STATISTICS IN MEDICINE
Statist. Med. **18**, 681–694 (1999)

MULTIPLE IMPUTATION OF MISSING BLOOD PRESSURE COVARIATES IN SURVIVAL ANALYSIS

S. VAN BUUREN[1,*], H. C. BOSHUIZEN[1] AND D. L. KNOOK[2]

[1] *TNO Prevention and Health, Department of Statistics, P.O. Box 2215, 2301 CE Leiden, The Netherlands*
[2] *Leiden University, Department of Internal Medicine, Section of Gerontology, P.O. Box 9600, 2300 RC Leiden, The Netherlands*

SUMMARY

This paper studies a non-response problem in survival analysis where the occurrence of missing data in the risk factor is related to mortality. In a study to determine the influence of blood pressure on survival in the very old (85+ years), blood pressure measurements are missing in about 12·5 per cent of the sample. The available data suggest that the process that created the missing data depends jointly on survival and the unknown blood pressure, thereby distorting the relation of interest. Multiple imputation is used to impute missing blood pressure and then analyse the data under a variety of non-response models. One special modelling problem is treated in detail; the construction of a predictive model for drawing imputations if the number of variables is large. Risk estimates for these data appear robust to even large departures from the simplest non-response model, and are similar to those derived under deletion of the incomplete records.

1. INTRODUCTION

A problem in survival analysis occurs when data are missing on one or more risk factors. The standard response to this problem is to simply exclude these individuals from the analysis. Apart from being a waste of costly collected data, this practice could lead to invalid results if the excluded group is a selective subsample from the entire sample.

We were confronted with such a problem in the analysis of the relation between blood pressure (BP) and mortality in persons over 85 years of age.[1] The main interest of that study was to determine the influence of measures of health on the relation between mortality and BP in the elderly. It has been found that, in this age group, low BP is associated with increased mortality.[2,3] This has raised concerns as to whether prescription of anti-hypertensive drugs could inadvertently shorten life. To uncover the mechanism that governs the effect, the idea was to investigate whether the observed relation could be attributed to differences in health between different BP groups. If so, BP is more likely to be a symptom than a cause of bad health, which would diminish concerns about possible life-shortening side-effects of the

* Correspondence to: S. van Buuren, TNO Prevention and Health, Department of Statistics, P.O. Box 2215, 2301 CE Leiden, The Netherlands. E-mail: S.vanBuuren@pg.tno.nl

CCC 0277–6715/99/060681–14$17.50

Received November 1996
Accepted June 1998

hypertension treatment. More in particular, the scientific interest focused on the comparison of two models, A and B:

A. the relation between mortality and BP adjusted for age and sex;
B. the relation between mortality and BP adjusted for age, sex and health.

The main analysis consists of a Cox regression of mortality on BP, adjusted for age, sex and health. The detailed comparison of model A and B as well as its implications is reported elsewhere.[4]

The analysis is based on a data set in which approximately 12·5 per cent of the blood pressure measurements are missing. As will be indicated in Section 2, we suspected that individuals with lower blood pressures and higher mortality risks had fewer BP measurements. Excluding the incomplete cases from the analysis could thus produce deflated mortality estimates for the lower blood pressure groups, thereby yielding a distorted impression of the influence of BP on survival. The present paper reports our strategy to handle this problem.

Several approaches for dealing with incomplete covariates in survival analysis exist.[5–8] These methods all rely on the assumption that the non-response probabilities do not depend on any unobserved information, that is, that the data are missing at random (MAR).[9] Since this is a dubious assumption with our data, we used an alternative approach based on multiple imputation.[10,11] The idea is to create a small number (m) of completed matrices in which the missing values have been replaced by plausible values. The number of imputations needed depends on the amount of missing information, but is usually quite small, often 3 or 5. The variability among the m imputations reflects the uncertainty about the hypothetically observed, but unknown, value. Under quite general conditions, it has been shown that (i) if the complete data model leads to valid inferences in the absence of non-response and (ii) if the imputation procedure is proper with respect to the non-response mechanism, then multiple imputation yields valid inferences. The term 'proper' refers to a set of technical conditions that delineate the class of distributions from which imputations can be created (see pp. 118–119 of Rubin[10]). A more accessible description can be found in Schafer's book.[12]

Of particular interest is that multiple imputation allows display of the sensitivity of the inferences to different mechanisms that could have created the non-response. There is no need to assume one 'true' response model and stick to that. Several plausible mechanisms can be tried. If none of these mechanisms changes the relation of interest, then inference is robust against the specified causes of the non-response. On the other hand, if the results do depend on the specific form of the non-response model, then more precise statements can be made regarding the exact conditions under which the obtained results apply.

The present paper focuses on a number of practical aspects the analyst encounters when dealing with missing data problems: (i) what information should be used for choosing between different non-response mechanisms; (ii) how to choose a useful set of imputation predictors from a large set of variables; (iii) how to generate the actual imputations when the variables are of mixed type; and (iv) how to specify the different models for the non-response. We introduce a convenient and quite general regression switching scheme for generating the actual imputations.

2. DESCRIPTION OF THE PROBLEM

2.1. The Leiden 85 + Cohort

The cohort under study[13,14] consists of 1236 citizens of Leiden who were 85 years or older on 1 December 1986. These individuals were visited by a physician between January 1987 and May

1989. A full medical history, information on current use of drugs, a venous blood sample, and other health-related data were obtained. Blood pressure (BP) was routinely measured during the visit. Apart from a few individuals who were bedridden, BP was measured while seated. A mercury manometer was used and BP was rounded to the nearest 5 mmHg. Measurements were usually taken near the end of the interview. The mortality status of each individual on 1 March 1994 was retrieved from administrative sources. The cohort will be referred to as the 'Leiden 85+ Cohort'.

Of the original cohort, a total of 218 died before they could be visited, 59 people did not want to participate (some because of health problems), 2 emigrated and 1 was erroneously not interviewed, so 956 individuals were visited. In an early analysis, we found that the effect of analysing a subsample from the entire cohort can be accounted for by taking the date of the home visit as the start of the observation period, and adjusting the analysis for sex and age.[1] This type of selection will therefore not be considered here.

2.2. Factors that affect the measurement of blood pressure

BP was not measured for 121 individuals, sometimes because of time constraints, or sometimes because the investigator did not want to place any additional burden on the respondent. In some cases, it was reported that the subject was too ill to be measured. Table I indicates that BP was measured less frequently for very old people and for those with health problems. Also, BP was measured more often if it was suspected that the BP was too high, for example if the respondent indicated a previous diagnosis of hypertension, or if the respondent used any medication that lowered blood pressure. The non-response rate for BP also varies during the data collection period of the study. The rate gradually increases during the first seven months of the sampling period from 5 to 40 per cent of the cases, and then suddenly drops to a fairly constant level of 10–15 per cent. A complicating factor here is that the sequence in which the respondents were interviewed was not random. High risk groups, that is, elderly in hospitals and nursing homes and those over 95 years, were visited first.

Figure 1 displays survival curves for two groups in our study: one group with observed BP measures ($n_{\text{obs}} = 835$) and one with missing BP ($n_{\text{mis}} = 121$). These curves have been obtained as baseline hazards after fitting a proportional hazards model adjusted for age, sex and type of residence, and stratified by the missingness indicator. People without BP measures apparently have higher mortality rates. Figure 1 suggests that eliminating the incomplete data will overestimate the true survival of the cohort. Moreover, if the process that causes the missing data depends jointly on survival and the unknown BP, then the estimate of scientific interest to us, namely the relative mortality risks of subcohorts of different BP levels, can be biased. Problems in model A would occur if, in the conditional distribution given age and sex, the relation between BP and mortality is different for those with BP measured than for those without. It is, however, impossible to demonstrate this from the data, because BP data are missing in the second group.

Table II shows the proportion of people for whom BP was not measured, cross-classified by three-year survival and history of hypertension. Of all persons who died within three years and who have no history of hypertension, more than 19 per cent have no BP measurement. The rate for the other categories is about 9 per cent. This suggests that a relatively large group of individuals wihtout hypertension and with high mortality risk is missing from the sample for which BP is known. In that case, confounding by selection could occur in the sense that an

Table I. Some variables that have different distributions in the response ($n = 835$) and non-response groups ($n = 121$). Shown are percentages. Significance levels correspond to the χ^2-test

Variable	Observed BP	Missing BP
Age (year)	$p < 0{\cdot}001$	
85–89	63	48
90–94	32	34
95+	6	18
Type of residence	$p < 0{\cdot}001$	
Independent	52	35
Home for elderly	35	54
Nursing home	13	12
Activities of daily living (ADL)	$p < 0{\cdot}001$	
Independent	73	54
Dependent on help	27	46
History of hypertension	$p = 0{\cdot}06$	
No	77	85
Yes	23	15
Uses diuretics	$p = 0{\cdot}03$	
No	55	67
Yes	45	33

analysis that uses only the complete cases underestimates the mortality of the lower and normal BP groups. Note that this reasoning is somewhat tentative as it relies on the use of hypertension history as a proxy for BP. If true, however, we would expect that more of the lower BP measures are missing. Thus, selection might blur the effect of BP on mortality.

2.3. Response mechanisms for blood pressure

A first step to account for the missingness is to specify a number of plausible response mechanisms. Let Y be an $n \times p$ matrix of partially observed outcome variables in a sample of size n. In the present case, $p = 4$, where the columns $Y_1, \ldots, Y_4$ are:

Y_1 systolic blood pressure (mmHg);
Y_2 diastolic blood pressure (mmHg);
Y_3 survival or censoring time (in days since the home visit);
Y_4 censoring indicator (0, 1).

Let Z denote an $n \times q$ matrix of observed covariates. For the blood pressure problem we have $q = 31$, where the columns $Z_1, \ldots, Z_{31}$ are:

Z_1 age;
Z_2 sex;
$Z_{3,\ldots,31}$ health measures.

Health was measured by 29 variables, including mental state, presence of handicaps, being dependent in activities of daily living (ADL), history of cancer, and so on. Until Section 3.3 it is assumed that Z is completely observed for everyone in the sample.

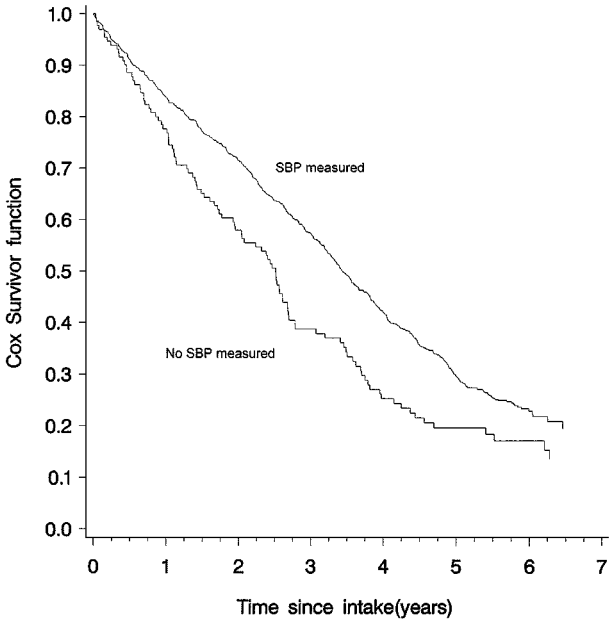

Figure 1. Survival curves obtained by fitting a proportional hazards model adjusted for age, sex and type of residence and stratified by the presence of a systolic blood pressure measurement ($n_{obs} = 835$, $n_{mis} = 121$)

Table II. Proportion of people for whom no BP was measured, cross-classified by three-year survival and previous hypertension history. Shown are proportions per cell (number of cases with missing BP/total cell count)

Survived > 3 years	History of previous hypertension			
	no		yes	
yes	8·7%	(34/390)	8·1%	(10/124)
no	19·2%	(69/360)	9·8%	(8/82)

The primary scientific interest centres on modelling the conditional densities $p(Y_3, Y_4 \mid Y_1, Z)$ and $p(Y_3, Y_4 \mid Y_2, Z)$, where the problem is that risk factors Y_1 and Y_2 are incomplete. Y_3 and Y_4 are fully observed, apart from censoring. Let Y_{obs} and Y_{mis} denote the observed and missing parts of Y, so $Y = (Y_{obs}, Y_{mis})$. Let R be an $n \times p$ binary matrix indicating the elements of Y that are observed ($R_{ij} = 1$ if Y_{ij} is observed).

The response mechanism models the probability that Y is observed as a function of observed and unobserved data, and is written as a conditional density $p(R = 1 \mid Y_{obs}, Y_{mis}, Z)$. Different assumptions concerning the relation between R on the one hand and Y_{obs}, Y_{mis} and Z on the other define different types of response mechanisms. We now explain in what way response mechanisms are relevant to the blood pressure problem. Three types of mechanisms will be distinguished: missing completely at random (MCAR); missing at random (MAR), and not missing at random

(NMAR).[9] The response mechanisms apply to blood pressure, that is, to R_1 and R_2, and not to R_3 and R_4, which are complete.

(i) *Mechanism MCAR*. The simplest response mechanism to consider is $p(R = 1) = r$, where r is the proportion of responses estimated by $r = n_{\text{obs}}/n = 835/956 = 0{\cdot}87$. The model stipulates that non-response of BP is randomly distributed. It will be clear that this model is not realistic in the Leiden 85+ Cohort study. Figure 1 shows that systematic differences in mortality exist, which would not be expected if the missing data were really MCAR.

(ii) *Mechanism*: *MAR on* Y_{obs}. Figure 1 suggests that the probability of BP measurement depends on survival, as in $p(R \mid Y_3, Y_4)$. In particular, short-term survivors have more missing BP data. This is not surprising since elderly in poorer condition (that is, people with reduced survival probabilities) are less likely to be measured. Since the response probability depends on survival, this factor must be taken into account when correcting for non-response.

(iii) *Mechanism*: *MAR on Z*. The probability of non-response could also be related to age, sex, type of residence, drug use and health status (see Table I). This mechanism $p(R \mid Z)$ is plausible in the Leiden 85+ Cohort study, so covariates Z will have to be considered when making non-response corrections.

(iv) *Mechanism NMAR*. Table II suggests that people with low BP are missing more frequently. Here, the probability of non-response is related to the BP, which (unfortunately) is unobserved sometimes. The mechanism is typified by $p(R \mid Y_{\text{obs}}, Y_{\text{mis}})$. Properly accounting for this mechanism requires external information about the distribution of Y_{mis} that is typically beyond the data.

An adequate treatment of non-response will mix the four mechanisms. Both MAR models can be usefully combined into one MAR mechanism $p(R \mid Y_{\text{obs}}, Z)$ that conditions upon all observed data. This is the basic model that will be used as a starting point in the next section. In addition, some alternatives regarding $p(Y_{\text{mis}})$ will be investigated to display the sensitivity of the results under various NMAR assumptions.

3. MULTIPLE IMPUTATION

Multiple imputation will be applied to account for the non-response. The main tasks to be accomplished in multiple imputation are:

1. Specify the posterior predictive density $p(Y_{\text{mis}} \mid X, R)$, where X is a set of predictor variables, given the non-response mechanism $p(R \mid Y, Z)$ and the complete data model $p(Y, Z)$.
2. Draw imputations from this density to produce m complete data sets.
3. Perform m complete-data analyses (Cox regression in our case) on each completed data matrix.
4. Pool the m analyses results into final point and variance estimates.

Simulation studies have shown that the required number of repeated imputations m can be as low as three for data with 20 per cent of missing entries.[10] In the following we use $m = 5$, which is a conservative choice.

3.1. Specification of the imputation model

The specification of the imputation model is the most complex step in multiple imputation. We first deal with the situation in which the response mechanism is MAR. In this case, no explicit

non-response model is needed,[9] and only the posterior $p(Y_{\text{mis}} \mid X)$ needs to be specified. The imputation model that forms the statistical basis for creating imputations involves two modelling choices: the form of model (linear, polynomial, logistic, etc.), and the set of predictors X that enter the model.

We use linear regression imputation with the 'closest predictor' option as in Rubin (reference 10, p. 168). This method models missing BP as a linear combination of predictor variables X, finds the complete case whose BP estimate is closest to that of the current incomplete case, and takes the observed BP from the former as the imputed BP value for the latter. A linear model may seem a rather simplistic choice here. Note, however, that the only function of the imputation model is to provide ranges of plausible values. Neither the form of the model nor the parameters estimates are particularly interesting. Unless the uncertainty associated with the missing entries is small, the exact form of the functional part of the model is largely immaterial.

A second choice concerns the selection of predictor variables. As a general rule, using all available information yields multiple imputations that have minimal bias and maximal certainty. This principle implies that the number of predictors should be as large as possible. It has been observed that including as many predictors as possible tends to make the MAR assumption more plausible, thus reducing the need to make special adjustments for NMAR mechanisms.[12,15] However, the full data set of the Leiden 85+ Cohort contains several hundred variables, all of which can potentially be used to generate imputations. It is not feasible (because of multicollinearity and computational problems) to include all these variables. It is also not necessary. The increase in explained variance in linear regression is typically negligible after the best, say, 15 variables have been included. For imputation purposes, it is expedient to select a suitable subset of data that contains no more than 15 to 25 variables. The strategy we followed for selecting predictor variables from a large data base consists of four steps:

1. Include all variables that appear in the complete-data model. Failure to do so may bias the complete-data analysis, especially if the complete-data model contains strong predictive relations. In particular, this means that Y_{obs} and Z are always part of the set of predictors.
2. In addition, include the variables that appear in the response model. Factors that are known to have influenced the occurrence of missing data (stratification, reasons for non-response) are to be included on substantive grounds. Other variables of interest are those for which the distributions differ between the response and non-response groups. These can be found by inspecting their correlations with the response indicator of the target variable (that is, the variable to be imputed). If the magnitude of this correlation exceeds a certain level, then the variable is included. The included set of variables is identified by U.
3. In addition, include variables that explain a considerable amount of variance of the target variable. Such predictors help to reduce the uncertainty of the imputations. They are crudely identified by their correlation with the target variable. The selected set of variables is identified by V.
4. Remove from the sets U and V those variables that have too many missing values within the subgroup of incomplete cases. A simple indicator is the percentage of observed cases within this subgroup, the percentage of usable cases.

The complete set of predictor variables is now given by $X = [Y_{\text{obs}}, Z, U, V]$.

The selection procedure was applied to data from the Leiden 85+ Cohort. Table III contains a summary of the selected predictors. Columns 1 and 2 give the correlation of each variable with systolic and diastolic BP, respectively. Column 3, labelled $r(R_1)$, provides the correlation with the

Table III. Summary of variables that are used for imputation. Columns 2 and 3 contain the correlations of the row variables with systolic and diastolic blood pressure. Column 4 lists the correlation with the response indicator for systolic blood pressure. The percentage of usable cases is equal to the percentage of the observed data of the row variable within the subgroup of cases ($n = 121$) that have missing systolic blood pressure data

Variable	r(SBP)	r(DBP)	$r(R_1)$	% usable cases
Y: Incomplete and outcome variables				
Systolic BP	1·00	0·59		
Diastolic BP	0·59	1·00		
Survival date	0·18	0·14	0·12	100
Censoring flag	0·13	0·11	0·08	100
Z: Covariates (model A)				
Sex	−0·10	−0·10	−0·04	100
Age	−0·11	−0·11	−0·14	100
U: Variables related to the non-response				
Type of residence	−0·21	−0·15	−0·08	100
ADL	−0·24	−0·11	−0·14	98
Previous hypertension	0·16	0·14	0·06	90
Uses diuretics	−0·04	−0·03	0·06	85
Year of interview	0·18	0·09	0·18	100
Year of blood sample	0·17	0·11	0·16	89
V: Prediction variables				
Serum albumin	0·24	0·18	0·02	67
Cognition (MMSE)	0·24	0·18	0·07	78
Current hypertension	0·23	0·17	0·01	83
Current/previous hypertension	0·22	0·19	0·04	83
Survival year	0·21	0·15	0·14	100
In (survival date)	0·20	0·15	0·09	100
Score GHQ	−0·19	−0·18	−0·01	83
Serum cholesterol	0·17	0·17	0·12	65
Fraction erythrocytes	0·17	0·20	0·08	70
Treated by specialist	−0·16	−0·11	0·02	100
Haemoglobin	0·15	0·18	0·08	70
Haematocrit	0·11	0·18	0·10	70

ADL: activities of daily living
MMSE: mini-mental-state examination

response indicator of systolic BP. This correlation has zero expectation under MCAR. Column 4 gives the percentage of observed values (out of 121) that can potentially be used to impute BP. Step 1 of the procedure includes the variables that appear in complete data model A: blood pressure; survival; sex and age. Step 2 adds a number of variables found to be related to the non-response (see Table I): type of residence; ADL-dependency; previous hypertension; use of diuretics; year of the interview and blood sample. Step 3 selects all variables whose absolute correlation with BP (systolic or diastolic) exceeds 0·15. The logarithm of the survival time was included as a potential predictor so that multiplicative relations between survival time and the covariates could be modelled by additive models. Step 4 removes variables with percentages of usable cases lower than 50 per cent. The total number of variables thus selected is equal to 24.

3.2. Drawing imputations, univariate case

Let $\theta = (\beta, \log \sigma)$ where β is a regression weight and σ is its standard deviation. We take $\log(\sigma)$ instead of σ since this enables the use of the conventional non-informative prior. The posterior predictive density can be written as

$$p(Y_{\text{mis}} \mid X, R) = \int p(Y_{\text{mis}} \mid X, R, \theta) p(\theta \mid X, R) \, \mathrm{d}\theta.$$

The standard procedure for creating multiple imputations consists of two steps: first, draw a value of θ^* from $p(\theta \mid X, R)$; second, draw a value Y^*_{mis} from its conditional posterior distribution given θ^*, that is, from $p(Y_{\text{mis}} \mid X, R, \theta = \theta^*)$. Repeating these steps m times yields m draws from the posterior distribution of Y_{mis}, which are to be used as the actual multiple imputations.

We first deal with the situation where Y_{mis} is univariate and where X is complete. Let Y_{obs} be the complete part of the variable to be imputed, and Y_{mis} denote the incomplete component. Let X_{obs} denote the predictors for n_{obs} individuals with observed BP, and let X_{mis} denote the complement of n_{mis} cases with missing BP. Let r be the number of predictors to be used for Y, which is $24 - 1 = 23$ our case. Given that Y is modelled by a linear regression model, and assuming the conventional uniform prior for $\theta \propto c$, the algorithm for creating m multiple imputations Y_{mis} is as follows:[10]

1. Calculate $W = (X'_{\text{obs}} X_{\text{obs}})^{-1}$, $\hat{\beta} = W X'_{\text{obs}} Y_{\text{obs}}$, and $\hat{Y}_{\text{obs}} = X_{\text{obs}} \hat{\beta}$.
2. Draw a random variable g from the χ^2-distribution with d.f. $= n_{\text{obs}} - r$.
3. Calculate $\sigma^2_* = (Y_{\text{obs}} - \hat{Y}_{\text{obs}})'(Y_{\text{obs}} - \hat{Y}_{\text{obs}})/g$.
4. Draw an r-dimensional Normal random vector $D \sim \text{N}(0, I_r)$, where I_r is the identity matrix of order r.
5. Calculate $\hat{\beta}_* = \hat{\beta} + \sigma_* W^{1/2} D$, where $W^{1/2}$ is the triangular square root of W obtained by the Cholesky decomposition.
6. Calculate predicted values $\hat{Y}_{\text{mis}} = X_{\text{mis}} \hat{\beta}_*$.
7. For each missing value $i = 1, \ldots, n_{\text{mis}}$ find the respondent whose $\hat{Y}_{\text{obs}}$ is closest to $\hat{Y}_{\text{mis},i}$ and take Y_{obs} of this respondent as the imputed value of i.
8. Repeat steps 2–7 m times to create $Y^{(1)}_{\text{mis}}, Y^{(2)}_{\text{mis}}, \ldots, Y^{(m)}_{\text{mis}}$.

Step 1 obtains $\hat{\beta}$ and $\hat{Y}_{\text{obs}}$ from the observed data by linear regression. Steps 2–5 provide a random draw from the posterior distribution of β. The idea in steps 6 and 7 is to borrow imputions from similar, but complete cases. The index of similarity between cases is the distance between their predictive means for BP (that is, their $\hat{Y}$-values) when predicted from the observed data. This technique is robust to substantial departures from the linear model and yields imputations that are always in the metric of the observed data. Note that the algorithm not only incorporates uncertainty due to deviations around the regression line (steps 2 and 3), but also reflects the variation of the regression line itself due to finite sampling (step 4).

3.3. Drawing imputations, multivariate case

Missing data are usually multivariate. It is convenient to split the multivariate problem into a series of univariate problems, and solve the multivariate case by iteration. For example, suppose that the data are multivariate Normal, then it is possible to generate imputations from this

distribution by applying an iterative algorithm that draws samples from a sequence of univariate linear regression.[12] We follow a slightly different approach, where we specify a set of conditional distributions, one for each incomplete variable. We do not explicitly assume a particular form for the multivariate distribution as in Schafer, but do assume that a multivariate distribution exists, and that draws from it can be generated by Gibbs sampling the conditional distributions.[16,17] Kennickell applied a similar idea to special patterns of missing data.[18] The present paper uses this method for general missing data patterns.

First, each incomplete entry is initialized by filling in a random draw from the marginal distribution of Y_{obs}. Then, Y_1 is imputed by the elementary procedure conditional on all other data (observed and imputed combined), then Y_2 conditional on all other data (using the most recent imputations for Y_1), and so on, until all incomplete variables in Y, Z, U and V have been imputed. Subsequently, start a second pass through the data, using all imputations created during the first pass, and so on. The set of imputations that are created after the 20th pass are used to derive the first complete data matrix. This whole procedure is executed m times in parallel, thus producing m completed data sets. We call this method regression switching.

Note that additional variability enters into step 1 of the elementary procedure. This reflects the fact that information is missing from the predictors. The method is a Gibbs sampler. Under quite general conditions the draws converge to the appropriate multivariate posterior density $p(Y_{\text{mis}} \mid Y_{\text{obs}}, X, R)$. It is, however, not always certain that the posterior actually exists. It is possible that the specification of two conditional distributions $p(Y_1 \mid Y_2)$ and $p(Y_2 \mid Y_1)$ are incompatible, so that no joint distribution $p(Y_1, Y_2)$ exists. Since there is no distribution to converge to, the algorithm will then alternate between isolated conditional distributions. In the linear case, this is probably more the exception than the rule. The subject of incompatible conditionals is, however, still an open research problem. Brand[19] studied the performance of a variety of regression switching algorithms based on possibly incompatible conditionals. It appears that these methods work very well when evaluated by classic frequentistic criteria.

The number of iterations (20) is much lower than is common in modern Markov chain simulation techniques, that often require thousands of iterations. In regression switching, the posterior distributions of the regression coefficients absorb the uncertainty in the predictors. The main question now is whether 20 steps are enough to stabilize these posteriors. Regression switching is reminiscent of HOMALS-like algorithms,[20] which usually convergence fast during the first few iterations. We therefore expect that not much will happen to the coefficients after, say, 10 iterations. Also, note that the elementary procedure creates imputations that are already statistically independent. No iterations need to be wasted for achieving independence between successive draws, as is typical for MCMC methods. To check convergence, we increased the number of iterations to 50, but did not find appreciable differences. Brand's simulation study successfully used just five iterations.[19]

3.4. δ-adjustment

Thus far we have assumed that the non-response mechanism is MAR. Section 2 discusses the possibility that the mechanism is NMAR, even after conditioning on Z. This section explains a simple adaptation of the switching method that can be used to adjust the imputations. The adjustment is independent of X and represents a relatively crude way of incorporating the idea that the non-responders are expected to have lower blood pressures. Though more advanced techniques based on selection[21] or pattern-mixture models[22] could also be applied here, it is

useful to see the effects of the crude δ-adjustment before addressing the added complexities that such models would bring. The primary function of the δ-adjustment is to investigate the robustness of the MAR assumption against violations. Studying the effect of varying δ on the complete data analysis helps to determine whether the relation between BP and mortality is affected by the non-response, and if so, at what point.

Suppose that the distribution of BP for the entire sample is known, but that actual data are only available for a subset of individuals. Table IV presents a numerical example of a specification of the response mechanism, where the probability of missingness $p(R = 0 \mid \mathrm{BP})$ varies between 0 and 35 per cent and depends on BP. One can apply Bayes rule to calculate the BP distribution of the responders $p(\mathrm{BR} \mid R = 1)$ and of the non-responders $p(\mathrm{BP} \mid R = 0)$. Both distributions are approximately Normal, but differ in location by $\delta = 151{\cdot}6 - 138{\cdot}6 = 13$ mmHg. This suggests that, in the absence of predictors, one can generate an imputation by subtracting an amount δ from a random draw from $p(\mathrm{BP} \mid R = 1)$.

Incorporating this idea into the regression switching method involves the addition of a location term to the imputation model as $Y_1 = X\beta + (1 - R_1)\delta + \varepsilon$. Here R_1 is the binary response indicator of systolic BP, and δ is a constant that is specified in advance by the imputer. This model postulates a mean difference in excess of that induced by X of δ units between responders and non-responders.

The non-response adjustment is applied to systolic BP only. Because SBP and DBP are correlated, both are imputed simultaneously in the same run. During the first iteration, imputations for SBP are decreased by δ points. These imputations are subsequently used for imputing third variables, amongst others DBP, which in turn are used to re-impute SBP during the second pass. Thus, the effect of the δ-adjustment on SBP automatically carries over to DBP, and it is even somewhat amplified by iteration. Since it is expected that the blood pressure of the non-responders is lower, δ is chosen as 0, -5, -10, -15 and -20 mmHg. The model reduces to the MAR case if $\delta = 0$.

3.5. Pooling

The computation of the final estimate for complete-data model parameters Q given the m completed data sets follows the standard rules (Rubin,[10] p. 76). Suppose that $\hat{Q}_i$ is a k-dimensional column vector containing the estimates of interest obtained by analysing the ith imputed data set $(i = 1, \ldots, m)$. Let U_i denote the corresponding $k \times k$ matrix of covariances among the estimates. The combined point estimate is then equal to $\hat{Q} = \sum_i^m \hat{Q}_i/m$. The combined covariance matrix is $T = U + (1 + m^{-1})B$, where $U = \sum_i^m U_i/m$, and where $B = \sum_i^m (\hat{Q}_i - \hat{Q})'(\hat{Q}_i - \hat{Q})/(m - 1)$. For large samples, the 95 per cent confidence interval for Q is estimated as $\hat{Q} \pm 1{\cdot}96\sqrt{T}$. Relative risk estimates and their confidence limits in the proportional hazards model can be obtained as $\exp(\hat{Q})$ and $\exp(\hat{Q} \pm 1{\cdot}96\sqrt{T})$.

4. RESULTS

Table V contains the mean blood pressure under various models. As expected, the mean blood pressure under MAR (with $\delta = 0$) is lower than the mean of the observed data, though the difference is small: 1·8 mmHg (SBP) and 1·3 mmHg (DBP). Decreases beyond this are due to the δ-adjustment.

Table VI contains relative mortality risks for different blood pressure strata. These are estimated by a classic proportional hazards model, corrected for age and sex. This corresponds

Table IV. Numerical example of an NMAR non-response mechanism, when there are more missing data for lower blood pressures

Class midpoint of Systolic BP (mmHg)	$p(R=0\|\text{BP})$	$p(\text{BP})$	$p(\text{BP}\|R=1)$	$p(\text{BP}\|R=0)$
100	0·35	0·02	0·01	0·06
110	0·30	0·03	0·02	0·07
120	0·25	0·05	0·04	0·10
130	0·20	0·10	0·09	0·16
140	0·15	0·15	0·15	0·19
150	0·10	0·30	0·31	0·25
160	0·08	0·15	0·16	0·10
170	0·06	0·10	0·11	0·05
180	0·04	0·05	0·05	0·02
190	0·02	0·03	0·03	0·00
200	0·00	0·02	0·02	0·00
Mean (mmHg)		150	151·6	138·6

Table V. Mean and standard deviation of the observed and imputed blood pressures. The statistics of imputed BP are pooled over $m = 5$ multiple imputations

	N	δ	SBP		DBP	
			Mean	SD	Mean	SD
Observed BP	835		152·9	25·7	82·8	13·1
Imputed BP	121	0	151·1	26·2	81·5	14·0
	121	−5	142·3	24·6	78·4	13·7
	121	−10	135·9	24·7	78·2	12·8
	121	−15	128·6	25·0	75·3	12·9
	121	−20	122·3	25·2	74·0	12·1

to model A of the introduction. It was expected that multiple imputation would raise the risk estimates in comparison with the analysis based on the complete cases, but the results do not confirm this. Only slight differences in mortality exist, even among non-response models with very different δ's. It appears that, for this application, risk estimates are insensitive to the missing data and the various non-response models used to deal with them.

5. CONCLUSION

A critical point in our application is the poor prediction of blood pressure (multiple r^2 (SBP) = 0·24 and r^2 (DBP) = 0.17). The generated imputations thus are quite uncertain and contain considerable residual noise. The increase of precision of risk estimates under the ignorable model is therefore, at best, remote. This situation is not atypical, as low r^2 is common in epidemiological

Table VI. Relative mortality risks (with 95 per cent confidence interval) estimates by the classic proportional hazards model corrected for age and sex. Estimates are relative to the reference group (SBP: 145–160, DBP: 75–80). The top row in each table contains the estimates from the complete case (CC) analysis. The bottom five rows are based on multiple imputation, where δ refers to the parameter in the δ-adjustment

δ	Systolic blood pressure (mmHg)					n
	<125	125–140	165–180	185–200	>200	
CC	1·76 [1·36–2·28]	1·48 [1·19–1·84]	1·11 [0·87–1·42]	1·14 [0·85–1·54]	0·89 [0·51–1·57]	835
0	1·71 [1·34–2·20]	1·47 [1·20–1·80]	1·12 [0·89–1·42]	1·15 [0·87–1·53]	0·97 [0·56–1·71]	956
−5	1·69 [1·25–2·29]	1·40 [1·14–1·73]	1·07 [0·82–1·39]	1·09 [0·82–1·46]	0·98 [0·57–1·67]	956
−10	1·73 [1·33–2·25]	1·47 [1·17–1·85]	1·09 [0·83–1·43]	1·11 [0·82–1·51]	0·95 [0·54–1·67]	956
−15	1·67 [1·30–2·14]	1·46 [1·18–1·80]	1·07 [0·82–1·39]	1·09 [0·80–1·47]	0·91 [0·51–1·60]	956
−20	1·69 [1·35–2·11]	1·42 [1·14–1·76]	1·10 [0·86–1·40]	1·10 [0·82–1·48]	0·91 [0·51–1·62]	956
δ	Diastolic blood pressure (mmHg)					
	<65	65–70	85–90	95–100	>100	n
CC	1·82 [1·33–2·48]	1·21 [0·95–1·55]	0·96 [0·78–1·19]	0·87 [0·67–1·13]	0·81 [0·55–1·20]	830
0	1·78 [1·34–2·37]	1·19 [0·95–1·50]	0·97 [0·78–1·19]	0·90 [0·70–1·14]	0·83 [0·58–1·19]	956
−5	1·63 [1·21–2·19]	1·09 [0·87–1·38]	0·87 [0·71–1·06]	0·80 [0·62–1·01]	0·81 [0·55–1·19]	956
−10	1·83 [1·32–2·53]	1·22 [0·96–1·55]	0·97 [0·78–1·20]	0·88 [0·68–1·14]	0·81 [0·55–1·20]	956
−15	1·69 [1·25–2·28]	1·20 [0·94–1·53]	0·94 [0·76–1·15]	0·83 [0·64–1·08]	0·76 [0·51–1·13]	956
−20	1·62 [1·23–2·13]	1·20 [0·92–1·56]	0·92 [0·75–1·14]	0·85 [0·66–1·09]	0·78 [0·53–1·15]	956

studies. Even if the imputed blood pressures are lowered by enlarging δ, the risk estimates hardly change. One factor contributing to the apparent stability is the moderate amount of missing date (12·5 per cent). Another factor might be that differences in mortality between responders and non-responders are simply too small to exert a serious impact on the estimates. It is, however, difficult to know this beforehand.

The method used to generate the imputations contains some compromises that could mask the relevant effects. Note that predictor selection is optimized for SBP, and that the same univariate technique is used for each incomplete predictor. Sharper imputations require separate modelling of each incomplete predictor. Though possible, this introduces additional complications in the algorithm. We are currently working on a more extensive methodology that allows the user to specify a separate imputation model for each incomplete variable. This work builds on the strategy put forth in this paper. We expect that this leads to a flexible and generally useful methodology for creating sharp imputations in multivariate missing data.

It is known that complete-case methods yield valid inferences when missingness depends on the regressors.[23,24] The blood pressure problem is more complicated, however, because missingness depends on survival, the outcome measure. The options for a proper analysis of the data are limited in this case, and complete-case analysis does not provide appropriate estimates in general. The fact that the results of multiple imputation and complete-case analysis are similar does not imply that complete-case analysis had been appropriate in the first place. However, given that we now know that the missing data hardly influence the risk estimates, more elaborate analyses (for example, model B) of the same material will probably continue to yield valid inferences if only the complete cases are taken into account.

REFERENCES

1. Boshuizen, H. C., Izaks, G. J., van Buuren, S. and Ligthart, G. J. *Bloeddruk en Sterfte bij Hoogbejaarden* [*Blood Pressure and Mortality in the Oldest Old*], TNO-rapport PG 95.032, Leiden, 1995.
2. Rajala, S., Haavisto, M., Heikinheimo, R. and Mattila, K. 'Blood pressure and mortality in the very old', (letter) *Lancet*, (**ii**), 520–521 (1983).
3. Glynn, R. J., Field, T. S., Rosner, B., Hebert, P. R., Taylor, J. O. and Hennekens, C. H. 'Evidence for a positive linear relation between blood pressure and mortality in eldery people', *Lancet*, (**i**); **345**, 825–829 (1995).
4. Boshuizen, H. C., Izaks, G. J., van Buuren, S. and Ligthart, G. J. 'Blood pressure and mortality in the oldest old', *British Medical Journal*, **316**, 1780–1784 (1998).
5. Schluchter, M. D. and Jackson, K. L. 'Log-linear analysis of censored survival data with partially observed covariates', *Journal of the American Statistical Association*, **84**, 42–52 (1989).
6. Lin, D. Y. and Ying, Z. 'Cox regression with incomplete covariate measurements', *Journal of the American Statistical Association*, **88**, 1341–1349 (1993).
7. Robins, J. M., Rotnitzky, A. and Zhao, L. P. 'Estimation of regression coefficients when some regressors are not always observed', *Journal of the American Statistical Association*, **89**, 846–866 (1994).
8. Ahn, H. and Loh, W.-J. 'Tree-structured proportional hazards modelling', *Biometrics*, **50**, 471–485 (1994).
9. Rubin, D. B. 'Inference and missing data', *Biometrika*, **63**, 581–592 (1976).
10. Rubin, D. B. *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York, 1987.
11. Rubin, D. B. 'Multiple imputation after 18+ years (with discussion)', *Journal of the American Statistical Association*, **91**, 473–518 (1996).
12. Schafer, J. L. *Analysis of Incomplete Multivariate Data*, Chapman & Hall, London, 1997.
13. Lagaay, A. M. 'The Leiden 85-plus study: A population based comprehensive investigation of the oldest old', University of Leiden, Academic Thesis, 1991.
14. Lagaay, A. M., van der Meij, J. C. and Hijmans, W. 'Validation of medical history taking as part of a population based survey in subjects aged 85 and over', *British Medical Journal*, **304**, 1091–1092 (1992).
15. Rubin, D. B., Stern, H. S. and Vehovar, V. 'Handling Don't Know survey responses: The case of the Slovenian plebiscite', *Journal of the American Statistical Association*, **90**, 822–828 (1995).
16. Gelfand, A. E. and Smith, A. F. M. 'Sampling-based approaches to calculating marginal densities', *Journal of the American Statistical Association*, **85**, 398–409 (1990).
17. Tanner, M. A. *Tools for Statistical Inference*, 2nd edn, Springer, New York, 1993.
18. Kennickell, A. B. 'Imputation of the 1989 survey of consumer finances: Stochastic relaxation and multiple imputation', in *ASA 1991 Proceedings of the Section on Survey Research Methods*, ASA, Alexandria, 1991, pp. 1–10.
19. Brand, J. P. L. 'Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets', Academic thesis, Erasmus University, Rotterdam, 1998.
20. Gifi, A. *Nonlinear Multivariate Analysis*, Wiley, New York, 1990.
21. Heckman, J. 'The common structure of statistical models of truncation, sample selection, and limited dependent variables, and a simple estimator for such models', *Annals of Economic and Social Measurement*, **5**, 475–492 (1976).
22. Little, R. J. A. 'Pattern-mixture models for multivariate incomplete data', *Journal of the Amerian Statistical Association*, **88**, 125–134 (1993).
23. Little, R. J. A. 'Regression with missing X's: A review', *Journal of the American Statistical Association*, **87**, 1227–1237 (1992).
24. Greenland, S. and Finkle, W. D. 'A critical look at methods for handling missing covariates in epidemiologic regression analyses', *American Journal of Epidemiology*, **142**, 1255–1264 (1995).

