Insights into the use of Bayesian models for informative missing data

Alexina Mason[1*], Nicky Best[1], Ian Plewis[2] and Sylvia Richardson[1]

[1] *Department of Epidemiology and Biostatistics, Imperial College London, UK*

[2] *Social Statistics, University of Manchester, UK*

SUMMARY

Many studies are affected by missing data, which complicates subsequent analyses for researchers. Here, we are concerned with missing outcomes generated by a missingness mechanism that is informative. In this case, ad hoc approaches are not suitable and if we wish to adequately model this type of missing data, we need to use 'statistically principled' methods. We investigate one of these methods, Bayesian full probability modelling, in which a joint model consisting of a model of interest and a model for the informative missing data mechanism is specified.

Using simulated data, we explore the performance of Bayesian methods, finding that the addition of a model of missingness generally improves the overall fit of the model of interest leading to better prediction, but that the estimates of parameters of interest can be adversely affected by skewness in the response variable. The effective number of parameters, p_D, is a measure of the ratio of the information in the likelihood to that of the posterior. We consider the use of the scaled p_D of the model of missingness as a diagnostic that indicates the amount of informativeness in the missing data given our assumptions. We find that it is useful for indicating how far our missing data departs from missing at random, but that it should not be used for choosing the 'best' model of missingness. These points are illustrated with two real examples, which analyse test score data from the 1958 British

birth cohort study and data from a clinical trial. Copyright © 2000 John Wiley & Sons, Ltd.

1. INTRODUCTION

Missing data is commonly encountered in many types of studies and is generally an unavoidable nuisance, which can lead to biased and inefficient inference if ignored or handled inappropriately. An extensive literature has built up on the topic and the various approaches have been catalogued and reviewed in papers [1, 2], as well as detailed in comprehensive textbooks [3, 4, 5, 6].

The appropriateness of a particular approach is dependent on the mechanism that leads to the missing data, and Rubin [7] developed a framework for inference from incomplete data that is still widely used. Following Rubin, missing data are generally classified into three types: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Informally, MCAR occurs when the missingness does not depend on observed or unobserved data, in the less restrictive MAR it depends only on the observed data, and when neither MCAR or MAR hold, the data are MNAR.

A common ad hoc approach is complete-case analysis, in which individuals whose information is incomplete are discarded. Although this method has the advantage of simplicity, it is generally inappropriate as it leads to loss of precision and, unless the missing data mechanism is MCAR, to bias. By contrast, 'statistically principled' methods seek to combine information in the observed data with assumptions about the missing value mechanism, and account for the uncertainty introduced by the missing data.

One such method entails building a joint model including a model of interest and a model of missingness. The Bayesian approach to modelling informative missing responses that we discuss

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

uses such joint models. In addition to allowing the incorporation of realistic assumptions about missingness, it has the advantage of enabling coherent model estimation. Also, because the models are constructed in a modular way, they are relatively easy to adapt to explore a range of assumptions about the missingness mechanism. This is important as often the missing data mechanism is unknown and the data alone cannot determine whether we have MAR or MNAR missingness, making sensitivity analysis essential. In recent years Markov chain Monte Carlo (MCMC) methods have provided a way of analysing complex Bayesian models [8, 9], and examples of Bayesian methods for non-ignorable missing data have begun to appear [10, 11].

Despite the increasing use of Bayesian joint models for informative missing data, there has been little written on how the addition of the model of missingness affects the estimation of the model of interest parameters and how Bayesian diagnostics should be interpreted. To this end, we explore the use of Bayesian full probability modelling for data with missing response values which are assumed to be informative, comparing its performance with complete-case analysis.

We start by using simulated data to gain a basic understanding of the performance of joint models, before applying our methods to real datasets. The models that we use are described in Section 2, and the data are introduced in Section 3. In Section 4, after discussing model evaluation, we describe our investigation using simulated data. To provide context, we start with a look at the deficiencies of complete-case analysis and then discuss what improvements can be expected from a joint model. In particular, we consider how critical are the strength of the relationship in the model of interest and the adequacy of the model of missingness. We show that our joint model works better for symmetric than asymmetrically distributed data, so selecting an appropriate transformation of the response is important but difficult in the

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

presence of missing values. We finish this investigation with a look at the interpretation of possible diagnostics that can help determine whether a missing not at random assumption is reasonable. Our methods are applied to two real examples in Section 5 and we conclude with a discussion in Section 6.

2. BAYESIAN FULL PROBABILITY MODELLING OF INFORMATIVE MISSING DATA

Let $\boldsymbol{y} = (y_i)$ denote a dataset for i=1,...,n individuals, and partition $\boldsymbol{y}$ into observed, $\boldsymbol{y}^{obs}$, and missing, $\boldsymbol{y}^{mis}$, values, i.e. $\boldsymbol{y} = (\boldsymbol{y}^{obs}, \boldsymbol{y}^{mis})$. Now define $\boldsymbol{m} = (m_i)$ to be a binary indicator variable such that

$$m_i = \begin{cases} 0: \quad y_i \quad \text{observed} \\ 1: \quad y_i \quad \text{missing} \end{cases} \tag{1}$$

and let $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ be unknown parameters. The joint distribution of the complete data is $f(\boldsymbol{y}, \boldsymbol{m}|\boldsymbol{\beta}, \boldsymbol{\theta}) = f(\boldsymbol{y}^{obs}, \boldsymbol{y}^{mis}, \boldsymbol{m}|\boldsymbol{\beta}, \boldsymbol{\theta})$, which can be factorised as

$$f(\boldsymbol{y}^{obs}, \boldsymbol{y}^{mis}, \boldsymbol{m}|\boldsymbol{\beta}, \boldsymbol{\theta}) = f(\boldsymbol{m}|\boldsymbol{y}^{obs}, \boldsymbol{y}^{mis}, \boldsymbol{\beta}, \boldsymbol{\theta}) f(\boldsymbol{y}^{obs}, \boldsymbol{y}^{mis}|\boldsymbol{\beta}, \boldsymbol{\theta}). \tag{2}$$

This can be simplified to

$$f(\boldsymbol{y}^{obs}, \boldsymbol{y}^{mis}, \boldsymbol{m}|\boldsymbol{\beta}, \boldsymbol{\theta}) = f(\boldsymbol{m}|\boldsymbol{y}^{obs}, \boldsymbol{y}^{mis}, \boldsymbol{\theta}) f(\boldsymbol{y}^{obs}, \boldsymbol{y}^{mis}|\boldsymbol{\beta}) \tag{3}$$

if we assume that $\boldsymbol{m}|\boldsymbol{y}, \boldsymbol{\theta}$ is conditionally independent of $\boldsymbol{\beta}$, and $\boldsymbol{y}|\boldsymbol{\beta}$ is conditionally independent of $\boldsymbol{\theta}$, which is usually reasonable in practice. This factorisation of the joint distribution is known as a selection model [4, 1] and underpins Bayesian full probability modelling of missing data in which a joint model is specified for the relationship of interest, $f(\boldsymbol{y}|\boldsymbol{\beta})$, and the missing data mechanism, $f(\boldsymbol{m}|\boldsymbol{y}, \boldsymbol{\theta})$. In Bayesian full probability modelling, the joint posterior distribution, which is the basis for all Bayesian inference, is estimated

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

simultaneously for both unknown parameters and missing data, so all sources of uncertainty are properly taken into account.

In this paper, we restrict ourselves to the case where only the response has missing values, and assume that the missingness mechanism is non-ignorable. We consider the simple but nonetheless informative case where the model of interest is a linear regression with a univariate outcome y_i and a vector of covariates $x_{1i}, \ldots, x_{pi}$, for i=1,...,n individuals, i.e.

$$y_i \sim N(\mu_i, \sigma^2), \tag{4}$$

$$\mu_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ki}$$

and the model of missingness has the form

$$m_i \sim Bernoulli(p_i), \tag{5}$$

$$logit(p_i) = \theta_0 + \theta_1 y_i$$

where m_i is a binary missing value indicator for y_i. Note that it is assumed that the parameter θ_1 captures the dependence of the missingness on the outcome.

We wish to estimate all the parameters in this joint model, but it is not obvious where the information for estimating the model of missingness parameters will come from. One possibility is to place strong priors on the $\boldsymbol{\theta}$ parameters [10], which is similar to a sensitivity analysis. Here, we are not following this approach, but instead try to learn about the missingness mechanism from the data using a combination of the distributional assumptions of the model of interest and the proposed functional form of the model of missingness.

We shall refer to a joint model of this form, consisting of a model of interest and a missingness model, run with missing response values, as JM. For simulated datasets where the missing response values are known, we also run a joint model of the same form but with a full set

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

of response values, which we shall call TARGET. This is used for bench marking the results from the simulated datasets, as it gives targets for the fit of both the model of interest and the missingness model. In addition, we run the model of interest (Equation 4) on complete cases only, referred to as CC, providing a comparison of the model of interest fit with a commonly used method.

Vague priors are specified for the unknown parameters of the model of interest: the $\boldsymbol{\beta}$ parameters are assigned N(0,10000) priors and the precision, $\frac{1}{\sigma^2}$, a Gamma(0.001,0.001) prior. Following Wakefield [12] and Jackson et al. [13], we specify a logistic(0,1) prior for θ_0 and a weakly informative N(0,0.68) prior for θ_1, which corresponds to an approximately flat prior on the scale of p_i.

3. DATA

To explore the performance of Bayesian missing data models, we use a variety of simulated and real datasets as described below.

3.1. Simulated multivariate Normal data (MVNsim)

50 datasets each with 1000 records comprising a response, y, and a single covariate, x, are simulated from a multivariate Normal distribution, s.t.

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right). \tag{6}$$

For these datasets the true values of the parameters of our model of interest are $\beta_0 = 1$ and $\beta_1 = 0.5$. We then delete some of the responses, y, according to different models of missingness described in Section 4. This simple setup is useful to highlight key characteristics of the performance of JM.

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

3.2. NCDS test score data (NCDSsim and NCDSreal)

Our first real example is taken from the National Child Development Study (NCDS), a longitudinal study which follows all those living in Great Britain who were born in one week in March 1958. It is multi-disciplinary, with domains of interest including health, family background and education. Response patterns for the different domains vary [14], with unusually low response for education in sweep 3 when the cohort are 16 years old (affected by change of school leaving age). For sweep 3, 87% of the target sample responded, of which 82% provided data on educational attainments. We use a subset of this educational data, using models proposed by Goldstein [15] for investigating the effects of social class on educational attainment as a starting point. Goldstein initially fitted a pair of linear equations which regressed 11-year test scores on 7-year test scores, and 16-year test scores on both 7 and 11-year test scores, modelling reading and mathematics separately. Social class, based on the occupation of the child's father, was incorporated as an additive variable in a further model. It is thought that the missingness in the 16-year test score may be informative, with individuals more likely to have not taken the test if they were likely to perform poorly.

We restrict our attention to the mathematics test scores and use this NCDS data in two ways. Firstly, we take fully observed subsets of the data and simulate missingness in the response variable for use in our investigation (NCDSsim), and secondly we apply our proposed methods to all the collected data including individuals with unknown response values (NCDSreal). Although we are working with educational test scores, such scores are typical of data arising from medical and epidemiological studies, as well as social science applications.

We construct NCDSsim as 10% samples from the 10,312 NCDS individuals with complete observations for the test scores at ages 11 and 16. The sizes of the subsets vary slightly, as

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

each subset is created by sampling from the 10,312 cohort members completely at random with a 0.1 probability of inclusion. Some of the responses are then deleted according to different missingness criteria, as described in Section 4. Our model of interest regresses the test score at age 16 on the test score at age 11.

Figure 1 shows that the relationship between the mathematics test scores at each age is approximately linear, and that the distributions of the test scores at 11 and 16 are asymmetric, with lower scores more prevalent than higher scores.

Figure 1 here

For Section 5 on applications to real data, similar models with additional covariates are run on NCDSreal, which includes individuals with unknown response values (although individuals with unknown covariates are still excluded).

3.3. Antidepressant trial data (HAMD)

Our second real example uses data from a six centre clinical trial comparing three treatments of depression, which were previously analysed by Diggle and Kenward (DK) [16] and Yun et al. [17]. 367 subjects were randomised to one of three treatments and rated on the Hamilton depression score (HAMD) on five weekly visits, the first before treatment, week 0, and the remaining four during treatment, weeks 1-4. The higher the HAMD score, the more severe the depression. Some subjects dropped out of the trial from week 2 onwards, with approximately one third lost by the end of the study. DK found evidence of informative missingness given their modelling assumptions, and we examine the evidence provided by Bayesian models.

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

4. INVESTIGATING THE PERFORMANCE OF BAYESIAN JOINT MODELS

Having introduced our data, we now investigate how well Bayesian joint models perform when data have missing responses generated by an informative missingness mechanism. Firstly we look at what happens to the parameter estimates and fit of the model of interest when we ignore the missingness and perform a complete-case analysis. We then compare these results with those obtained when a model of missingness is added to the model of interest, and discuss the improvements.

Our models are run for 15,000 iterations including 10,000 burn-in, with three chains initialised using diffuse starting values. Both variables are centred and standardised, which is recommended good practice to improve mixing in MCMC estimation [18]. For each run we have looked at the Gelman-Rubin convergence statistic [19] for the individual parameters, and have assumed convergence if all of these are less than 1.05 and a visual inspection of the trace plots is satisfactory. On this basis, all the runs discussed in this paper converged unless stated otherwise and have been run using the WinBUGS software [20].

4.1. Model evaluation

As part of the assessment of our models, we look at the bias and efficiency of the parameter estimates. We define the percentage bias of a parameter estimate as

$$\% \text{ bias} = \frac{(\hat{\beta} - \tilde{\beta}_F)}{\tilde{\beta}_F} \times 100 \tag{7}$$

where $\tilde{\beta}_F$ is the parameter estimate based on the full dataset (modelled by TARGET) and $\hat{\beta}$ is the parameter estimate for some other model, i.e. JM or CC. In each case, the parameter estimates are taken to be the posterior means. Note that this is slightly different to the usual definition of bias as the expectation of the difference between a parameter estimate and its

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

theoretical true value. The efficiency of a parameter estimate is defined to be the width of the 95% interval given by fitting TARGET divided by the 95% interval from some other model (JM or CC). The 95% intervals are calculated from the 2.5 and 97.5 percentiles of the posterior distribution of the parameter.

Additionally, we use mean square error (MSE), i.e.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - E(y_i|\boldsymbol{\beta}))^2, \tag{8}$$

as a measure of overall fit for comparing the performance of CC and JM. We calculate this quantity using the posterior means of the $\boldsymbol{\beta}$ parameters as plug-ins.

4.2. What are the deficiencies of complete-case analysis?

The deficiencies of complete-case analysis are well known, see for example Little and Rubin [4, chap. 3]. Nevertheless, we begin our investigation by reviewing these using the MVNsim and NCDSsim data introduced in Section 3, to explore the extent to which complete-case analysis introduces bias in practice.

4.2.1. MVNsim For each MVNsim dataset, we impose three forms of missingness on y, using the equation $p_i = \phi_0 + \phi_1 y_i$ with varying values of $\boldsymbol{\phi}$, where p_i is the probability of being missing. The resulting linear missingness is an intuitively simple setup, but we will have to model this using the linear logit specified by Equation 5 to ensure that the probabilities lie in the range [0,1]. A logistic transformation of a linear line gives a sigmoid curve which is essentially linear for probabilities between 0.2 and 0.8, but non-linear outside this range [21]. We shall refer to the three forms of missingness as MCAR, posMNAR and negMNAR, defined

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

as follows.

MCAR: the probability of being missing is set to 0.5 for all values of y, i.e. $\phi_0 = 0.5$, $\phi_1 = 0$.

posMNAR: linear missingness with a steep positive gradient, such that the probability of being missing for the lowest value of y is 0 and the probability of being missing for the highest value of y is 1, i.e. ϕ_1 is positive.

negMNAR: linear missingness with a steep negative gradient, such that the probability of being missing for the lowest and highest values of y are 1 and 0 respectively, i.e. ϕ_1 is negative.

The design of this simulation means that while the gradient of missingness, ϕ_1, is always zero for MCAR, it varies slightly across the 50 replicates for posMNAR (0.13 to 0.18) and negMNAR (-0.18 to -0.13) as it is dependent on the range of the generated responses. The percentage of missingness has a mean of 50% for all forms of missingness, with a range of 43-58%. A complete-case analysis was performed on each form of missingness for each dataset by running CC using WinBUGS as described in Section 2. To get a target fit, TARGET (see Section 2) was also run.

We know that complete-case analysis assumes that the missingness mechanism is MCAR, and therefore expect bias in our parameter estimates for MNAR missingness. The extent and pattern of this bias is shown by the CC points (black crosses) in the β_0 and β_1 bias plots of Figure 2. (We will discuss the JM points (open green circles) in Section 4.3.1.) The MCAR estimates for β_0 and β_1 look unbiased. By contrast, on average, CC under-estimates the slope of the model of interest, β_1, for posMNAR and negMNAR by similar amounts. When high responses are more likely to be lost (posMNAR) β_0 is always substantially under-estimated, while when low responses are more likely to be lost (negMNAR) CC always substantially

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

over-estimates β_0. The % bias for β_0 is about three times the % bias for β_1.

Figure 2 here

As expected, the deterioration in the overall fit compared to TARGET, as measured by MSE, is slight for MCAR, but has a mean of 8-9% for the two forms of MNAR missingness (Figure 2).

4.2.2. NCDSsim Having seen how MNAR missingness biases the intercept of the model of interest either up or down, depending on whether low or high responses are more likely to be lost, but always biases the slope downwards, we carried out a further simulation using NCDSsim. We simulate the missingness, with varying gradients, ϕ_1, and proportions of missingness. The gradients varied from -0.23 to 0.23 including 0 (which is equivalent to a MCAR mechanism), and the percentage of missingness from 4.9% to 96.5%. As before the models TARGET and CC are run for each dataset.

The % differences between the CC and TARGET estimates of β_1 are plotted against the gradient of missingness, ϕ_1, as black crosses for four levels of missingness in the top panel in Figure 3. (The JM points (open green circles) which are also shown will be discussed in Section 4.3.2.) When ϕ_1 is positive (individuals are more likely to be missing if they have high test scores at age 16) CC always under-estimates β_1, apart from a few datasets with gradient close to 0. The magnitude of this under-estimation increases with ϕ_1 and the percentage of missing values. From our multivariate Normal simulations we could expect to see a similar degree of under-estimation when ϕ_1 is negative, but in fact CC sometimes under-estimates and sometimes over-estimates β_1, with over-estimation more likely, and the bias is less for negative ϕ_1 compared to positive ϕ_1. The NCDSsim datasets differ from our MVNsim datasets in that both the response and covariate distributions are skewed (see Figure 1), and we will explore

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

the implications of this asymmetry later.

Figure 3 here

Our findings for the β_0 bias are as expected, as shown in the bottom panel in Figure 3. Note that absolute, rather than percentage, bias is shown for β_0, because TARGET β_0 is very close to 0 and percentages would be unstable. CC increasingly over-estimates β_0 as ϕ_1 becomes increasingly negative and increasingly under-estimates β_0 as ϕ_1 becomes increasingly positive at all levels of missingness.

As regards MSE (graph not shown), the overall fit of CC compared to TARGET deteriorates as ϕ_1 becomes steeper or the percentage of missing values increases.

4.3. What improvements can we expect from a joint model?

We have seen that with missing responses, complete-case analysis results in biased parameter estimates unless the missingness is MCAR. Further, the direction of this bias is affected by the shape of the distribution of the original data (observed and missing) in addition to the shape of the missingness pattern. We now investigate the extent to which these biases are removed by adding a model of missingness to our model of interest. Again we start by looking at the impact on the simulated MVNsim data before examining the more realistic NCDSsim data.

4.3.1. MVNsim A third model, JM, was run as described in Section 2. Looking at the JM points in Figure 2, we see that the bias in the $\boldsymbol{\beta}$ estimates is almost eliminated and the overall fit of our model of interest is close to the TARGET model. Interestingly, for MNAR missingness, the estimate of β_1 is always higher for JM than CC, resulting in a reduced β_1 difference from TARGET for most but not all repetitions.

So far we have concentrated on the bias of the parameter estimates, but the efficiency of these

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

estimates is also of interest. The loss of records for CC results in efficiency of approximately 70% for the estimates of both parameters. Similar efficiency is achieved for the JM estimate of β_1, with the additional information from the partially observed cases being offset by greater uncertainty about the $\boldsymbol{\beta}$ parameters introduced by the missing response values. However, for the JM estimate of β_0, efficiency is further reduced to just over 40%. Hence running a joint model does not provide gains of efficiency.

4.3.2. NCDSsim JM was run for NCDSsim. The dataset with the highest percentage of missingness is excluded from the results as it failed to converge. Looking again at Figure 3, we see that, consistent with our findings for MVNsim, JM always pulls β_1 upwards from the CC estimate (apart from MCAR or almost MCAR), not always giving an improved estimate. However JM consistently produces an estimate for β_0 which is much closer to the target from TARGET, correcting both under-estimation and over-estimation. The addition of a model of missingness leads to an improvement in the overall fit (MSE) for all but the shallowest gradients (results not shown).

4.4. How critical is the strength of the relationship in the model of interest?

It is known that selection models can be sensitive to the correct specification of both parts of the joint model [22]. We now explore the sensitivity of our findings to a related issue, the strength of the relationship in the model of interest.

Because MVNsim is simulated data, we know that our assumption of a linear relationship between the covariate and response is correct. Our findings so far are based on a true correlation between the response and covariate of 0.5. To investigate how the strength of this relationship impacts our results, we repeated the simulation using correlations of 0.1, 0.25, 0.75 and 0.9

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

(results not shown). We found that as the relationship between the variables gets stronger, the CC bias is reduced for both β_0 and β_1, there is less variation between replicates and the efficiency of the estimation of the JM β_0 increases towards the CC level. Reassuringly, JM seems to correct the bias in the parameter estimates regardless of the strength of the relationship. However, as the correlation gets weaker we start to encounter MCMC convergence problems, with 10 of the 150 JM runs failing to converge with the 0.25 correlation and 110 of the 150 JM runs not converging when the correlation is 0.1. This suggests that JM identifiability is driven by a strong relationship in the model of interest.

4.5. How critical is the adequacy of the model of missingness?

We now turn our attention to the other part of the joint model, and consider the adequacy of the model of missingness. Our results are potentially affected by two sources of error in our model of missingness: the use of a linear logit model to approximate a linear relationship and failure to fit the 'best possible linear logit' (the 'best possible linear logit' is assumed to be the linear logit fitted by TARGET).

To gain insight into the relative importance of the two sources of error, we repeated the MVN simulations using exactly the same datasets with the same missing responses either (i) replacing the linear logit with the exact equation which was originally used to select the missing responses, i.e. $p_i = \phi_0 + \phi_1 y_i$ or (ii) retaining the linear logit for the model of missingness, but fixing its $\boldsymbol{\theta}$ parameters to the posterior means of θ_0 and θ_1 that were estimated by TARGET. Our results (not shown) suggest that in Bayesian joint modelling the use of a linear logit adequately models linear missingness, but there are some benefits to improving the fit of the model of missingness if possible. This might be achieved by the use of informative priors to

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

incorporate additional knowledge on the shape of the missingness model.

Having established that a linear logit is a good choice for modelling linear missingness, we investigated what happens when we use an incorrect model of missingness. We again repeated the JM analysis of the MVNsim datasets, this time using a restricted linear missingness model in which θ_1 is restricted to positive values for negMNAR and negative values for posMNAR. The missingness model slope parameter, θ_1, is now estimated to be close to 0, and our joint model no longer removes the bias in the model of interest parameter estimates, producing similar estimates to CC. This can be seen from the linear (blue) and restricted linear (red) points labelled "no" (indicating no transformation) in Figure 4. (The remaining points in this graph will be discussed later.)

Figure 4 here

We also ran the MVN simulations using the quadratic logistic equation $logit(p_i) = \theta_0 + \theta_1 y_i + \theta_2 y_i^2$ as the model of missingness. About 20% of the repetitions failed to converge, and those which did converge failed to correct the bias in the parameter estimates or reduce MSE as well as the linear logistic equation (see the green points in Figure 4). Hence the missingness model needs to be a good approximation of the true missingness mechanism in order to reduce bias and MSE in the model of interest.

4.6. How critical is the error distribution in the model of interest?

In Section 4.3 we found that JM was much better at correcting bias in the β_1 estimate for symmetric MVNsim than skewed NCDSsim, and we now consider the reasons for this. In setting up our model of interest, we have assumed that the errors follow a Normal distribution. This assumption is crucially used by JM when it fills in the missing responses, in a way that will

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

attempt to correct any skewness in the observed responses given their covariates [4, chap. 15]. For MVNsim this is fine, because all the skewness in the observed responses stems from the missingness mechanism and so JM does well. By contrast, the NCDS response distribution is already skewed and JM now tries to compensate for the combined effects of the original skewness and the added skewness from the imposed missingness. It has no way of distinguishing between the two sources of skewness, so cannot limit its correction to the skewness from the missingness mechanism as required.

To better understand what is happening, we transform our original MVNsim data and then impose MCAR, posMNAR and negMNAR linear missingness as defined in Section 4.2.1 on this transformed data. We use three transformations, namely square (sq), square root (sr) and log, and run CC, TARGET and JM. JM is run twice, once with a linear model of missingness and once with the restricted model of missingness described in Section 4.5. JM failed to converge for a few repetitions, mainly for the square transformation. Using the converged runs, the performance of these models in terms of the mean % bias of the parameters of the model of interest and the MSE is compared for the transformed and untransformed data in Figure 4.

We start by considering the JM with the linear model of missingness (blue symbols in Figure 4). For the transformed data our model of interest has an incorrect error distribution, and the addition of an adequate model of missingness reduces the MSE but the bias in the individual model of interest parameter estimates may not be removed or even reduced. For β_0, the bias is removed if the skewness from the transformation and the missingness are in the same direction (negMNAR missingness and square transform, or posMNAR missingness and log or square root transform, indicated by "S" label), but only reduced if the two sources of skewness are in conflict (indicated by "C" label). As regards β_1, if the two sources of skewness

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

are in the same direction JM increases the bias, but if they are in opposite directions then JM reduces the bias. If we choose an incorrect model of missingness for JM (red symbols in Figure 4), then there are only slight changes in the parameter estimates from CC, but they may result in a small deterioration in the fit of the model as measured by MSE.

This provides an explanation of the performance of JM with NCDSsim, which has a positively skewed response. When the gradient of the imposed missingness, ϕ_1, is positive we add negative skewness which is in conflict with the original skewness. From our findings from MVNsim we expect JM to reduce the bias in β_1, which is confirmed by Figure 3. For negative ϕ_1 the two sources of skewness are in the same direction and as expected the β_1 bias is generally increased.

The distributional skewness and the skewness attributable to informative non-response must be in the same direction for the bias in β_1 to be reduced. However, we have no way of verifying the size or direction of either skewness from the data.

4.7. What diagnostics are available?

For complete data, the Deviance Information Criterion (DIC) is widely used for Bayesian model comparison. With missing data, DIC can be constructed in different ways [23, 6, 24], and its use and interpretation are not straightforward. One option, is a conditional DIC, which treats the missing data as additional parameters [23]. WinBUGS automatically generates a conditional DIC, giving separate values for the model of interest and model of missingness. The model of interest values are based on the records with observed responses only. An alternative construction is based on the observed data likelihood, which differs from a conditional DIC in the model of missingness part, which is evaluated by integrating over the missing data

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

rather than by conditioning on it. Mason et al. [24] propose a strategy for comparing selection models by combining information from two measures taken from these different constructions of the DIC. A DIC based on the observed data likelihood is used to compare joint models with different models of interest but the same model of missingness, and a comparison of models with the same model of interest but different models of missingness is carried out using the model of missingness part of a conditional DIC. In this paper we focus on the measure of complexity, p_D, calculated for the model of missingness part of the conditional DIC and consider its interpretation as an indicator of departure from the MAR assumption in Bayesian selection models.

Spiegelhalter et al. [25] point out that p_D can be thought of as a measure of the ratio of the information in the likelihood about the parameters to the information in the posterior (likelihood plus prior). So for a model with uninformative prior distributions on all the parameters, all the information will come from the data and p_D can be interpreted as approximately the true number of parameters in the model. For a model with strong prior information about the parameters, p_D will be much smaller than the actual number of such parameters. We are particularly interested in possible interpretations of p_D for the model of missingness. In this case, the data are the missing value indicators, m_i, for which we have specified a Bernoulli likelihood (Equation 5), and the missing outcomes of interest, y_i, are treated as unknown parameters together with the regression coefficients $\boldsymbol{\theta}$. If the missingness mechanism is MAR, then by definition, m_i contains no information about y_i, and so p_D should simply reflect the information in the data about $\boldsymbol{\theta}$. However, if the mechanism is MNAR, we expect m_i to be informative about y_i (assuming a well-specified model), and hence p_D to be

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

higher. If we define scaled p_D as

$$\text{scaled } p_D = \frac{p_D - \text{number of coefficients}}{\text{number of missing observations}} \tag{9}$$

where the 'number of coefficients' is the dimension of $\boldsymbol{\theta}$, this allows us to assess the information being derived per missing observation without being influenced by the total number of missing observations. Note that, strictly speaking, Equation 9 only holds if we have uninformative priors on $\boldsymbol{\theta}$, since with informative priors, each coefficient will contribute less than 1 to p_D. Below, we carry out an empirical investigation of the estimation and interpretation of scaled p_D for the model of missingness.

For a Bernoulli model the deviance is given by

$$Deviance = -2\sum_{i=1}^{n} \left(m_i \log(p_i) + (1 - m_i)\log(1 - p_i)\right), \tag{10}$$

where m_i and p_i are as defined by Equation 5 if the 'link' is taken to be a logit function. The version calculated by WinBUGS uses plug-ins defined by the stochastic parameters in the likelihood, i.e. it calculates $\text{logit}(\hat{p}_i) = \hat{\theta}_0 + \hat{\theta}_1 \hat{y}_i$ (where $\hat{y}_i = y_i$ for $m_i = 0$) using the posterior means $\hat{\theta}_0 = E(\theta_0)$ and $\hat{\theta}_1 = E(\theta_1)$ as the plug-ins, assuming prior distributions were specified on θ_0 and θ_1. It is possible to get negative p_D values when the posterior distribution for a parameter is skewed, and we find that these plug-ins sometimes lead to negative p_Ds for the missingness model. To attempt to alleviate this problem, we have calculated the posterior means of the $\text{logit}(p_i)$ and used these as our plug-ins, which is the canonical parameterisation and tends to be more symmetric [25]. We now consider possible interpretations of the scaled p_D for the model of missingness.

4.7.1. Relationship between scaled p_D and the gradient of missingness, ϕ_1 Using NCDSsim, Figure 5 shows how scaled p_D increases as the magnitude of the gradient of missingness

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

increases, and that for similar gradients of missingness, scaled p_D tends to decrease as the percentage of missingness increases. In particular, when $\phi_1 = 0$ (i.e. missingness is MAR), scaled $p_D \approx 0$, whereas scaled $p_D > 0$ for models with informative missingness ($|\phi_i| > 0$). If we replace the gradient of missingness by the fitted slope of the model of missingness, θ_1, which is on the logit scale, we find a sharper version of the same relationship. This is consistent with findings from our MVNsim simulation (see the left plot in Figure 6, the right plot will be discussed later).

Figure 5 here

Figure 6 here

4.7.2. Relationship between p_D and reduction in MSE from CC to JM From the black crosses in Figure 7 we see that the percentage reduction in MSE from CC to JM increases as scaled p_D increases, so scaled p_D is correlated with the improvement in overall fit, as measured by MSE, from CC to JM. Since we have also seen that scaled p_D is correlated with the gradient of missingness, one interpretation is that this reflects the amount of information in the missingness model that can be used to improve the fit of the model of interest. The further our missing data is from MAR, the higher scaled p_D tends to be, and the greater the potential for extracting information from the joint model.

Figure 7 here

The purple, blue, green and red circles in Figure 7 show the mean model of missingness scaled p_D against the mean % reduction in MSE from CC to JM taken over the 50 replicates from the MVN simulations with 0.25, 0.5, 0.75 and 0.9 correlation respectively. There are two points for each simulation, one for posMNAR and one for negMNAR, which are always close together. The mean percentage of missingness for these simulations is 50%, and so they have been

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

placed in both the 25-50% missing and 50-75% missing panels, where they generally reinforce the pattern seen with NCDSsim. The exception is the simulations with 0.25 correlation, which have higher scaled p_D for the level of MSE reduction than seen with the NCDS simulations. So there is some evidence that the relationship between scaled p_D and reduction in MSE from CC to JM is affected by the strength of the correlation between response and covariate.

Further simulations, MVN(n100) and MVN(n10000), suggest that sample size also affects the relationship, despite having attempted to adjust for sample size by scaling (see triangles in Figure 7). The means for the MVN simulations using log, square and square root transforms of the response (shown as brown, pink and light green squares) are positioned within or close to the black crosses, which suggests that the relationship is not affected by transforming the response. As a final experiment, the NCDS simulation was rerun with the responses artificially dichotomised and a logistic regression model of interest fitted, and an equivalent plot to Figure 7 shows similar shape and variable range. This provides some evidence that the relationship is robust to the choice of model of interest.

To summarise, this research suggests that Figure 7 is not data or model specific and provides some idea of the magnitude of scaled p_D in certain circumstances, although the number of data points, percentage of missing data and strength of the relationship of the model of interest all have some effect. When scaled p_D is close to zero this is consistent with the data being MAR, and we expect that a joint model will not change the fit of our model of interest very much, but if it is bigger than about 0.1 then we expect the joint model to make a substantial difference.

4.7.3. Relationship between scaled p_D and the change in β_1 between JM and CC For MVNsim, the right graph in Figure 6 plots the model of missingness scaled p_D against the difference in

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

the β_1 estimates between JM and CC. This provides evidence that scaled p_D is indicative of the size of the change of the slope parameter estimate in our model of interest.

The left plot in Figure 8 shows a similar relationship for NCDSsim. The right plot suggests that scaled p_D is also indicative of the size of the change in the % bias of β_1. However, it tells us nothing about the direction of this change, JM β_1 could be closer or further away from the TARGET β_1 than the CC β_1. This ties in with our findings in Section 4.6, as positive ϕ_1 adds skewness in the opposite direction to the original skewness, thus reducing the β_1 bias (red circles), while for negative ϕ_1 both sources of skewness are in the same direction, so increasing the β_1 bias (blue crosses). So p_D is not helpful in determining whether the estimation of the model of interest slope parameter has improved, but can be interpreted as an indicator of the magnitude of effect of adding a missingness model on its estimation.

Figure 8 here

5. APPLICATIONS

We now apply our Bayesian joint models in a more realistic setting, again assuming that the missingness mechanism is non-ignorable, using two real data examples.

5.1. NCDSreal example

For NCDSreal, we consider a model of interest with multiple covariates, using mathematics test score at 7 and social class at age 11, in addition to the mathematics test score at 11 that we have been using in our NCDS simulations. This is one of the models used by Goldstein [15]. Following Goldstein, we aggregate social class into three groupings: non-manual workers (social classes I, II and III non-manual), skilled and semi-skilled manual workers (social classes

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

III manual and IV) and unskilled manual workers (social class V). As we are focussing on the impact of missing response, we use only the records from the full NCDS dataset in which all the covariates are fully observed, leaving 10,944 records for JM, of which 25% must be discarded for fitting CC. Our model of missingness is the linear logit model specified by Equation 5, which has no additional covariates.

The investigation of NCDSsim suggests that the response should be transformed. However, choosing a transform is difficult because using only the observed data requires making assumptions about the missing data that cannot be justified from the data at hand. A possible approach is to carry out a sensitivity analysis to explore the impact of using different transforms. So, we run CC and JM five times, with the response transformed according to a Box Cox power transformation [26], i.e.

$$y = \begin{cases} \frac{(y+\lambda_2)^{\lambda_1}-1}{\lambda_1} & : \quad \lambda_1 \neq 0 \\ \log(y+\lambda_2) & : \quad \lambda_1 = 0 \end{cases} \tag{11}$$

with λ_2 set to 2 to ensure that $y + \lambda_2$ is always positive and λ_1 taking values of 0 to 1 at 0.25 intervals. The observed data suggests that the response is normalised when λ_1 is a half. The parameter estimates change monotonically as λ_1 changes, and so the results for only three of the runs (no transform, square root transform and log transform) are shown in Table I.

Table I here

The addition of a missingness model results in a small decrease in the constant parameter, but the other parameter estimates for the model of interest are very similar for CC and JM, regardless of the transformation of the response. The θ_1 parameter from the model of missingness provides evidence that lower test scores are more likely to be missing, which intuitively seems reasonable. We might also interpret the θ_1 estimates as evidence

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

of informative missingness, but scaled p_D for the model of missingness is 0.006 for the joint models run with and without transforming the response, which contradicts this. The explanation lies in the high correlation between the covariates and response in this dataset. In this longitudinal example, age 11 score is a good proxy for age 16 score (0.77 correlation), so we could alternatively have fit a MAR model of missingness, using the age 11 score as the regressor. The model of interest can now be estimated separately from the model of missingness, and in this case the model of missingness p_D will just be the number of $\boldsymbol{\theta}$ parameters and hence scaled p_D should be 0 by definition.

5.2. HAMD example

In our second example, using the clinical trial data described in Section 3, exploratory plots indicate a downwards trend in the HAMD score over time. So for our model of interest, we follow DK and regress HAMD against time, allowing a quadratic relationship and a different intercept for each centre, s.t.

$$\begin{aligned} y_{iw} &= \mu_{iw} + \delta_{iw} \\ \mu_{iw} &= \beta_{c(i)} + \eta_{t(i)} w + \xi_{t(i)} w^2 \end{aligned} \tag{12}$$

where i=individual, t=treatment (1,...,3), c=centre (1,...,6) and w=week (0,...,4). $c(i)$ and $t(i)$ denote the centre and treatment of individual i respectively. The δ_{iw}s follow a second-order autoregressive process defined by

$$\begin{aligned} &\delta_{i0} = \epsilon_{i0}; \;\; \delta_{i1} = \alpha_1 \delta_{i0} + \epsilon_{i1}; \;\; \delta_{iw} = \alpha_1 \delta_{i(w-1)} + \alpha_2 \delta_{i(w-2)} + \epsilon_{iw}, \;\; w \geq 2 \\ &\epsilon_{iw} \sim N(0, \sigma^2). \end{aligned} \tag{13}$$

We assign vague priors to the unknown parameters: giving the regression coefficients N(0,10000) priors and the precision ($\frac{1}{\sigma^2}$) a Gamma(0.001,0.001) prior.

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

We specify our model of missingness to be

$$logit(p_{iw}) = \theta_0 + \theta_1 y_{i(w-1)} + \theta_2 (y_{iw} - y_{i(w-1)}) \tag{14}$$

and assign a logistic prior to θ_0 and weakly informative Normal priors to θ_1 and θ_2 as discussed in Section 2. This form of the logit allows dependence on the previous week's HAMD score, i.e. the severity of the subject's depression, and the change in the HAMD score, which reflects the successfulness of the treatment. CC and JM are run for 110,000 iterations with 100,000 burn-in, and the parameter estimates are shown in Table II. The downwards impact of the addition of the missingness model can be seen from the mean response profiles for CC (solid lines) and JM (dashed lines) shown in Figure 9.

Table II here

Figure 9 here

Turning our attention to the model of missingness, we find the θ_1 estimate is close to zero suggesting that the level of the HAMD score is not highly associated with drop-out. However, the negative θ_2 estimate indicates that change in the HAMD score is informative with individuals more likely to drop-out if their HAMD score goes down, i.e. their treatment is successful. From our previous investigation, we interpret the model of missingness scaled p_D of 0.11 as providing evidence of informative missingness.

Allowing for informative missingness using Equation 14 affects prediction of HAMD scores, but not conclusions about differences in treatments. However, by adjusting our model of missingness to incorporate separate $\boldsymbol{\theta}$ for each treatment, we allow treatment to directly affect the missingness process which is more likely to impact these conclusions. To investigate this, we run a joint model with separate $\boldsymbol{\theta}$ for each treatment which we shall denote by JM* and show as dotted lines in Figure 9. We now get a higher model of missingness scaled p_D, 0.28, and

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

increased differences between treatments. A comparison of these joint models with alternative selection models using different measures of DIC is given by Mason et al. [24].

As a sensitivity analysis, we rerun CC and JM, using power ($y^{1.5}$), square root and log transformations of the HAMD scores. There is greater evidence for informative missingness with the power transformation (scaled $p_D = 0.41$) but less for the square root and log transformations (0.04 and 0.02 scaled p_D respectively). In this case, analysis (not shown) of the HAMD scores in weeks 0 and 1, which are fully observed, can help with the choice of an appropriate transformation of the response, and suggests no transform is needed. However, this demonstrates how our conclusions about informative missingness can be affected by our choice of transform.

6. DISCUSSION

Our simulation studies have shown that adding the *correct* model of missingness to a model of interest specified with the *correct* error distribution, will successfully remove the bias in the parameter estimates of the model of interest and improve the overall fit of the model of interest as measured by MSE. We have found that the joint model still gives an improvement even if the relationship of interest is relatively weak.

However, as shown by Kenward [22], selection models can be sensitive to the correct specification of both parts of the joint model, and assume that the same model structure is appropriate for both observed and missing individuals. Unfortunately these assumptions are not testable from the data, so we have examined the consequences of getting these assumptions wrong.

If we specify an incorrect model of missingness, then little further harm is done to the

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

fit of the model of interest (compared to complete case analysis), but the potential benefits from using a joint model are reduced or lost, as demonstrated by our use of restricted linear missingness models for MVNsim. Although the specification of a quadratic missingness model seems attractive, given that it encompasses the linear model, the added complexity results in greater difficulty in achieving convergence and a small deterioration in the model of interest parameter estimation and overall fit compared to the correct model.

By contrast, the effects of misspecifying the error distribution of the model of interest are much less predictable. Indeed, the shape of the error distribution is crucially used by JM to fill in the missing responses in an attempt to reproduce the distributional shape specified for the response. In this case, there are two sources of skewness in our model, (i) attributable to skewness in the responses after adjusting for covariates and (ii) resulting from the missingness mechanism. As Skinner comments in discussion of Diggle and Kenward [16], disentangling informative non-response and distributional skewness is difficult. The bias in the β_0 parameter is still mostly removed and the MSE reduced with a joint model even with a misspecified error distribution for the model of interest. However, the estimation of β_1 is not so robust and the behaviour of the joint model depends on whether or not the two sources of skewness are in conflict. If they are in conflict, we get a reduction rather than removal of the bias, but if they are in the same direction the bias is greater than in a complete case analysis. This suggests that joint models need careful interpretation if our primary concern is the estimation of the relationship between the response and a particular covariate rather than predicting the response based on several covariates.

The scaled p_D in the model of missingness can be used to get some idea of how far the missing data departs from MAR given that the other assumptions are correct, but is also

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

affected by the size of the dataset, the proportion of missing data and the strength of the relationship in the model of interest. However, higher values should not necessarily be taken as an indication of 'better' model of interest parameter estimates, but of the magnitude of effect of the missingness model on their estimation.

In the studies described, we have imposed or assumed a linear missingness pattern on the response, but other patterns of missingness exist and may have a different impact. For example, if all of the responses above or below a certain threshold are missing, then the β_1 bias from complete-case analysis is potentially much more serious. There is no certainty that our findings will continue to hold in these circumstances.

Given the uncertainties, it is clear that sensitivity analysis is crucial to see how conclusions are affected by varying the key assumptions relating to both the model of interest and the model of missingness, in particular the choice of transform for the response and the form of the missingness model. External information would be very useful for informing these choices. A Bayesian framework has the flexibility to carry out necessary sensitivities relatively easily, and also offers the possibility of incorporating external information or expert knowledge via prior distributions, and in the case of informative missing responses, it is clear that seeking to include prior knowledge will carry great benefits.

ACKNOWLEDGEMENTS

Financial support: this work was supported by an ESRC PhD studentship (Alexina Mason). Nicky Best and Sylvia Richardson would like to acknowledge support from ESRC: RES-576-25-5003 and RES-576-25-0015. The authors are grateful to Mike Kenward for useful discussions and providing the clinical trial data analysed in this paper.

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

REFERENCES

1. Schafer JL, Graham JW. Missing Data: Our View of the State of the Art. *Psychological Methods* 2002; **7**(2):147–177.
2. Ibrahim JG, Chen MH, Lipsitz SR, Herring AH. Missing-Data Methods for Generalized Linear Models: A Comparative Review. *Journal of the American Statistical Association* 2005; **100**(469):332–346.
3. Schafer JL. *Analysis of Incomplete Multivariate Data*. 1st edn., Chapman & Hall, 1997.
4. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. 2nd edn., John Wiley and Sons, 2002.
5. Molenberghs G, Kenward MG. *Missing Data in Clinical Studies*. 1st edn., John Wiley and Sons, 2007.
6. Daniels MJ, Hogan JW. *Missing Data In Longitudinal Studies Strategies for Bayesian Modeling and Sensitivity Analysis*. Chapman & Hall, 2008.
7. Rubin DB. Inference and Missing Data. *Biometrika* 1976; **63**(3):581–592.
8. Gilks WR, Richardson S, Spiegelhalter DJ. *Markov Chain Monte Carlo in Practice*. 1st edn., Chapman & Hall, 1996.
9. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*, chap. 11. 2nd edn., Chapman & Hall, 2004; 283–309.
10. Best NG, Spiegelhalter DJ, Thomas A, Brayne CEG. Bayesian Analysis of Realistically Complex Models. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* 1996; **159**(2):323–342.
11. Carpenter J, Pocock S, Lamm CJ. Coping with missing data in clinical trials: A model-based approach applied to asthma trials. *Statistics in Medicine* 2002; **21**:1043–1066.
12. Wakefield J. Ecological inference for 2 x 2 tables. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* 2004; **167**(3):385–445.
13. Jackson C, Best N, Richardson S. Improving ecological inference using individual-level data. *Statistics in Medicine* 2006; **25**:2136–2159.
14. Plewis I, Calderwood L, Hawkes D, Nathan G. National Child Development Study and 1970 British Cohort Study Technical Report: Changes in the NCDS and BCS70 Populations and Samples over Time. *Technical report, 1st edition*, Institute of Education, University of London 2004.
15. Goldstein H. Some Models for Analysing Longitudinal Data on Educational Attainment (with discussion). *Journal of the Royal Statistical Society, Series A (General)* 1979; **142**(4):407–442.
16. Diggle P, Kenward MG. Informative Drop-out in Longitudinal Data Analysis (with discussion). *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 1994; **43**(1):49–93.

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

17. Yun SC, Lee Y, Kenward MG. Using hierarchical likelihood for missing data problems. *Biometrika* 2007; **94**(4):905–919.
18. Gilks WR, Richardson S, Spiegelhalter DJ. *Markov Chain Monte Carlo in Practice*, chap. 6. 1st edn., Chapman & Hall, 1996; 89–114.
19. Brooks S, Gelman A. General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics* 1998; **7**:434–455.
20. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS – A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing* 2000; **10**:325–337.
21. Collett D. *Modelling Binary Data*. 2nd edn., Chapman & Hall, 2003.
22. Kenward MG. Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. *Statistics in Medicine* 1998; **17**:2723–2732.
23. Celeux G, Forbes F, Robert CP, Titterington DM. Deviance Information Criteria for Missing Data Models. *Bayesian Analysis* 2006; **1**(4):651–674.
24. Mason A, Richardson S, Best N. Using DIC to compare selection models with non-ignorable missing responses. *Technical report*, Imperial College London 2009. Available at `www.bias-project.org.uk`.
25. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 2002; **64**(4):583–639.
26. Box GEP, Cox DR. An Analysis of Transformations. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 1964; **26**(2):211–252.

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

Figure 1. NCDS mathematics test scores (subset with observed values of both scores)

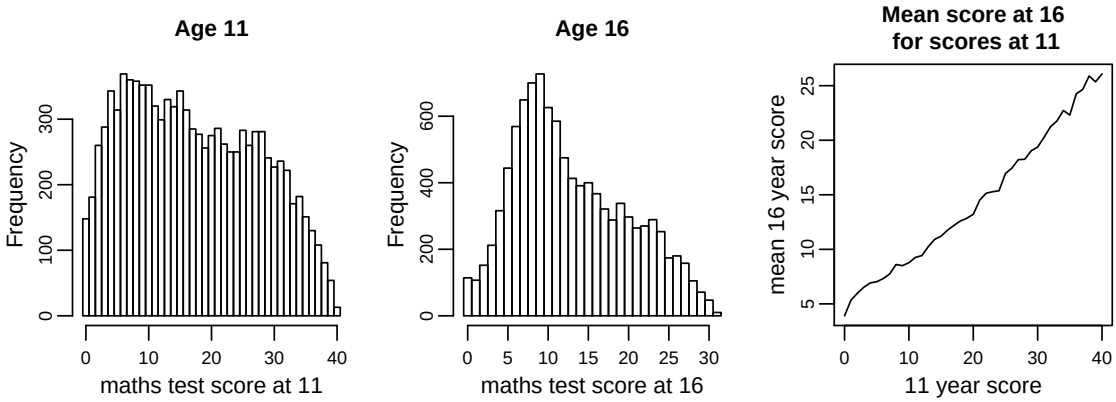
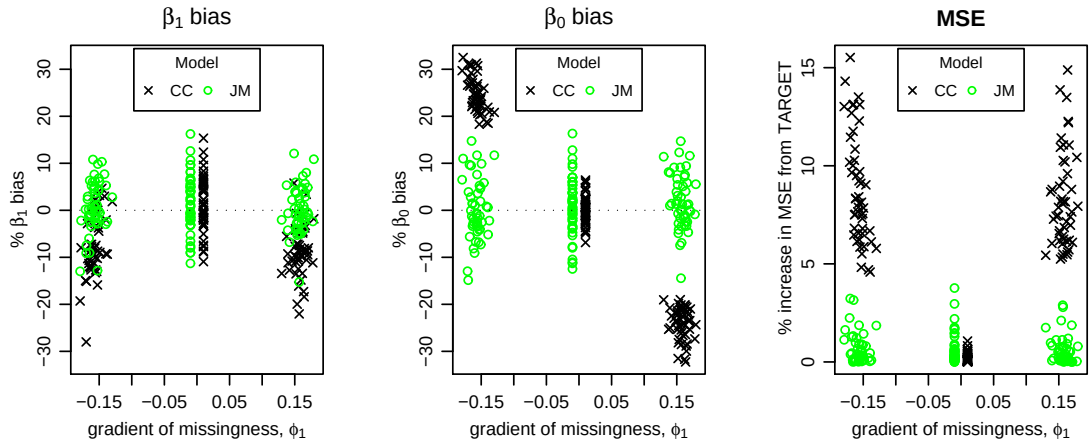


Figure 2. MVNsim data: performance of CC and JM compared with the TARGET generated targets

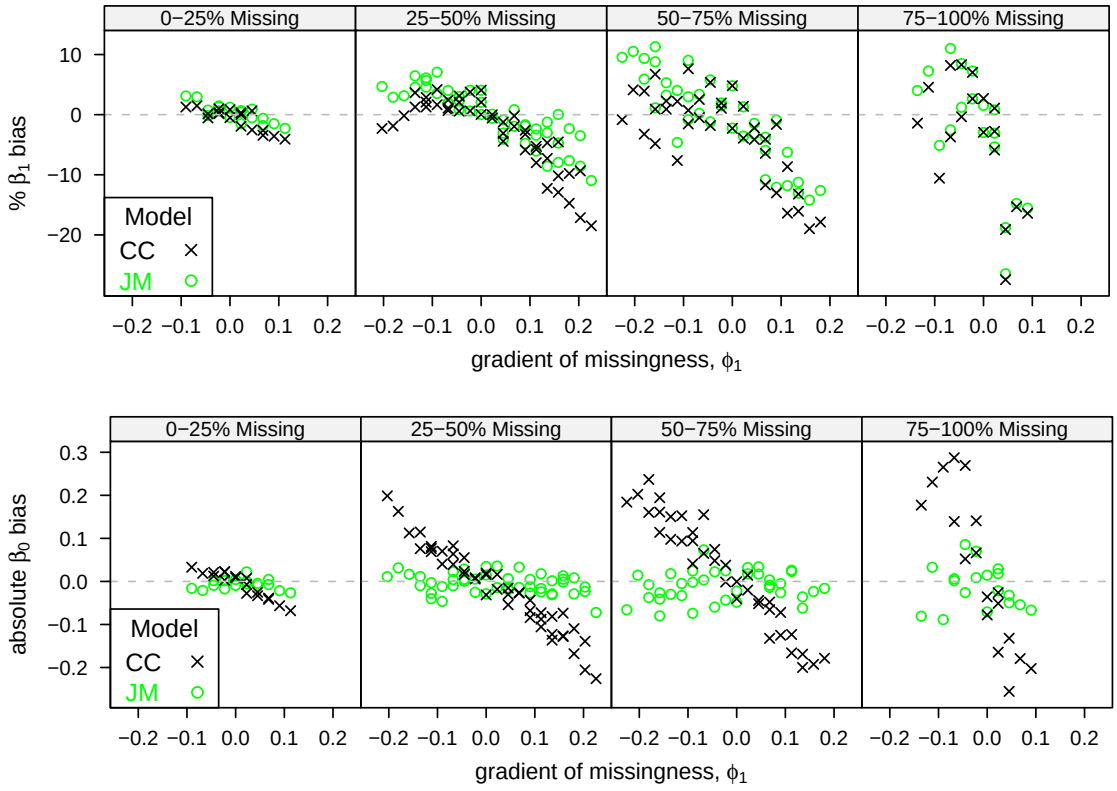


The points grouped on the left of each graph correspond to the 50 negMNAR runs ($-0.18 \leq \phi_1 \leq -0.13$), the points in the middle correspond to the 50 MCAR runs (ϕ_1=0, but CC and JM offset for clarity) and the points grouped on the right correspond to the 50 posMNAR runs ($0.13 \leq \phi_1 \leq 0.18$).

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

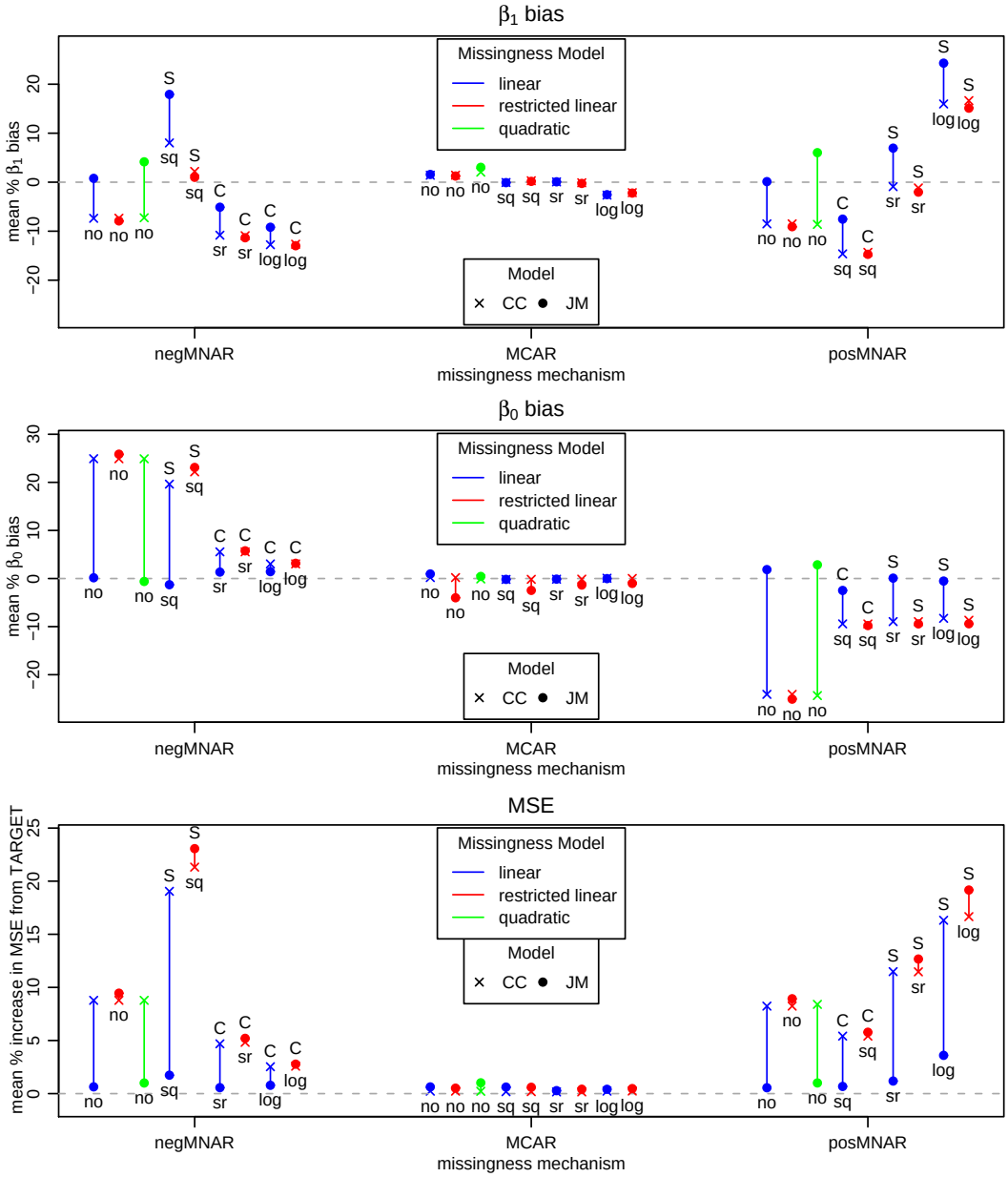
Figure 3. NCDSsim data: performance of CC and JM compared with the TARGET generated targets



Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

Figure 4. MVNsim data: impact of different transforms and missingness models



(1) The set of repetitions varies, as only converged runs are included. (2) In the Missingness Model legend, restricted linear indicates θ_1 is restricted to positive values for negMNAR and negative values for posMNAR. (3) The transformation used is indicated beneath each pair of points (no=none, sq=square, sr=square root and log=log). (4) A letter above a pair of points indicates whether the skewness from the transformation and the missingness are in the same direction (S) or in conflict (C). (5) The length of the line joining a dot and cross indicates the size of the change in the mean % β_i bias (top two plots) or the change in the increase in MSE from TARGET (bottom plot). If the dot is closer to the zero line than the cross, then JM performs better than CC for the plotted measure. Our target is for the dot to lie on the zero line.

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

Figure 5. NCDSsim data: the relationship of Scaled p_D with the gradient of missingness

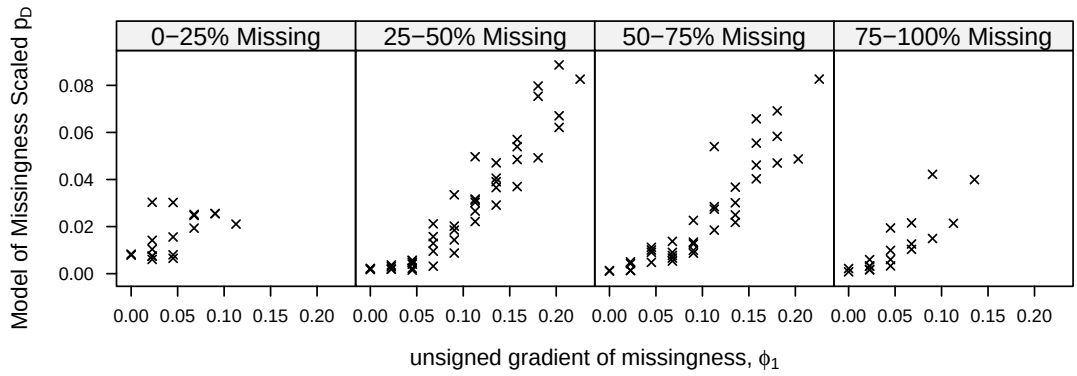
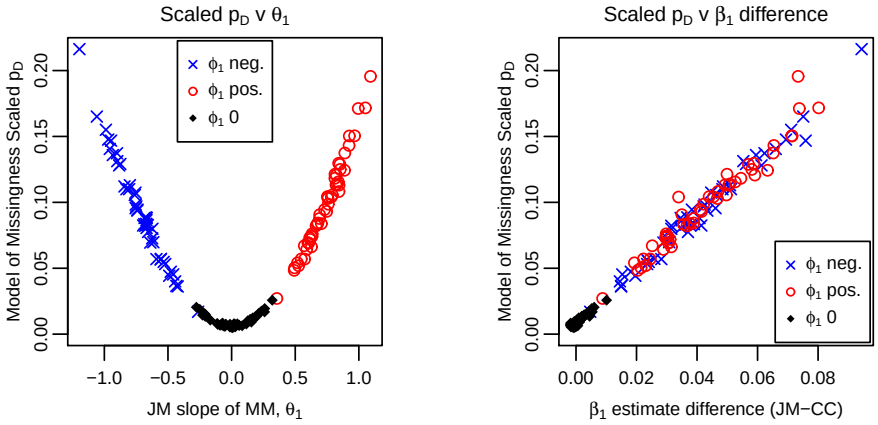


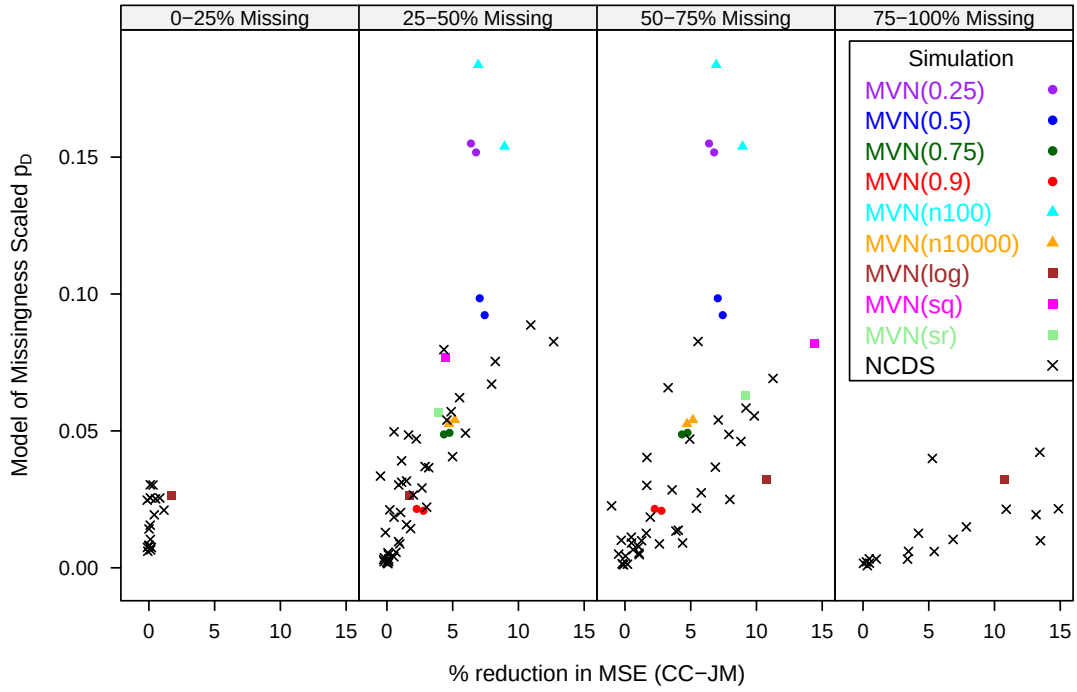
Figure 6. MVNsim data: the relationship of Scaled p_D with β_1 and θ_1



Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

Figure 7. The relationship of Scaled p_D to the % reduction in MSE from CC to JM

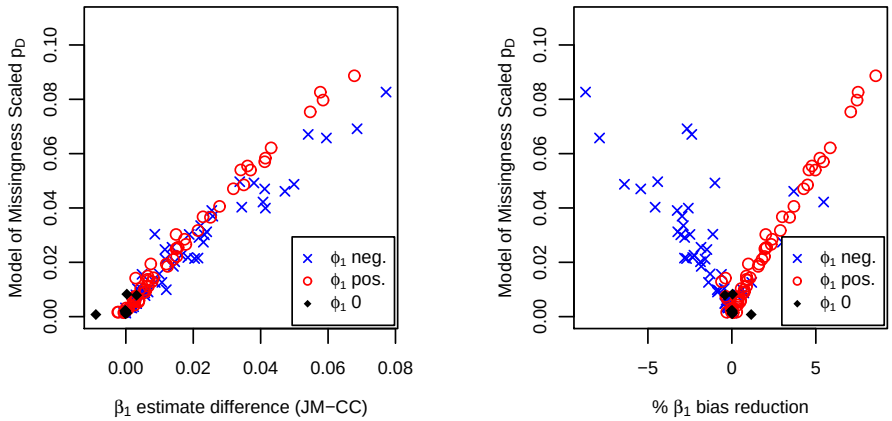


The MVN points are means across multiple repetitions and are shown in two panels if they are close to a boundary (all simulations except MVN(sq) and MVN(sr)). MVN(0.25), MVN(0.5), MVN(0.75) and MVN(0.9) are all simulations with 1000 records and the correlation shown in brackets; MVN(n100) and MVN(n10000) are both simulations with correlation 0.5, but have 100 and 10,000 records respectively; MVN(log), MVN(sq) and MVN(sr) are simulations with correlation 0.5, but the response was transformed using the log, square and square root transform respectively. Two points are shown for each MVN simulation, one for posMNAR and one for negMNAR. All unconverged runs are excluded from the calculations.

Copyright © 2000 John Wiley & Sons, Ltd.

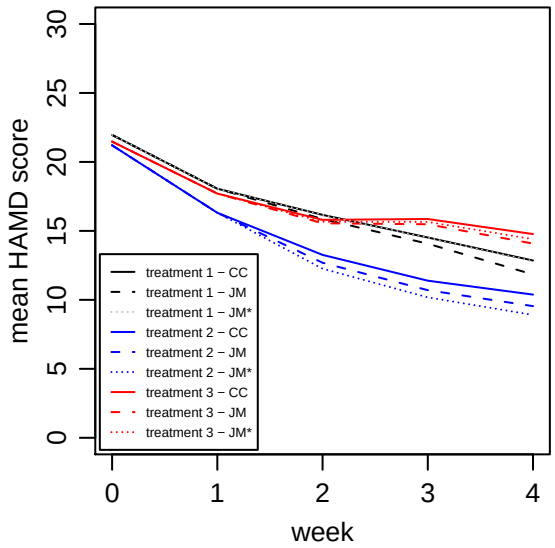
Prepared using simauth.cls

Figure 8. NCDSsim data: the relationship of Scaled p_D to the slope in the model of interest



% bias reduction is the absolute percentage reduction in bias from CC to JM, and is calculated as % bias reduction $= abs(CC$ % bias$) - abs(JM$ % bias). Positive numbers indicate that JM is doing better in terms of bias than CC, while negative numbers indicate that JM is doing worse.

Figure 9. HAMD example: modelled mean response profiles



JM* has a separate model of missingness for each treatment.
The CC and JM* lines for treatment 1 are almost coincident.

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

Table I. NCDSreal: Model of Interest (MoI) and Model of Missingness (MoM) parameter estimates

Parameter	Box-Cox λ_1	CC		JM		% diff[a]
β_0	0	0.60	(0.58,0.62)	0.56	(0.54,0.58)	-6.6
MoI	0.5	0.82	(0.80,0.84)	0.77	(0.75,0.79)	-5.7
intercept	1	1.12	(1.10,1.15)	1.06	(1.04,1.09)	-5.6
β_1	0	0.38	(0.37,0.40)	0.39	(0.38,0.40)	1.4
MoI slope	0.5	0.50	(0.49,0.51)	0.51	(0.49,0.52)	1.2
(for age 11 test score)	1	0.69	(0.68,0.71)	0.70	(0.68,0.72)	1.2
β_2	0	0.05	(0.04,0.06)	0.05	(0.04,0.06)	-0.1
MoI slope	0.5	0.06	(0.05,0.07)	0.06	(0.05,0.07)	-0.4
(for age 7 test score)	1	0.07	(0.06,0.09)	0.07	(0.06,0.09)	-0.6
β_3	0	-0.09	(-0.11,-0.07)	-0.09	(-0.11,-0.07)	-0.1
MoI slope	0.5	-0.13	(-0.15,-0.10)	-0.13	(-0.15,-0.10)	-0.6
(social class skilled & semi-skilled)	1	-0.18	(-0.22,-0.15)	-0.18	(-0.22,-0.15)	-0.3
β_4	0	-0.15	(-0.19,-0.11)	-0.16	(-0.20,-0.11)	3.5
MoI slope	0.5	-0.18	(-0.23,-0.14)	-0.19	(-0.24,-0.14)	2.8
(for social class unskilled)	1	-0.24	(-0.30,-0.17)	-0.24	(-0.31,-0.18)	3.2
θ_0	0			-0.90	(-0.96,-0.85)	
MoM	0.5			-0.87	(-0.93,-0.81)	
intercept	1			-0.86	(-0.93,-0.80)	
θ_1	0			-0.41	(-0.50,-0.32)	
MoM slope	0.5			-0.35	(-0.43,-0.27)	
(for response)	1			-0.26	(-0.32,-0.20)	

Table shows the posterior mean, with the 95% interval in brackets.
[a] % difference in parameter estimate from CC to JM.

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls

Table II. HAMD example: parameter estimates

Parameter[a]	CC		JM		% diff[b]
β_1	21.73	(20.64,22.90)	21.76	(20.61,22.88)	0.1
β_2	22.46	(21.45,23.48)	22.42	(21.35,23.52)	-0.2
β_3	19.36	(18.36,20.38)	19.42	(18.39,20.45)	0.3
β_4	23.94	(22.90,25.01)	24.02	(22.88,25.18)	0.3
β_5	20.70	(19.67,21.75)	20.78	(19.72,21.88)	0.4
β_6	20.81	(19.77,21.89)	20.71	(19.69,21.81)	-0.5
η_1	-3.50	(-4.31,-2.64)	-3.45	(-4.27,-2.62)	-1.5
η_2	-5.31	(-6.18,-4.51)	-5.56	(-6.45,-4.67)	4.6
η_3	-3.71	(-4.53,-2.91)	-3.71	(-4.51,-2.92)	0.1
ξ_1	0.33	(0.12,0.53)	0.26	(0.03,0.47)	-22.4
ξ_2	0.65	(0.44,0.85)	0.65	(0.45,0.86)	0.8
ξ_3	0.52	(0.32,0.72)	0.48	(0.28,0.68)	-7.8
θ_0			-3.19	(-3.80,-2.62)	
θ_1			0.04	(0.00,0.09)	
θ_2			-0.14	(-0.27,-0.02)	

Table shows the posterior mean, with the 95% interval in brackets.
[a] as specified by Equations 12 and 14.
[b] % difference in parameter estimate from CC to JM.

Copyright © 2000 John Wiley & Sons, Ltd.

Prepared using simauth.cls