HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2017 November 30.

Published in final edited form as:

Stat Med. 2016 November 30; 35(27): 5029–5039. doi:10.1002/sim.7040.

A Note on Posterior Predictive Checks to Assess Model Fit for Incomplete Data

Dandan Xu[1], **Arkendu Chatterjee**[2], and **Michael J. Daniels**[3,*]

[1]Department of Statistics, University of Florida, Gainesville, FL 32611

[2]Novartis, East Hanover, NJ 07936

[3]Department of Statistics & Data Sciences and Department of Integrative Biology, The University of Texas, Austin, TX 78712

Abstract

We examine two posterior predictive distribution based approaches to assess model fit for incomplete longitudinal data. The first approach assesses fit based on replicated complete data as advocated in Gelman et al. (2005). The second approach assesses fit based on replicated observed data. Differences between the two approaches are discussed and an analytic example is presented for illustration and understanding. Both checks are applied to data from a longitudinal clinical trial. The proposed checks can easily be implemented in standard software like (Win)BUGS/JAGS/Stan.

Keywords

Extrapolation factorization; Missing data; Nonignorable missing data; Model diagnostics; Posterior predictive distribution

1 Introduction

The posterior predictive distribution for replicated data $\boldsymbol{y}_{\text{rep}}$ under a data model, $p(\boldsymbol{y}|\boldsymbol{\theta})$ with prior $p(\boldsymbol{\theta})$ is given by

$$p\left(\boldsymbol{y}_{\text{rep}}|\boldsymbol{y}\right)=\int p\left(\boldsymbol{y}_{\text{rep}}|\theta\right)p\left(\theta|\boldsymbol{y}\right)d\theta,$$

where $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ is the posterior distribution of $\boldsymbol{\theta}$. Samples from the posterior predictive distribution are replicates of the data generated by the model. In this paper we will discuss approaches for Bayesian model checking for models for incomplete data (so $\boldsymbol{y}$ is not completely observed) based on the posterior predictive distribution. We first review the relevant literature on posterior predictive checks.

* mjdaniels@austin.utexas.edu.

Supplementary Materials

The supplementary materials contain detailed derivations of the probabilities in Section 3 and WinBUGS code to sample from the posterior distribution of the parameters and to compute the posterior predictive checks in Sections 3 and 4. R code for the figures in Section 3 are available as a separate file.

1.1 Model Fit for Complete Data

Rubin et al. [1] first used the posterior predictive distribution of a statistic to calculate the tail-area probability corresponding to the observed value of a test statistic. Meng [2] called this probability a posterior predictive p-value. In following, we will refer to it as posterior predictive probability due to its problematic interpretation as a p-value [3]. This probability is a measure of discrepancy between the observed data and the posited modeling assumptions as measured by a summary quantity $T(\cdot)$. The posterior predictive distribution of $T(\boldsymbol{y}_{\text{rep}})$ can identify problems when the wrong model is fitted on the data and compared with (the distribution of) $T(\boldsymbol{y})$. For the assumed model, the posterior predictive approach provides a reference distribution. The fit of the model to the data is assessed by comparing the posterior predictive distribution of $T(\boldsymbol{y}_{\text{rep}})$ with $T(\boldsymbol{y})$. Meng [2] formally defined this probability as,

$$P\left(T\left(\boldsymbol{y}_{\text{rep}}\right)>T\left(\boldsymbol{y}\right)|\boldsymbol{y}\right)=\int\int I\left\{\left[T\left(\boldsymbol{y}_{\text{rep}}\right)>T\left(\boldsymbol{y}\right)\right]\right\}p\left(\boldsymbol{y}_{\text{rep}}|\theta\right)p\left(\theta|\boldsymbol{y}\right)d\boldsymbol{y}_{\text{rep}}d\theta. \quad (1)$$

In the Bayesian formulation this approach also allows the use of a parameter dependent test statistic, called a discrepancy statistic [2, 4]. For a discrepancy, $D(\boldsymbol{y}; \boldsymbol{\theta})$, the reference distribution can be computed from the joint distribution of $(\boldsymbol{y}_{\text{rep}}, \boldsymbol{\theta})$,

$$p\left(\boldsymbol{y}_{\text{rep}},\theta|\boldsymbol{y}\right)=p\left(\boldsymbol{y}_{\text{rep}}|\theta\right)p\left(\theta|\boldsymbol{y}\right).$$

However locating the realized value of $D(\boldsymbol{y}; \boldsymbol{\theta})$ within the reference distribution is not feasible since $D(\boldsymbol{y}; \boldsymbol{\theta})$ depends on the unknown $\boldsymbol{\theta}$. This complication has led authors to use the tail area probability of D under its posterior reference distribution. Gelman et al. [5] constructed a probability as in (1) but eliminated the dependence on unknown $\boldsymbol{\theta}$, by integrating out $\boldsymbol{\theta}$ with respect to its posterior distribution. The tail area probability of the discrepancy statistic is then given by

$$P\left(D\left(\boldsymbol{y}_{\text{rep}};\theta\right)\geq D\left(\boldsymbol{y};\theta\right)|\boldsymbol{y}\right)$$

This is analogous to the posterior predictive probability in (1). The choice of test or discrepancy statistic is clearly very important and often reflects the inferential interests. In general, these checks are called posterior predictive checks.

1.2 Incomplete Data Model Fit

1.2.1 Notation and Review—To introduce posterior predictive checks for incomplete longitudinal data, we need to first introduce some notation and concepts. Let $\boldsymbol{Y}_i : i = 1, \ldots, n$ denote the J-dimensional longitudinal response vector (with components $Y_{ij} : j = 1, \ldots, J$) for individual i and $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$. Let $\boldsymbol{R}$ be the vector, ordered as $\boldsymbol{Y}$, of observed data indicators; i.e., $R_{ij} = I\{Y_{ij}$ is observed$\}$ and let $\boldsymbol{Y}_{\text{obs}}$ be $\{Y_{ij} : r_{ij} = 1\}$. The *full data* is given as $(\boldsymbol{y}, \boldsymbol{r})$; the *observed data* as $(\boldsymbol{y}_{\text{obs}}, \boldsymbol{r})$. The extrapolation factorization of the full data model is,

$$p(\boldsymbol{y}, \boldsymbol{r}|\omega) = p(\boldsymbol{y}_{\text{mis}}, \boldsymbol{y}_{\text{obs}}, \boldsymbol{r}|\omega) = p(\boldsymbol{y}y_{\text{mis}}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}, \omega_E)\, p(\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}|\omega_O), \quad (2)$$

where $\boldsymbol{\omega}_E$ indexes the conditional distribution of missing responses given observed data (the extrapolation distribution) and $\boldsymbol{\omega}_O$ indexes the distribution of the observed data. The parameters $\boldsymbol{\omega}_E$ and $\boldsymbol{\omega}_O$ are both functions of $\boldsymbol{\omega}$ and can be overlapping, often with $\boldsymbol{\omega}_E$ containing a subset of $\boldsymbol{\omega}_O$ (see Section 3 for an example). Inference about the full data distribution, $p(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{\omega})$, and the full data response model, $p(\boldsymbol{y}|\boldsymbol{\theta}(\boldsymbol{\omega}))$, clearly requires unverifiable assumptions about the extrapolation distribution $p(\boldsymbol{y}_{\text{mis}}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}, \boldsymbol{\omega}_E)$ for which the observed data provide no information. Sensitivity parameters are functions of $\boldsymbol{\omega}_E$ [6] and are used to introduce (external) information about the missing data mechanism.

As an example, consider a bivariate normal model (similar to the data example in Section 4). With only missingness in the second measurement, the extrapolation distribution is $p(y_2|y_1, r = 0)$ and its parameters are $\boldsymbol{\omega}_E$; the observed data distribution consists of the following four components: $p(y_1|r = 0)$, $p(y_1|r = 1)$, $p(r)$, and $p(y_2|y_1, r = 1)$ and $\boldsymbol{\omega}_O$ are the parameters of these four distributions.

1.2.2 Current status quo—Gelman et al. [7] proposed an extension of the posterior predictive approach to the setting of missing and latent data. To assess the fit of the model they defined a test statistic $T(\cdot)$, which is a function of the complete data. The 'missing' data, either truly missing or latent, were filled in at each iteration using data augmentation. The test statistic was compared to the test statistic computed based on replicated complete data. Graphical approaches were implemented for model checking along with the calculation of posterior predictive probabilities.

1.2.3 Problems with status quo—In our setting of missing data, in particular non-ignorable missingness, the current checks based on replicated complete data are problematic in that the checks will provide different evidence about the fit of the model to the observed data by varying sensitivity parameters (which are not identified by the observed data). This is an issue since sensitivity parameters a necessary component for the analysis of missing data [8]. We provide more details in Section 2.

1.3 Layout of the paper

In Section 2, we provide further details on posterior predictive checks for incomplete longitudinal data, point out the problems with replicated complete data checks of Gelman et al. [7], and propose an alternative. In Section 3, we provide an analytic example of the two approaches. We demonstrate the checks on a data example in Section 4. Finally, we provide conclusions, recommendations, and extensions in Section 5.

2 Posterior Predictive Checks for Incomplete Data

In this section we explore posterior predictive checks for incomplete longitudinal data. To implement these checks, we will sample from the posterior predictive distribution, $p(\boldsymbol{y}^{\text{rep}}, \boldsymbol{r}^{\text{rep}}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r})$, though the complete data checks will ignore $\boldsymbol{r}^{\text{rep}}$. When we model the missing

data mechanism, as in nonignorable missingness, we can compute data summaries based on replicates of observed data and replicates of complete data. Gelman et al. [7] proposed doing checks using complete data. However Gelman considered more general settings that include latent variables (as missing data), ignorable missingness and nonignorable missingness. We focus on the nonignorable case for which we will argue complete data checks are not appropriate (at least for assessing model fit to the observed data).

2.1 Complete Data Replications

We now review replicated complete data and the corresponding posterior predictive checks. The 'data' for these checks are sampled from $p(\boldsymbol{y}^{\text{rep}}, \boldsymbol{r}^{\text{rep}}, \boldsymbol{y}_{\text{mis}}|\boldsymbol{\omega}, \boldsymbol{y}_{\text{obs}}, \boldsymbol{r})$ (and $\boldsymbol{r}^{\text{rep}}$ is ignored). For each sample from the above distribution, the complete data is defined as $(\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}})$ and the replicated complete data are $\boldsymbol{y}^{\text{rep}}$; the dimension of $\boldsymbol{y}^{\text{rep}}$ is the same as $(\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}})$.

To assess the fit, we choose a test statistic of interest, $T_c(\cdot)$, which we evaluate at each sample of complete data, $(\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}})$ and replicated complete data, $\boldsymbol{y}^{\text{rep}}$. We then compute the following posterior predictive probability

$$P\left(T_c\left(\boldsymbol{Y}^{\text{rep}}\right) > T_c\left(\boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}}\right)|\boldsymbol{y}_{\text{obs}}, \text{r}\right) = \int\int I\left\{\left[T_c\left(\boldsymbol{y}^{\text{rep}}\right) > T_c\left(\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}\right)\right]\right\} p\left(\boldsymbol{y}^{\text{rep}}|\omega\right) p\left(\boldsymbol{y}_{\text{mis}}|\omega, \boldsymbol{y}_{\text{obs}}, \boldsymbol{r}\right) p\left(\omega|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}\right) d\boldsymbol{y}^{\text{rep}} d\boldsymbol{y}_{\text{mis}} d\omega.$$

In the above, we have implicitly integrated over $\boldsymbol{r}^{\text{rep}}$.

2.2 Observed Data Replications

We now introduce replicated observed data. These are sampled from $p(\boldsymbol{y}^{\text{rep}}, \boldsymbol{r}^{\text{rep}}|\boldsymbol{\omega}, \boldsymbol{y}_{\text{obs}}, \boldsymbol{r}$ and defined as

$$\boldsymbol{y}_{\text{obs}}^{\text{rep}} = \left\{y_j^{\text{rep}} : r_j^{\text{rep}} = 1\right\},$$

i.e., the components of $\boldsymbol{y}^{\text{rep}}$ (the replicated complete datasets) for which the corresponding replicated missing data indicators, $\boldsymbol{r}^{\text{rep}}$ are equal to one. Note that samples of $\boldsymbol{r}^{\text{rep}}$ will not exactly match $\boldsymbol{r}$. As such, it is good to use a statistic that is 'normalized' based on the dimension of the (replicated) observed data (e.g., a mean).

To assess the fit, we choose a test statistic of interest, $T_o(\cdot)$, which we evaluate at $\boldsymbol{y}_{\text{obs}}$ and each sample from the posterior predictive distribution of $\boldsymbol{y}_{\text{obs}}^{\text{rep}}$. We then compute the following posterior predictive probability

$$P\left(T_o\left(\boldsymbol{Y}_{\text{obs}}^{\text{rep}}\right) > T_o\left(\boldsymbol{Y}_{\text{obs}}\right)|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}\right) = \int\int I\left\{\left[T_o\left(\boldsymbol{y}_{\text{obs}}^{\text{rep}}\right) > T_o\left(\boldsymbol{y}_{\text{obs}}\right)\right]\right\} p\left(\boldsymbol{y}^{\text{rep}}|\boldsymbol{r}^{\text{rep}}, \omega, \boldsymbol{y}_{\text{obs}}, \boldsymbol{r}\right) p\left(\omega|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}\right) d\boldsymbol{y}^{\text{rep}} dF\left(\boldsymbol{r}^{\text{rep}}|\omega, \boldsymbol{y}_{\text{obs}}, \boldsymbol{r}\right) d\omega,$$

(3)

where $y_{\text{obs}}^{\text{rep}}$ is a function of ($\boldsymbol{y}^{\text{rep}}$, $\boldsymbol{r}^{\text{rep}}$), F is the conditional cdf for $\boldsymbol{r}^{\text{rep}}$, and typically $p(\boldsymbol{y}^{\text{rep}}|\boldsymbol{r}^{\text{rep}}, \boldsymbol{\omega}, \boldsymbol{y}_{\text{obs}}, \boldsymbol{r}) = p(\boldsymbol{y}^{\text{rep}}|\boldsymbol{r}^{\text{rep}}, \boldsymbol{\omega})$. Computational details for complete and observed data replications and checks are given in the next subsection.

2.3 Posterior computations

Here, we provide details on the steps for generating both complete and observed data replications.

At iteration k,

1. Sample $\boldsymbol{\omega}^{(k)}$ from the observed data posterior, $p(\boldsymbol{\omega}^{(k)}|\boldsymbol{y}_{\text{obs}},\boldsymbol{r})$ or from the data augmented posterior, $p\left(\omega^{(k)}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}^{(k-1)}, \boldsymbol{r}\right)$ whichever is simpler (the later uses data augmentation explicitly)
2. Sample $\boldsymbol{y}_{\text{mis}}^{(k)}$ from $p\left(\boldsymbol{y}_{\text{mis}}^{(k)}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}, \omega^{(k)}\right)$; this step is only needed for the complete data replications, but can simplify step 1 for the observed data replications via data augmentation.
3. Sample replicated data from $p\left(\boldsymbol{y}^{\text{rep}^{(k)}}, \boldsymbol{r}^{\text{rep}^{(k)}}|\omega^{(k)}\right)$
   (a) Complete data replication: Keep $\boldsymbol{y}^{\text{rep}^{(k)}}$
   (b) Observed data replication: Keep $\left(\boldsymbol{y}^{\text{rep}^{(k)}}, \boldsymbol{r}^{\text{rep}^{(k)}}\right)$
4. Compute summary quantities
   (a) Complete data replication: Compute the summary quantities $T_c(\cdot)$ and $I\{T_c(\boldsymbol{y}^{\text{rep}(k)}) > T_c(\boldsymbol{y}^{(k)})\}$, where $\boldsymbol{y}^{(k)} = \left(\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}^{(k)}\right)$
   (b) Observed data replication: Compute the summary quantities $T_o(\cdot)$ and $I\left\{T_o\left(\boldsymbol{y}_{\text{obs}}^{\text{rep}^{(k)}}\right) > T_o\left(\boldsymbol{y}_{\text{obs}}\right)\right\}$ where $\boldsymbol{y}_{\text{obs}}^{\text{rep}^{(k)}} = \left\{y_j^{\text{rep}^{(k)}} : r_j^{\text{rep}^{(k)}} = 1\right\}$

For the complete data replications, estimate the desired probability using the empirical average of $I\left\{T_c\left(\boldsymbol{y}^{\text{rep}^{(k)}}\right) > T_c\left(\boldsymbol{y}^{(k)}\right\}\right.$ across all iterations. For the observed data replications, estimate the desired probability using the empirical average of $I\left\{T_o\left(\boldsymbol{y}_{\text{obs}}^{\text{rep}^{(k)}}\right) > T_o\left(\boldsymbol{y}_{\text{obs}}\right)\right\}$ across all iterations.

2.4 Issues

Posterior predictive checks based on replications of complete data have some advantages. For example, under ignorable missingness, which does not require explicit specification of the missing data mechanism, it is not a problem to create replications of the complete data. But of course, in that situation, it is not possible to assess the joint fit of the observed responses *and* the missingness indicators and checks condition on the observed missingness

indicators. Obviously, this approach loses some power versus a setting with no missing data since the missing data are filled in (via data augmentation) under the assumed model; thus, slightly biasing the checks in favor of the model. However, in the setting of nonignorable missingness, the checks have a fatal flaw as they are *not* invariant to the extrapolation distribution [9] and are not in the spirit of sensitivity analysis (an essential part of the analysis of missing data as documented in a recent NRC report [8]). To be more explicit, two models with the same fit to the observed data, i.e., the same $p(\mathbf{y}_{obs}, \mathbf{r}|\boldsymbol{\omega}_O)$ but different (implicit) extrapolation distributions, $p(\mathbf{y}_{\text{mis}}|\mathbf{y}_{\text{obs}}, \mathbf{r}, \boldsymbol{\omega}_E)$ can give different conclusions on model fit using posterior predictive checks based on replicated complete data, i.e., the values of posterior predictive probabilities change with different extrapolation distributions. This is not a desirable property for a check designed to assess model fit to the *observed* data.

Checks based on replicated observed data satisfy the property of invariance to the extrapolation distribution and provide the same conclusions as the extrapolation distribution (i.e., sensitivity parameters are) is varied (unlike checks based on replicated complete data), i.e., different nonignorable models with the same fit to the observed data. In addition, they can assess any features of the joint distribution of $(\mathbf{y}_{\text{obs}}, \mathbf{r})$ as desired. However, one would surmise that some power is lost relative to checks with no missing data (and possibly complete data replication checks) given a lack of one-to-one correspondence between the observed data responses and the replicated observed data responses; for example, Y_{ij} might be observed in the actual data, but is not necessarily 'observed' in the replicated data.

We will explore these checks further in practice using a simple analytical example in the next section.

3 Analytical Example

In this section we examine analytically the behavior of the posterior predictive probability based on observed data and complete data replications under a simple, illustrative model: a mixture model for a bivariate response given by

$$\begin{aligned} Y_1|R_2{=}k &\sim N\left(\mu{+}\xi I\left\{k{=}1\right\}, \sigma^2_{k+1}\right) : k{=}0,1 \\ Y_2|Y_1, R_2{=}0 &\sim N\left(\alpha{+}\Delta{+}\phi Y_1, \tau^2\right) \\ Y_2|Y_1, R_2{=}1 &\sim N\left(\alpha{+}\phi Y_1, \tau^2\right) \\ R_2 &\sim Ber\left(\eta\right), \end{aligned} \tag{4}$$

where there is only missingness in Y_2 and [$Y_2|Y_1$, $R_2 = 0$] is the extrapolation distribution. Here, $\omega_O = \left(\mu, \xi, \sigma_1^2, \sigma_2^2, \alpha, \phi, \tau^2, \eta\right)$ and $\omega_E = \left(\alpha, \phi, \tau^2, \Delta\right)$; note that we have set some of the parameters in the extrapolation distribution equal to corresponding observed data parameters. Δ is the only parameter that *only* appears in the extrapolation distribution. Δ is a sensitivity parameter that measures departures from MAR; Δ = 0 corresponds to MAR. The purpose of this is to better understand and illustrate the behavior of both checks described in Section 2 in a simple setting.

3.1 Derivation of posterior predictive probabilities

Let n be the total number of subjects and n_1 the number of subjects with $r_2 = 1$. Assume the data are sorted so that all the units with missing Y_2 are at the end. Let

$$\bar{y}_{2,\text{obs}}=\frac{\sum_{i=1}^{n_1} y_{2i}}{n_1}, \bar{y}_{1,\text{obs}}=\frac{\sum_{i=1}^{n_1} y_{1i}}{n_1} \text{ and } \bar{y}_{1,\text{mis}}=\frac{\sum_{i=n_1+1}^{n} y_{1i}}{n-n_1}.$$

We specify a (diffuse) normal prior on the regression parameter, a, $a \sim N(0, \nu^2)$, and to simplify the below derivation, we assume the remaining parameters are known. The test statistic for the complete data replication, $T_c(\cdot)$, and the observed data replication, $T_o(\cdot)$ are defined as the corresponding means of Y_2 (details in what follows). For the observed data replications approach, $T_o(\cdot)$ is evaluated at the observed data at time 2, $T_o(\boldsymbol{y}_{\text{obs}}) = \bar{Y}_{2,\text{obs}}$, and the replicated observed data at time 2, $T_o(\boldsymbol{y}^{\text{rep}}_{\text{obs}}) = \bar{Y}^{\text{rep}}_{2,\text{obs}}$. The corresponding (posterior predictive) tail area probability can be approximated as

$$\begin{aligned} P\left(\bar{y}^{\text{rep}}_{2,\text{obs}} > \bar{y}_{2,\text{obs}} | \boldsymbol{y}_{\text{obs}}, \boldsymbol{r}_2\right) &\approx \Phi\left(\frac{-\frac{\tau^2}{\tau^2+n_1\nu^2}\sqrt{n_1}\left(\bar{y}_{2,\text{obs}}-\phi(\mu+\xi)\right)-\frac{n_1\nu^2}{\tau^2+n_1\nu^2}\phi\sqrt{n_1}\left(\bar{y}_{1,\text{obs}}-(\mu+\xi)\right)}{\sqrt{\frac{n_1}{n\eta}\left(\tau^2+\phi^2\sigma_2^2\right)+\frac{n_1\tau^2\nu^2}{\tau^2+n_1\nu^2}}}\right) \\ &= \Phi\left(\frac{-o_p(1)A-O_p(1)\phi B}{\sqrt{O_p(1)\left(\tau^2+\phi^2\sigma_2^2\right)+O_p(1)\tau^2}}\right) \\ &= \Phi\left(\frac{m}{\sqrt{V}}\right), \end{aligned} \tag{5}$$

where

$$A=\sqrt{n_1}\left(\bar{y}_{2,\text{obs}}-\phi\left(\mu+\xi\right)\right), \quad B=\sqrt{n_1}\left(\bar{y}_{1,\text{obs}}-\left(\mu+\xi\right)\right), m=-o_p\left(1\right)A-O_p\left(1\right)B \text{ and}$$

$$V=\sqrt{O_p\left(1\right)\left(\tau^2+\phi^2\sigma_2^2\right)+O_p\left(1\right)\tau^2}.$$

For the complete data replication approach, $T_c(\cdot)$ is evaluated at the completed data at time 2, $T_c(\boldsymbol{y}) = \bar{Y}_{2,\text{com}}$ and the replicated complete data at time 2, $T_c(\boldsymbol{y}^{\text{rep}}) = \bar{Y}^{\text{rep}}_{2,\text{com}}$. The corresponding tail area probability can be approximated as

$$\begin{aligned} &P\left(\bar{y}^{\text{rep}}_{2,\text{com}} > \bar{y}_{2,\text{com}} | \boldsymbol{y}_{\text{obs}}, \boldsymbol{r}_2\right) \\ &\approx \Phi\left(\frac{m+\sqrt{\frac{n}{n_1}}\frac{n_1-n\eta}{\sqrt{n}}(\Delta-\phi\xi)+\sqrt{\frac{n-n_1}{n_1}}\phi\sqrt{n-n_1}\left(\mu-\bar{y}_{1,\text{mis}}\right)}{\sqrt{V+\frac{n\eta}{n_1}(1-\eta)(\Delta-\phi\xi)^2+\tau^2\left(\frac{2n-n_1}{n_1}-\frac{n_1}{n\eta}\right)+\phi^2\sigma_2^2\left(\frac{n\eta}{n_1}-\frac{n_1}{n\eta}\right)+\phi^2\sigma_1^2\frac{n-n\eta}{n_1}}}\right) \\ &= \Phi\left(\frac{m+O_p(1)\frac{1}{\sqrt{\eta}}(\Delta-\phi\xi)C-O_p(1)\sqrt{\frac{1-\eta}{\eta}}\phi D}{\sqrt{V+O_p(1)(1-\eta)(\Delta-\phi\xi)^2+O_p(1)\frac{2(1-\eta)}{\eta}\tau^2+o_p(1)\phi^2\sigma_2^2+O_p(1)\frac{1-\eta}{\eta}\phi\sigma_1^2}}\right), \end{aligned} \tag{6}$$

where $C=\frac{n_1-n\eta}{\sqrt{n}}$ and $D=\sqrt{n-n_1}\left(\bar{y}_{1,\text{mis}}-\mu\right)$. Details on the derivation of both these probabilities can be found in the supplementary materials.

3.2 Comparison and implications

We will examine two main properties of the posterior predictive checks: 'power' and 'Type I error'. In what follows, we will define 'power' as the ability of the posterior predictive check to detect model departures or inadequacies. Good 'power' results when the posterior predictive probabilities approach either zero or one when there are model departures/ inadequacies. On the other hand, 'Type I error' refers to the situation of having probabilities approaching zero or one when a correct model is specified.

Note that given a fixed sample size, n, as $\eta \to 0$ (more and more missing data), the term $\sqrt{n_1}\xi$ will be smaller so that both approaches will have less "power" to detect model departures. As an illustration, Figure 1 shows the number of times that the checks detect model departures (based on the posterior predictive probability being less than 0.05 or greater than 0.95) at various values of η when the model is misspecified with parameter values based on the data example in Section 4. Both approaches detect less model departures (less power) as η increases.

We first examine the posterior probability for the complete data replication approach. Since C is $O_p(1)$, (6) can be written as $\Phi\left(\frac{m+c(\Delta)}{\sqrt{V+v(\Delta)}}\right)$, where $c(\Delta)$ and $v(\Delta)$ are functions of Δ. So the probability using complete data depends on the sensitivity parameter, Δ, which indicates the check is *not* invariant to the extrapolation distribution. In addition, as $|\Delta| \to \infty$, the denominator and numerator in (6) are both $O_p(\Delta)$, so the probability can be approximated by $\Phi\left(\frac{\frac{1}{\sqrt{n}}C}{\sqrt{1-\eta}}\right)$ or $\Phi\left(-\frac{\frac{1}{\sqrt{n}}C}{\sqrt{1-\eta}}\right)$, which only depends on n_1 but no other data. Figure 2 plots the posterior predictive probabilities using both the observed data replication approach and complete data replication approach at various values of Δ when one model is correctly specified and the other misspecified. The probabilities using the observed data stay the same across different Δ, but the probabilities using the complete data approach change with Δ and eventually converge to $\Phi\left(\frac{\frac{1}{\sqrt{n}}C}{\sqrt{1-\eta}}\right)$ as $\Delta \to +\infty$ and $-\Phi\left(\frac{\frac{1}{\sqrt{n}}C}{\sqrt{1-\eta}}\right)$ as $\Delta \to -\infty$

One specific departure from the model fit is if we assume $\xi = 0$, but in truth, $\xi \neq 0$. The term B in the posterior probability for the observed data replications is now $\sqrt{n_1}\left(\bar{y}_{1,\text{obs}}-\mu\right)=\sqrt{n_1}\left(\bar{y}_{1,\text{obs}}-(\mu+\xi)\right)+\sqrt{n_1}\xi$. In (5), the dominating term in the numerator, $\sqrt{n_1}\xi=O_p(\sqrt{n_1})$ will drive the probabilities to zero or one. In (6), although for a given Δ, the probability will go to zero or one when $n_1 \to \infty$, the complete replicate

approach will have less power than the observed replicate approach because (6) has a larger denominator than (5). Furthermore, its power is inversely correlated with |Δ|, i.e. the larger |Δ| is, the less power the complete data replication approach has. In the extreme case, when |Δ| is of higher order than $\sqrt{n_1}$, the probability will only depend on n_1 as shown earlier, which results in no power at all. Figure 3 shows the number of times that the checks detect model departures (the posterior predictive probability is less than 0.05 or greater than 0.95) at various values of Δ when the model is misspecified using parameter values based on the data example. The observed data replication approach identifies more model departures than the complete data replication approach. Also, the complete data replication approach identifies less as |Δ| increases.

In the simple case when $\sigma_1 = \sigma_2$ and Δ is set to $\varphi\xi$, the probability in (5) is approximately $\Phi\left(\frac{-\phi B}{\sqrt{2\tau^2+\phi^2\sigma_2^2}}\right)$ (6) is approximately $\Phi\left(\frac{-\phi B-\sqrt{\frac{1-\eta}{\eta}}\phi D}{\sqrt{\frac{1}{\eta}\left(2\tau^2+\phi^2\sigma_2^2\right)}}\right)$. When the correct model is specified, both (5) and (6) are Φ(*H*) where $H \sim N\left(0, \frac{\phi^2\sigma_2^2}{2\tau^2+\phi^2\sigma_2^2}\right)$ since both B and D are $N\left(0, \sigma_2^2\right)$ and they are independent (though given the data the two probabilities are not necessarily the same). So the two approaches will have the same Type I error. When $\sigma_1 = \sigma_2$ but $\Delta \neq \varphi\xi$, the complete replicate approach will have larger Type I error, since it has larger variance than *H*. Figure 4 shows the number of times that the checks falsely detect model departures (based on the posterior predictive probability being less than 0.05 or greater than 0.95) at various values of Δ when the model is correctly specified in an example where $\sigma_1 = \sigma_2$. The observed data replication approach identifies the same number of model departures as the complete data replication approach when $\Delta = \varphi\xi$. The complete data replication approach identifies more (larger Type I error) as |Δ| increases. In general, the complete data replication approach has larger Type I error especially when |Δ| is large. There are exceptions, for example, when $\Delta = \varphi\xi$ and $\sigma_1 < \sigma_2$, the observed approach has larger Type I error.

4 Data example

We illustrate the checks using the model in Section 3 on data from a randomized clinical trial. The objective of the trial was to examine the effects of recombinant human growth hormone therapy for building and maintaining muscle strength in the elderly. The study, which we will refer to as GH, enrolled 161 participants and randomized them to one of four treatments arms. Various muscle strength measures were recorded at baseline, 6 months, 12 months. We focus on mean quadriceps strength, measured as the maximum foot-pounds of torque that can be exerted against resistance provided by mechanical device. We will focus on two of the treatment groups, Exercise + Growth Hormone (EG) and Exercise + Placebo (EP), denoted as $Z = 1$ and $Z = 2$. Of the 78 randomized to these two arms, only 53 had complete follow-up (and the missingness was monotone); see Table 1. For illustration, we focus on the month 0 (baseline) and month 12 measures. As such, $\boldsymbol{Y} = (Y_1, Y_2)^T$ is quad strength measured at months 0 and 12. The corresponding observed data indicators are $\boldsymbol{R} = (R_1, R_2)^T$. In this data, the baseline quad strength is always observed, so $P(R_1 = 1) = 1$.

Let ($\boldsymbol{y}_{\text{obs},i}$, $\boldsymbol{r}_i$, z_i) be the observed data for subject i, $i = 1, \ldots, 78$. We fit the data from both treatment groups to the model introduced in Section 3. The parameters of the observed data model, $\boldsymbol{\omega}_O$, are given diffuse priors. The test statistic for each treatment group is defined as the corresponding means of Y_2. For example, for the treatment group EG, the test statistics used in the observed data replication approach are $T_o^{EG}(\boldsymbol{y}_{\text{obs}}) = \frac{\sum_{i=1}^{78} y_{2i} I(r_{2i}=1) I(z_i=1)}{\sum_{i=1}^{78} I(r_{2i}=1)(z_i=1)}$ and $T_o^{EG}(\boldsymbol{y}_{\text{obs}}^{\text{rep}}) = \frac{\sum_{i=1}^{78} y_{2i}^{\text{rep}} I(r_{2i}^{\text{rep}}=1) I(z_i=1)}{\sum_{i=1}^{78} I(r_{2i}^{\text{rep}}=1) I(z_i=1)}$, and the test statistics used in the complete data replication approach are $T_c^{EG}(\boldsymbol{y}_{\text{obs}}) = \frac{\sum_{i=1}^{78} y_{2i} I(z_i=1)}{\sum_{i=1}^{78} I(z_i=1)}$ and $T_c^{EG}(\boldsymbol{y}_{\text{obs}}^{\text{rep}}) = \frac{\sum_{i=1}^{78} y_{2i}^{\text{rep}} I(z_i=1)}{\sum_{i=1}^{78} I(z_i=1)}$. We calculate and compare the posterior predictive probabilities using complete and observed replications at various values (0, 5, 10, 20, −5, −10 and −20) of the sensitivity parameter Δ. Note previous analyses (e.g., [6]) considered negative Δ's up to a value of −20 as dropouts were thought to be doing worse than completers. However, for illustration here, we also consider positive values of Δ.

Table 2 shows the marginal means of the responses $E(Y_1)$ and $E(Y_2)$ estimated from the model for different Δ. $E(Y_2)$ changes with the sensitivity parameter. The posterior predictive probabilities using observed and complete replicated datasets at various values of the Δ are shown in Table 3. As shown in Section 3, the posterior predictive probabilities using observed replicated datasets are invariant to the sensitivity parameter Δ; those using complete replicated datasets change dramatically with Δ. Also using observed replicated datasets seems to have more power (posterior predictive probability of 0.08 for EG) to detect model departure than using complete replicated datasets (posterior probability probability ≥ 0.2 for EG) given the observed differences in means at month 12 in Table 1.

Given the relatively poor fit, we tried to improve the fit by not assuming the parameters were the same for EP and EG. In particular, we use the same model but with separate parameters for EP and EG. Table 4 has the marginal means for the responses for EP and EG and Table 5 has posterior predictive probabilities in the new model. Clearly model fit is improved with posterior probabilities very close to 1/2.

5 Conclusion

We have proposed a convenient way to assess the fit of the Bayesian models in the presence of incomplete data using posterior predictive checks; such checks can easily be implemented in WinBUGS/JAGS/Stan (see the supplementary materials for WinBUGS code for the model and checks from Sections 3 and 4). Both approaches (based on either complete or observed data replications) not surprisingly result in less power than if we actually had complete data. And the analytical example and data example showed how sensitivity parameters, Δ in our development in Sections 3 and 4, can have a large impact on the assessment of model fit for complete replication approaches. The fact that the observed replications satisfy the invariance to the extrapolation distribution un- like the complete replications which arguably, is a necessary property [9], leads us to recommend checks based on the replicated observed

data as the preferred approach for nonignorable missingness even at the potential loss of power in some settings. Our approach, using the observed replications, separates the fit of the model to the observed data from the (subjective) reasonableness of the imputations. These two pieces correspond respectively to the two components in the extrapolation distribution.

Clearly, further work needs to be done to better understand the behavior and operating characteristics of these checks based on replicated observed data in various settings with nonignorable missingness and potentially causal inference settings for which checks should also share a property similar to the invariance to the extrapolation distribution. However, comparing the complete and observed data replications by simulation will not serve a useful purpose since only the latter have the desired invariance property.

We do note that the complete replications can be useful to assess the 'reasonableness' of imputed missing response (as in one of the examples in [7]), but not to assess model fit based on observed data. The other important message is that methods used for latent data are often not valid for (nonignorably) missing data (other than similar computational algorithms). We see the same idea in the recommendations for DIC in [10] which do not coincide with those in [11] and the fact that different distributions for latent variables provide differential fits to the observed data unlike different values for sensitivity parameters (in the extrapolation distribution) for nonignorable missngness.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The last author was partially funded by NIH grants CA85295 and CA183854.

References

[1]. Rubin DB. Bayesianly justifiable and relevant frequency calculations for the applied statistician. The Annals of Statistics. 1984; 12(4):1151–1172.

[2]. Meng XL. Posterior predictive p-values. The Annals of Statistics. 1994; 22(3):1142–1160.

[3]. Robins JM, van der Vaart A, Ventura V. Asymptotic distribution of p values in composite null models. Journal of the American Statistical Association. 2000; 95(452):1143–1156.

[4]. Zellner A. Bayesian and non-Bayesian analysis of the regression model with multivariate Student-t error terms. Journal of the American Statistical Association. 1976; 71(354):400–405.

[5]. Gelman A, Meng XL, Stern H. Posterior predictive assessment of model fitness via realized discrepancies. Statistica sinica. 1996; 6(4):733–760.

[6]. Daniels, MJ.; Hogan, JW. Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis. CRC Press; 2008.

[7]. Gelman A, Van Mechelen I, Verbeke G, Heitjan DF, Meulders M. Multiple imputation for model checking: completed-data plots with missing and latent data. Biometrics. 2005; 61(1):74–85. [PubMed: 15737080]

[8]. National Research Council. The Prevention and Treatment of Missing Data in Clinical Trials. The National Academies Press; 2010.

[9]. Daniels MJ, Chatterjee AS, Wang C. Bayesian model selection for incomplete data using the posterior predictive distribution. Biometrics. 2012; 68(4):1055–1063. [PubMed: 22551040]

[10]. Wang C, Daniels MJ. A note on MAR, identifying restrictions, model comparison, and sensitivity analysis in pattern mixture models with and without covariates for incomplete data. Biometrics. 2011; 67(3):810–818. [PubMed: 21361893]

[11]. Celeux G, Forbes F, Robert CP, Titterington DM. Deviance information criteria for missing data models. Bayesian Analysis. 2006; 1(4):651–673.

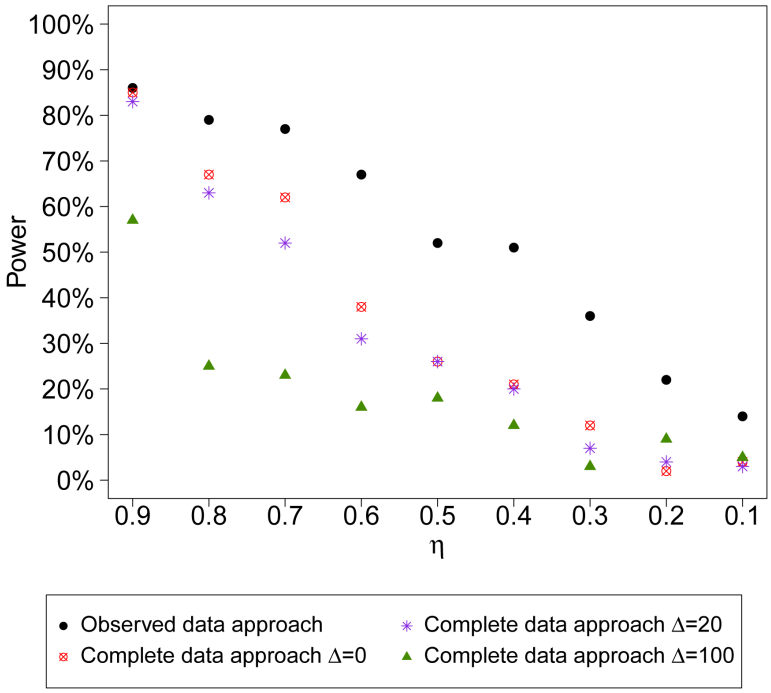

Figure 1.

'Power' of detecting model departures using the observed data replication approach and complete data replication approach (percent of times the posterior predictive probability < 0.05 or < 0.95 in 100 replicated datasets) at various values of η. The true data model has parameters $n = 100$, $\mu = 62$, $\xi = 9$, $\sigma_1 = 28$, $\sigma_2 = 24$, $\alpha = 16$, $\varphi = 0.9$, $\tau = 16$. The model is misspecified by assuming $\xi = 0$ when performing model checks. $\Delta = 0, 20, 100$ are used for complete data approach.

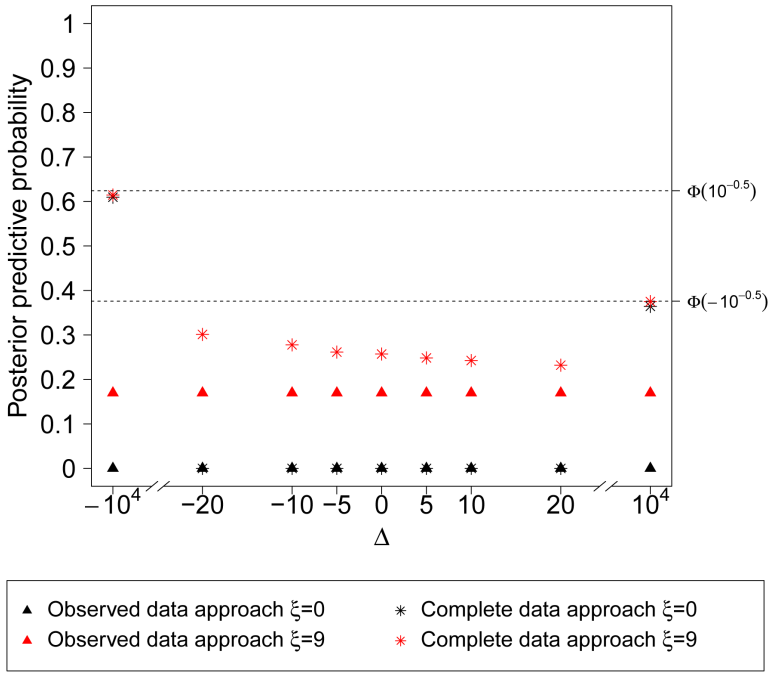

Figure 2.

Posterior predictive probabilities using the observed data replication approach and complete data replication approach at various values of Δ. The true data model has parameters n = 1000, μ = 62, ξ = 9, σ_1 = 28, σ_2 = 24, α = 16, φ = 0.9, τ = 16, η = 0.5. The data example has n_1 = 495, so $\frac{\frac{1}{\sqrt{n}}C}{\sqrt{1-\eta}} = -10^{-0.5}$. When performing model checks, one model is correctly specified and the other misspecified by assuming ξ = 0.

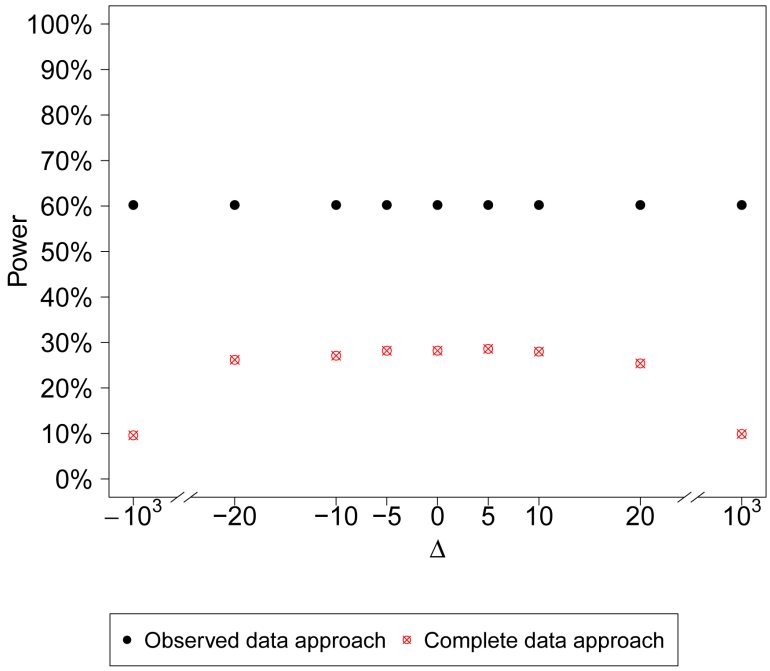

Figure 3.

'Power' of detecting model departures using the observed data replication approach and complete data replication approach (percent of times the posterior predictive probability < 0.05 or > 0.95 in 1000 replicated datasets) at various values of Δ. The true data model has parameters $n = 100$, $\mu = 62$, $\xi = 9$, $\sigma_1 = 28$, $\sigma_2 = 24$, $\alpha = 16$, $\varphi = 0.9$, $\tau = 16$, $\eta = 0.5$. The model is misspecified by assuming $\xi = 0$ when performing model checks.

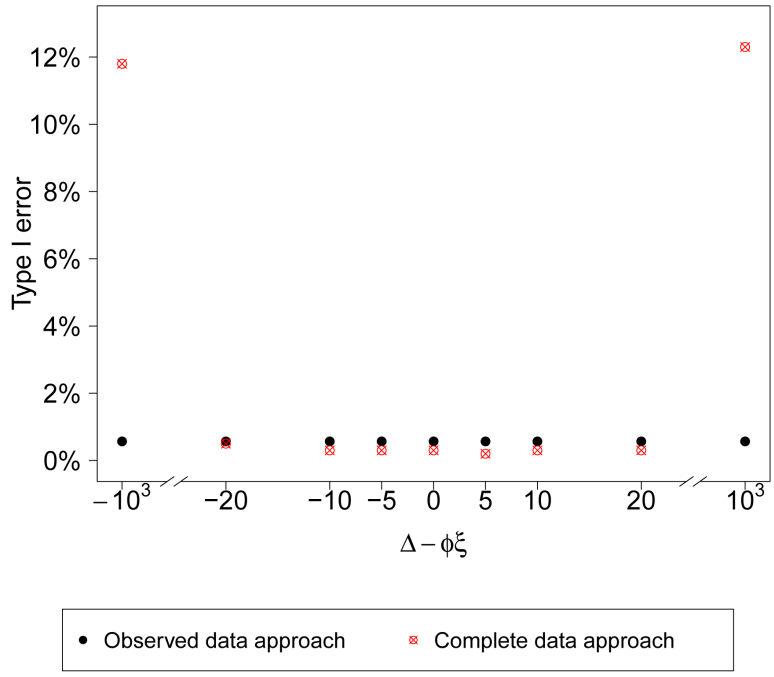

Figure 4.

Type I error using the observed data replication approach and complete data replication approach (percent of times the posterior predictive probability < 0.05 or > 0.95 in 1000 replicated datasets) at various values of Δ when model is correctly specified and $\sigma_1 = \sigma_2$. The true data model has parameters $n = 500$, $\mu = 62$, $\xi = 9$, $\sigma_1 = \sigma_2 = 28$, $\alpha = 16$, $\varphi = 0.9$, $\tau = 25$, $\eta = 0.7$.

Table 1

Growth hormone trial: sample means (standard deviations) stratified by treatment group

Treatment	R_2	n	Y_1	Y_2
EG	0	16	58 (23)	
	1	22	78 (24)	88 (32)
	All	38	69 (25)	88 (32)
EP	0	9	70 (35)	
	1	31	65 (24)	72 (21)
	All	40	66 (26)	72 (21)
All	0	25	62 (28)	
	1	53	70 (24)	79 (27)
	All	78	68 (26)	79 (27)

Table 2

Marginal means of responses estimated from the model for different Δ

	Δ =						
Response	**0**	**5**	**10**	**20**	**−5**	**−10**	**−20**
Y_1	67.6	67.6	67.6	67.6	67.6	67.6	67.6
Y_2	76.6	78.2	79.9	83.1	75.0	73.4	70.1

Table 3

Posterior predictive probabilities using *observed* data replications and *complete* data replications for both treatments

Treatment	Type	0	5	10	Δ = 20	−5	−10	−20
EG	observed	0.08	0.08	0.08	0.08	0.08	0.08	0.08
EG	complete	0.30	0.27	0.24	0.20	0.33	0.36	0.43
EP	observed	0.84	0.84	0.84	0.84	0.84	0.84	0.84
EP	complete	0.69	0.72	0.75	0.80	0.66	0.62	0.55

Table 4

Marginal means of responses estimated from the model for different Δ

		Δ =						
Treatment	**Response**	**0**	**5**	**10**	**20**	**−5**	**−10**	**−20**
EG	Y_1	69.3	69.3	69.3	69.3	69.3	69.3	69.3
EG	Y_2	79.3	81.4	83.5	87.7	77.1	75.0	70.8
EP	Y_1	65.9	65.9	65.9	65.9	65.9	65.9	65.9
EP	Y_2	73.3	74.5	75.6	78.0	72.1	70.9	68.5

Table 5

Posterior predictive probabilities using *observed* data replications and *complete* data replications for both treatments

Treatment	Type	0	5	10	Δ = 20	−5	−10	−20
EG	observed	0.50	0.50	0.50	0.50	0.50	0.50	0.50
EG	complete	0.50	0.50	0.50	0.50	0.49	0.49	0.49
EP	observed	0.50	0.50	0.50	0.50	0.50	0.50	0.50
EP	complete	0.50	0.50	0.50	0.51	0.50	0.50	0.49