The Statistician (1993) 42, pp. 307–313

THEORY AND METHODS

Data analysis using hot deck multiple imputation

MARIE REILLY

Department of Mathematics and Statistics, Dublin Institute of Technology, Kevin Street, Dublin 8, Republic of Ireland

Abstract. Covariate data which are missing or measured with error form the subject of a growing body of statistical literature. Parametric methods have not been widely adopted, quite possibly due to the necessity of specifying the form of a 'nuisance function' not required for complete data analysis, and the non-robustness of the methods to mis-specification. A non-parametric counterpart of multiple imputation, known as 'hot deck', was proposed by Rubin (1987) and has been used by the Census Bureau to complete public-use databases. However, inference using this method has not been possible due to the distribution theory not being available.

Recently, it has been shown that the hot deck estimator has the same asymptotic distribution as the 'mean score' estimator, so that inference using hot deck is now possible. The method is intuitively appealing and easily implemented. Furthermore, it accommodates missingness which depends on outcome, which is an important generalization of many currently available methods. In this paper, the hot deck multiple imputation method is explained, its asymptotic distribution presented and its application to data analysis demonstrated by an example.

1 Introduction

Many problems in applied statistics involve the estimation of the strength of association between an outcome Y and various covariates. It is common in practice to have a number of observations for which one or more of the covariates are missing. This is particularly true of epidemiological and clinical studies, where it is not uncommon to have a very large number of covariates. We will denote the incomplete covariates by X and the remaining covariates by Z, so that our objective is to estimate β in the conditional likelihood $f_\beta(Y|X, Z)$ when some of the X values are missing. We will assume that X is 'missing at random' (MAR) (Little & Rubin, 1987), i.e. missingness may depend on Y and on Z but not on X. The data may be missing in the usual sense, or may be missing due to the failure to recognize an important covariate until late in the study. The covariate may also be missing by design, for example in a study where X is expensive or difficult to measure. We consider here the situation where the outcome Y and auxiliary covariates Z are categorical, and the covariate X is available for a small subsample of subjects, hereafter referred to as the validation sample. The ability to draw inferences from such data has important consequences. In the simplest case where the missingness is random, the analysis of only those cases which are complete will yield unbiased estimates, but an analysis which uses the additional information in incomplete cases could improve the precision of the estimates. More importantly, if the missingness is not random, a complete case analysis is inappropriate, so that alternative methods are necessary.

Parametric methods of dealing with such data require the specification of the form of the conditional density $f_\theta(X|Z)$, which is not required for complete data analysis. Further discouragement arises from the non-robustness to mis-specification (Carroll *et al.*, 1984). The estimated likelihood method (Pepe & Fleming, 1991) overcomes these problems, but does not accommodate missingness, which depends on outcome. Since parametric multiple imputation (Rubin, 1987) accommodates missingness depending on outcome, its non-parametric counterpart, hot deck multiple imputation, offers an alternative approach to this problem. The basic idea of the hot deck method is that for subjects whose X is missing, an X is imputed by simply choosing at random from among validation sample members with matching Y and Z. This 'completed' data set is then analyzed using standard methods to

yield an estimate $\hat{\beta}_k$. This 'filling in' followed by standard analysis is repeated a number of times and the estimates are averaged to give the overall estimate $\hat{\beta}_{HD}$. Though it has been used by the US Census Bureau to complete public-use databases, the hot deck method has not been used for statistical inference as its distribution theory has been unavailable.

It has now been shown (Reilly, 1991) that the hot deck estimator has the same asymptotoic distribution as the mean score estimator (Pepe *et al.*, 1993; Reilly & Pepe, 1993). This common distribution has been derived, so that inference is now possible using hot deck multiple imputation. This is a conceptually simple method which can be implemented with standard software. Furthermore, the asymptotic relative efficiency has been studied, and for random missingness the method offers a good gain in efficiency over an analysis of complete cases only. For non-random missingness, it offers a means of valid inference.

The asymptotic distribution of the hot deck estimator is presented in Section 2. The choice of the number of imputations is discussed in Section 3 and an expression for the variance for a finite number of imputations is derived. Finally, Section 4 presents an example of the use of the method for data analysis.

2 Asymptotic results

Since missingness can depend on Y and Z, the validation sample can be thought of as a stratified random sample, the strata being the cells defined by the categorical Y and Z. The total sample size is denoted by n, and the validation sample size by n_V, while the number of validation members in the (Z, Y) cell is n_{VZY}. The probability that an individual takes values Z and Y is denoted by p_{ZY}, while this probability is $p_{ZY|V}$ for a validation sample member. The probability that an individual will be assigned to the validation sample is p_V, while this probability is $p_{V|ZY}$ for an individual with values (Z, Y). Similar notation is used for the non-validation sample $\bar{V}$. Under some regularity conditions on the density function $f_\beta(Y|X, Z)$, and assuming that p_V and p_{ZY} are non-negligible, it has been shown (Reilly, 1991; Reilly & Pepe, 1993) that the mean score estimator and the hot deck estimator for an infinite number of imputations have the same asymptotic distribution, which is given by

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{D} N(0, I_V^{-1} + I_V^{-1} \Sigma I_V^{-1}) \tag{1}$$

$I_V = E[-\frac{\partial}{\partial\beta} S_\beta(Y|X, Z)]$, where $S_\beta(Y|X, Z) = \frac{\partial}{\partial\beta} \log f_\beta(Y|X, Y)$ and

$$\Sigma = \sum_{Z,Y} \frac{(p_{\bar{V}|ZY})(p_{ZY})}{p_{V|ZY}} \operatorname{Var}(S_\beta(Y|X, Z)|Y, Z) \tag{2}$$

where summation is taken over all (Z, Y) combinations. I_V is the usual information term for a complete case (i.e. a validation member) and if the covariate data were complete, a standard likelihood analysis would give an estimate with variance I_V^{-1}/n. The hot deck estimator has a variance somewhat larger, since imputation introduces correlation between the validation and non-validation score components. However, a study of the asymptotic relative efficiency (Reilly & Pepe, 1993) shows that hot deck has good efficiency even when up to 50% of the X data is missing. It is very easily shown that for missingness which is completely at random, MCAR (Little & Rubin, 1987), hot deck offers a gain in precision over an analysis of only those subjects whose covariate data are complete. The gain is largest when the auxiliary covariate is highly informative about the missing X as would be expected. For non-informative Z, there is no gain in precision but neither is there any penalty for having imputed the missing covariates. The variance of the estimates (given by equation (1)) can be consistently estimated by

$$\hat{V}(\hat{\beta}) = \frac{1}{n}(\hat{I}_V^{-1} + \hat{I}_V^{-1} \hat{\Sigma} \hat{I}_V^{-1})$$

where $\hat{\beta}$ denotes the hot deck estimate of β, $\hat{\Sigma}$ denotes the consistent estimate of Σ obtained by replacing the components of equation (2) by their estimates, i.e.

$$\hat{\Sigma} = \sum_{ZY} \frac{(n_{ZY})(n_{\bar{Y}ZY})}{n(n_{VZY})} \hat{\mathrm{V}}\mathrm{ar}[S_{\hat{\beta}}(Y|X, Z)|Y, Z]$$

and $\hat{I}_V$ is a consistent estimate of I_V, which is motivated by noting that

$$I_V = E[-\tfrac{\partial}{\partial\beta} S_\beta(Y|X, Z)] = E[I_\beta(Y|X, Z)]$$
$$= \sum_{ZY} p_{ZY} E[I_\beta(Y|X, Z)|Y, Z)$$

and replacing the components on the right-hand side by their estimates to give

$$\hat{I}_V = \frac{1}{n}\sum_{ZY} n_{ZY}\left(\sum_{V^{ZY}} \frac{I_{\hat{\beta}}(Y|X, Z)}{n_{VZY}}\right)$$

where V^{ZY} denotes validation sample members with values Z and Y.

3 How many imputations?

For K imputations, the hot deck estimate is found by averaging K 'completed data' estimates, and will be denoted as $\bar{\beta}_K$. If the number of imputations was increased indefinitely, then the resultant 'ideal' hot deck estimate for the data set, which will be denoted by $\bar{\beta}_\infty$, is asymptotically equivalent to the mean score estimator. The task is to choose K large enough so that $\bar{\beta}_K \approx \bar{\beta}_\infty$.

For a given set of data, $\beta_1, \ldots, \beta_K$ can be regarded as a simple random sample from the infinite population of completed data estimates. So $\bar{\beta}_K$ is a sample mean, which is an estimate of the population mean $\bar{\beta}_\infty$. The sample variance V_K of $\beta_1, \ldots, \beta_K$ can be used as an estimate of the population variance. Therefore, with a probability of approximately 0·997

$$|\bar{\beta}_K - \bar{\beta}_\infty| \leq 3\sqrt{(V_K/K)}$$

If $\bar{\beta}_\infty$ and its variance were available, they could be used to construct a 90% confidence interval for the true parameter β as follows:

$$\bar{\beta}_\infty \pm 1{\cdot}65\sqrt{(V(\bar{\beta}_\infty))}$$

If $\bar{\beta}_K$ is used in place of $\bar{\beta}_\infty$, then the interval becomes

$$\bar{\beta}_K \pm 1{\cdot}65\sqrt{(V(\bar{\beta}_\infty))}$$

For this interval to give a coverage close to the nominal value, we need to choose K large enough so that $|\bar{\beta}_K - \bar{\beta}_\infty|$ is negligible compared to $\sqrt{(V(\bar{\beta}_\infty))}$. If K is chosen large enough so that

$$\sqrt{(V_K/K)} < 0{\cdot}05\ \sqrt{(V(\bar{\beta}_\infty))} \tag{3}$$

then the nominal 90% confidence interval can be shown (Reilly, 1991) to have a coverage of at least 87%. By using $\hat{V}(\bar{\beta}_K)$ to estimate $V(\bar{\beta}_\infty)$, we have an adaptive algorithm which permits the implementation of hot deck analysis with an effectively infinite number of imputation cycles.

A simulation study was done to compare the performance of this adaptive algorithm to a hot deck analysis using a fixed number of imputations. Since parametric multiple imputation generally uses between 2 and 10 imputations (Rubin, 1987, p. 15), it was decided to record the hot deck estimates and their variances for 3, 5 and 10 imputations. These estimates were then compared to the results for the 'effectively infinite' number of imputation cycles defined above.

For the purposes of the simulation study, a dichotomous outcome variable Y was modelled as a function of a covariate X, using the logistic function:

$$P[Y = 1 \mid X] = \frac{\exp^{\alpha + \beta X}}{1 + \exp^{\alpha + \beta X}}$$

Values of 0 and 1 were used for β corresponding to odds ratios of 1 and 2·7, and α was set to 0 or 1. The covariate X was generated as a standard normal random variable, and Z was chosen as the dichotomous variable indicating whether X plus some random normal error is positive.

$$Z = I[X + \varepsilon > 0], \qquad X \sim N(0, 1), \quad \varepsilon \sim N(0, \sigma^2)$$

By increasing the magnitude of σ^2 (relative to 1), Z can be made more or less informative about X: values of 0·25, 1 and 2 were used for σ. Two kinds of sampling strategy were used to obtain the validation sample: (i) simple random sampling, where the validation sample is a random sample of the overall sample and (ii) a balanced design, where equal numbers of validation members are chosen in each (Z, Y) stratum. Since it is of interest to assess these methods for use with data where the missingness is substantial, validation fractions of 0·2 and 0·5 are considered, and sample sizes in the range of 100 to 500 were used. For each investigation, 400 realizations of data were generated in accordance with the above models. The score equations were solved using the Newton–Raphson iterative procedure to obtain estimates. This work was done on IBM-compatible 386 and 486 machines, using the GAUSS (1988) programming language. The results for a total sample size of 100 with a balanced validation sample of 20 are presented in Table 1.

Table 1. Comparison of adaptive hot deck with 3, 5 and 10 imputation cycles; balanced design. 'Mean' and 'var' are the simulation mean and variance. The adaptive estimates are denoted by $\hat{\beta}_\infty$, and the three estimates in the $\hat{\beta}_{HD}$ columns are the hot deck estimates for 3, 5 and 10 imputations. The last two columns record the average number of imputations required by the adaptive algorithm and the % of simulations where more than 10 imputations were needed

		$n = 100$, $\sigma = 0{\cdot}25$						$n = 100$, $\sigma = 2$					
α	β	Mean $\hat{\beta}_\infty$	Mean $\hat{\beta}_{HD}$	Var $\hat{\beta}_\infty$	Var $\hat{\beta}_{HD}$	Ave imp	% > 10	Mean $\hat{\beta}_\infty$	Mean $\hat{\beta}_{HD}$	Var $\hat{\beta}_\infty$	Var $\hat{\beta}_{HD}$	Ave imp	% > 10
0	0	− 0·023	− 0·022 − 0·022 − 0·020	0·166	0·170 0·168 0·165	31	86	0·016	0·006 0·015 0·015	0·419	0·437 0·428 0·420	39	95
1	0	− 0·006	− 0·013 − 0·010 − 0·006	0·165	0·167 0·164 0·165	31	88	0·020	0·027 0·026 0·023	0·382	0·401 0·394 0·386	41	95
0	1	1·137	1·137 1·137 1·136	0·217	0·225 0·221 0·220	31	95	1·385	1·381 1·384 1·381	0·987	1·015 1·031 0·983	32	97
1	1	1·168	1·160 1·164 1·166	0·302	0·298 0·305 0·303	27	93	1·368	1·371 1·367 1·365	0·730	0·791 0·763 0·740	34	97

For the series of models studied here, there appeared to be little or no gain in using more than 5 or 10 imputation cycles. This is so even for situations where more than 10 imputation cycles were required to satisfy (3) on almost all realizations. It may of course be possible to find models where the combination of model parameters, missingness and strength of association between X and Z results in large number of imputation cycles being necessary. But it is interesting that for the scenarios presented here, such is not the case. A similar investigation by Rubin (1987, p. 114) for parametric multiple imputation indicates that for a number of reasonable models, 2 or 3 imputation cycles are adequate when the amount of

missing information is small. In the present study, although there are situations where the missing information is substantial, it is encouraging to note that there is little gain in precision over a simple hot deck procedure using 5 or 10 imputation cycles.

In applying hot deck multiple imputation, the user may wish to use some adaptive algorithms such as that outlined above, or simply impute until the estimate and its standard error have stabilized to within some desired precision. However, as we have seen, such methods may not offer a substantial gain in precision over the application of some small fixed number of imputations, which latter strategy may be preferred by many users. This is particularly true where a large data set is involved, since the computational demand of multiple imputation is equivalent to many full data analyses. For a small number of imputations, a simple correction to the asymptotic variance expression can be made as follows:

$$V[\sqrt{n}(\bar{\beta}_K - \beta)] = V[\sqrt{n}(\bar{\beta}_\infty - \beta)] + V[\sqrt{n}(\bar{\beta}_K - \bar{\beta}_\infty)]$$
$$= nV(\bar{\beta}_\infty) + nV(\bar{\beta}_K - \bar{\beta}_\infty)$$

where the covariance term can be seen to be zero by conditioning on the data. This last expression can be written as

$$nV(\bar{\beta}_\infty) + nE[V(\bar{\beta}_K) \mid \text{data}]$$

The first term is the asymptotic variance formula evaluated at $\bar{\beta}_K$ while the second term (the 'correction term') can be estimated by $1/K$ times the sample variance of the completed data estimates $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$. This correction term is very similar to that used by parametric multiple imputation (see Rubin, 1987). Simulation studies indicated that the hot deck estimator and this variance estimate behave well in small and moderate samples.

4 Data example

To illustrate the hot deck method, we use data from a study of graft-versus-host disease in 97 female patients, who received bone marrow transplants from female sibling donors (see Pepe & Fleming, 1991). The aim of the study was to evaluate the effect of a prophylactic regimen, where a combination of methotrexate and cyclosporine is compared to methotrexate alone, cyclosporine alone, or a combination of methotrexate, cyclosporine and prednisone. Other factors considered important were patient age and isolation of the patient in a laminar air flow post-transplant. However, late in the study it was realized that previous pregnancy of the donor was a potentially important risk factor. This variable had not been recorded for 66 study subjects, but was subsequently obtained for the remaining 31.

The occurrence of acute graft-versus-host disease (GVHD) is the dichotomous outcome variable Y of interest. The complete covariates are patient age (0 = 20–29, 1 = 30–40), laminar air flow (LAF, 0 = no, 1 = yes) and prophylactic regimen (1 = MTX + CSP, 0 = other) and are represented by Z. The incomplete covariate X denotes donor pregnancy status (0 = no, 1 = yes). On examining the data, it was found that there were two (Y, Z) strata (corresponding to patients with and without GVHD who had LAF = 1 and PROPHYLAXIS = 0) with no validation members. For this reason, and because LAF was found to be non-significant in a previous analysis of this data (Pepe & Fleming, 1991), it was decided to exclude LAF from the analysis.

A naïve logistic regression analysis of the 66 subjects on whom we have complete data gave the results presented in Table 2. This analysis is valid only if the validation sample is a random sample of all the study subjects, an assumption that is challenged by the observation that missingness is significantly associated with age, both overall ($p = 0{\cdot}006$) and for patients without GVHD ($p = 0{\cdot}04$).

The hot deck method provides a means of incorporating the information in the non-validation members, thus correcting for possible bias in the complete data estimates

Table 2. Results of logistic regression analysis of the 66 complete cases in the donor pregnancy data

Risk factor	Estimate	Standard error
Donor pregnancy (yes/no)	1·247	0·669
Patient age	− 0·004	0·670
Prophylaxis (MTX + CSP/other)	− 1·716	0·844

due to non-random missingness and improving the precision of the estimates. Table 3 provides the means and standard errors for the hot deck analysis of these data using 3, 10 and 100 imputations (the adaptive criterion (3) was not satisfied after 100 imputations for this data set). These results enable us to conclude that both prophylaxis and donor pregnancy are strong independent risk factors for GVHD. We note that the hot deck analysis with 10 imputations gives results which are in very good agreement with the results for 100 imputations for this small data set. Furthermore, the results for the hot deck analysis with only 3 imputations are remarkably good.

Table 3. Estimates (and standard errors) for hot deck (HD) analyses with 3, 10 and 100 imputations for the donor pregnancy example

Risk factor	HD (3 imps)	HD (10 imps)	HD (100 imps)
Donor pregnancy	2·516 (0·829)	2·583 (0·831)	2·620 (0·771)
Patient age	− 0·299 (0·526)	− 0·252 (0·527)	− 0·244 (0·513)
Prophylaxis	− 2·126 (0·729)	− 2·097 (0·730)	− 2·094 (0·695)

5 Conclusions

Non-parametric multiple imputation was proposed some years ago (Rubin, 1987), but has not been used by applied statisticians, as a variance expression for the estimator was not available. It has been shown (Reilly, 1991; Reilly & Pepe, 1993) that this intuitively appealing estimator has the same asymptotic distribution as another non-parametric estimator known as 'mean score'. This common distribution has been derived, making inference possible. A series of simulation studies indicates that the estimator behaves well when missingness is substantial (or the sample size is small), a property not enjoyed by parametric multiple imputation.

A major advantage of hot deck mutiple imputation is that it can be implemented with any standard software once the user can program the imputation step. Furthermore, it generalizes many existing methods by accommodating missingness which can depend on outcome and on the complete covariates.

Further work is needed to extend the method to continuous complete covariates. Hot deck offers an appealing method of improving precision by incorporating all the available information in the data, and more importantly, it provides a method of valid analysis in the case of non-random missingness. This conceptually simple method may well prove a popular choice among applied statisticians faced with the problem of missing covariate data.

Acknowledgements

This work was supported in part by Scientific Research Grant No. SC/92/238 from EOLAS, Ireland, and a Seed Funding grant from the Dublin Institute of Technology.

References

CARROLL, R. J. *et al.* (1984) On errors in variables for binary regression models, *Biometrika*, 71, pp. 19–25.
GAUSS SYSTEM VERSION 2.0 (1988) (Kent, WA, Aptech Systems).
LITTLE, R. J. A. & RUBIN, D. B. (1987) *Statistical Analysis with Missing Data* (New York, Wiley).
PEPE, M. S. & FLEMING, T. R. (1991) Non parametric methods of dealing with mismeasured covariate data, *Journal of the American Statistical Association*, 86, pp. 413.108–113.
PEPE, M. S., REILLY, M. & FLEMING, T. R. (1993) Auxillary outcome data and the mean score method, submitted to *Journal of Statistical Planning and Inference*.
REILLY, M. (1991) Semi-parametric Methods of Dealing with Missing or Surrogate Covariate Data. PhD dissertation, Biostatistics, University of Washington, Seattle, WA 98195, USA.
REILLY, M. & PEPE, M. S. (1993) A mean score method for missing and auxillary covariate data in regression models, submitted to *Biometrika*.
RUBIN, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys* (New York, Wiley).