of alcohol may increase the risk to be an Alzheimer patient. Table 1(B) shows results for the complete dataset obtained via multiple imputation (five imputed values for each missing data). Estimates with a significant p-values are: genotypes 4/2, 4/3, 4/4 (in agreement with the medical literature); and, like for simple imputation, having a diseased sister and not making moderate consumption of alcohol.

3. Multiple imputation for simulated datasets

It is natural to wonder if results of a logistic regression model applied to complete datasets obtained by multiple imputation are affected by the percentage of missing information. So, starting from the complete dataset obtained in Section 2 through simple imputation, we consider “smoking” and “alcohol” as the only incomplete variables and then simulate datasets for combinations of percentages of missing values on the two variables varying from 5% to 30%. We present only comparisons of p-values, since estimates and standard errors are very similar in the two cases and with respect to Table 1(B). Table 2 does not report the p-value of the intercept, but includes p-values for all other possible risk factors, even those not affected by missing data. This is because we want to check the effects of missing values over the two risk factors considered on the significance of all parameters in the model. When generating the vectors of imputed values for alcohol and smoke, the set of predictors is made up by all variable (complete and incomplete), as explained in Section 1. We take five imputed values for each missing data.

Variable “sex” presents a p-value that decreases when passing from low to high fractions of missing values, resulting significant for the 10-30%, 10-20% and 30-30% combinations. P-values associated with variable “years of school” goes up and down with respect to the percentages of missing data, but they are always not significant. The different genotype combinations of ApoE result to be always significant. The parameter for the health status of the two brothers and the second sister is always not significant while it is always significant for the first sister. The parameter for “smoking” is never significant; the parameter for “alcohol consumption” is always highly significant. The parameter as-

sociated with “mother’s health status” is never significant, whilst that associated with “father’s health status” is significant for the combination 10-30%.

Table 2: *P-values of simulations (≤ 0.05 boldface)*

	percentages of missing data							
	5-5%	5-10%	5-20%	5-30%	10-10%	10-30%	20-30%	30-30%
sex1	0.309	0.232	0.098	0.055	0.179	0.028	0.016	0.021
scol1	0.800	0.875	0.759	0.470	0.929	0.462	0.602	0.568
gen2	0.047	0.042	0.028	0.020	0.050	0.013	0.027	0.023
gen3	0.034	0.032	0.024	0.016	0.033	0.010	0.019	0.016
gen4	0.014	0.012	0.009	0.008	0.011	0.004	0.006	0.005
gen5	0.004	0.004	0.003	0.002	0.004	0.001	0.002	0.001
gen6	0.002	0.002	0.001	0.001	0.002	0.000	0.001	0.001
smk1	0.935	0.716	0.517	0.728	0.422	0.494	0.065	0.226
alc1	0.001	0.000	0.000	0.003	0.001	0.001	0.000	0.005
F1	0.167	0.153	0.198	0.126	0.132	0.037	0.074	0.070
M1	0.194	0.209	0.258	0.187	0.207	0.233	0.258	0.265
B11	0.565	0.664	0.770	0.427	0.765	0.606	0.733	0.651
B21	0.643	0.592	0.553	0.880	0.535	0.861	0.624	0.735
S11	0.025	0.030	0.016	0.007	0.029	0.004	0.011	0.009
S21	0.312	0.259	0.270	0.175	0.243	0.092	0.138	0.123

4. Conclusions

Imputation techniques allow to use common software by reconstructing complete datasets; they do not discard information, but can improperly lead to wrong conclusions. In the Alzheimer study, the significance of estimates for both techniques is very similar, what changes is the magnitude of the estimates by using multiple imputation. In fact, estimates have lower values, probably due to an “averaging” effect over the five imputed values. In that way we loose information about risk factors with marginal effects, but are pushed to consider significant what is really significant.

Our simulations show another risk: significance is strictly linked to the percentage of missing values. For high fractions of missing data, p-values generally decrease and this can bring to wrong conclusions. This suggests that imputation is not a *panacea*: there does not exist a “best solution” to manage missing data. The proper treatment of missing data depends upon the goals of the study, and sensitivity analysis to reflect uncertainty is always advisable.

References

- Little R. and Rubin D. (1987) *Statistical Analysis with Missing Data*, John Wiley & Sons, New York.
- Rubin D. (1996) Multiple imputation after 18+ years, *Journal of the American Statistical Association*, 91, 434, 473–489.
- Rubin D. and Schafer J. (1998) Multiple imputation for missing-data problems, [url:www.stat.psu.edu/jls](http://www.stat.psu.edu/jls).
- Van Buuren S. and Oudshoorn C. (2000) *Multivariate Imputation by Chained Equations - MICE v.1 Users’s manual*, [url:web.inter.nl.net/users/S.van.Buuren/mi/html/about.htm](http://web.inter.nl.net/users/S.van.Buuren/mi/html/about.htm).
- Viviano L. (2002) *Alzheimer: fattori di rischio e familiarità. Uno studio per dati categoriali su unità dipendenti*, Ph.D. thesis.