

Bayesian Missing Data Models with the MCMC Procedure

Overview

Missing data are nearly always an issue in longitudinal studies. This example uses the MCMC procedure to fit Bayesian logistic regression models to analyze air pollution data with missing values. A typical missing data strategy is to eliminate observations with missing data, but the resulting inference might not be valid if the observations with missing covariates are substantially different from the observations with complete data.

Analysis

Researchers studied the effects of air pollution on respiratory disease in children. The response variable (Y) represented whether a child exhibited wheezing symptoms; it was recorded as 1 for symptoms exhibited and 0 for no symptoms exhibited. City of residency (X1) and maternal smoking status (X2) were the explanatory variables. The variable X1 was coded as 1 if the child lived in the more polluted city, Steel City, and 0 if the child lived in Green Hills. The X2 variable was the number of cigarettes the mother reported that she smoked per day.

```
data air;
  input y x1 x2;
  datalines;
0 0 0
0 0 0
0 1 0
0 0 0
0 0 11
0 1 7

... more lines ...

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

The MEANS and FREQ procedures provide summary statistics for the response and proposed covariates. Figure 0.1 displays frequencies for the city of residency variable, and Figure 0.2 displays summary statistics for the number of cigarettes smoked per day by the child's mother. Figure 0.3 displays frequencies for the response Y. The covariates X1 and X2 are missing for 4% and 8% of the 390 subjects in the study, respectively.

Figure 0.1 PROC FREQ Summary of City of Residency

The FREQ Procedure	
x1	Frequency
Green Hills	201
Steel City	172
Frequency Missing = 17	

Figure 0.2 PROC MEANS Summary of Cigarettes Smoked

The MEANS Procedure				
Analysis Variable : x2				
N	Mean	Std Dev	Minimum	Maximum
360	4.7666667	5.2604585	0	17.0000000

Figure 0.3 PROC FREQ Summary of Wheezing Symptom Variable

y	Frequency
No Symptoms	289
Symptoms	101

The outputs from the FREQ and MEANS procedures enable you to see a simple summary of the subjects and also to quantify the missingness in the data.

Little and Rubin (2002) discuss many different classifications of missing data. This example illustrates nonignorable missing at random (NMAR) data. Data are said to be NMAR when the missingness depends on the value that would have been observed. NMAR is the most general type of missing data and occurs frequently in longitudinal studies. You should make inference only after the model that is assumed to be correct is specified for the missing data mechanism. The results are then sensitive to those assumptions.

Suppose you want to fit a Bayesian logistic regression model for whether the subject develops wheezing symptoms with density as follows:

$$Y_i \sim \text{binary}(p_i) \quad (1)$$

$$\text{logit}(p_i) = \beta_0 + \beta_1 \cdot X1_i + \beta_2 \cdot X2_i \quad (2)$$

for the $i = 1, \dots, 390$ subjects.

The complete data likelihood function for each of the subjects is

$$p(Y_i | \beta_0, \beta_1, \beta_2, X1_{mis,i}, X2_{mis,i}, X1_{obs,i}, X2_{obs,i}) = \text{binary}(p_i) \quad (3)$$

where $p(\cdot|\cdot)$ denotes a conditional probability density. The binary density is evaluated at the specified value of Y_i and corresponding mean parameter p_i as defined in Equation 2. The three parameters in the complete data likelihood are β_0 , β_1 , and β_2 , which correspond to an intercept, adjustment for living in Steel City, and slope for maternal smoking, respectively.

Next, the covariates $X1$ and $X2$ are written in terms of whether they were missing ($X1_{mis}$ and $X2_{mis}$) or observed ($X1_{obs}$ and $X2_{obs}$). Your goal is to make inferences from the observed data likelihood

$$p(Y_i | \beta_0, \beta_1, \beta_2, X1_{obs,i}, X2_{obs,i}) \quad (4)$$

by multiplying the conditional distribution $p(X1_{mis,i}, X2_{mis,i} | X1_{obs,i}, X2_{obs,i})$ by Equation 3 and integrating over the missing observations. To make inferences from Equation 4, you need to specify a distribution for the missing covariates, $p(X1_{mis,i}, X2_{mis,i} | X1_{obs,i}, X2_{obs,i}, \alpha)$, where α represents the hyperparameters in the missing data distributions. In missing data problems, you can model the missing data mechanism to fit the true missing data mechanism any way you choose. Suppose you specify a joint distribution of $X1$ and $X2$ in terms of the product of a conditional and marginal distribution; that is

$$p(X1_{mis}, X2_{mis} | \alpha) = p(X1_{mis} | X2_{mis}, \alpha_{10}, \alpha_{11}) p(X2_{mis} | \alpha_{20})$$

For this example, say $p(X1_{mis,i} | X2_{mis,i}, \alpha_{10}, \alpha_{11})$ is a logistic regression and $p(X2_{mis,i} | \alpha_{20})$ is a Poisson distribution. You treat the missing covariates as parameters, and you place prior distributions on them and their hyperparameters.

Suppose you place the following prior distributions on the three regression parameters, the missing covariates, and hyperparameters:

$$\pi(\beta_0), \pi(\beta_1), \pi(\beta_2) = \text{normal}(0, \sigma^2 = 10) \quad (5)$$

$$p(X1_{mis,i} | X2_i, \alpha_{10}, \alpha_{11}) = \text{binary}(p_{c,i}) \quad (6)$$

$$\pi(\alpha_{10}), \pi(\alpha_{11}) = \text{normal}(0, \sigma^2 = 10) \quad (7)$$

$$\text{logit}(p_{c,i}) = \alpha_{10} + \alpha_{11} \cdot X2_i \quad (8)$$

$$p(X2_{mis,i} | \alpha_{20}) = \text{Poisson}(e^{\alpha_{20}}) \quad (9)$$

$$\pi(\alpha_{10}), \pi(\alpha_{11}), \pi(\alpha_{20}) = \text{normal}(0, \sigma^2 = 2) \quad (10)$$

where $\pi(\cdot)$ indicates a prior distribution.

The following DATA step creates two variables that identify and index the number of missing covariates. NMISS_X1 increments each time X1 has a missing value, and NMISS_X2 increments each time X2 has a missing value. At the end of the data set, macro variables &NMISS_X1 and &NMISS_X2 are also created to store the total number of missing X1 and X2 covariates, respectively. The %PUT statements enable you to verify that the macro variables have been created successfully.

```
data pollution;
  set air end=eof;
  nmiss_X1 + nmiss(X1);
  nmiss_X2 + nmiss(X2);
  obs = _n_;
  if eof then do;
    call symputx('nmiss_X1', nmiss_X1);
    call symputx('nmiss_X2', nmiss_X2);
  end;
run;

%put &nmiss_X1; %put &nmiss_X2;
```

Using Bayes' theorem, the observed data likelihood function and prior distributions determine the posterior distribution of β_0 , β_1 , and β_2 as follows:

$$\pi(\beta_0, \beta_1, \beta_2 | Y, X1_{obs}, X2_{obs}) \propto p(Y | \beta_0, \beta_1, \beta_2, X1_{obs}, X2_{obs}, X1_{mis}, X2_{mis}) \prod_{i=1}^{n_{X1}} p(X1_{mis,i} | X2_i, \alpha_{10}, \alpha_{11}) \prod_{i=1}^{n_{X2}} p(X2_{mis,i} | \alpha_{20}) \pi(\beta_0) \pi(\beta_1) \pi(\beta_2) \pi(\alpha_{10}) \pi(\alpha_{11}) \pi(\alpha_{20})$$

where n_{X1} and n_{X2} are the macro variables &NMISS_X1 and &NMISS_X2, respectively. PROC MCMC obtains samples from the desired posterior distribution. You do not need to specify the exact form of the posterior distribution.

The odds ratio for comparing Steel City to Green Hills can be written as follows:

$$OR_{X1} = \exp(\beta_1)$$

The odds ratio is useful for interpreting how the odds of developing a wheeze change for a child living in the more polluted city. Similarly, the odds ratio for the maternal smoking effect is written as follows:

$$OR_{X2} = \exp(\beta_2)$$

The following SAS statements use the complete data likelihood function, missing data model assumptions, and prior distributions to fit the Bayesian logistic regression model. The SEED= option

specifies a seed for the random number generator (the seed guarantees the reproducibility of the random stream). The NMC= option specifies the number of posterior simulation iterations. The MISSING=AC specifies that observations with missing values should be included in the analysis. The MONITOR= option outputs analysis on selected symbols of interest in the program.

```
ods graphics on;
proc mcmc data=pollution seed=1181 nmc=10000 missing=ac
  monitor=(beta0 beta1 beta2 orx1 orx2);
  array x1m[&nmiss_x1];
  array x2m[&nmiss_x2];

  parms beta0 -1 beta1 0.1 beta2 0.01;
  parms alpha10 0 alpha11 0 alpha20 0;
  parms x1m1-x1m&nmiss_x1 0;
  parms x2m1-x2m&nmiss_x2 0;

  prior beta: alpha1: ~ normal(0,var=10);
  prior alpha20 ~ normal(0,var=2);
  prior x2m: ~ poisson(exp(alpha20));

  if x2 = . then x2 = x2m[nmiss_x2];

  if obs = 1 then lpm = 0;
  if x1 = . then do;
    x1 = x1m[nmiss_x1];
    p_c = logistic(alpha10 + alpha11*x2);
    lpm = lpm + lpdfbern(x1,p_c);
  end;
  prior x1m: ~ dgeneral(lpm);

  p = logistic(beta0 + beta1*x1 + beta2*x2);
  model y ~ binary(p);

  beginprior;
    orx1 = exp(beta1);
    orx2 = exp(beta2);
  endprior;
run;
ods graphics off;
```

Each of the two ARRAY statements associates a name with a list of variables and constants. The ARRAY statements enable easy referencing of the variables and corresponding values in the procedure. They also create the X1M and X2M parameters for the observations' missing covariates in X1 and X2, respectively. The macro variables &NMISS_X1 and &NMISS_X2 specify the number of missing values for X1 and X2, respectively.

The PARMs statements specify the parameters in the model, create blocks with the respective parameters, and assign initial values to each of them. The PRIOR statements specify priors for all the parameters. The notations beta:, alpha:, X1M:, and X2M: in the PRIOR statements are shorthand for all variables that start with 'beta,' 'alpha,' 'X1M,' and 'X2M,' respectively. The shorthand notation is not necessary, but it keeps your code succinct.

The first IF statement checks whether the covariate X2 is missing. If missing, X2 is replaced with the parameter X2M, which is sampled from the Poisson distribution given in Equation 9. The second IF statement resets the value of the prior log likelihood (LPM) to zero at the top of the data set (that is, when the data set variable OBS is 1). The third IF statement checks whether the value of X1 is missing for each observation. If it is missing, then the three statements inside the DO loop are executed: The missing values of X1 are replaced with the corresponding sampled values of X1M. Then the P_C assignment statement calculates $p_{c,i}$ according to Equation 8. As the MCMC procedure cycles through each observation and finds a missing value, the LPM statement cumulatively adds the logarithm of the probability mass function for the binary model given in Equation 6. Note that X1M is conditional on X2, but not on whether X2 is the sampled or observed value.

The PRIOR statement for the X1M parameters uses the DGENERAL distribution. The letter “D” stands for discrete because the covariate X1 is binary. The GENERAL function indicates that you are using a SAS statement to construct a nonstandard density or distribution. The argument is an expression that takes the value of the logarithm function of the prior or likelihood distribution. The variable LPM is the expression for the sum of the logarithm of probability mass functions given in Equation 6. The DGENERAL function assigns this prior distribution to the X1M missing value parameters.

The P assignment statement calculates p_i in the logistic model, as given in Equation 2. The MODEL statement specifies the complete data likelihood function for Y, as given in Equation 3.

The BEGINPRIOR and ENDPRIOR statements reduce unnecessary observation-level computations. The statements inside the BEGINPRIOR and ENDPRIOR statements create a block of statements that are run only once per iteration rather than for each observation at each iteration. This enables a quick update of the symbols enclosed in the statements. The statements within the BEGINPRIOR and ENDPRIOR block calculate the odds ratios for the two covariates in the model.

Figure 0.4 displays diagnostic plots to assess whether the Markov chains have converged.

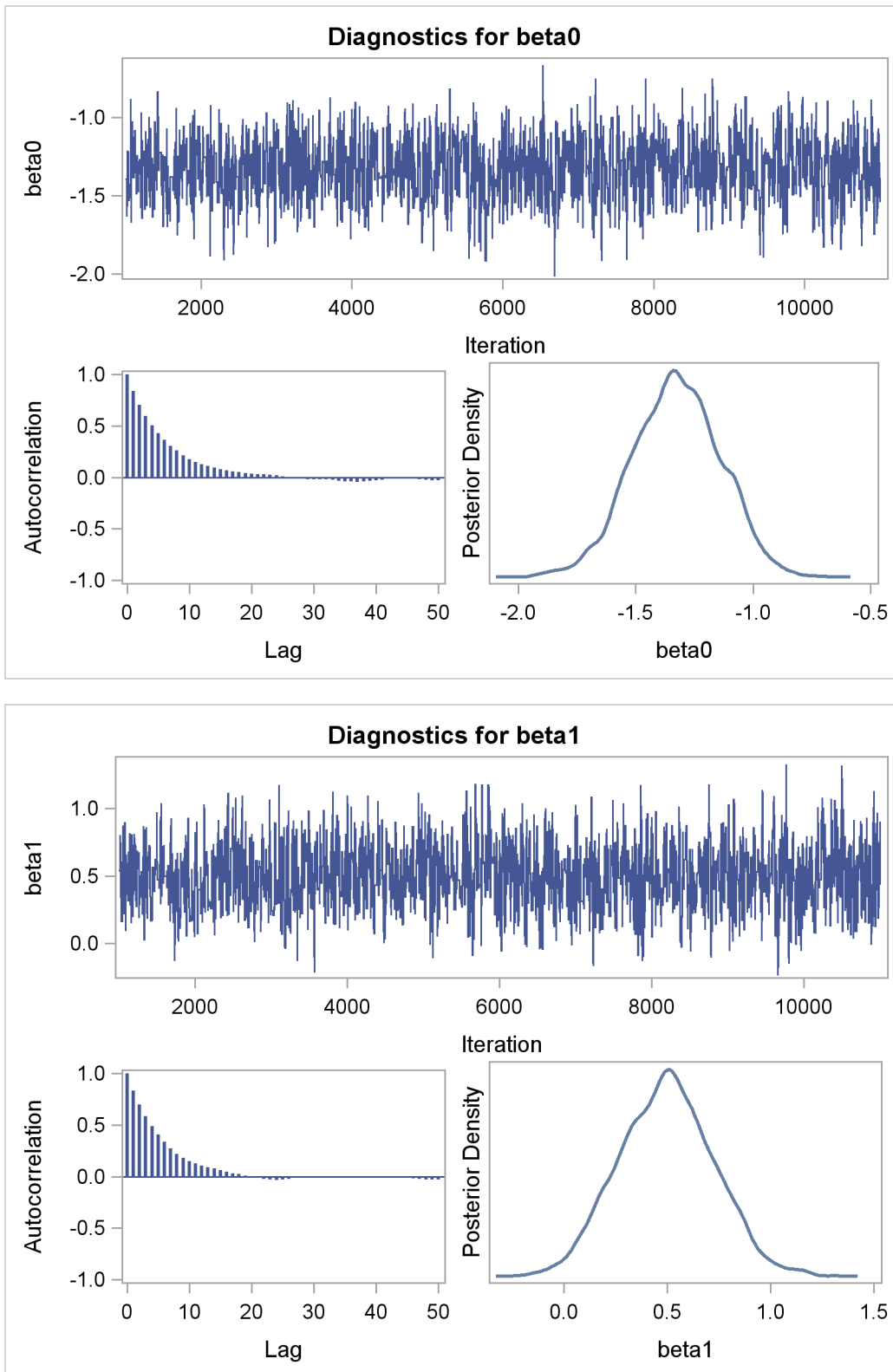
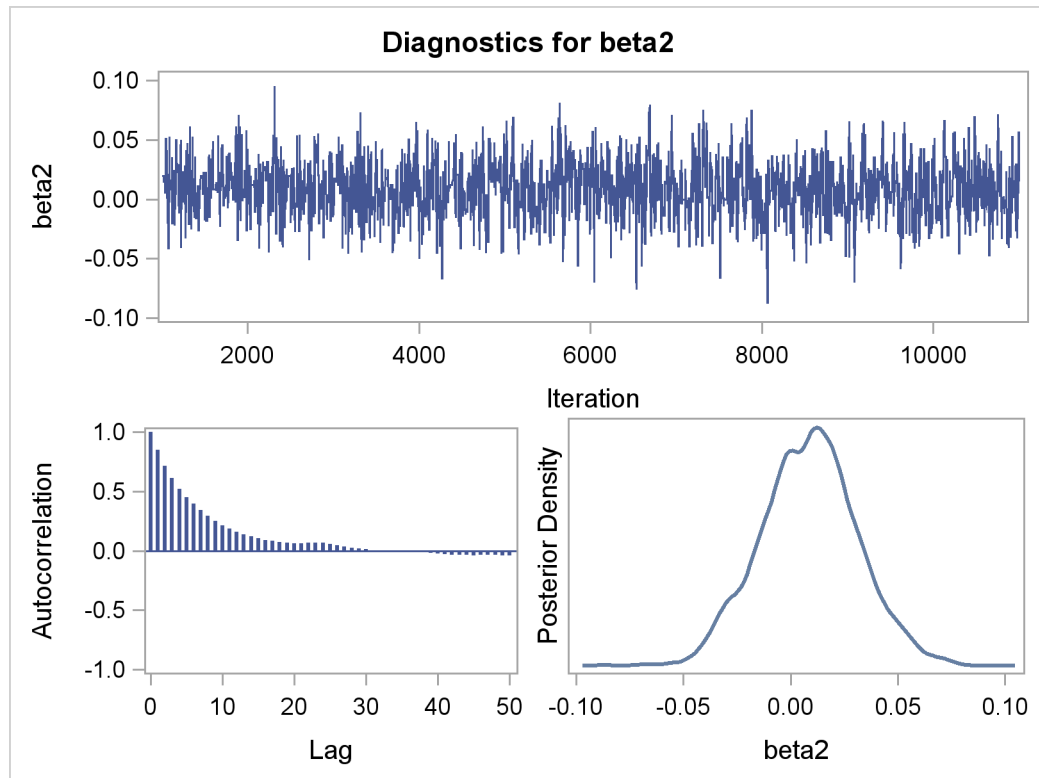
Figure 0.4 Bayesian Logistic Model Diagnostic Plots for β_0 , β_1 , and β_2 

Figure 0.4 *continued*

The trace plots show that the mean of the Markov chain is constant over the graph and is stabilized. The chain was able to traverse the support of the target distribution, and the mixing is good. The trace plots also show that the Markov chain appears to have reached stationary distributions.

The autocorrelation plots indicate low autocorrelation and efficient sampling. The kernel density plots show smooth, unimodal posterior marginal distributions for each parameter.

Figure 0.5 displays a number of convergence diagnostics, including Monte Carlo standard errors, autocorrelations at selected lags, Geweke diagnostics, and the effective sample sizes.

Figure 0.5 Bayesian Logistic Model MCMC Convergence Diagnostics

The MCMC Procedure			
Monte Carlo Standard Errors			
Parameter	MCSE	Standard Deviation	MCSE/SD
beta0	0.00634	0.1862	0.0341
beta1	0.00736	0.2275	0.0324
beta2	0.000826	0.0226	0.0366
orx1	0.0126	0.3914	0.0321
orx2	0.000834	0.0228	0.0366

Figure 0.5 *continued*

Posterior Autocorrelations				
Parameter	Lag 1	Lag 5	Lag 10	Lag 50
beta0	0.8390	0.4279	0.1762	-0.0306
beta1	0.8353	0.4088	0.1474	-0.0301
beta2	0.8492	0.4517	0.2147	-0.0390
orx1	0.8313	0.3983	0.1414	-0.0234
orx2	0.8492	0.4520	0.2155	-0.0394

Geweke Diagnostics		
Parameter	z	Pr > z
beta0	-0.7828	0.4338
beta1	-1.3553	0.1753
beta2	2.0812	0.0374
orx1	-1.7241	0.0847
orx2	2.0430	0.0410

Effective Sample Sizes			
Parameter	ESS	Correlation Time	Efficiency
beta0	861.5	11.6071	0.0862
beta1	955.0	10.4708	0.0955
beta2	746.7	13.3916	0.0747
orx1	972.3	10.2854	0.0972
orx2	745.9	13.4070	0.0746

Figure 0.6 displays the “Number of Observations” table. It lists the number of observations read from the DATA= data set and the number of observations used in the analysis. If the MISSING=AC statement were omitted from the PROC MCMC statement, the number of observations used in the analysis would be less than the number of observations read because only complete cases would be used in the analysis.

Figure 0.6 Bayesian Logistic Model Number of Observations

The MCMC Procedure	
Number of Observations Read	390
Number of Observations Used	390

Figure 0.7 displays summary and interval statistics for each parameter’s posterior distribution.

Figure 0.7 Posterior Model Summary of Bayesian Logistic Regression

The MCMC Procedure						
Posterior Summaries						
Parameter	N	Mean	Standard Deviation	Percentiles		
				25%	50%	75%
beta0	10000	-1.3257	0.1862	-1.4543	-1.3279	-1.2006
beta1	10000	0.4975	0.2275	0.3403	0.4999	0.6457
beta2	10000	0.00859	0.0226	-0.00619	0.00917	0.0233
orx1	10000	1.6879	0.3914	1.4053	1.6486	1.9074
orx2	10000	1.0089	0.0228	0.9938	1.0092	1.0236

Posterior Intervals					
Parameter	Alpha	Equal-Tail Interval		HPD Interval	
beta0	0.050	-1.6967	-0.9743	-1.7016	-0.9827
beta1	0.050	0.0648	0.9477	0.0511	0.9153
beta2	0.050	-0.0348	0.0532	-0.0348	0.0532
orx1	0.050	1.0670	2.5797	1.0349	2.4643
orx2	0.050	0.9658	1.0547	0.9639	1.0527

The odds ratio for X1 is the multiplicative change in the odds of a child wheezing in Steel City compared to the odds of the child wheezing in Green Hills. The estimated odds ratio (ORX1) value is 1.6879 with a corresponding 95% credible interval of (1.0670, 2.5797). City of residency is a significant factor in a child's wheezing status. The estimated odds ratio for X2 is the multiplicative change in the odds of developing a wheeze for each additional reported cigarette smoked per day. The odds ratio of ORX2 indicates that the odds of a child developing a wheeze is 1.0089 times higher for each reported cigarette a mother smokes. The corresponding 95% credible interval is (0.9658, 1.0547). Since this interval contains the value 1, maternal smoking is not considered to be a influential effect.

The MCMC procedure is best suited for models with relatively few parameters although there are no formal limitations on the number of parameters you can specify. Data sets with many missing observations will be computationally expensive with the MCMC procedure and might have a long run time.

References

Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, Second Edition, Hoboken, NJ: Wiley.