

Paper 436-2013

# Missing No More: Using the MCMC Procedure to Model Missing Data

Fang Chen, SAS Institute Inc.

## ABSTRACT

Missing data are often a problem in statistical modeling. The Bayesian paradigm offers a natural model-based solution for this problem by treating missing values as random variables and estimating their posterior distributions. This paper reviews the Bayesian approach and describes how the MCMC procedure implements it.

Beginning with SAS/STAT<sup>®</sup> 12.1, PROC MCMC automatically samples all missing values and incorporates them in the Markov chain for the parameters. You can use PROC MCMC to handle various types of missing data, including data that are missing at random (MAR) and missing not at random (MNAR). PROC MCMC can also perform joint modeling of missing responses and covariates.

## INTRODUCTION

Missing data problems arise frequently in practice and are caused by many circumstances. For example, study subjects might fail to answer questions on a questionnaire, data can be lost, covariate measurements might be unavailable, a survey might not receive enough responses, and so on. The impact of missing data on statistical inference is potentially important, especially in cases where the subjects that have missing data differ systematically from those that have complete data. Coherent estimation and valid inference require adequate modeling of the missing values; simply discarding the missing data can lead to biased results.

Traditionally there are a number of approaches to tackle the missing data problem. The expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977), is a general iterative algorithm that can be used to find the maximum likelihood estimates (MLEs) in missing data problems. The algorithm is most useful when maximization from the complete data likelihood is straightforward and maximization based on the observed data likelihood is difficult. The multiple imputation (MI) algorithm (Rubin 1987) constructs multiple complete data sets by filling each missing datum with plausible values, and then it obtains parameter estimates by averaging over multiple data sets. Proper imputation leads to valid large sample inferences and yields estimators with good large-sample properties (Little and Rubin 2002, Chapter 10). There are also semiparametric approaches, such as weighted estimating equations (WEE) that do not rely on distributional assumptions of the missing values (Robins, Rotnitzky, and Zhao 1994; Lipsitz, Ibrahim, and Zhao 1999). They are computationally efficient and robust, and they can yield estimates that are consistent in more relaxed settings. For a more comprehensive treatment of missing data analysis, see Little and Rubin (2002).

The Bayesian paradigm offers an alternative model-based solution, in which the missing values are treated as unknown parameters and are estimated accordingly. From an estimation point of view, introducing additional parameters adds limited complexity to the problem, because these missing value parameters are simply an additional layer of variables that can be sampled sequentially in a Markov chain Monte Carlo (MCMC) simulation. As a result, you can obtain the posterior distributions of the incomplete data given the observed data (for example, for prediction purposes). More importantly, the Bayesian approach takes into account the uncertainty about the missing values and enables you to estimate the posterior marginal distributions of the parameters of interest conditional on observed (and partially observed) data. The Bayesian approach offers a principled way of handling missing data that is also flexible; it uses all available information in the analysis and can also satisfy different modeling requirements as a result of various types of missingness assumptions that you might have about the data.

This paper is organized as follows. The section “[NOTATION](#)” introduces notation that is used throughout the paper. The section “[TYPES OF MISSING DATA MODELS](#)” covers the definitions of the most commonly used missing data models. The section “[MISSING DATA ANALYSIS IN PROC MCMC](#)” discusses the enhancements to PROC MCMC in SAS/STAT 12.1 to handle missing data. The section “[EXAMPLES](#)” shows three missing data analysis examples: a bivariate normal model with partial missing data, an air pollution data analysis with missing covariates, and a selection model approach to model nonignorable missing data.

## NOTATION

Let  $\mathbf{Y} = (y_1, \dots, y_n)$ , where  $\mathbf{Y}$  denotes the response variable, which can be multidimensional.  $\mathbf{Y}$  consists of two portions,

$$\mathbf{Y} = \{\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}\}$$

where  $\mathbf{Y}_{\text{obs}}$  denotes the observed data and  $\mathbf{Y}_{\text{mis}}$  denotes the missing values. In a SAS data set, a missing value is usually coded by using a single dot (.).

The sampling distribution that governs the missing response variable is assumed to have the generic form,

$$y_i \sim f(y_i | x_i, \theta)$$

where  $f(\cdot)$  is a known distribution (which is the usual likelihood in your model),  $x_i$  are the covariates, and  $\theta$  is the parameter of interest, which could be multidimensional.

Let  $\mathbf{R}_Y = (r_1, \dots, r_n)$  be the missing value indicator, also called the missingness random variable, for any missing response variable, where  $r_{y_i} = 1$  if  $y_i$  is missing and  $r_{y_i} = 0$  otherwise.  $\mathbf{R}$  is known when the  $\mathbf{Y}$  are known.

Covariates are denoted by  $\mathbf{X} = (x_1, \dots, x_n)$ , which can also contain missing values:

$$\mathbf{X} = \{\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}}\}$$

Covariates are typically considered to be fixed constants. But in analysis that involves missing covariate data,  $\mathbf{X}$  becomes a random variable over which you must specify a probability distribution. This probability distribution, which is similar to the likelihood function of the response variable, gives rise to the values of the covariate,

$$x_i \sim \pi(x_i | u_i, \gamma)$$

where  $u_i$  can be another set of covariates and  $\gamma$  can be a parameter vector that can overlap with  $\theta$ .

Similar to  $\mathbf{R}_Y$ , you have a missing value indicator,  $\mathbf{R}_X$ , for missing covariate data. In practice, you construct an  $\mathbf{R}$  for each variable that has missing values. To simplify notation in the remainder of the paper,  $\mathbf{R}$  refers to either a response or covariate missing value indicator. The missing mechanism refers to the probability of observing a missing value; it is a statistical model for  $\mathbf{R}$ .

## TYPES OF MISSING DATA MODELS

Generally speaking, there are three types of missing data models (Rubin 1976). This section reviews the definitions.

- Data are said to be *missing completely at random* (MCAR) if the probability of a missing value is independent of any observation in the data set. This approach assumes that both the observed and unobserved data are random samples from the same data-generating mechanism. In data that have only missing response values, MCAR assumes that the probability of observing a missing  $y_i$  is independent of other  $y_j$ , for  $j \neq i$ , and is independent of other covariates  $x_i$ . For data that have missing covariates, you must create a model for the covariate. An MCAR analysis assumes that missingness is independent of any unobserved data (response and covariate), conditional on observed values of the covariate. Under the MCAR assumption, you can discard any missing observations and proceed with the analysis by using only the observed data. This type of analysis is called a complete-case (CC) analysis. If the assumption of MCAR does not hold, then a CC analysis is biased. If the MCAR assumption holds, a CC analysis is less efficient than an analysis that uses the full data. The CC analysis is still unbiased as long as any covariates that are required for MCAR are included in the analysis.
- The *missing at random* (MAR) approach assumes that the missing observations are no longer random samples that are generated from the same sampling distribution as the observed values. Hence the missing values must be modeled. Specifically, for data that have only missing response values, an MAR

analysis assumes that the probability of a missing value can depend on some observed quantities but does not depend on any unobserved quantities. In the MAR approach, you can model the probability of observing a missing  $y_i$  by using  $x_i$ , but the probability is independent of the unobserved data value (which would be the actual  $y_i$  value). When missing covariate data are present, MAR assumes that missingness is independent of unobserved data, conditional on both observed and modeled covariate data and on observed response data.

MAR assumes that responses that have similar observed characteristics (covariates  $x_i$ , for example) are comparable and that the missing values are independent of any unobserved quantities. This also implies that the missing mechanism can be ignored and does not need to be taken into account as part of the modeling process. MAR is sometimes referred to as *ignorable missing*; it is not the missing values but the missing mechanism that can be ignored.

- The most general and most complex missing data scenario is *missing not at random* (MNAR). In an MNAR model, the probability that a value is missing can depend not only on other observed quantities but also on the unobserved observations (the would-have-been values). The missing mechanism is no longer ignorable, and a model for  $\mathbf{R}$  is required in order to make correct inferences about the model parameters. MNAR is sometimes referred to as *nonignorable missing*.

In an MNAR model, you specify a joint likelihood function over  $\mathbf{R}$  and  $\mathbf{Y}$ :  $f_{\mathbf{R},\mathbf{Y}}(r_i, y_i | x_i, \theta)$ . This joint distribution can be factored in two ways: as a selection model and as a pattern-mixture model.

- The *selection model* (Rubin 1976; Heckman 1976; Diggle and Kenward 1994) factors the joint distribution  $r_i$  and  $y_i$  into a marginal distribution for  $y_i$  and a conditional distribution for  $r_i$ ,

$$f(r_i, y_i | x_i, \theta) \propto f(y_i | x_i, \alpha) \cdot f(r_i | y_i, x_i, \beta)$$

where  $\theta = (\alpha, \beta)$ ,  $f(r_i | y_i, x_i, \alpha)$ , is usually a binary model with a logit or probit link that involves regression parameters  $\alpha$ , and  $f(y_i | x_i, \beta)$ , also known as the *outcome model* or the *response model*, is the sampling distribution that generates  $y_i$  with model parameters  $\beta$ .

To some, the selection model is a natural way of decomposing the joint distribution, with one marginal model for the response variable and one conditional model that describes the missing mechanism. In addition, MAR analysis does not require the consideration of the conditional model on  $\mathbf{R}$ . When MNAR analysis is required, adding the conditional model is an easy and natural extension.

- The *pattern-mixture model* (Glynn, Laird, and Rubin 1986; Little 1993) factors the opposite way: as a marginal distribution for  $r_i$  and a conditional distribution for  $y_i$ ,

$$f(r_i, y_i | x_i, \theta) \propto f(y_i | r_i, x_i, \delta) \cdot f(r_i | x_i, \gamma)$$

where  $\theta = (\gamma, \delta)$ .

You can use the marginal model to model different patterns of the missing mechanism  $\mathbf{R}$ . And you can build meaningful models for subsets of the response variable conditional on different missing patterns. On the other hand, you must always model  $\mathbf{R}$  in a pattern-mixture model.

For more in-depth discussions and comparisons of selection and pattern-mixture models, see Little (2009) and references therein.

## MISSING DATA ANALYSIS IN PROC MCMC

Prior to and including SAS/STAT 9.3, PROC MCMC performs a complete-case analysis by default when the data contain missing values. This means that PROC MCMC discards any records that contain missing values prior to the analysis. To model missing values in these versions of PROC MCMC, you create a parameter for each missing value, specify its prior distribution, associate each missing value with its parameter, and calculate the likelihood by using both observed data and parameters in place of missing values. Coding requires more bookkeeping, and the Markov chain tends to converge more slowly.

Beginning with SAS/STAT 12.1, the procedure models missing values whenever it can. PROC MCMC still discards observations that have missing or partially missing values when you specify the MISSING=CC option.

To model missing values, you must declare the variable in a MODEL statement, which has the following syntax:

```
MODEL variable-list ~ distribution < options > ;
```

The distribution is the usual likelihood function when the MODEL statement is applied to a response variable; the distribution becomes the prior distribution for a covariate. You can think of the prior distribution as a sampling distribution that generates the covariate of interest. The prior distribution can be a stand-alone distribution, such as the normal distribution with unknown mean, or a more complex model that involves additional regression covariates.

During the setup stage, PROC MCMC identifies all missing values in the input data set that are in *variable-list* in the MODEL statement and creates a separate internal variable for each missing value. This internal variable, which is called a missing data variable, becomes an additional parameter in the model. The name of the missing data variable name is created by concatenating the data set variable name with the observation index. At each iteration, PROC MCMC automatically samples each missing data variable from its conditional posterior distribution, just as the procedure does for all parameters in the model. For a response variable, the posterior distribution is the same as the likelihood function. Direct sampling algorithms are often used to draw these samples. For covariates, the posterior distribution is the product of its prior distribution and its contribution to the likelihood function. PROC MCMC resorts to scenario-specific sampling algorithms to draw these samples.

PROC MCMC models missing values only for variables that are specified in the MODEL statement. Missing covariate data require that the covariate be modeled and that the covariate variable be specified in the MODEL statement. If a covariate has missing values and the covariate is not modeled, records that contain missing values are discarded before the analysis.

PROC MCMC supports partial missing data in a multivariate normal distribution (MVN; see the section “[Example 1: Bivariate Normal with Partial Missing Data](#)” for an example), in an autoregressive multivariate normal distribution (MVNAR), and in the GENERAL and DGENERAL functions. PROC MCMC samples the partial missing values from their posterior distributions, conditional on the nonmissing variable values. SAS/STAT 12.1 does not support partial missing data in Dirichlet, inverse Wishart, and multinomial distributions (records that have partial missing values in these distributions are discarded). Alternatively, you can factor a joint distribution into marginal and conditional distributions, and use multiple MODEL statements to handle partial missingness in multidimensional distributions.

The *options* in the MODEL statement apply only when there are missing values in the *variable-list*. The *options* and brief descriptions are listed in [Table 1](#).

**Table 1** Options for the MODEL Statement

Option	Description
<b>INITIAL=</b>	specifies the initial values of the missing data, which are used to start the Markov chain. The default starting value is the average of nonmissing values of the variable.
<b>MONITOR=</b>	outputs analysis for selected missing data variables. You can choose the list by names or indices, or you can have the procedure randomly select for you.
<b>NAMESUFFIX=</b>	specifies how to create the names of the missing data variables. You can construct the names by observation indices or by the sequence of appearance of the missing values in the data set.
<b>NOOUTPOST</b>	suppresses the output of the posterior samples of the missing data variables.

## EXAMPLES

This section presents three missing data analysis examples that use PROC MCMC: a bivariate normal data set that contains partial missing data, a logistic model that has missing covariates, and a repeated measurements MNAR model for missing response data.

### Example 1: Bivariate Normal with Partial Missing Data

Murray (1977) proposes a simple bivariate normal sample problem with missing data. It is an artificial problem, but it presents unique estimation challenges that were later studied by many, including Tanner and Wong (1987) and Tan, Tian, and Ng (2010). The data set consists of 12 observations from a bivariate normal distribution. Both variables have missing values. Table 2 displays the data set Binorm.

**Table 2** Bivariate Normal Data

<b>x1</b>	1	1	-1	-1	2	2	-2	-2	.	.	.	.
<b>x2</b>	1	-1	1	-1	.	.	.	.	2	2	-2	-2

The variables **x1** and **x2** are assumed to have zero means, correlation coefficient  $\rho$ , and variances  $\sigma_1^2$  and  $\sigma_2^2$ . These assumptions translate to the following covariance matrix:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

with the determinant of  $\sigma_1^2\sigma_2^2(1 - \rho^2)$ . The prior distribution on  $\Sigma$  is an inverse Wishart distribution (Box and Tiao 1973),

$$\pi(\Sigma) \propto |\Sigma|^{-\frac{p+1}{2}} = \left(\sigma_1\sigma_2\sqrt{1-\rho^2}\right)^{-(p+1)}$$

where  $p$  is the dimension of the multivariate normal distribution, which is 2 in this example.

A complete-case analysis uses the first four pairs of data, which would produce a rather simple posterior estimate of  $\rho$  of 0: two pairs ((1,1) and (-1,-1)) have correlation 1, and two pairs ((1,-1) and (-1,1)) have correlation -1. Treating this as an MCAR analysis is equivalent to assuming that the unobserved values are either (nearly) perfectly positively or perfectly negatively correlated with the observed values of 2 and -2, and that the missing values can take on only values that do not differ much from these. Further, by discarding partially observed values, you lose substantial information in estimating the two variance terms.

The following statements perform an MAR analysis on the bivariate normal data set:

```
proc mcmc data=binorm nmc=20000 seed=17 outpost=postout
  diag=none plots=none;
  array x[2] x1 x2;
  array mu[2] (0 0);
  array sigma[2, 2];

  parms rho 0.5 / slice;
  parms sig1 1 sig2 1;
  if (sig1 > 0 and sig2 > 0) then
    lprior = -3 * log(sig1 * sig2 * sqrt(1-rho*rho));
  else
    lprior = .;
  prior rho sig1 sig2 ~ general(lprior);
  sigma[1,1] = sig1*sig1;
  sigma[1,2] = rho * sig1 * sig2;
  sigma[2,1] = sigma[1,2];
  sigma[2,2] = sig2*sig2;
  model x ~ mvn(mu, sigma);
run;
```

The PARMs statements specify three model parameters. The SLICE option in the first PARMs statement selects the slice algorithm (Neal 2003) to sample the **rho** parameter. Although computationally more expensive, the slice sampler often performs well in drawing samples from a nonnormal-like distribution, such as a correlation parameter.

The IF-ELSE statements specify the positive support for the standard deviation parameters **sig1** and **sig2**. The purpose of the constraint is to prevent the sampling of **sig1** and **sig2** that are both negative, which is mathematically allowable according to the definition of the  $\Sigma$  matrix. The LPRIOR= assignment statement

and the PRIOR statement specify a joint prior distribution on the logarithm scale<sup>1</sup> on **rho**, **sig1**, and **sig2**. Elements of the **sigma** covariance matrix are constructed by using additional programming statements as functions of the model parameters. Finally, the MODEL statement rounds up the specification of the likelihood function.

Figure 1 shows the missing data information table. The variable **x1** has four missing values, and the variable **x2** also has four. The Observation Indices column lists the observation indices that have missing values.

**Figure 1** Missing Data Information Table of the Bivariate Normal Analysis

Missing Data Information Table						
Variable	Number of Missing Obs	Observation Indices				Sampling Method
<b>x1</b>	4	9	10	11	12	Direct
<b>x2</b>	4	5	6	7	8	Direct

In this example, the conditional distributions of the missing values are known and are univariate normal. When  $x_1$  is known,

$$\pi(x_2|\rho, \sigma_1, \sigma_2, x_1) \sim N\left(\rho \frac{\sigma_2}{\sigma_1} x_1, \sigma_2^2(1 - \rho^2)\right)$$

When  $x_2$  is known,

$$\pi(x_1|\rho, \sigma_1, \sigma_2, x_2) \sim N\left(\rho \frac{\sigma_1}{\sigma_2} x_2, \sigma_1^2(1 - \rho^2)\right)$$

PROC MCMC uses direct sampling methods to draw the missing values from these conditional univariate normal distributions.

Figure 2 shows the posterior summary statistics of the model parameters.

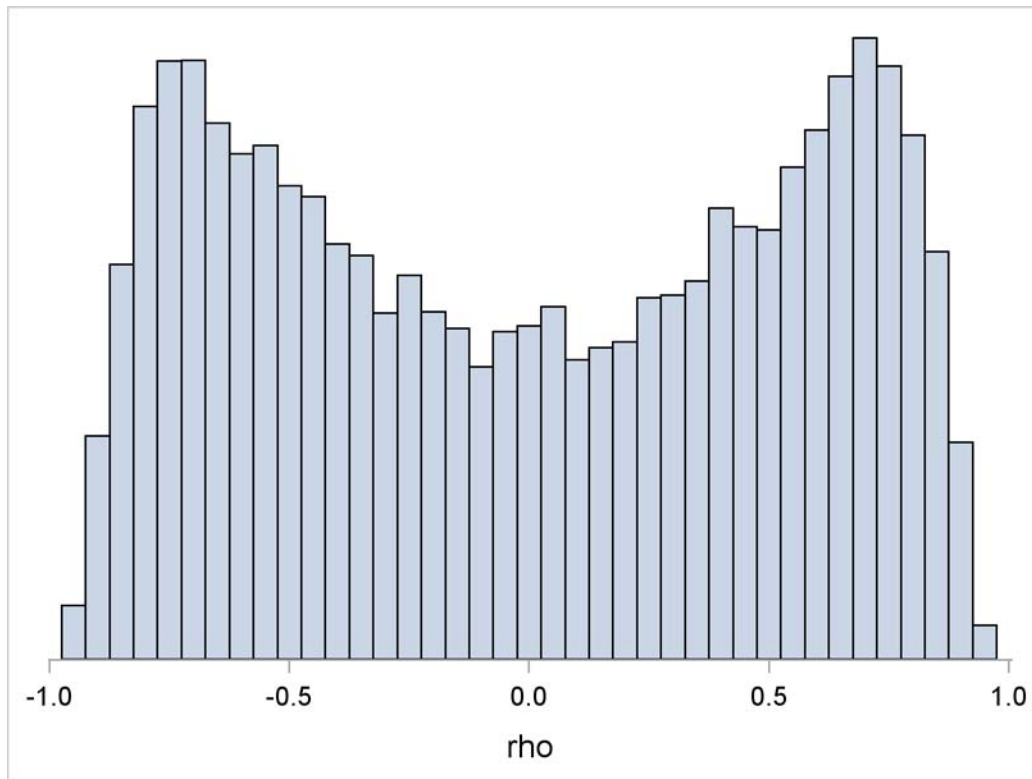
**Figure 2** Posterior Summary Statistics of the Bivariate Normal Analysis

The MCMC Procedure						
Posterior Summaries						
Parameter	N	Mean	Standard Deviation	Percentiles		
				25%	50%	75%
<b>rho</b>	20000	-0.00184	0.5582	-0.5283	-0.00452	0.5305
<b>sig1</b>	20000	1.6555	0.4458	1.3387	1.5702	1.8713
<b>sig2</b>	20000	1.6306	0.4444	1.3336	1.5523	1.8524
Posterior Intervals						
Parameter	Alpha	Equal-Tail Interval		HPD Interval		
<b>rho</b>	0.050	-0.8573	0.8544	-0.8591	0.8519	
<b>sig1</b>	0.050	1.0263	2.7404	0.9828	2.6014	
<b>sig2</b>	0.050	1.0059	2.6416	0.9478	2.4860	

<sup>1</sup>The GENERAL function requires that any user-specified distributions be expressed on the logarithm scale to ensure greater stability in numerical calculation.

The mean estimate of **rho** is close to 0, with a standard deviation of 0.56. This estimate is misleading because the posterior distribution is not unimodal, as shown in Figure 3.

**Figure 3** Posterior Density of  $\rho$



By default, estimates of the missing values are not displayed. To obtain posterior estimates of these variables, you can either use the MONITOR= option in the MODEL statement to output posterior analysis, or use postprocessing macros, such as the %POSTSUM and %POSTINT autocall macros<sup>2</sup>, as follows:

```
%postsum(data=postout, var=x:);
%postint(data=postout, var=x:);
```

On the other hand, the posterior distributions of the missing values are nicely unimodal (the output is not displayed).

### Example 2: Missing Covariates

This example analyzes an air pollution data set to illustrate how you can use PROC MCMC to model missing covariates. The data set is simulated to be similar to the data from the six-cities longitudinal study of health effects on respiratory disease of children (Ware et al. 1984), which was analyzed by Ibrahim and Lipsitz (1996), among others. The following statements generate the Air data set:

<sup>2</sup>For a complete list of available autocall postprocessing macros, see Chapter 56, "The MCMC Procedure" (*SAS/STAT User's Guide*).



```

data Air;
  input y city smoke @@;
  datalines;
0 0 0 0 0 0 0 1 0 0 0 0 0 0 11 0 1 7
0 0 8 0 1 10 0 1 9 0 0 0 1 1 6 0 1 10
0 1 12 0 0 . 0 0 0 0 1 0 0 1 7 1 1 15

... more lines ...

0 . 0 1 0 0 1 1 10 0 . 4 1 1 16 0 . 13
;

```

The three variables are as follows:

- **y**: wheezing symptoms of a child (1 for symptoms exhibited; 0 otherwise)
- **city**: city of residence (1 for Steel City; 0 Green Hills)
- **smoke**: maternal cigarette smoking, measured in cigarettes per day

There are a total of  $n = 390$  subjects, with 17 cases missing **city** and 30 cases missing **smoke**. In one case both **city** and **smoke** are missing, but there are no missing values in the response variable **y**.

The following logistic regression models the wheezing status:

$$\text{logit}(p(\mathbf{y}_i = 1 | \mathbf{city}_i, \mathbf{smoke}_i)) = \beta_0 + \beta_1 \cdot \mathbf{city}_i + \beta_2 \cdot \mathbf{smoke}_i$$

To model the missing covariates, you can consider a joint distribution for **city** and **smoke** to be of the form

$$[\mathbf{city}, \mathbf{smoke}] = [\mathbf{smoke} | \mathbf{city}] \cdot [\mathbf{city}]$$

where **[city]** is a marginal binary model of the city of residence,

$$\text{logit}(p(\mathbf{city}_i)) = \phi$$

and **[smoke | city]** is a count model that estimates the number of cigarettes smoked daily. A quick PROC FREQ call (output not displayed) shows a large number (181 out of 390) of subjects who did not smoke (0 cigarettes per day):

```

proc freq data=air;
  table smoke / nopercents;
  where smoke eq 0;
run;

```

The results suggest that there are two subpopulations of mothers, those who smoked and those who didn't. The usual choice to model count data is Poisson regression, which is inadequate here. A more sensible alternative is a two-component mixture distribution—the zero-inflated Poisson (ZIP) model—which can capture the smoking patterns in the subjects more accurately,

$$\pi(\mathbf{smoke}_i | \mathbf{city}_i) = \eta p_s + (1 - \eta) \text{Poisson}(\mu)$$

where

$$\begin{aligned}
 p_s &= \begin{cases} 1 & \text{if } \mathbf{smoke}_i = 0 \\ 0 & \text{if } \mathbf{smoke}_i \neq 0 \end{cases} \\
 \mu &= \exp(\alpha_0 + \alpha_1 \cdot \mathbf{city}_i) \\
 0 &\leq \eta \leq 1
 \end{aligned}$$

The regression parameters are given a normal prior with large variance, and  $\eta$ , the weight parameter, is given a uniform prior from 0 to 1. The following PROC MCMC statements analyze the Air data set that has missing covariates:



```

proc mcmc data=air seed=1181 nmc=20000 plots=none outpost=airout propcov=quanew;
  parms beta0-beta2 phi alpha0 alpha1 eta;
  prior beta: phi alpha: ~ normal(0,var=10);
  prior eta ~ uniform(0, 1);

  p1 = logistic(phi);
  model city ~ binary(p1);
  mu = exp(alpha0 + alpha1 * city);
  den = eta * (smoke eq 0) + (1-eta) * pdf("poisson", smoke, mu);
  if den > 0 then lprior = log(den);
  else lprior = -1e200;
  model smoke ~ dgeneral(lprior, lower=0);
  p = logistic(beta0 + beta1*city + beta2*smoke);
  model y ~ binary(p);
run;

```

The first MODEL statement specifies a marginal binary distribution for **city**. The MU= programming statement defines the mean component in the Poisson model; the DEN= programming statement calculates the ZIP density on the density scale. The IF-ELSE statements prevent the program from taking a logarithm of 0 value. The second MODEL statement uses the DGENERAL function to complete the nonstandard discrete prior specification for **smoke**. The third MODEL statement specifies the likelihood for the response variable in **y**, with a binary probability that is a function of both **city** and **smoke**.

Figure 4 displays histograms of the observed **smoke** variable and the simulated values of all missing **smoke** values. The simulated values show reasonable agreement with the observed **smoke** values, although the weight on the 0 value is a bit underestimated. This disagreement is not completely unexpected; total agreement would indicate that the missing data model is MCAR and that the missing covariate data are interchangeable with those that are observed. In an MAR model, the missing covariate data might depend on other observed data, whose characteristics play an important role in determining the plausible outcome of these missing values.

**Figure 4** Comparison of Observed and Simulated **smoke** Values

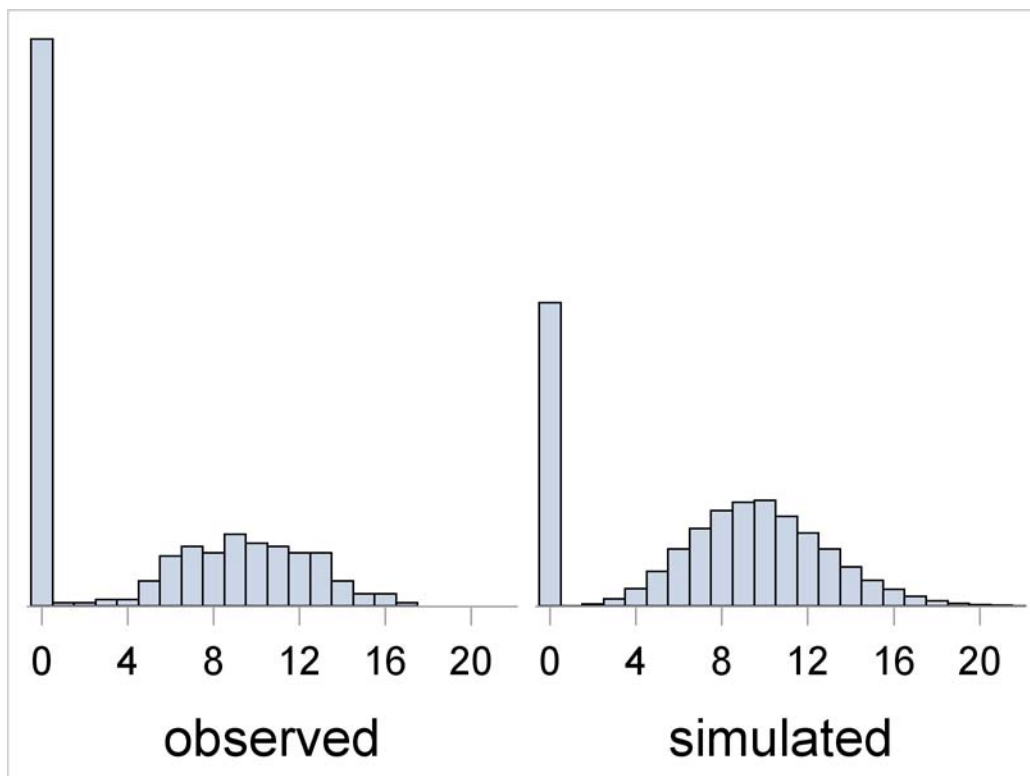


Figure 5 displays the posterior point and interval estimates of the model parameters in the AIR data analysis.

Figure 5 Posterior Estimates

The MCMC Procedure						
Posterior Summaries						
Parameter	N	Mean	Standard Deviation	Percentiles		
				25%	50%	75%
beta0	20000	-1.3721	0.2066	-1.5104	-1.3666	-1.2324
beta1	20000	0.4783	0.2386	0.3186	0.4780	0.6324
beta2	20000	0.0165	0.0221	0.00239	0.0160	0.0312
phi	20000	-0.1575	0.1084	-0.2319	-0.1602	-0.0852
alpha0	20000	2.2595	0.0343	2.2368	2.2610	2.2825
alpha1	20000	0.00879	0.0503	-0.0253	0.0100	0.0416
eta	20000	0.4853	0.0264	0.4673	0.4855	0.5034

Posterior Intervals					
Parameter	Alpha	Equal-Tail Interval		HPD Interval	
beta0	0.050	-1.7791	-0.9715	-1.7706	-0.9688
beta1	0.050	0.0106	0.9531	0.0182	0.9560
beta2	0.050	-0.0282	0.0601	-0.0260	0.0616
phi	0.050	-0.3638	0.0581	-0.3564	0.0633
alpha0	0.050	2.1920	2.3257	2.1919	2.3252
alpha1	0.050	-0.0894	0.1078	-0.0917	0.1038
eta	0.050	0.4335	0.5358	0.4326	0.5336

Estimates for both **beta1** and **beta2** are positive—0.48 and 0.017, respectively—indicating that living in Steel City and increased daily smoking by the mothers worsen the wheezing symptoms of a child. However, the posterior interval estimates of **beta2** include 0, and the effect is not significant, with a posterior probability of  $\Pr(\text{beta2} > 0 \mid \text{Data}) = 77.7\%$ . The following statements calculate the probability:

```
data Prob;
  set airout;
  Indicator = (beta2 > 0);
run;

proc means data = Prob(keep=Indicator) mean;
run;
```

The Poisson regression does not have much predictive power of **city** (**alpha1**) on **smoke**. This is as expected, because the city of residence probably wouldn't have much effect on a smoking habit.

The weight parameter **eta** estimate is close to 0.5, which agrees with the observed data that about 50% of mothers did not smoke. In the right panel of Figure 4, note that the **eta** variable is much less than 50%. This is because the parameter **eta** models the overall rate of nonsmokers in the data (both observed and unobserved), and the zero component estimate in the histogram represents the estimated probability of nonsmokers among those who have missing **smoke** values. The discrepancy suggests that subjects whose smoking status is missing are more likely to be smokers. In a CC analysis, an MCAR model cannot infer from this subpopulation a difference in **eta**. The MAR analysis uses the information that would have been discarded—that is, children whose mothers have a missing smoking status probably have worse wheezing symptoms than children whose mothers reported their smoking status—and derives a more informative conclusion.

Figure 6 displays a density comparison plot of the regression coefficients between an MAR model and an MCAR analysis. The following statements perform an MCAR analysis:

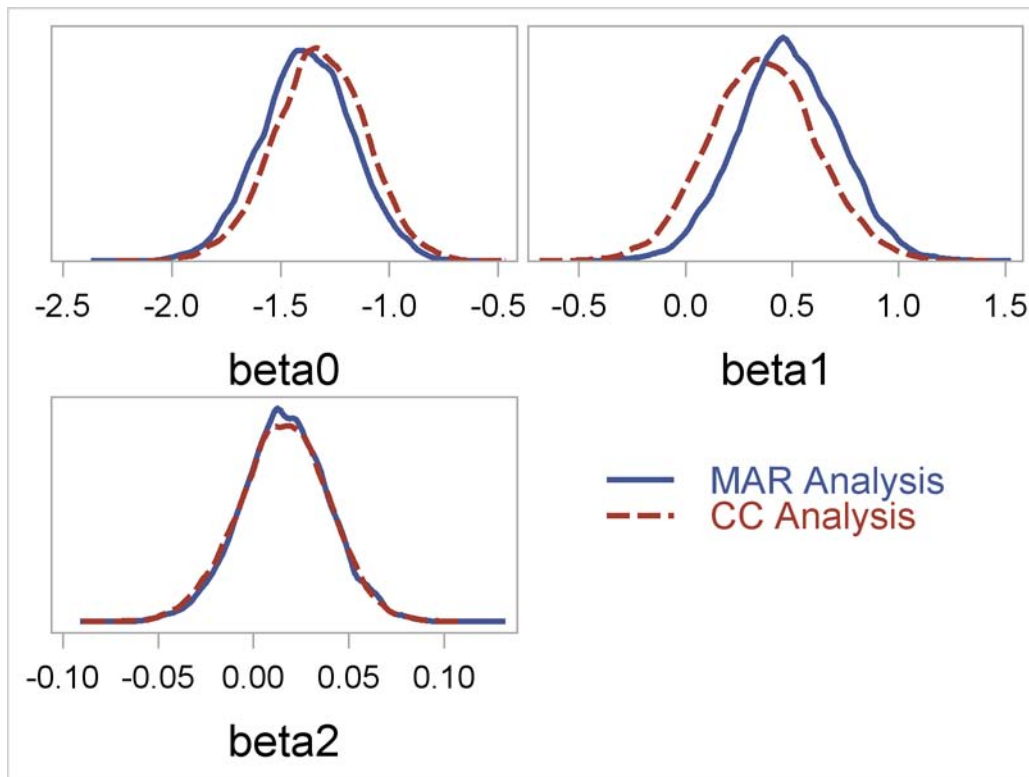
```

proc mcmc data=air seed=1181 nmc=50000 outpost=airoutcc
  stats=none diag=none plots=none propcov=quanew;
  parms beta0 -1 beta1 0.1 beta2 .01;
  prior beta: ~ normal(0,var=10);
  p = logistic(beta0 + beta1*city + beta2*smoke);
  model y ~ binary(p);
run;

```

The model is a simple logistic regression with no modeling of any missing values. In Figure 6, solid blue lines indicate density estimates from the MAR analysis; dashed red lines indicate the MCAR analysis.

**Figure 6** Comparison Plots of Complete-Case versus MAR Analysis of the Air Data Set



The MAR analysis has a minor impact on the posterior estimates, most noticeably on the **city** effect, where the MAR model shows a stronger effect on a child's wheezing status than the MCAR model shows.

### Example 3: Selection Model

This example illustrates a selection model approach to nonignorable missing data. The statistical model is a repeated measurements model that fit a two-arm clinical trial data set over a period of four weeks.

The data set that is used in this example is based on a clinical trial that was originally reported by Goldstein et al. (2004), who conducted a double-blind study that compared antidepressants. The Drug Information Association (DIA) working group on missing data made this data set available at [www.missingdata.org.uk](http://www.missingdata.org.uk). To avoid implications for marketed drugs, all patients in this data set who took medications are grouped into a single **DRUG** group, and only a subset of those on active treatment in the original trial are included. There remain 171 subjects<sup>3</sup> in the data set: 88 were in a control group, and 83 were given some forms of medication. The presentation of the analysis follows Mallinckrodt et al. (2013). The following statements create the Selection data set:

<sup>3</sup>One record, subject 3618, was removed from the original DIA data set because the subject had an intermediate missing data pattern. The treatment of such subjects, while feasible using PROC MCMC, complicates the flow of the example without adding useful information and is not presented here.

```

data Selection;
input PATIENT baseval change1-change4 r1-r4 THERAPY $ POOLINV $ last wkMax;
datalines;
1503 32 -11 -12 -13 -15 0 0 0 0 DRUG 006 4 4
1507 14 -3 0 -5 -9 0 0 0 0 PLACEBO 006 4 4
1509 21 -1 -3 -5 -8 0 0 0 0 DRUG 006 4 4
1511 21 -5 -3 -3 -9 0 0 0 0 PLACEBO 006 4 4
1513 19 5 . . . 0 1 1 1 DRUG 006 1 2

... more lines ...

4909 28 -4 6 0 5 0 0 0 0 PLACEBO 999 4 4
;

```

The following variables are included in the data set:

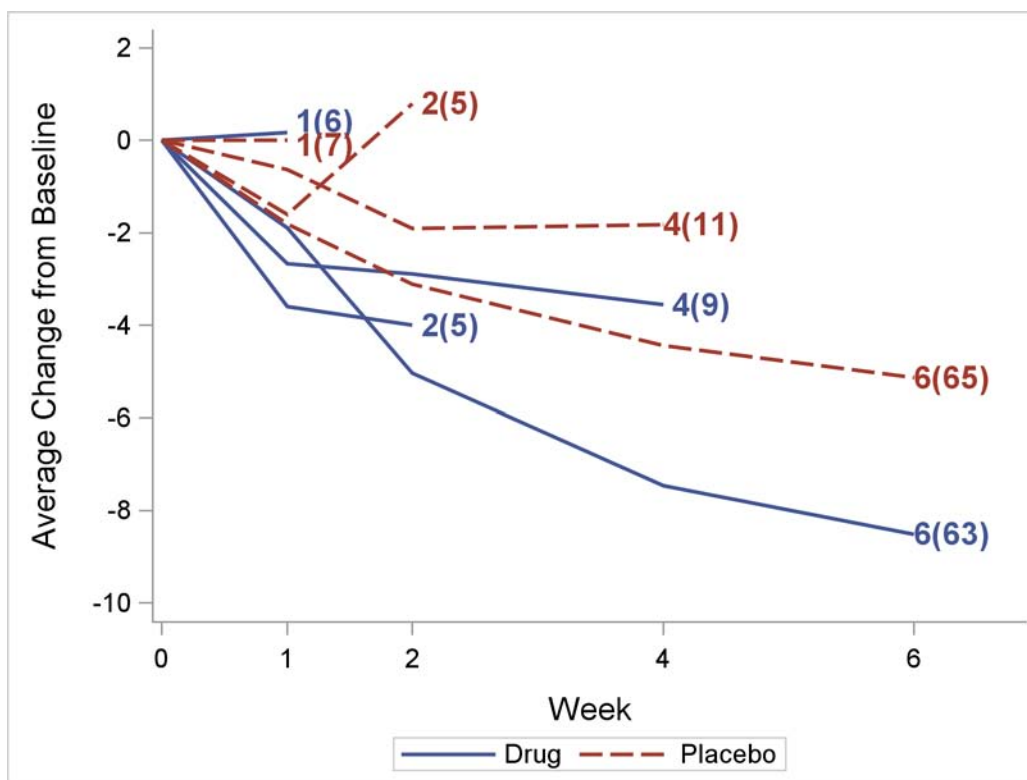
- **patient**: patient ID
- **baseval**: baseline assessment on the Hamilton 17-item rating scale for depression (HAMD<sub>17</sub>, Hamilton 1960) taken at the beginning of the study (week 0)
- **change1–change4**: change in HAMD<sub>17</sub> at weeks 1, 2, 4, and 6. Missing values, represented by dots, indicate that the subjects had dropped out; no measurements are available for these weeks. A lower HAMD<sub>17</sub> score indicates less depression (“good”). Large negative scores from **baseval** indicate success.
- **r1–r4**: missing data indicator for each of the **change** variables
- **therapy**: treatment (**DRUG** versus **PLACEBO**)
- **poolinv**: blocking information (groups formed by pooling investigator)
- **last**: week index to last nonmissing **change** value. It is the last visitation week of the patient.
- **wkMax**: maximum number of weeks to be included in the analysis

Every patient had a baseline measure of HAMD<sub>17</sub> score at the beginning of the study, but not every subject completed the trial. Of the 84 patients in the **DRUG** group, 21 patients dropped out early; of the 88 patients in the **PLACEBO** group, 23 patients dropped out early. The completion rates were about the same for both groups.

The input data set is organized by subjects, with multiple response variables across each observation. This input format, as opposed to combining all variables into one column, is necessary to account for any dependence structures across the repeated measurements. PROC MCMC assumes that observations are independent of one another, and it does not support a REPEATED statement (like the one in PROC MIXED) that enables you to estimate covariance structures on subjects. By putting all repeated measurements in each observation, you can use the MODEL statement to infer dependence by specifying a joint distribution (such as a multivariate normal distribution) over all of them.

### Selection Mechanism

Figure 7 plots the average change in the HAMD<sub>17</sub> score from baseline for the two groups of patients. The graph is organized according to the number of weeks that the patients stayed in the trial. Solid blue lines represent patients in the **DRUG** group; dashed red lines represent patients in the **PLACEBO** group. There are four solid lines, one for each group of **DRUG** patients who stayed until the current week. Similarly, there are four dashed lines, for the four **PLACEBO** groups. The first number at the end of each line indicates the last visitation week. The number in parentheses is the sample size for that group of patients. The majority of patients completed the study.

**Figure 7** Average Mean Change for Patients Who Completed the Trial and Patients Who Dropped Out

This graph suggests that dropout probabilities might be correlated with the level of improvement (change in scores) that a patient experienced at the last visit. Patients who failed to see an improvement (flat or upswinging lines) might not return for more treatments. On the other hand, the two groups that completed the study saw a steady decrease in their HAMD<sub>17</sub> scores, suggesting that feeling better increased their chances of staying in the trial.

Classic MAR analyses are often used in such situations. They assume that, conditional on the data that have been observed so far, whether or not a subject withdraws is unrelated to his or her potential future values, which might be unobserved if he or she decides to drop out. (One such value could be how well the subject feels at the current week.) One way to explore sensitivity to this assumption is to model the missingness mechanism in such a way that the probability of withdrawal depends on unobserved future data values. And a *selection model* can incorporate this dependency that the missing mechanism has on the would-be values of the response variables.

One feature of the study is that the missing pattern is monotonic, meaning that a subject who leaves the study does not return. That is, if  $\mathbf{r3}_i$  for subject  $i$  is 1 in week 3,  $\Pr(\mathbf{r4}_i = 1) = 1$ . You do not need to model the missing data indicator variables for subjects after they leave the study. Instead, the missing mechanism for each variable is modeled conditionally:  $\mathbf{r2}_i$  is modeled conditional upon  $\mathbf{r1}_i = 0$ ;  $\mathbf{r3}_i$  is modeled conditional upon  $\mathbf{r2}_i = \mathbf{r1}_i = 0$ . The variable **wkMax** keeps track of the maximum number of visitation weeks that should be included in the analysis. The variable **wkMax** is 4 for subjects who complete the study; it is the last observed visitation week plus 1 in other cases.

### Statistical Model

This section describes the model that fits the response variables and the parametric selection model that fits the missing data indicators, given the responses.

Let  $i$  be the subject index,  $j = \{1, 2, 3, 4\}$  be the week index,  $k = \{1, 2\}$  ( $1=\text{DRUG}$ ,  $2=\text{PLACEBO}$ ) be the treatment index, and  $l$  be the block index (**poolinv**). You can model the response variable  $\mathbf{change}_i = \{\mathbf{change}_{j,i}\}$  by using a multivariate normal distribution with mean  $\boldsymbol{\mu}_i = (\mu_{1i}, \mu_{2i}, \mu_{3i}, \mu_{4i})$  and covariance matrix  $\boldsymbol{\Sigma}$ . The mean parameter for the  $i$ th subject in the  $j$ th week takes on the regression model,

$$\mu_{ji} = m_{kj} + \beta_j \cdot (\text{baseval}-18) + \gamma_l$$

where  $m_{kj}$  is the treatment effect for the  $j$ th week,  $\beta_j$  is the slope effect for the  $j$ th week, and  $\gamma_l$  is the intercept effect for each grouped patient, indicated by the variable **poolinv**.

All parameters are given a flat prior:

$$\begin{aligned}\pi(m_{kj}) &\propto 1 \\ \pi(\beta_j) &\propto 1 \\ \pi(\gamma_l) &\propto 1\end{aligned}$$

Group intercept  $\gamma_l$  with a flat prior makes the design matrix rank deficient. A common strategy is to use a *corner point* constraint by setting one of the redundant  $\gamma_l$  parameters to 0 and reducing the total number of  $\gamma$  parameters by 1. You do not need to put constraints on the  $m_{.k}$  parameters because the model does not have overall (multidimensional) intercepts over the weeks.

The selection model uses a logistic regression that models the dropout probabilities. The covariates include previous and current (possibly missing) response variables, because you want to infer conditional on not having withdrawn earlier if either of the following might affect a patient's willingness to continue the trial: score improvement from previous visitation or the current state of wellness. This type of selection model is referred as the Diggle-Kenward selection model (Diggle and Kenward 1994; Daniels and Hogan 2008). Because there is an expected treatment effect, you want the logistic regression to include separate intercepts and regression coefficients for each treatment group.

The variable **change1** does not contain any missing values, making all **r1** values equal to 1 and therefore irrelevant to the analysis. For  $j = \{2, \dots, \text{wkMax}\}$  and  $k = \{1, 2\}$ , you can fit the following logistic regression to model the selection mechanism to each subject  $i$ :

$$\begin{aligned}\mathbf{r}_{kji} &\sim \text{binary}(q_{kji}) \\ q_{kji} &= \text{logistic}(\phi_{k1} + \phi_{k2} \cdot \text{change}_{(j-1)i} + \phi_{3k} \cdot \text{change}_{ji})\end{aligned}$$

The variable **wkMax** is the upper limit of  $j$  to exclude unwanted missing data indicators from the analysis.

In the model, the covariance matrix takes on an inverse-Wishart prior distribution, and the rest of the parameters are assigned flat priors.

## Analysis

The following PROC MCMC statements fit the selection model:

```
proc mcmc data=selection nmc=50000 seed=17 outpost=seleout
  diag=none plots=none monitor=(beta m phi);
  array Change[4] Change1-Change4;
  array mu[4];
  array Sigma[4,4];
  array S[4,4];
  array beta[4];
  array M[2,4] m1-m8;
  array phi[2,3] phi1-phi6;

  begincnst;
  call identity(s);
  endcnst;

  parms beta: 0;
  parms m1-m8 0;
  parms phi1-phi6 0;
  parms Sigma;
```

```

prior beta: m1-m8 phi: ~ general(0);
prior Sigma ~ iwish(4, S);

random gamma ~ general(0) subject=poolinv zero=first init=0;
do i=1 to 4;
  if therapy eq "DRUG" then do;
    mu[i] = m[1,i] + gamma + beta[i]*(baseval-18);
  end; else do;
    mu[i] = m[2,i] + gamma + beta[i]*(baseval-18);
  end;
end;
model Change ~ mvn(mu, Sigma);

/* selection mechanism */
array r[4] r1-r4;
llike = 0;
do i = 2 to wkMax;
  if therapy eq "DRUG" then do;
    p = logistic(phi[1,1] + phi[1,2] * change[i-1] + phi[1,3] * change[i]);
  end; else do;
    p = logistic(phi[2,1] + phi[2,2] * change[i-1] + phi[2,3] * change[i]);
  end;
  llike = llike + lpdfbern(r[i], p);
end;
model r2 r3 r4 ~ general(llike);
run;

```

The ARRAY statements allocate variables for the multidimensional response variables and parameters in the model. The CALL IDENTITY statement sets the hyperparameter **s** in the inverse Wishart distribution to an identity matrix. The PARMs and PRIOR statements declare model parameters and their prior distributions. Multiple PARMs statements are used here to break the joint updating of parameters into smaller blocks, which can increase the efficiency of the Metropolis algorithm.

The RANDOM statement on **gamma** with the ZERO=FIRST option is a “trick” that simplifies the coding of the corner constraint on the **poolinv**-level intercepts in the outcome model. Instead of creating a full-rank design matrix and fitting group intercepts that way, you use the RANDOM statement and declare **gamma** to be a random intercept for each cluster. By specifying ZERO=FIRST, you set the **gamma** parameter from the first group to 0, making the model identifiable. The ensuing programming statements calculate the mean parameter for each of the **change** variables, depending on the THERAPY variable. The MODEL statement completes the specification of the outcome model.

The next part of the program specifies the selection model. The DO loop sums over the logarithm of Bernoulli likelihood for the missing data variables, from week 2 to week **wkMax**. Three **phi** variables are parameters in the **DRUG** model, and the other three in the **PLACEBO** model. The Bernoulli probabilities depend on score changes from the previous week (**change[i-1]**) and the current week (**change[i]**). The MODEL statement declares a joint log likelihood for the indicators.

Figure 8 displays the posterior point and interval estimates of the model parameters.



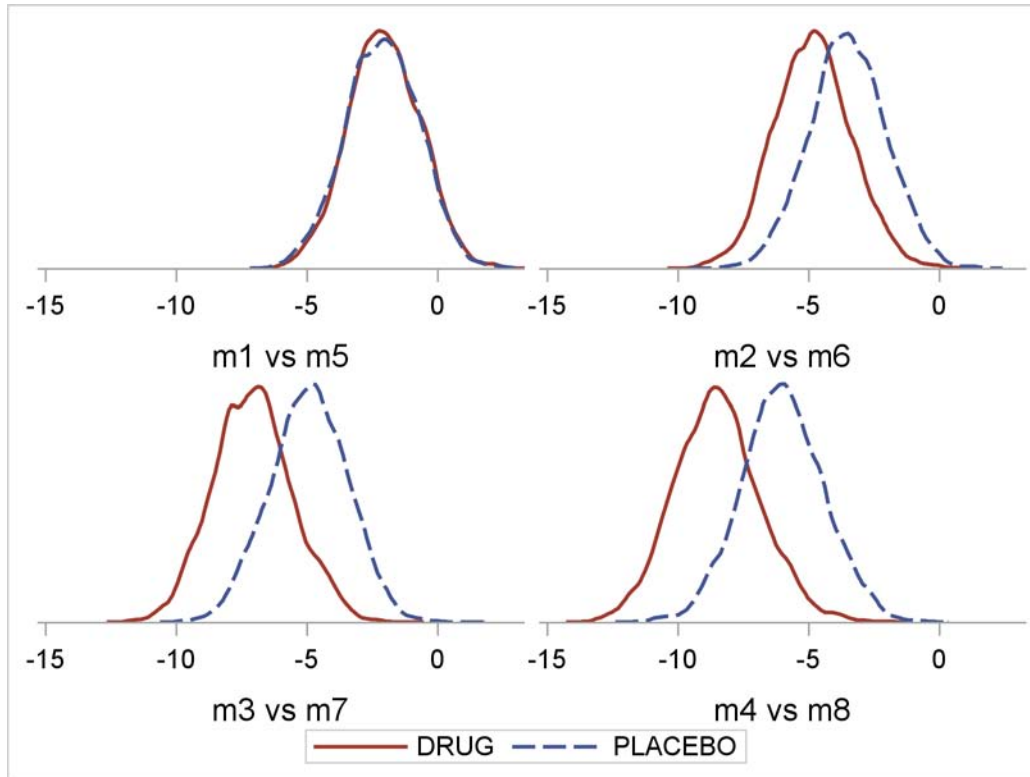
Figure 8 Posterior Estimates

The MCMC Procedure						
Posterior Summaries						
Parameter	N	Mean	Standard Deviation	Percentiles		
				25%	50%	75%
beta1	20000	-0.2928	0.0672	-0.3351	-0.2930	-0.2478
beta2	20000	-0.3352	0.0803	-0.3889	-0.3348	-0.2824
beta3	20000	-0.4187	0.0859	-0.4764	-0.4205	-0.3603
beta4	20000	-0.3415	0.0952	-0.4058	-0.3432	-0.2769
m1	20000	-2.0061	1.5214	-3.0716	-2.0648	-0.9333
m2	20000	-4.6596	1.6558	-5.7892	-4.6681	-3.5386
m3	20000	-6.8650	1.7675	-8.0493	-6.9568	-5.7561
m4	20000	-8.1468	1.9500	-9.4452	-8.2191	-7.0033
m5	20000	-2.0501	1.5025	-3.0566	-2.0922	-1.0279
m6	20000	-3.4138	1.5624	-4.4474	-3.4265	-2.3732
m7	20000	-4.8623	1.5913	-5.8970	-4.9037	-3.8210
m8	20000	-5.8467	1.7306	-6.9902	-5.8501	-4.7823
phi1	20000	-2.8049	0.8471	-3.1386	-2.5611	-2.2284
phi2	20000	0.0685	0.1938	-0.0281	0.1021	0.1988
phi3	20000	-0.0344	0.2253	-0.1825	-0.0630	0.0799
phi4	20000	-2.7389	0.5704	-3.0720	-2.6730	-2.3207
phi5	20000	0.1990	0.1005	0.1349	0.2039	0.2642
phi6	20000	-0.1512	0.1218	-0.2356	-0.1586	-0.0725

Posterior Intervals					
Parameter	Alpha	Equal-Tail Interval		HPD Interval	
beta1	0.050	-0.4267	-0.1560	-0.4189	-0.1507
beta2	0.050	-0.4910	-0.1780	-0.4949	-0.1825
beta3	0.050	-0.5878	-0.2468	-0.5952	-0.2570
beta4	0.050	-0.5335	-0.1577	-0.5271	-0.1529
m1	0.050	-4.8502	1.0461	-4.9816	0.8794
m2	0.050	-7.8433	-1.2709	-8.0195	-1.5218
m3	0.050	-10.1592	-2.9939	-10.3581	-3.3984
m4	0.050	-11.7193	-3.8123	-12.1661	-4.3531
m5	0.050	-5.0003	0.8244	-4.9506	0.8417
m6	0.050	-6.4752	-0.3347	-6.5470	-0.4608
m7	0.050	-8.0527	-1.7476	-7.8236	-1.5813
m8	0.050	-9.3365	-2.5106	-9.1787	-2.4775
phi1	0.050	-5.1737	-1.7732	-4.6620	-1.5826
phi2	0.050	-0.3572	0.3755	-0.3506	0.3789
phi3	0.050	-0.4275	0.4451	-0.4330	0.4351
phi4	0.050	-4.0633	-1.8456	-3.8179	-1.7330
phi5	0.050	-0.00616	0.3914	-0.00677	0.3890
phi6	0.050	-0.3753	0.0986	-0.3744	0.0991

The  $\beta$  parameters are all negative, indicating approximately the same declining rate from the **baseval** for all four weeks. Negative **m** values show improvements in the average of the depression scores. It is easier to see pairwise comparison plots of the posterior densities of the **m** variables, as displayed in Figure 9.

**Figure 9** Density Comparison Plots of the Posterior Distributions of  $m$ 

The X-axis ranges in all density plots in [Figure 9](#) are set to be the same to ensure unbiased comparison. The treatment difference at week 1 is negligible because the two posterior distributions for  $m_{\text{drug},1}$  (**m1**) and  $m_{\text{placebo},1}$  (**m5**) are identical. This is consistent with the data shown in [Figure 7](#), because the majority of the patients (such as the two large groups of subjects who completed the study) report approximately the same amount of score decline in week 1. The difference becomes more significant as the trial goes on, with the **DRUG** group declining at a faster pace. At the end of the trial, the two distributions,  $m_{\text{drug},4}$  (**m4**) and  $m_{\text{placebo},4}$  (**m8**), have the most significant deviation from each other.

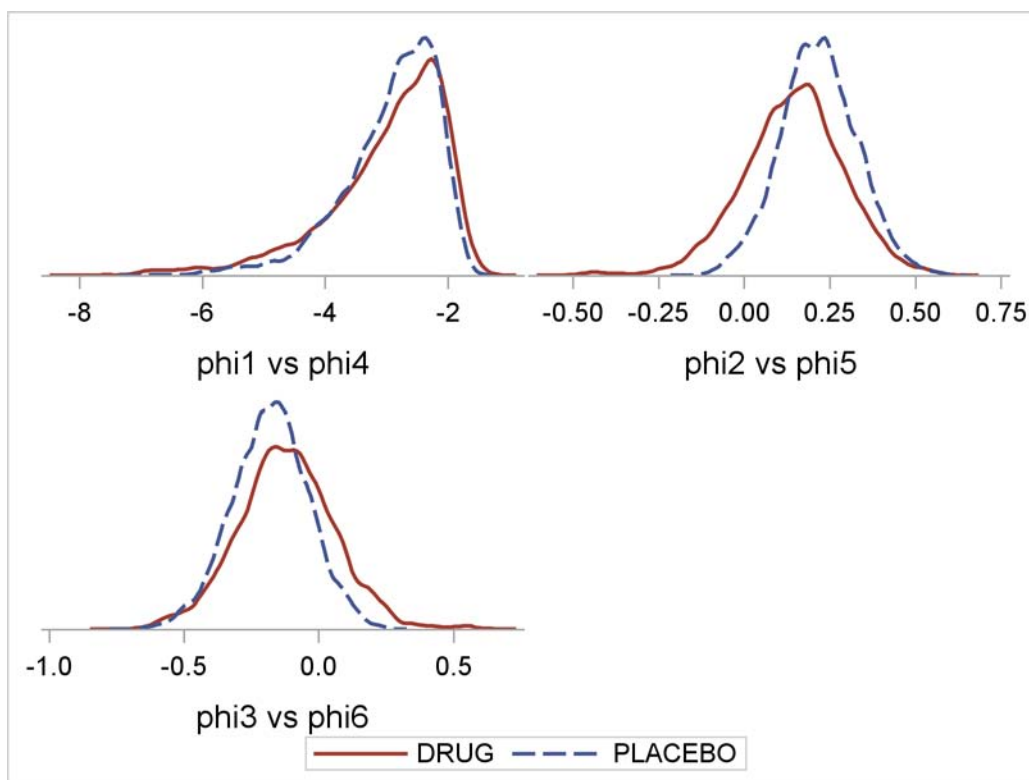
**Figure 10** Density Comparison Plots of the Posterior Distributions of  $\phi$ 

Figure 10 displays pairwise comparison plots of the posterior densities of the **phi** variables, which estimate the change in the probability of dropouts, given the score changes in the last and the current potentially unobserved week. The posterior mean estimates for  $\phi_{\text{drug},2}$  (**phi2**) and  $\phi_{\text{placebo},2}$  (**phi5**) are 0.13 and 0.21, respectively. The positive values suggest that because the patients felt worse (increase in HAMD<sub>17</sub> score) in their previous visit, they were more likely to drop out. The increase in dropout probability is slightly higher for the **PLACEBO** group than for the **DRUG** group. On the other hand, the posterior mean estimates for  $\phi_{\text{drug},2}$  (**phi3**) and  $\phi_{\text{placebo},2}$  (**phi6**) are both negative,  $-0.1$  and  $-0.19$ , respectively. This suggests that the patients were more likely to continue the trial if they felt better in the current week.

One potential focus of interest is the treatment difference (endpoint contrast) between the **DRUG** and **PLACEBO** groups at week 6, the last week of the trial. You estimate the treatment difference by finding the posterior distribution of  $m_{\text{drug},4} - m_{\text{placebo},4}$ , given the data. You can calculate the difference either in the PROC MCMC program or by using the following DATA step:

```
data diffs;
  set seleout;
  diff4=m4-m8;
  p0 = (diff4 > 0);
run;
%postsum(data=diffs, var=diff4 p0);
%postint(data=diffs, var=diff4 p0);
```

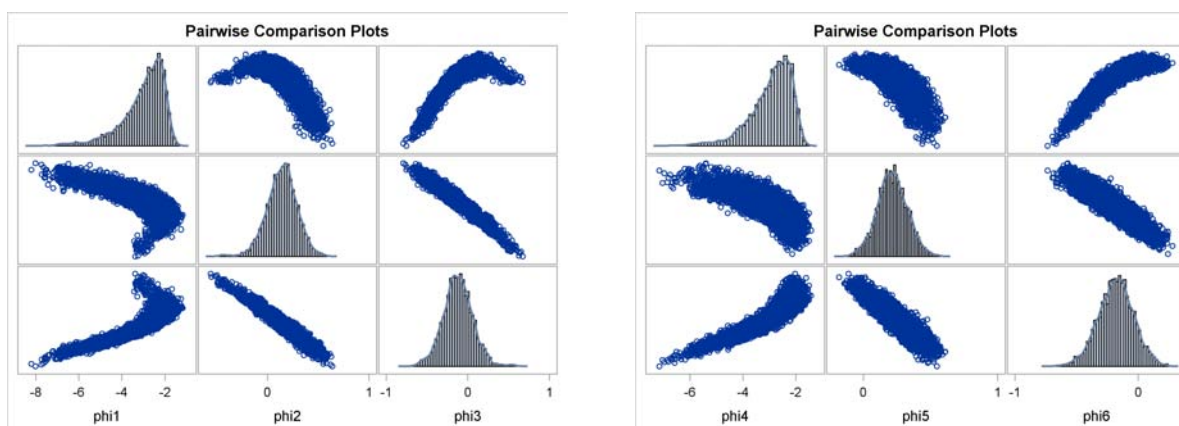
The **diff4** estimate is  $-2.50$  (output not displayed), with a small posterior probability (0.027) that it is greater than 0. This agrees with the graph in Figure 7, where the **DRUG** group patients have a lower HAMD<sub>17</sub> score than the **PLACEBO** group patients at the end of the trial.

### Sensitivity Analysis

The selection model in this section is not without complications. First, a certain amount of optimism is built into this selection model, which aims to estimate regression coefficients for effects from both the previous week and the current week; the latter set of effects actually depend on observations that you do not even see when patients drop out. In addition, the  $\phi$  parameters themselves are in fact difficult to estimate (Molenberghs

and Kenward 2007) because they have nonlinear posterior correlation structures among the parameters; Figure 11 shows pairwise scatter plots of the  $\phi$  estimate in the **DRUG** group (left panel) and the **PLACEBO** group (right panel). This makes the convergence of the chain potentially difficult to achieve.

**Figure 11** Pairwise Scatter Plots of the  $\phi$  Parameters from the **DRUG** and the **PLACEBO** Groups



In practice, you want to perform some type of sensitivity analysis (Kenward 1998; Molenberghs et al. 2001; Jansen et al. 2006). One idea is to fix the values of  $\phi_{\text{drug},2}$  and  $\phi_{\text{placebo},2}$  (Mallinckrodt et al. 2013), the parameters that model the nonignorable missing part of the data, and then examine the sensitivity of estimates from the outcome models, such as the endpoint treatment difference. You can use the BY statement in PROC MCMC to facilitate the estimation of models that have varying  $\phi_{\text{drug},2}$  and  $\phi_{\text{placebo},2}$  values.

The following statements produces seven copies of the same **selection** data set, each with a different combination of  $\phi_{\text{drug},2}$  (**phi3**) and  $\phi_{\text{placebo},2}$  (**phi6**) values:

```
data PhiVals;
  phi3=0;    phi6=0;    model=1; output;
  phi3=0;    phi6=0.2; model=2; output;
  phi3=0;    phi6=-0.2; model=3; output;
  phi3=0;    phi6=-0.4; model=4; output;
  phi3=0.2;  phi6=0;    model=5; output;
  phi3=-0.2; phi6=0;    model=6; output;
  phi3=-0.4; phi6=0;    model=7; output;
run;

data bysele;
  set selection;
  do i = 1 to num;
    set PhiVals nobs=num point=i;
    if model = i then output;
  end;
run;

proc sort data=bysele;
  by phi3 phi6;
run;
```

The program that is required here is very similar to the previous PROC MCMC program, except that now you use a BY statement to repeat the analysis multiple times for different **phi3** and **phi6** combinations. This program has only four **phi** parameters.

The following statements use fixed **phi3** and **phi6** values to fit a series of selection models and then calculate the endpoint contrasts in each scenario:

```
ods output postsummaries=ps postintervals=pi;
proc mcmc data=bysele nmc=50000 seed=176 outpost=byseleout
  diag=none plots=none monitor=(diff4 p0);
```

```

by phi3 phi6;
array Change[4] Change1-Change4;
array mu[4];
array Sigma[4,4];
array S[4,4] ;
array beta[4] ;
array M[2,4] m1-m8;
array phi[2,3] phi1-phi6;
begincnst;
call identity(s);
endcnst;
parms beta: 0 ;
parms m1-m8 0;
parms (phi1 phi2 phi4 phi5) 0;
parms Sigma ;
prior beta: m1-m8 phi1 phi2 phi4 phi5 ~ general(0);
prior Sigma ~ iwish(4, S);
random gamma ~ general(0) subject=poolinv zero=first init=0;
do i=1 to 4;
  if therapy eq "DRUG" then do;
    mu[i] = m[1,i] + gamma + beta[i]*(baseval-18);
  end; else do;
    mu[i] = m[2,i] + gamma + beta[i]*(baseval-18);
  end;
end;
model Change ~ mvn(mu, Sigma);
array r[4] r1-r4;
phi[1,3] = phi3;   phi[2,3] = phi6;
llike = 0;
do i = 2 to wkMax;
  if therapy eq "DRUG" then do;
    p = logistic(phi[1,1] + phi[1,2] * change[i-1] + phi[1,3] * change[i]);
  end; else do;
    p = logistic(phi[2,1] + phi[2,2] * change[i-1] + phi[2,3] * change[i]);
  end;
  llike = llike + lpdfbern(r[i], p);
end;
model r2 r3 r4 ~ general(llike);
beginnodata;
  diff4=m4-m8;
  P0 = (diff4 > 0);
endnodata;
run;

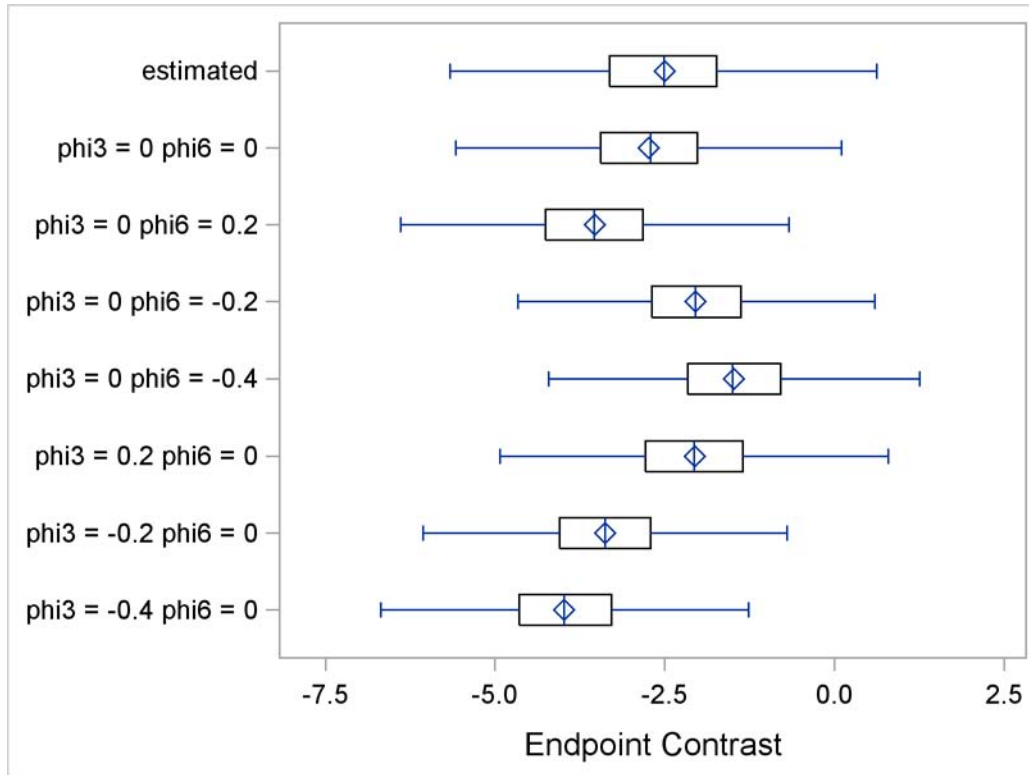
```

Table 3 displays the posterior mean, standard deviation estimates of endpoint contrasts, and the probability that they are greater than 0 for different values of **phi3** and **phi6**. The first row is from the model where **phi3** and **phi6** are estimated.

**Table 3** Sensitivity Analysis Results

<b>phi3</b>	<b>phi6</b>	mean	s.d.	Pr(diff > 0)
-0.12297	-0.1808	-2.50	1.22	0.027
0	0	-2.74	1.06	0.005
0	0.2	-3.53	1.04	0.001
0	-0.2	-2.05	1.01	0.018
0	-0.4	-1.48	1.02	0.069
0.2	0	-2.06	1.06	0.026
-0.2	0	-3.38	1.00	0.0002
-0.4	0	-3.98	1.02	0

Figure 12 shows side-by-side box plots of the posterior distributions of the endpoint contrasts. Only the interquartile ranges are displayed to avoid crowding due to outliers.

**Figure 12** Comparing Selection Models with Different  $\phi$  Values

The estimated model (first box plot) has a mean estimate that is similar to that of the MAR model (second box plot,  $\mathbf{phi3=0}$ ,  $\mathbf{phi6=0}$ ), albeit with a larger measure of uncertainty. Among models with fixed inputs, three cases (the third, seventh, and eighth box plots) have smaller endpoint contrast estimates than the MAR model (shifting to the left). These are the models in which  $\mathbf{phi3}$  is less than  $\mathbf{phi6}$ , reflecting a belief that patients in the **DRUG** group were less likely to drop out than their counterparts in the **PLACEBO** group if they felt improvement in the current week. This assumption translates to stronger treatment effect estimates. The opposite is true for the other three cases (the fourth, fifth, and sixth box plots), where  $\mathbf{phi3}$  takes larger values than  $\mathbf{phi6}$ . The sensitivity analysis illustrates that a small perturbation to the selection mechanism could affect inference, although the effect is relatively mild.

## CONCLUSION

To model missing values in previous releases of the MCMC procedure, you had to manually create a parameter for each missing value, specify its prior distribution, and associate each missing value in the data set variables with its corresponding parameters in order to complete the analysis. Coding was less intuitive and more difficult, and the PROC MCMC run tended to converge more slowly. In SAS/STAT 12.1, PROC MCMC improves these conditions; it offers a complete Bayesian solution by automatically identifying all missing values in the model and incorporating the sampling of these values as part of the Markov chain. The MODEL statement handles missing response variables, missing covariates variables, and partial missingness. You can use the procedure to model all three major types of missing data models: MCAR, MAR, and MNAR. Even in the most complex missing data scenario, the MNAR case, you can choose to model the missing data by using the selection or pattern-mixture models.

## REFERENCES

- Box, G. E. P. and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, New York: John Wiley & Sons.
- Daniels, M. J. and Hogan, J. W. (2008), *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*, Boca Raton, FL: Chapman & Hall/CRC.

- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Series B*, 39, 1–38.
- Diggle, P. and Kenward, M. G. (1994), "Informative Drop-Out in Longitudinal Data Analysis," *Applied Statistics*, 43, 49–73.
- Glynn, R. J., Laird, N. M., and Rubin, D. B. (1986), "Selection Modelling versus Mixture Modelling with Nonignorable Nonresponse," in H. Wainer, ed., *Drawing Inferences from Self-Selected Samples*, 115–142, New York: Springer.
- Goldstein, D. J., Lu, Y., Detke, M. J., Wiltse, C., Mallinckrodt, C., and Demitrack, M. A. (2004), "Duloxetine in the Treatment of Depression: A Double-Blind Placebo-Controlled Comparison with Paroxetine," *Journal of Clinical Psychopharmacology*, 24, 389–399.
- Hamilton, M. (1960), "A Rating Scale for Depression," *Journal of Neurology, Neurosurgery, and Psychiatry*, 23, 56–62.
- Heckman, J. (1976), "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*, 475–492.
- Ibrahim, J. G. and Lipsitz, S. R. (1996), "Parameter Estimation from Incomplete Data in Binomial Regression When the Missing Data Mechanism Is Nonignorable," *Biometrics*, 52, 1071–1078.
- Jansen, I., Hens, N., Molenberghs, G., Aerts, M., Verbeke, G., and Kenward, M. G. (2006), "The Nature of Sensitivity in Monotone Missing Not at Random Models," *Computational Statistics and Data Analysis*, 50, 830–858.
- Kenward, M. G. (1998), "The Effect of Untestable Assumptions for Data with Nonrandom Dropout in Longitudinal Data Analysis," *Statistics in Medicine*, 17, 2723–2732.
- Lipsitz, S. R., Ibrahim, J. G., and Zhao, L. P. (1999), "A Weighted Estimating Equation for Missing Covariate Data with Properties Similar to Maximum Likelihood," *Journal of the American Statistical Association*, 94, 1147–1160.
- Little, R. J. A. (1993), "Pattern-Mixture Models for Multivariate Incomplete Data," *Journal of the American Statistical Association*, 88, 125–134.
- Little, R. J. A. (2009), "Selection and Pattern-Mixture Models," in G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs, eds., *Longitudinal Data Analysis*, Boca Raton, FL: Chapman & Hall/CRC.
- Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, 2nd Edition, Hoboken, NJ: John Wiley & Sons.
- Mallinckrodt, C., Roger, J., Chuang-Stein, C., Molenberghs, G., Lane, P. W., O'Kelly, M., Ratitch, B., Xu, L., Gilbert, S., Mehrotra, D., Wolfinger, R., and Thijs, H. (2013), "Missing Data: Turning Guidance into Action," In review.
- Molenberghs, G. and Kenward, M. G. (2007), *Missing Data in Clinical Studies*, New York: John Wiley & Sons.
- Molenberghs, G., Verbeke, G., Thijs, H., Lesaffre, E., and Kenward, M. G. (2001), "Influence Analysis to Assess Sensitivity of the Dropout Process," *Computational Statistics and Data Analysis*, 37, 93–113.
- Murray, G. D. (1977), "Discussion of Paper by Dempster, Laird, and Rubin," *Journal of the Royal Statistical Society, Series B*, 39, 27–28.
- Neal, R. M. (2003), "Slice Sampling," *Annals of Statistics*, 31, 705–757.
- Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the Royal Statistical Society, Series B*, 89, 846–866.
- Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592.
- Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley & Sons.



- Tan, M. T., Tian, G. L., and Ng, K. W., eds. (2010), *Bayesian Missing Data Problems: EM, Data Augmentation, and Noniterative Computation*, New York: Chapman & Hall/CRC.
- Tanner, M. A. and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–540.
- Ware, J. H., Dockery, S. A., III, Speizer, F. E., and Ferris, B. G., Jr. (1984), "Passive Smoking, Gas Cooking, and Respiratory Health of Children Living in Six Cities," *American Review of Respiratory Diseases*, 129, 366–374.

## ACKNOWLEDGMENT

The author would like to thank James Roger of the London School of Hygiene and Tropical Medicine for his valuable comments and suggestions.

## CONTACT INFORMATION

The MCMC procedure requires SAS/STAT 9.2 and later. In SAS/STAT 12.1, PROC MCMC can automatically model missing values. Complete documentation for the MCMC procedure, in both PDF and HTML format, can be found on the Web at <http://support.sas.com/documentation/onlinedoc/stat/indexproc.html>.

You can find additional coding examples at <http://support.sas.com/rnd/app/examples/index.html>.

Your comments and questions are valued and encouraged. Contact the author at:

Fang Chen  
SAS Institute Inc.  
SAS Campus Drive, Cary, NC 27513  
E-mail: [fangk.chen@sas.com](mailto:fangk.chen@sas.com)

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.