COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2021.1921799

Full conditional distributions for Bayesian multilevel models with additive or interactive effects and missing data on covariates

Roy Levy[a] and Craig K. Enders[b]

[a]T. Denny Sanford School of Social & Family Dynamics, Arizona State University, Tempe, Arizona, USA;
[b]Psychology Department, University of California Los Angeles, Los Angeles, California, USA

ABSTRACT

Missing data are a common occurrence in analyses of multivariate data, including in multilevel modeling. Bayesian approaches to handling missing data in multilevel modeling have garnered increasing attention, either on their own or in service of multiple imputation. However, these applications are largely confined to specific models or missingness patterns. The current work provides a coherent account of Bayesian analysis of multilevel models in the presence of missing data on the outcomes, level-1 predictors, and level-2 predictors, that covers the main aspects of the models and missingness. In doing so, this work provides a grounding for estimation in fully Bayesian approaches that employ Gibbs sampling, and provides an account of how to generate the imputations in the first phase of a multiple imputation approach.

ARTICLE HISTORY
Received 12 September 2020
Accepted 20 April 2021

KEYWORDS
Bayesian; Gibbs sampling; Missing data; Multilevel modeling; Multiple imputation

Missing data are a common occurrence in analyses of multivariate data, including in multilevel modeling. In behavioral, social, education, and medical research, and related fields, strategies for handling missing data in multilevel settings have garnered increasing attention, including Bayesian and multiple imputation approaches, typically under the assumption that the data are missing at random (MAR) (Schafer and Yucel 2002; Gelman and Hill 2007; Yucel 2008; Goldstein, Carpenter, and Browne 2014; van Buuren 2011; Bartlett et al. 2015; Erler et al. 2016; Grund, Lüdtke, and Robitzsch 2016, 2018; Lüdtke, Robitzsch, and Grund 2017; Enders, Keller, and Levy 2018; Speidel, Drechsler, and Sakshaug 2018; Erler et al. 2019; Enders, Du, and Keller 2020). Multiple imputation can be thought of as having three phases: (1) imputation, (2) analysis, and (3) pooling. The imputation phase of multiple imputation can be framed as an instance of Bayesian inference for the missing data, yielding completed datasets which are subsequently analyzed with complete-data analyses, the results of which are then pooled.

Bayes' theorem may be conceptually expressed as

$$p(Unknown \mid Known) \propto p(Known, Unknown) = p(Known \mid Unknown)p(Unknown)$$

where $p(Unknown \mid Known)$ is the posterior distribution of all the unknown entities given known entities; $p(Known, Unknown)$ is the joint distribution of known and unknown entities; $p(Known \mid Unknown)$ is the conditional distribution of known entities given the unknown entities, which serves as the likelihood function for the unknown entities; and $p(Unknown)$ is the prior distribution for the unknown entities. In the current work, unknown entities include model parameters

CONTACT Roy Levy roy.levy@asu.edu T. Denny Sanford School of Social & Family Dynamics, Arizona State University, P.O. Box 873701, Tempe, AZ 85287-3701, USA.

and also missing data. Known entities include observed data and values of other entities specified by the analyst.

Bayesian inference supports two broad strategies for modeling the presence of missing data: a fully Bayesian approach, and a multiple-imputation approach. In a fully Bayesian approach, the focus is often on the parameters of the analytic model of interest (e.g., a multilevel model relating predictors at various levels to an outcome). A fully Bayesian analysis yields the joint posterior for parameters of the analytic model of interest (e.g., the parameters in a multilevel model relating predictors at various levels to an outcome) and the missing data. In a simulation-based estimation context, a series of draws from this posterior are obtained for both the model parameters and the missing data. In a fully Bayesian approach, the focus is often on just the parameters of the analytic model of interest, and the draws for the missing data are ignored. In a multiple imputation approach, the draws for the missing data are viewed as the imputations, and the draws for the model parameters are ignored; the imputations for the missing data yield completed datasets which are then analyzed in the second phase of multiple imputation.

Bayesian approaches to inference in the presence of missing data and/or imputing values for missing data in multilevel modeling have garnered increasing attention, but these applications are largely confined to specific models or missingness patterns. As examples, Schafer and Yucel (2002) considered multilevel models, but did not consider interactions and missingness was confined to the response variable. Kim, Sugar, and Belin (2015) and Lüdtke, Robitzsch, and West (2020) considered models with interactions, but only at a single level. Goldstein et al. (2009) treated multilevel models, but do not consider interactions. Erler et al. (2016) considered multilevel models with interactions, but focus on the relevant distribution for missing values, and do not provide the form of the relevant distributions for the model parameters.

Practitioners and methodologists alike benefit from treatments that attempt to bring these various situations under a single broad umbrella. Following existing efforts pushing toward such generality (Goldstein, Carpenter, and Browne 2014; Erler et al. 2016), the current work offers a coherent account of Bayesian analysis of multilevel models in the presence of data that are MAR, covering the main aspects of the models, allowing for interactions varying missingness patterns. The current work attempts to provide such an account under the assumption of normally distributed variables.

More specifically, for a series of models, we set out the posterior distribution under a Bayesian analysis, and derive the *full conditional distributions* under the Bayesian model. Let $Unknown_r$ denote the r^{th} component of the collection of unknowns. Let $Unknown_{\text{-}r}$ denote the remaining unknown components, that is, all the unknowns *except* $Unknown_r$. The full conditional distribution is then $p(Unknown_r \mid Unknown_{\text{-}r}, Known)$.

In doing so, this work contributes to the literature in four principal ways. First, it gives a coherent account of what goes on in a fully Bayesian analysis that seeks inference regarding the missing data and the parameters of the model for a larger class of models than has typically been presented in the literature. Thus, this paper provides a unifying framework that generalizes solutions given for specific models such as those reviewed above.

Second, by providing this account, it provides a grounding for estimation in fully Bayesian models via Gibbs sampling (Gelfand and Smith 1990), a popular approach to Markov chain Monte Carlo estimation that iteratively samples from the full conditional distributions. For a distribution with R components, an iteration for the Gibbs sampler proceeds by drawing from $p(Unknown_1 \mid Unknown_{\text{-}1}, Known), \ldots, p(Unknown_R \mid Unknown_{\text{-}R}, Known)$, where, for each full conditional, the current values in the chain for the unknowns on the right-hand side of the conditioning bar are used as the values for the unknowns. Subsequent iterations proceed in the same way. The limiting distribution of this sequence of draws converges to the desired posterior distribution of unknowns given knowns. Gibbs sampling is implemented in software such as JAGS

(Plummer 2017) and BUGS (Spiegelhalter et al. 2007), and has been shown to be advantageous over popular imputation-based methods in certain contexts (Erler et al. 2016).

Third, for analysts adopting a multiple imputation approach, it provides an account of how to generate the imputations in accordance with a full Bayesian model. This stands in contrast with popular multiple imputation approaches which separately specify imputation models as conditional distributions for the variable with missing values given other variables as following normal-theory linear regressions. These may suffice in simple models. However, in models with interaction effects, these approaches can yield imputation models that are incompatible, in the sense that there may not a joint distribution that corresponds to the set of conditional distribution (e.g., Liu et al. 2014; Bartlett et al. 2015; Chen and Ip 2015). In the current work, we derive the full conditional distributions as mentioned above. The Hammersley-Clifford theorem states that, under mild conditions, the complete set of full conditional distributions fully determine the joint distribution (Besag 1974; Robert and Casella 2004). This property underlies the use of Gibbs sampling (Gelfand and Smith 1990; Robert and Casella 2004), and in the current context ensures compatibility of the imputation models. Put somewhat casually, common imputation approaches begin with separately-specified conditional models, which may not correspond to the joint (posterior) distribution (i.e., they may not be compatible). In this work, we take the joint (posterior) distribution as the starting point and derive the full conditional distributions from it; the set of these uniquely define the full joint (posterior) distribution, and are therefore compatible. As such, this work provides a grounding for software aimed at producing imputations for subsequent analyses, such as Blimp (Keller and Enders 2019), REALCOM (Carpenter, Goldstein, and Kenward 2011), and the R packages 'jomo' (Quartagno and Carpenter 2018) and 'mdmb' (Robitzsch and Lüdtke 2021), among others.

When analytical formulations of the full conditional distributions are intractable, a more flexible but more computationally expensive Metropolis sampling step may be employed (Hastings 1970). This is commonly done in single-level and multilevel models with interactions or other nonlinear terms where there is missingness in the predictors. As examples, Goldstein, Carpenter, and Browne (2014) considered a single-level model with interactions, and Enders, Du, and Keller (2020) considered a multilevel model, and both employed Metropolis steps to sample from the distribution for missing predictors. Similarly, Lüdtke, Robitzsch, and West (2020) and Robitzsch and Lüdtke (2021) employed Metropolis-Hastings to sample from the distribution for missing predictors in single-level and multilevel models. In this work, the analytical forms for the full conditional distributions are developed for several scenarios in single-level and multilevel models with missing data, including those with interactions or other nonlinear terms. Thus the fourth main contribution of this work is to show that, for such models, the more computationally expensive Metropolis sampling steps are not required, as the full conditional distributions follow known forms.

This paper is organized as follows. The following section presents foundational results for Bayesian analyses with normal distributions, which are invoked throughout. In the sections that follow, we build up from a single-level model with no missingness to models with missingess on the outcome or the predictors. We start with the single-level model because it is easier to see the strategies involved in those contexts. These form the building blocks for our ultimate goal, which is a two-level model with missingness on the outcome, level-1 predictors, and level-2 predictors. For each model, the model and posterior distribution is expressed, and then the full conditionals are developed. For completeness and transparency in the full conditionals, we include conditioning on the hyperparameters that define the prior distributions. When the hyperparameters are specified in advance, they are not random variables in the sense of other model parameters and the data. However, including them serves to highlight that they are involved in the computations.

In a sense, the results presented here may be regarded as an application of the already established relationships and properties of Bayesian modeling, reviewed in the next section. However,

the details of how they play out may not be obvious, particularly when the situation involves interactions among covariates, missing data, and interactions among covariates with missing data. Deriving the analytical forms of the full conditional distributions for the model parameters and missing data, and presenting them in a single, coherent way serves the four principal intended contributions. A discussion concludes the paper.

1. Foundations

The current work focuses on linear models assuming normal distributions for the outcome and predictors at each level. Among other advantages, the normality assumption allows us to give analytic expressions for all Gibbs sampler steps, thus eliminating the need for more computationally expensive Metropolis sampler (Hastings 1970). Models with discrete variables may be viewed as extensions of these models, through the use of underlying latent variable formulations (Chib and Greenberg 1998; Carpenter and Kenward 2013). Accordingly, the current work may serve as a foundation for extensions to multilevel models for such variables (e.g., see Enders, Du, and Keller 2020; Enders, Keller, and Levy 2018; Goldstein et al. 2009). Many of the specifications involve normal distributions that afford conditional conjugacy relationships through the application of standard Bayesian results for models assuming normality; see Lindley and Smith (1972) and Rowe (2003) for extensive treatments of these results. As it is central to many of the full conditional distributions developed below, we state some key properties. In what follows, we purposefully employ notation that is in the main different than what we will use in defining the multilevel models in the sections that follow. We do this because, as we will see, the derivations involve viewing model parameters as akin to the data. Presenting this in a different notation aids in avoiding ambiguity or confusion later on.

The first foundational result comes from Lindley and Smith (1972). Let

$$\mathbf{z} \sim N(\mathbf{A}\boldsymbol{\theta}, \mathbf{C}_1) \tag{1}$$

denote the conditional distribution of a column vector of observed values $\mathbf{z}$ given a vector of unknown parameters $\boldsymbol{\theta}$, which are assigned a prior distribution

$$\boldsymbol{\theta} \sim N(\mathbf{D}\boldsymbol{\eta}, \mathbf{C}_2),$$

where $\boldsymbol{\eta}$ is a known vector and $\mathbf{A}$, $\mathbf{D}$, $\mathbf{C}_1$, and $\mathbf{C}_2$, are known matrices. The posterior distribution for $\boldsymbol{\theta}$ is then

$$\boldsymbol{\theta} \mid \mathbf{A}, \mathbf{C}_1, \boldsymbol{\eta}, \mathbf{D}, \mathbf{C}_2, \mathbf{y} \sim N\left(\mathbf{G}\dot{\boldsymbol{\beta}}, \mathbf{G}\right) \tag{2}$$

where

$$\dot{\boldsymbol{\beta}} = \mathbf{C}_2^{-1}\mathbf{D}\boldsymbol{\eta} + \mathbf{A}'\mathbf{C}_1^{-1}\mathbf{z} \quad \text{and} \quad \mathbf{G} = \left(\mathbf{C}_2^{-1} + \mathbf{A}'\mathbf{C}_1^{-1}\mathbf{A}\right)^{-1}$$

The next foundational result obtains from a similar setup. Suppose $\mathbf{C}_1 = c_1\mathbf{I}$ where $\mathbf{I}$ is an identity matrix. Viewing (1) as a regression structure, this then embodies an assumption of homogeneity of variance. Suppose $\boldsymbol{\theta}$ is known and c_1 is unknown and assigned a prior distribution

$$c_1 \sim \text{Inverse-Gamma}\left(\nu_0/2, \nu_0\sigma_{c_1}^2/2\right)$$

where ν_0 and $\sigma_{c_1}^2$ are known scalars. The posterior distribution for c_1 is then

$$c_1 \mid \boldsymbol{\theta}, \nu_0, \sigma_{c_1}^2, \mathbf{z}, \mathbf{A} \sim \text{Inverse-Gamma}\left(\frac{\nu_0 + n}{2}, \frac{\nu_0\sigma_{c_1}^2 + SS}{2}\right) \tag{3}$$

where n is the number of elements in $\mathbf{z}$ and

$$SS = (\mathbf{z} - \mathbf{A}\boldsymbol{\theta})'(\mathbf{z} - \mathbf{A}\boldsymbol{\theta}).$$

The next set of main results comes from Rowe (2003), who extended this situation to that of a matrix of observations. Rowe situates this in a multivariate regression context where for each subject i there is a vector of P dependent variables predicted by an intercept and a vector of Q predictor variables. In the following we slightly expand Rowe's notation and continue with the notational scheme introduced above. The regression model may be expressed as

$$\mathbf{z}_i = \boldsymbol{\theta}\mathbf{A}_i + \boldsymbol{\varepsilon}_i, \text{ for } i = 1, ..., n$$

where

- $\mathbf{z}_i = (z_{i1}, \ldots, z_{iP})'$ denotes an $(P \times 1)$ vector of values from case i on P outcome variables.
- $\mathbf{A}_i = (1, A_{i1}, \ldots, A_{iQ})'$ denotes a $[(Q+1) \times 1]$ vector containing values along Q predictor variables for case i, along with a leading 1. $\mathbf{A}_i$ may be thought of as an *augmented* predictor vector in that it appends a 1 to the usual vector of Q predictors, treating the regression intercept as a slope multiplied by a "predictor" that has a value of 1 for each subject.
- $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}'_{11} \\ \vdots \\ \boldsymbol{\theta}'_{P} \end{pmatrix} = \begin{bmatrix} \theta_{10} & \theta_{11} & \cdots & \theta_{1Q} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{P0} & \theta_{P1} & \cdots & \theta_{PQ} \end{bmatrix}$ denotes a $[P \times (Q+1)]$ *augmented* coefficient matrix which appends the column of regression intercepts to the usual matrix of regression coefficients. The p^{th} row in $\boldsymbol{\theta}$ contains the augmented vector of coefficients for predicting the p^{th} element of $\mathbf{z}$.
- $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iP})'$ denotes an $(P \times 1)$ vector of errors for case i on P outcome variables.

Collecting all of the data together in matrices, the model may be expressed as

$$\mathbf{Z} \mid \boldsymbol{\theta}, \mathbf{A}, \mathbf{E} = \mathbf{A}\boldsymbol{\theta}' + \mathbf{E},$$

where in addition to $\boldsymbol{\theta}$ defined above

- $\mathbf{Z} = \begin{pmatrix} \mathbf{z}'_{11} \\ \vdots \\ \mathbf{z}'_{m} \end{pmatrix} = \begin{bmatrix} z_{11} & \cdots & z_{1P} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nP} \end{bmatrix}$ denotes an $(n \times P)$ matrix containing values from n subjects on P outcome variables.
- $\mathbf{A} = \begin{pmatrix} \mathbf{A}'_1 \\ \vdots \\ \mathbf{A}'_n \end{pmatrix} = \begin{bmatrix} 1 & a_{11} & \cdots & a_{1Q} \\ 1 & \vdots & \ddots & \vdots \\ 1 & a_{n1} & \cdots & a_{nQ} \end{bmatrix}$ denotes an $[n \times (Q+1)]$ matrix containing values from n cases along Q predictor variables, along with a leading 1 for every case. $\mathbf{A}$ may be thought of as an *augmented* predictor matrix in that it appends a column of 1s to the usual matrix of Q predictors, treating the regression intercept as a slope multiplied by a "predictor" that has a value of 1 for every subject.
- $\mathbf{E} = \begin{pmatrix} \boldsymbol{\varepsilon}'_1 \\ \vdots \\ \boldsymbol{\varepsilon}'_n \end{pmatrix} = \begin{bmatrix} \varepsilon_{11} & \cdots & \varepsilon_{1P} \\ \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \cdots & \varepsilon_{nP} \end{bmatrix}$ denotes an $(n \times P)$ matrix of errors from n cases on P outcome variables.

Suppose

$$\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\varepsilon),$$

where $\mathbf{0}$ is a $(P \times 1)$ vector of 0s and $\boldsymbol{\Sigma}_\varepsilon$ is a $(P \times P)$ covariance matrix. The model may then be expressed in distributional form as

$$\mathbf{z}_i \,|\, \boldsymbol{\theta}, \mathbf{A}_i, \boldsymbol{\Sigma}_\varepsilon \sim N(\boldsymbol{\theta}\mathbf{A}_i, \boldsymbol{\Sigma}_\varepsilon), \text{ for } i = 1, ..., n.$$

Suppose $\boldsymbol{\Sigma}_\varepsilon$ is known and $\boldsymbol{\theta}$ is unknown, where

$$\boldsymbol{\omega} = vec(\boldsymbol{\theta}) \sim N(\boldsymbol{\omega}_\mathbf{0}, \mathbf{T}),$$

where "*vec*" is the vectorization operator that stacks that columns of its matrix argument, and $\boldsymbol{\omega}_\mathbf{0}$ and $\mathbf{T}$ are a known mean vector and covariance matrix, respectively. The posterior distribution for $\boldsymbol{\omega}$ is then

$$\boldsymbol{\omega} \,|\, \boldsymbol{\Sigma}_\varepsilon, \mathbf{z}, \mathbf{A}, \boldsymbol{\omega}_\mathbf{0}, \mathbf{T} \sim N(\mathbf{D}\dot{\boldsymbol{\omega}}, \mathbf{D}) \tag{4}$$

where

$$\mathbf{D} = \left[\mathbf{T}^{-1} + \mathbf{A}'\mathbf{A} \otimes \boldsymbol{\Sigma}_\varepsilon^{-1}\right]^{-1}, \; \dot{\boldsymbol{\omega}} = \mathbf{T}^{-1}\boldsymbol{\omega}_\mathbf{0} + \left(\mathbf{A}'\mathbf{A} \otimes \boldsymbol{\Sigma}_\varepsilon^{-1}\right)\hat{\boldsymbol{\omega}},$$

$\otimes$ is the Kronecker product, and

$$\hat{\boldsymbol{\omega}} = vec\left(\mathbf{z}'\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\right).$$

Suppose instead that $\boldsymbol{\theta}$ is known and $\boldsymbol{\Sigma}_\varepsilon$ is unknown, where

$$\boldsymbol{\Sigma}_\varepsilon \sim \text{Inverse-Wishart}(\mathbf{Q}, \nu).$$

The posterior distribution for $\boldsymbol{\Sigma}_\varepsilon$ is then

$$\boldsymbol{\Sigma}_\varepsilon \,|\, \boldsymbol{\theta}, \mathbf{z}, \mathbf{A}, \mathbf{Q}, \nu \sim \text{Inverse-Wishart}(\nu\mathbf{Q} + n\mathbf{S}, \nu + n), \tag{5}$$

where

$$\mathbf{S} = (\mathbf{z} - \mathbf{A}\boldsymbol{\theta}')'(\mathbf{z} - \mathbf{A}\boldsymbol{\theta}')/n.$$

2. Single-level regression, no missing data

We begin our treatment of different models with a single-level regression, with no missing data, and shift to a notation that will support natural expansions to multilevel models. Let $\mathbf{y} = (y_1, \ldots, y_n)$ denote the collection of outcome variables from n subjects. Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{iP})$ be the collection of P predictor variables for subject i. Further, let $\mathbf{x}$ be the full collection of P predictors from all n subjects. The basic linear regression model assuming exchangeability among the subjects specifies the conditional distribution of the outcomes as

$$p(\mathbf{y} \,|\, \beta_0, \boldsymbol{\beta}, \sigma_\varepsilon^2, \mathbf{x}) = \prod_{i=1}^{n} p(y_i \,|\, \beta_0, \boldsymbol{\beta}, \sigma_\varepsilon^2, \mathbf{x}_i),$$

where

$$y_i \,|\, \beta_0, \boldsymbol{\beta}, \sigma_\varepsilon^2, \mathbf{x}_i \sim N(\beta_0 + \beta_1 x_{i1} + \ \ldots \ + \beta_P x_{iP}, \sigma_\varepsilon^2).$$

We expand the model and notation to define an *augmented* set of predictors, denoted $\mathbf{x}_A$, which includes the original xs and any other deterministically defined elements that will serve as predictors. For example, in the model above $\mathbf{x}_A$ is the augmented predictor matrix obtained by combining an $(n \times 1)$ column vector of 1s to the predictor matrix $\mathbf{x}$. This notation accommodates transformations such as interactions. For example, suppose we have a model with two xs (x_1, x_2) and their interaction (as captured by the product x_1x_2) as predictors. We may write the model as

$$y_i \mid \beta_0, \boldsymbol{\beta}, \sigma_\varepsilon^2, \mathbf{x}_i \sim N\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}, \sigma_\varepsilon^2\right).$$

Defining the augmented set of predictors for subject i as $\mathbf{x}'_{Ai} = (1, x_{i1}, x_{i2}, x_{i1}x_{i2})$, we may compactly write the model as

$$y_i \sim N\left(\mathbf{x}'_{Ai}\boldsymbol{\beta}_A, \sigma_\varepsilon^2\right)$$

where $\boldsymbol{\beta}_A$ is the augmented vector of coefficients containing the coefficients for $\mathbf{x}_A$; in this case $\boldsymbol{\beta}_A = (\beta_0, \beta_1, \beta_2, \beta_3)$.

We can derive the full conditional distributions more compactly by working with a matrix representation of the regression model,

$$\mathbf{y} \mid \boldsymbol{\beta}_A, \sigma_\varepsilon^2, \mathbf{x} \sim N\left(\mathbf{x}_A\boldsymbol{\beta}_A, \sigma_\varepsilon^2\mathbf{I}\right), \tag{6}$$

where $\mathbf{x}_A$ is an augmented predictor matrix (with one row for each subject) and $\mathbf{I}$ in this case is an $(n \times n)$ identity matrix.

Note that this actually gives the conditional distribution of *some* of the data. The data in our regression situation include the predictors $\mathbf{x}$ as well as the outcomes $\mathbf{y}$. A fully Bayesian analysis that views the data as observed values of random variables includes a distributional specification for $\mathbf{x}$ as well as $\mathbf{y}$. If the values for $\mathbf{x}$ are fixed, then it can be viewed as if their probability $p(\mathbf{x})$ is known (Gelman et al. 2013) or alternatively as though they are not drawn from a density at all (Jackman 2009). More generally, letting $\boldsymbol{\Omega}$ denote the parameters that govern the distribution for $\mathbf{x}$, $p(\mathbf{x} \mid \boldsymbol{\Omega})$, the posterior distribution for the full model is then

$$p\left(\boldsymbol{\beta}_A, \sigma_\varepsilon^2, \boldsymbol{\Omega} \mid \mathbf{y}, \mathbf{x}\right) \propto p\left(\mathbf{y}, \mathbf{x} \mid \boldsymbol{\beta}_A, \sigma_\varepsilon^2, \boldsymbol{\Omega}\right) p\left(\boldsymbol{\beta}_A, \sigma_\varepsilon^2, \boldsymbol{\Omega}\right).$$

The first term on the right side is the conditional probability of *all* of the data. The second term is the prior distribution for all of the parameters. Assuming prior independence of $\left(\boldsymbol{\beta}_A, \sigma_\varepsilon^2\right)$ and $\boldsymbol{\Omega}$ allows for the factorization $p\left(\boldsymbol{\beta}_A, \sigma_\varepsilon^2, \boldsymbol{\Omega}\right) = p\left(\boldsymbol{\beta}_A, \sigma_\varepsilon^2\right) p(\boldsymbol{\Omega})$. It can be shown (e.g., Jackman 2009) that the posterior distribution can be then factored as

$$p\left(\boldsymbol{\beta}_A, \sigma_\varepsilon^2, \boldsymbol{\Omega} \mid \mathbf{y}, \mathbf{x}\right) = p\left(\boldsymbol{\beta}_A, \sigma_\varepsilon^2 \mid \mathbf{y}, \mathbf{x}\right) p(\boldsymbol{\Omega} \mid \mathbf{x}).$$

This implies that we can analyze the first term on the right side—the elements of the standard regression model—by itself with no loss of information. As a consequence, the distinction between $\mathbf{x}$ being fixed or stochastic is irrelevant in the Bayesian analysis of the model (Jackman 2009). Either way, the $p(\mathbf{x})$ and $\boldsymbol{\Omega}$ terms drop out of the model and subsequent analysis. As will be discussed in later sections, the situation is more complicated when $\mathbf{x}$ contains missing data.

Under a conditionally conjugate prior distribution the posterior distribution is given by

$$p\left(\boldsymbol{\beta}_A, \sigma_\varepsilon^2 \mid \mathbf{y}, \mathbf{x}\right) \propto \prod_{i=1}^{n} p\left(y_i \mid \boldsymbol{\beta}_A, \sigma_\varepsilon^2, \mathbf{x}_i\right) p(\boldsymbol{\beta}_A) p\left(\sigma_\varepsilon^2\right), \tag{7}$$

where

$$y_i \mid \boldsymbol{\beta}_A, \sigma_\varepsilon^2, \mathbf{x}_i \sim N\left(\mathbf{x}'_{Ai}\boldsymbol{\beta}_A, \sigma_\varepsilon^2\right) \text{ for } i = 1, \ldots, n,$$

$$\boldsymbol{\beta}_A \sim N(\boldsymbol{\gamma}, \boldsymbol{\tau}), \text{ and } \sigma_\varepsilon^2 \sim \text{Inv-Gamma}\left(\nu_0/2, \nu_0\sigma_{\varepsilon_0}^2/2\right).$$

2.1. Full conditional distributions

For the level-1 coefficients the full conditional can be expressed as

$$p\left(\boldsymbol{\beta}_A \mid \sigma_\varepsilon^2, \boldsymbol{\gamma}, \boldsymbol{\tau}, \mathbf{y}, \mathbf{x}\right) \propto p\left(\mathbf{y} \mid \boldsymbol{\beta}_A, \sigma_\varepsilon^2, \mathbf{x}\right) p\left(\boldsymbol{\beta}_A \mid \boldsymbol{\gamma}, \boldsymbol{\tau}\right).$$

The first term on the right-hand side is the normal distribution of the outcomes in (6). The second term is the normal prior distribution in (7). We can apply the results of the standard Bayesian theory for normal models described above in the Foundations section. Following (2), the full conditional distribution for the coefficients is

$$\boldsymbol{\beta}_A \mid \sigma_\varepsilon^2, \boldsymbol{\gamma}, \boldsymbol{\tau}, \mathbf{y}, \mathbf{x} \sim N\left(\boldsymbol{\mu}_{\boldsymbol{\beta}_A \mid \ldots}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}_A \mid \ldots}\right), \tag{8}$$

where

$$\boldsymbol{\mu}_{\boldsymbol{\beta}_{Aj} \mid \ldots} = \left(\boldsymbol{\tau}^{-1} + \mathbf{x}_A'\left(\sigma_\varepsilon^2\mathbf{I}\right)^{-1}\mathbf{x}_A\right)^{-1}\left(\boldsymbol{\tau}^{-1}\boldsymbol{\gamma} + \mathbf{x}_A'\left(\sigma_\varepsilon^2\mathbf{I}\right)^{-1}\mathbf{y}\right), \text{ and } \boldsymbol{\Sigma}_{\boldsymbol{\beta}_{Aj} \mid \ldots} = \left(\boldsymbol{\tau}^{-1} + \mathbf{x}_A'\left(\sigma_\varepsilon^2\mathbf{I}\right)^{-1}\mathbf{x}_A\right)^{-1}.$$

The ellipses in the subscripts in $\boldsymbol{\mu}_{\boldsymbol{\beta}_{Aj} \mid \ldots}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_{Aj} \mid \ldots}$ are meant as shorthand to indicate conditioning on all the relevant terms, namely, those that are expressed in the left-hand side of (8).

Turning to σ_ε^2, following (3), the full conditional distribution for σ_ε^2 is

$$\sigma_\varepsilon^2 \mid \boldsymbol{\beta}_A, \nu_0, \sigma_{\varepsilon_0}^2, \mathbf{y}, \mathbf{x} \sim \text{Inv-Gamma}\left(\frac{\nu_0 + n}{2}, \frac{\nu_0\sigma_{\varepsilon_0}^2 + SS(\mathbf{E})}{2}\right), \tag{9}$$

where

$$SS(\mathbf{E}) = \left(\mathbf{y} - \mathbf{x}_A\boldsymbol{\beta}_A\right)'\left(\mathbf{y} - \mathbf{x}_A\boldsymbol{\beta}_A\right).$$

A Gibbs sampler iteratively draws values from the full conditional distributions. That is, at any point in a chain, an iteration of the Gibbs sampler would take a draw for $\boldsymbol{\beta}_A$ from (8) using the current value for σ_ε^2, and then take a draw for σ_ε^2 from (9) using the just-drawn value for $\boldsymbol{\beta}_A$.The next iteration would proceed accordingly: take a draw for $\boldsymbol{\beta}_A$ from (8) using the just-drawn value for σ_ε^2, and then take a draw for σ_ε^2 from (9) using the just-drawn value for $\boldsymbol{\beta}_A$. And so on.

The following sections describe the full conditional distributions in situations that expand on this basic regression model, including different types of missingness and multilevel structures.

3. Single-level regression, missingness on the outcome

When some values for the outcome are missing, we may partition the full set of *potentially* observable outcomes $\mathbf{y}$ into two subsets, $\mathbf{y} = (\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}})$, where $\mathbf{y}_{\text{obs}}$ are the observed data and $\mathbf{y}_{\text{mis}}$ are the missing data. The posterior distribution is given by

$$p\left(\boldsymbol{\beta}_A, \sigma_\varepsilon^2, \mathbf{y}_{\text{mis}} \mid \mathbf{y}_{\text{obs}}, \mathbf{x}\right) \propto \prod_{i=1}^{n} p\left(y_i \mid \boldsymbol{\beta}_A, \sigma_\varepsilon^2, \mathbf{x}_i\right) p\left(\boldsymbol{\beta}_A\right) p\left(\sigma_\varepsilon^2\right).$$

The terms on the right-hand side are just those defined in (7).

Full conditional distributions

The full conditional distributions for the regression parameters $\left(\boldsymbol{\beta}_A, \sigma_\varepsilon^2\right)$ are just the same as in the fully observed data case; they are given in (8) and (9). The key difference is that now $\mathbf{y}$ is comprised of $\mathbf{y}_{\text{obs}}$ and $\mathbf{y}_{\text{mis}}$. In a Gibbs sampler, the values for $\mathbf{y}_{\text{mis}}$ will change from iteration to iteration, as draws are taken from its full conditional distribution.

Assuming exchangeability of subjects, the full conditional distribution for $\mathbf{y}_{\text{mis}}$ is given by

$$p\left(\mathbf{y}_{\text{mis}} \mid \boldsymbol{\beta}_A, \sigma_\varepsilon^2, \mathbf{y}_{\text{obs}}, \mathbf{x}\right) = p\left(\mathbf{y}_{\text{mis}} \mid \boldsymbol{\beta}_A, \sigma_\varepsilon^2, \mathbf{x}\right) = \prod_{i=1}^{n_{\text{mis}}} p\left(y_{i,\text{mis}} \mid \boldsymbol{\beta}_A, \sigma_\varepsilon^2, \mathbf{x}_i\right),$$

where $y_{i,\text{mis}}$ is the value for the outcome for subject i, which is missing, and n_{mis} are the number of subjects with missing values on y. Following the regression model, the full conditional distribution for each subject is

$$y_{i,\text{mis}} \mid \boldsymbol{\beta}_A, \sigma_\varepsilon^2, \mathbf{x}_i \sim N\left(\mathbf{x}'_{Ai}\boldsymbol{\beta}_A, \sigma_\varepsilon^2\right).$$

4. Single-level regression, missingness on a predictor

A number of previous works describe methods for incomplete covariates in single-level regression models (e.g., Ibrahim, Chen, and Lipsitz 2002; Bartlett et al. 2015; Kim, Sugar, and Belin 2015; Zhang and Wang 2017). In our framework, when some values for a predictor are missing we partition the full set of *potentially* observable predictors $\mathbf{x}$ into two subsets, $\mathbf{x} = (\mathbf{x}_{\text{obs}}, \mathbf{x}_{\text{mis}})$, where $\mathbf{x}_{\text{obs}}$ are the observed data and $\mathbf{x}_{\text{mis}}$ are the missing data. As $\mathbf{x}_{\text{mis}}$ are unknown, they will require a distributional specification, governed by parameters $\boldsymbol{\Omega}$. Assuming prior independence of $\left(\boldsymbol{\beta}_A, \sigma_\varepsilon^2\right)$ and $\boldsymbol{\Omega}$ allows for the factorization $p\left(\boldsymbol{\beta}_A, \sigma_\varepsilon^2, \boldsymbol{\Omega}\right) = p\left(\boldsymbol{\beta}_A, \sigma_\varepsilon^2, \sigma_\varepsilon^2\right)p(\boldsymbol{\Omega})$, and the posterior distribution can be then factored as

$$\begin{aligned} p\left(\boldsymbol{\beta}_A, \sigma_\varepsilon^2, \boldsymbol{\Omega}, \mathbf{x}_{\text{mis}} \mid \mathbf{y}, \mathbf{x}_{\text{obs}}\right) &\propto p\left(\boldsymbol{\beta}_A, \sigma_\varepsilon^2, \boldsymbol{\Omega}, \mathbf{x}_{\text{mis}}, \mathbf{y}, \mathbf{x}_{\text{obs}}\right) \\ &= p\left(\mathbf{y} \mid \boldsymbol{\beta}_A, \sigma_\varepsilon^2, \mathbf{x}_{\text{mis}}, \mathbf{x}_{\text{obs}}\right)p(\boldsymbol{\beta}_A)p\left(\sigma_\varepsilon^2\right)p(\mathbf{x}_{\text{mis}} \mid \boldsymbol{\Omega}, \mathbf{x}_{\text{obs}})p(\mathbf{x}_{\text{obs}} \mid \boldsymbol{\Omega})p(\boldsymbol{\Omega}). \end{aligned}$$

The first three terms on the right-hand side are just those defined in (7). The last three terms on the right-hand side are new, reflecting the need for a distributional specification for $\mathbf{x}$. In what follows we adopt a multivariate normal distribution for $\mathbf{x}$, in which case assuming exchangeability among subjects yields $\mathbf{x}_i \sim N(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$. (Note that $\mathbf{x}$ does not refer to the augmented set of predictors, which could include interactions. Thus, this distributional specification for $\mathbf{x}_i$ holds even if the substantive analysis includes interactions, because the product terms are not a part of $\mathbf{x}_i$.) In this case, we have $\boldsymbol{\Omega} = (\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$. Employing conditionally conjugate prior distributions for these parameters defines the last term on the right-hand side:

$$\boldsymbol{\mu}_x \sim N\left(\boldsymbol{\mu}_{\boldsymbol{\mu}_x}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}_x}\right), \boldsymbol{\Sigma}_x \sim \text{Inv-Wishart}(\boldsymbol{\Sigma}_{x0}, d).$$

For subject i, the distribution for the subject's observed values is

$$\mathbf{x}_{i\text{obs}} \mid \boldsymbol{\Omega} \sim N\left(\boldsymbol{\mu}_{x_{i\text{obs}}}, \boldsymbol{\Sigma}_{x_{i\text{obs}}}\right),$$

and the distribution of the subject's missing values is

$$\mathbf{x}_{i\text{mis}} \mid \boldsymbol{\Omega}, \mathbf{x}_{i\text{obs}} \sim N\left(\boldsymbol{\mu}_{x_{i\text{obs}}} + \boldsymbol{\Sigma}_{x_{i\text{obs}}x_{i\text{mis}}}\boldsymbol{\Sigma}_{x_{i\text{obs}}}^{-1}\left(\mathbf{x}_{i\text{mis}} - \boldsymbol{\mu}_{x_{i\text{obs}}}\right), \boldsymbol{\Sigma}_{x_{i\text{mis}}} - \boldsymbol{\Sigma}_{x_{i\text{mis}}x_{i\text{obs}}}\boldsymbol{\Sigma}_{x_{i\text{obs}}}^{-1}\boldsymbol{\Sigma}_{x_{i\text{obs}}x_{i\text{mis}}}\right), \tag{10}$$

where

$$\boldsymbol{\mu}_x = \begin{bmatrix} \boldsymbol{\mu}_{x_{i\text{mis}}} \\ \boldsymbol{\mu}_{x_{i\text{obs}}} \end{bmatrix}; \boldsymbol{\Sigma}_x = \begin{bmatrix} \boldsymbol{\Sigma}_{x_{i\text{mis}}} & \boldsymbol{\Sigma}_{x_{i\text{mis}}x_{i\text{obs}}} \\ \boldsymbol{\Sigma}_{x_{i\text{obs}}x_{i\text{mis}}} & \boldsymbol{\Sigma}_{x_{i\text{obs}}} \end{bmatrix}.$$

Here, $\boldsymbol{\mu}_{x_{i\text{obs}}}$ is the subvector of $\boldsymbol{\mu}_x$ corresponding to the variables for which subject i has observed values, $\boldsymbol{\mu}_{x_{i\text{mis}}}$ is the subvector of $\boldsymbol{\mu}_x$ corresponding to the variables for which subject i has missing values, $\boldsymbol{\Sigma}_{x_{i\text{obs}}}$ is the submatrix of $\boldsymbol{\Sigma}_x$ corresponding to the variables for which subject i has observed values, $\boldsymbol{\Sigma}_{x_{i\text{mis}}}$ is the submatrix of $\boldsymbol{\Sigma}_x$ corresponding to the variables for which subject i has missing values, $\boldsymbol{\Sigma}_{x_{i\text{mis}}x_{i\text{obs}}}$ contains the submatrix of $\boldsymbol{\Sigma}_x$ containing the covariances between pairings of variables for which subject i has observed values with variables for which subject i has missing values.

Note that, in essence, the distribution for the missing data in (10) is what emerges from regressing the predictors with missingness on the other predictors. Alternatives to this multivariate approach adopted here include a univariate approach and a conditional specification, discussed later.

4.1. Full conditional distributions

The full conditional distributions for the regression parameters $\left(\boldsymbol{\beta}_A, \sigma^2_\varepsilon\right)$ are just the same as in the fully observed data case; they are given in (8) and (9). The key difference is that now $\mathbf{x}$ is comprised of $\mathbf{x}_{\text{obs}}$ and $\mathbf{x}_{\text{mis}}$. In a Gibbs sampler, the values for $\mathbf{x}_{\text{mis}}$ will change from iteration to iteration, as draws are taken from its full conditional distribution.

4.1.1. Full conditional for the missing values of the predictor

The present description for the full conditional distribution for missing values on a predictor echoes that of Kim, Sugar, and Belin (2015). Assuming exchangeability of subjects, the full conditional distribution for x_{mis} may be factored as

$$p\left(\mathbf{x}_{\text{mis}} \mid \boldsymbol{\Omega}, \mathbf{y}, \mathbf{x}_{\text{obs}}, \boldsymbol{\beta}_A, \sigma^2_\varepsilon\right) = \prod_{i=1}^{n_{\text{mis}}} p\left(\mathbf{x}_{i\text{mis}} \mid \boldsymbol{\Omega}, y_i, \mathbf{x}_{i\text{obs}}, \boldsymbol{\beta}_A, \sigma^2_\varepsilon\right),$$

where n_{mis} is the number of subjects with missing data on the predictors. The implication is we have a full conditional distribution for each subject.

We describe the full conditional for the missing values in a univariate fashion, that is, in terms of each missing value separately. Extensions to multivariate full conditionals are possible, in particular to accommodate multiple cases with the same pattern of missing data at the same time, but the univariate approach facilitates later comparisons with alternative univariate and conditional specifications for the distributions of the observed variables. Let x_{is} denote the value for subject i on predictor s, which is missing, and let $\mathbf{x}_{i(\text{-}s)}$ denote the values for subject i on the remaining predictors. The full conditional for the missing value is

$$p\left(x_{is} \mid \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x, y_i, \mathbf{x}_{i(-s)}, \boldsymbol{\beta}_A, \sigma^2_\varepsilon\right) \propto p\left(y_i \mid \boldsymbol{\beta}_A, \sigma^2_\varepsilon, x_{is}, \mathbf{x}_{i(-s)}\right) p\left(x_{is} \mid \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x, \mathbf{x}_{i(-s)}\right). \tag{11}$$

The first term on the right-hand side is the conditional distribution of the outcome, given the predictors and level-1 model parameters. This can be written to isolate the missing predictor:

$$y_i \mid \boldsymbol{\beta}_A, \sigma^2_\varepsilon, x_{is}, \mathbf{x}_{i(-s)} \sim N\left(\mathbf{x}'_{Ai(-s)}\tilde{\boldsymbol{\beta}}_{A(-s)} + \tilde{\beta}_s x_{is}, \sigma^2_\varepsilon\right),$$

where $\tilde{\beta}_s$ is the *effective coefficient* for x_s, obtained by factoring x_s out of $\mathbf{x}'_{Ai}\boldsymbol{\beta}_A$, $\tilde{\boldsymbol{\beta}}_{A(-s)}$ are the remaining coefficients in the model, and $\mathbf{x}'_{Ai(-s)}$ are the remaining augmented predictors. For example, let us return to the case of a model with two xs (x_1, x_2) and their interaction (x_1x_2) as predictors. As before, we may write the model as

$$y_i \mid \boldsymbol{\beta}_A, \sigma^2_\varepsilon, x_{is}, \mathbf{x}_{i(-s)} \sim N\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}, \sigma^2_\varepsilon\right).$$

Now suppose there is missingness for subject i on x_1. In this case, factoring out x_1 yields the equivalent representation

$$y_i \mid \boldsymbol{\beta}_A, \sigma^2_\varepsilon, x_{is}, \mathbf{x}_{i(-s)} \sim N\left(\beta_0 + \beta_2 x_{i2} + (\beta_1 + \beta_3 x_{i2}) x_{i1}, \sigma^2_\varepsilon\right).$$

In this terms of the more general notation, $\tilde{\beta}_s = (\beta_1 + \beta_3 x_{i2})$ is the effective coefficient for x_1, $\tilde{\boldsymbol{\beta}}_{A(-s)} = (\beta_0, \beta_2)$, and $\mathbf{x}'_{Ai(-s)} = (1, x_{i2})$. In the full conditional for x_{i1}, all other terms are treated as known. We can therefore rewrite the conditional distribution of y_i as

$$\left[y_i - (\beta_0 + \beta_2 x_{i2})\right] \left| \boldsymbol{\beta}_A, \sigma^2_\varepsilon, x_{is}, \mathbf{x}_{i(-s)} \right. \sim N\left((\beta_1 + \beta_3 x_{i2}) x_{i1}, \sigma^2_\varepsilon\right).$$

Returning to the more general case and notation, we may write the conditional distribution of y_{i} as

$$\left[y_i - \mathbf{x}'_{Ai(-s)}\tilde{\boldsymbol{\beta}}_{A(-s)}\right] \left| \boldsymbol{\beta}_A, \sigma^2_\varepsilon, x_{is}, \mathbf{x}_{i(-s)} \right. \sim N\left(\tilde{\beta}_s x_{is}, \sigma^2_\varepsilon\right).$$

We can recognize this form as one that conforms to that laid out above in the Foundations section as the standard Bayesian analysis for normal distributions, where the mean of the normal distribution is defined by having the unknown quantity (here, x_{is}) multiplied by a known quantity (here, $\tilde{\beta}_s$).

The second term on the right-hand side of (11) is the distribution for the missing value conditional on the other predictors, which is a univariate version of that given in (10):

$$x_{is} \mid \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x, \mathbf{x}_{i(-s)} \sim N\left(\mu_{x_s} + \boldsymbol{\Sigma}_{x_s x_{(-s)}} \boldsymbol{\Sigma}^{-1}_{x_{(-s)} x_{(-s)}} \left(\mathbf{x}_{i(-s)} - \boldsymbol{\mu}_{x_{(-s)}}\right), \sigma^2_{x_s} - \boldsymbol{\Sigma}_{x_s x_{(-s)}} \boldsymbol{\Sigma}^{-1}_{x_{(-s)} x_{(-s)}} \boldsymbol{\Sigma}_{x_{(-s)} x_s}\right)$$

where

$$\boldsymbol{\mu}_x = \begin{bmatrix} \mu_{x_s} \\ \boldsymbol{\mu}_{x_{(-s)}} \end{bmatrix} \text{ with sizes } \begin{bmatrix} 1 \times 1 \\ (P-1) \times 1 \end{bmatrix};$$

$$\boldsymbol{\Sigma}_x = \begin{bmatrix} \sigma^2_{x_s} & \boldsymbol{\Sigma}_{x_s x_{(-s)}} \\ \boldsymbol{\Sigma}_{x_{(-s)} x_s} & \boldsymbol{\Sigma}_{x_{(-s)}} \end{bmatrix} \text{ with sizes } \begin{bmatrix} 1 \times 1 & 1 \times (P-1) \\ (P-1) \times 1 & (P-1) \times (P-1) \end{bmatrix}.$$

Following the standard Bayesian analysis for normal distributions, the full conditional distribution for the missing value on the predictors in (11) is

$$x_{is} \mid \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x, y_i, \mathbf{x}_{i(-s)}, \boldsymbol{\beta}_A, \sigma^2_\varepsilon \sim N\left(\mu_{x_{is} \mid \ldots}, \sigma^2_{x_{is} \mid \ldots}\right),$$

where

$$\mu_{x_{is} \mid \ldots} = \left(\frac{1}{\sigma^2_{x_s} - \boldsymbol{\Sigma}_{x_s x_{(-s)}} \boldsymbol{\Sigma}^{-1}_{x_{(-s)}} \boldsymbol{\Sigma}_{x_{(-s)} x_s}} + \frac{\tilde{\beta}^2_s}{\sigma^2_\varepsilon}\right)^{-1}$$
$$\left(\frac{\mu_{x_s} + \boldsymbol{\Sigma}_{x_s x_{(-s)}} \boldsymbol{\Sigma}^{-1}_{x_{(-s)}} \left(\mathbf{x}_{Ai(-s)} - \boldsymbol{\mu}_{x_{(-s)}}\right)}{\sigma^2_{x_s} - \boldsymbol{\Sigma}_{x_s x_{(-s)}} \boldsymbol{\Sigma}^{-1}_{x_{(-s)}} \boldsymbol{\Sigma}_{x_{(-s)} x_s}} + \frac{\tilde{\beta}_s \left(y_i - \mathbf{x}'_{Ai(-s)} \tilde{\boldsymbol{\beta}}_{A(-s)}\right)}{\sigma^2_\varepsilon}\right)$$

$$\text{and } \sigma^2_{x_{is} \mid \ldots} = \left(\frac{1}{\sigma^2_{x_s} - \boldsymbol{\Sigma}_{x_s x_{(-s)}} \boldsymbol{\Sigma}^{-1}_{x_{(-s)}} \boldsymbol{\Sigma}_{x_{(-s)} x_s}} + \frac{\tilde{\beta}^2_s}{\sigma^2_\varepsilon}\right)^{-1}.$$

4.1.2. Full conditional for the parameters that govern the distribution of the predictors

Turning to the parameters that govern the distribution of the predictors, the full conditional for $\boldsymbol{\Omega}$ may be expressed as

$$p(\boldsymbol{\Omega} \mid \mathbf{x}_{\text{mis}}, \mathbf{x}_{\text{obs}}) \propto p(\mathbf{x}_{\text{mis}}, \mathbf{x}_{\text{obs}} \mid \boldsymbol{\Omega}) p(\boldsymbol{\Omega}) = p\left(\mathbf{x}_{\text{mis}}, \mathbf{x}_{\text{obs}} \mid \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x\right) p(\boldsymbol{\mu}_x) p(\boldsymbol{\Sigma}_x)$$

and we may proceed with the means and the covariance matrix separately.

The full conditional for the means of the predictors is

$$p\left(\boldsymbol{\mu}_x \mid \mathbf{x}, \boldsymbol{\Sigma}_x, \boldsymbol{\mu}_{\boldsymbol{\mu}_x}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}_x}\right) \propto p\left(\mathbf{x} \mid \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x\right) p\left(\boldsymbol{\mu}_x \mid \boldsymbol{\mu}_{\boldsymbol{\mu}_x}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}_x}\right).$$

The first term is the normal distribution of the predictors. The second term is the normal prior distribution for the means of the predictors. Following the standard theory, the full conditional distribution is

$$\boldsymbol{\mu}_x \mid \mathbf{x}, \boldsymbol{\Sigma}_x, \boldsymbol{\mu}_{\boldsymbol{\mu}_x}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}_x} \sim N\left(\boldsymbol{\mu}_{\boldsymbol{\mu}_x \mid \ldots}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}_x \mid \ldots}\right),$$

where

$$\boldsymbol{\mu}_{\boldsymbol{\mu}_x|\ldots} = \left(\boldsymbol{\Sigma}_{\boldsymbol{\mu}_x}^{-1} + n\boldsymbol{\Sigma}_x^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_{\boldsymbol{\mu}_x}^{-1}\boldsymbol{\mu}_{\boldsymbol{\mu}_x} + n\boldsymbol{\Sigma}_x^{-1}\bar{\mathbf{x}}\right),\ \boldsymbol{\Sigma}_{\boldsymbol{\mu}_x|\ldots} = \left(\boldsymbol{\Sigma}_{\boldsymbol{\mu}_x}^{-1} + n\boldsymbol{\Sigma}_x^{-1}\right)^{-1},$$

and $\bar{\mathbf{x}}$ is the $(P \times 1)$ vector of means of the treated-as-known predictors over the n subjects.

The full conditional distribution of the covariance matrix of the predictors is

$$p\left(\boldsymbol{\Sigma}_x \mid \mathbf{x}, \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{x0}, d\right) \propto p\left(\mathbf{x} \mid \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x\right)p(\boldsymbol{\Sigma}_x \mid \boldsymbol{\Sigma}_{x0}, d).$$

The first term is the normal distribution of the predictors. The second term is the inverse-Wishart prior distribution. By the standard Bayesian analyses, the full conditional distribution is

$$\boldsymbol{\Sigma}_x \mid \mathbf{x}, \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{x0}, d \sim \text{Inv-Wishart}(d\boldsymbol{\Sigma}_{x0} + n\mathbf{S}_x, d + n),$$

where

$$\mathbf{S}_x = \frac{1}{n}\sum_i (\mathbf{x}_i - \boldsymbol{\mu}_x)\left(\mathbf{x}_i - \boldsymbol{\mu}_x\right)'.$$

5. Two-level regression with varying intercepts and slopes and level-2 covariates

5.1. Core model specification

Here we describe a two-level model with varying group-specific intercepts and slopes for groups (clusters) with covariates at level 2. An expansion of the notation is necessary. Let $y_{i(j)}$ denote the outcome for subject i in group j, and let $\mathbf{y}_j$ denote the vector of the $y_{i(j)}$s from group j. Similarly, let $\mathbf{x}_{i(j)}$ denote the variables that form the level-1 predictors for subject i in group j, and let $\mathbf{x}_j$ the collection of these from the subjects in group j. Let $\mathbf{x}_{Aj}$ denote the augmented matrix of predictors in group j, including $\mathbf{x}_j$ and any transformations of them (e.g., a vector of 1 s, an interaction). Let $\boldsymbol{\beta}_{Aj}$ denote the augmented vector of level-1 regression coefficients for group j, and let $\boldsymbol{\beta}_A = (\boldsymbol{\beta}_{A1}, \ldots, \boldsymbol{\beta}_{A1})'$ denote the matrix with the full collection of level-1 coefficients.

The level-1 model for group j is then

$$\mathbf{y}_j \mid \boldsymbol{\beta}_{Aj}, \sigma_\varepsilon^2, \mathbf{x}_j \sim N\left(\mathbf{x}_{Aj}\boldsymbol{\beta}_{Aj}, \sigma_\varepsilon^2\mathbf{I}\right) \text{ for } j = 1, \ldots, J. \tag{12}$$

Formulating the model at the individual level, the level-1 model for subject i who is a member of group j is

$$y_{i(j)} \mid \boldsymbol{\beta}_{Aj}, \sigma_\varepsilon^2, \mathbf{x}_{i(j)} \sim N\left(\mathbf{x}'_{Ai(j)}\boldsymbol{\beta}_{Aj}, \sigma_\varepsilon^2\right) \text{ for } i = 1, \ldots, n_j, j = 1, \ldots, J,$$

where $\mathbf{x}'_{Ai(j)}$ is row i of $\mathbf{x}_{Aj}$ containing the augmented set of predictors for subject i in group j. The second level of the model specifies regression structures for the level-1 coefficients with level-2 covariates denoted as νs. In the current development we allow the first P level-2 covariates to be the group-specific means of the level-1 predictors and specify additional level-2 covariates. $\mathbf{v}'_1 = (\mu_{j1}, \ldots, \mu_{jP}, \nu_{1(P+1)}, \ldots, \nu_{1Q})$ denotes the level-2 predictors for group j and

$$\mathbf{v} = \begin{pmatrix} \mathbf{v}'_1 \\ \vdots \\ \mathbf{v}'_j \end{pmatrix} = \begin{bmatrix} \mu_{11} & \cdots & \mu_{1P} & \nu_{1(P+1)} & \cdots & \nu_{1Q} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mu_{J1} & \cdots & \mu_{JP} & \nu_{J(P+1)} & \cdots & \nu_{JQ} \end{bmatrix},$$

denotes the full collection across all groups, with μ_{jp} denoting the mean for level-1 predictor p in group j. As with the level-1 predictors, we may define an augmented level-2 predictor matrix $\mathbf{v}_A$, with rows $\mathbf{v}'_{Aj}$ containing the augmented predictors for group j.

Letting S denote the number of level-1 coefficients in $\boldsymbol{\beta}_{Aj}$, the second level of the model specifies, for group j:

$$\boldsymbol{\beta}_{Aj} \,|\, \boldsymbol{\Gamma}_A, \boldsymbol{\tau}, \mathbf{v} = \begin{bmatrix} \beta_{0j} & \beta_{1j} & \cdots & \beta_{Sj} \end{bmatrix}' \,|\, \boldsymbol{\Gamma}_A, \boldsymbol{\tau}, \mathbf{v} \sim N\left(\boldsymbol{\Gamma}_A \mathbf{v}'_{Aj}, \boldsymbol{\tau}\right) \tag{13}$$

where, in addition to terms previously defined,

$$\boldsymbol{\Gamma}_A = \begin{bmatrix} \boldsymbol{\gamma}'_0 & \cdots & \boldsymbol{\gamma}'_s \end{bmatrix}'$$

is the full collection of augmented level-2 coefficients with each row containing the coefficients for the level-2 predictors in predicting the level-1 coefficients and $\boldsymbol{\tau}$ is a covariance matrix for the augmented level-1 coefficients. Importantly, the means of the level-1 predictors (the μ_js) can be within-group aggregates of the observed level-1 predictors, or they can be modeled as distinct parameters with distributions, akin to latent variables (Lüdtke et al. 2008, 2011; Shin and Raudenbush 2010). Following the recommendation of Grund, Lüdtke, and Robitzsch (2018), we use latent group means approach because doing so naturally accommodates unequal group sizes.

To accommodate latent means, we specify the distribution for the predictors as a group-specific multivariate normal distribution. For subject i in group j,

$$\mathbf{x}_{i(j)} = \begin{bmatrix} x_{i(j)1} & \cdots & x_{i(j)P} \end{bmatrix}' \sim N\left(\boldsymbol{\mu}_{xj}, \boldsymbol{\Sigma}_{xj}\right),$$

where $\boldsymbol{\mu}_{xj}$ contains the means of the level-1 predictors for group j (which, under the latent means specification adopted here, also are a part of $\mathbf{v}_j$) and $\boldsymbol{\Sigma}_{xj}$ is the covariance matrix for the level-1 predictors in group j.

The group-specific means and covariances require prior distributional specifications. The conditionally conjugate prior distribution for the covariance matrix is:

$$\boldsymbol{\Sigma}_{xj} \sim \text{Inv} - \text{Wishart}(\boldsymbol{\Sigma}_{x0}, d_x). \tag{14}$$

To specify a distribution for the means for the group, $\boldsymbol{\mu}_{xj}$, recall that the means are also used as predictors at level 2. We accomplish the specification of a distribution for the means via the distribution for the larger set of elements that form the predictors at level 2. For group j, we specify the predictors to follow a normal distribution:

$$\mathbf{v}_j = \begin{bmatrix} \mu_{xj1} & \cdots & \mu_{xjP} & v_{j(P+1)} & \cdots & v_{jQ} \end{bmatrix}' \sim N(\boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v). \tag{15}$$

Employing conditionally conjugate prior distributions for the parameters that govern the distribution of the level-2 predictors, we have

$$\boldsymbol{\mu}_v \sim N\left(\boldsymbol{\mu}_{\boldsymbol{\mu}_v}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}_v}\right) \tag{16}$$

$$\boldsymbol{\Sigma}_v \sim \text{Inv} - \text{Wishart}(\boldsymbol{\Sigma}_{v0}, d_v). \tag{17}$$

5.2. Incorporating missingness at level-1

Now suppose the we have missing values among the level-1 predictors. As in the single-level model, we will need to structure the distribution of the observed data, the missing data, and the parameters that govern the data. Let $\boldsymbol{\Omega}_{1j} = \left(\boldsymbol{\mu}_{xj}, \boldsymbol{\Sigma}_{xj}\right)$ denote the parameters governing the distribution of the level-1 predictors in group j, and let $\boldsymbol{\Omega}_1 = (\boldsymbol{\Omega}_{11}, \ldots, \boldsymbol{\Omega}_{1J})$ denote the full collection of those parameters. The joint distribution is then

$$\begin{aligned} p(\mathbf{x}_{\text{mis}}, \mathbf{x}_{\text{obs}}, \boldsymbol{\Omega}_1) &= p(\mathbf{x}_{\text{mis}} \,|\, \boldsymbol{\Omega}_1, \mathbf{x}_{\text{obs}}) p(\mathbf{x}_{\text{obs}} \,|\, \boldsymbol{\Omega}_1) p(\boldsymbol{\Omega}_1) \\ &= \prod_{j=1}^{J} \prod_{j=1}^{n_j} p\left(\mathbf{x}_{i(j)\text{mis}} \,|\, \boldsymbol{\mu}_{xj}, \boldsymbol{\Sigma}_{xj}, \mathbf{x}_{i(j)\text{obs}}\right) p\left(\mathbf{x}_{i(j)\text{obs}} \,|\, \boldsymbol{\mu}_{xj}, \boldsymbol{\Sigma}_{xj}\right) p\left(\boldsymbol{\mu}_{xj}, \boldsymbol{\Sigma}_{xj}\right) \end{aligned}$$

where the third term has already been specified via the conditionally conjugate priors in (14) and (15). The remaining terms are the distribution of the subject's observed values

$$\mathbf{x}_{i(j)\text{obs}} \mid \boldsymbol{\mu}_{xj}, \boldsymbol{\Sigma}_{xj} \sim N\left(\boldsymbol{\mu}_{x_{i(j)\text{obs}}}, \boldsymbol{\Sigma}_{x_{i(j)\text{obs}}}\right),$$

and the distribution of the subject's missing values

$$\begin{aligned}\mathbf{x}_{i(j)\text{mis}} \mid \boldsymbol{\mu}_{xj}, \boldsymbol{\Sigma}_{xj}, \mathbf{x}_{i(j)\text{obs}} \sim N\Big(&\boldsymbol{\mu}_{x_{i(j)\text{mis}}} + \boldsymbol{\Sigma}_{x_{i(j)\text{mis}}x_{i(j)\text{obs}}}\boldsymbol{\Sigma}^{-1}_{x_{i(j)\text{obs}}}\left(\mathbf{x}_{i(j)\text{obs}} - \boldsymbol{\mu}_{x_{i(j)\text{obs}}}\right),\\ &\boldsymbol{\Sigma}_{x_{i(j)\text{mis}}} - \boldsymbol{\Sigma}_{x_{i(j)\text{mis}}x_{i(j)\text{obs}}}\boldsymbol{\Sigma}^{-1}_{x_{i(j)\text{obs}}}\boldsymbol{\Sigma}_{x_{i(j)\text{obs}}x_{i(j)\text{mis}}}\Big),\end{aligned} \tag{18}$$

where

$$\boldsymbol{\mu}_{xj} = \begin{bmatrix}\boldsymbol{\mu}_{x_{i(j)\text{mis}}}\\ \boldsymbol{\mu}_{x_{i(j)\text{obs}}}\end{bmatrix} \text{ and } \boldsymbol{\Sigma}_{xj} = \begin{bmatrix}\boldsymbol{\Sigma}_{x_{i(j)\text{mis}}} & \boldsymbol{\Sigma}_{x_{i(j)\text{mis}}x_{i(j)\text{obs}}}\\ \boldsymbol{\Sigma}_{x_{i(j)\text{obs}}x_{i(j)\text{mis}}} & \boldsymbol{\Sigma}_{x_{i(j)\text{obs}}}\end{bmatrix}$$

are the mean vector and covariance matrix for group j, now arranged in terms of the subsets of the observed and missing data for subject i. Note that, in essence, the distribution for the missing data in (18) is what emerges from regressing the predictors with missingness on the other predictors, within group j. Alternatives to this multivariate approach adopted here include a univariate approach and a conditional specification, discussed later.

5.3. Incorporating missingness at level-2

Now suppose the we have missing values among the level-2 predictors. Indeed we necessarily will, if we view the latent group means as missing values. From this perspective, there will always be missing values at level 2, but we also allow for the possibility of missingness of the other vs at level 2. As above, we will need to structure the distribution of the observed data, the missing data, and the parameters that govern the data. Let $\boldsymbol{\Omega}_2 = (\boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v)$ abstractly denote the parameters governing the distribution of the level-2 predictors. The joint distribution is then

$$\begin{aligned}p(\mathbf{v}_{\text{mis}}, \mathbf{v}_{\text{obs}}, \boldsymbol{\Omega}_2) &= p(\mathbf{v}_{\text{mis}} \mid \boldsymbol{\Omega}_2, \mathbf{v}_{\text{obs}})p(\mathbf{v}_{\text{obs}} \mid \boldsymbol{\Omega}_2)p(\boldsymbol{\Omega}_2)\\ &= \prod_{j=1}^{J} p\left(\mathbf{v}_{j\text{mis}} \mid \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v, \mathbf{v}_{j\text{obs}}\right)p\left(\mathbf{v}_{j\text{obs}} \mid \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v\right)p(\boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v),\end{aligned}$$

where the third term has already been specified via the conditionally conjugate priors in (16) and (17). The remaining terms are the distribution of the group's observed values

$$\mathbf{v}_{j\text{obs}} \mid \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v \sim N\left(\boldsymbol{\mu}_{v_{j\text{obs}}}, \boldsymbol{\Sigma}_{v_{j\text{obs}}}\right),$$

and the distribution of the group's missing values

$$\begin{aligned}\mathbf{v}_{j\text{mis}} \mid \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v, \mathbf{v}_{j\text{obs}} \sim N\Big(&\boldsymbol{\mu}_{v_{j\text{mis}}} + \boldsymbol{\Sigma}_{v_{j\text{obs}}v_{j\text{mis}}}\boldsymbol{\Sigma}^{-1}_{v_{j\text{obs}}}\left(\mathbf{v}_{j\text{obs}} - \boldsymbol{\mu}_{v_{j\text{obs}}}\right),\\ &\boldsymbol{\Sigma}_{v_{j\text{mis}}} - \boldsymbol{\Sigma}_{v_{j\text{mis}}v_{j\text{obs}}}\boldsymbol{\Sigma}^{-1}_{v_{j\text{obs}}}\boldsymbol{\Sigma}_{v_{j\text{obs}}v_{j\text{mis}}}\Big),\end{aligned} \tag{19}$$

where

$$\boldsymbol{\mu}_v = \begin{bmatrix}\boldsymbol{\mu}_{v_{j\text{mis}}}\\ \boldsymbol{\mu}_{v_{j\text{obs}}}\end{bmatrix} \text{ and } \boldsymbol{\Sigma}_v = \begin{bmatrix}\boldsymbol{\Sigma}_{v_{j\text{mis}}} & \boldsymbol{\Sigma}_{v_{j\text{mis}}v_{j\text{obs}}}\\ \boldsymbol{\Sigma}_{v_{j\text{obs}}v_{j\text{mis}}} & \boldsymbol{\Sigma}_{v_{j\text{obs}}}\end{bmatrix}$$

are the mean vector and covariance matrix, now arranged in terms of the subsets of the observed and missing data for group j. Note that, in essence, the distribution for the missing data in (19) is what emerges from regressing the predictors with missingness on the other predictors. Alternatives to this multivariate approach adopted here include a univariate approach and a conditional specification, discussed later.

5.4. The complete model and posterior distribution

We have specified all the relevant terms. Pulling them together and restating them all in one place, the posterior distribution is

$$p\left(\boldsymbol{\beta}_A, \sigma_\varepsilon^2, \boldsymbol{\Gamma}_A, \boldsymbol{\tau}, \boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2, \mathbf{y}_{\text{mis}}, \mathbf{x}_{\text{mis}}, \mathbf{v}_{\text{mis}} \mid \mathbf{y}_{\text{obs}}, \mathbf{x}_{\text{obs}}, \mathbf{v}_{\text{obs}}\right) \tag{20}$$

$$\begin{aligned}
&\propto \prod_{j=1}^{J}\prod_{i=1}^{n_j} p\left(y_{i(j)} \mid \boldsymbol{\beta}_{Aj}, \sigma_\varepsilon^2, \mathbf{x}_{i(j)}\right) p\left(\boldsymbol{\beta}_{Aj} \mid \boldsymbol{\Gamma}_A, \boldsymbol{\tau}, \mathbf{v}_j\right) \\
&\quad \times p\left(\mathbf{x}_{i(j)\text{mis}} \mid \boldsymbol{\mu}_{xj}, \boldsymbol{\Sigma}_{xj}, \mathbf{x}_{i(j)\text{obs}}\right) p\left(\mathbf{x}_{i(j)\text{obs}} \mid \boldsymbol{\mu}_{xj}, \boldsymbol{\Sigma}_{xj}\right) p\left(\boldsymbol{\Sigma}_{xj}\right) \\
&\quad \times p\left(\mathbf{v}_{j\text{mis}} \mid \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v, \mathbf{v}_{j\text{obs}}\right) p\left(\mathbf{v}_{j\text{obs}} \mid \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v\right) p(\boldsymbol{\mu}_v) p(\boldsymbol{\Sigma}_v) \\
&\quad \times p\left(\sigma_\varepsilon^2\right) p(\boldsymbol{\Gamma}_A) p(\boldsymbol{\tau})
\end{aligned}$$

where, taking each term in turn,

$$\begin{gathered}
y_{i(j)} \mid \boldsymbol{\beta}_{Aj}, \sigma_\varepsilon^2, \mathbf{x}_{i(j)} \sim N\left(\mathbf{x}'_{Ai(j)}\boldsymbol{\beta}_{Aj}, \sigma_\varepsilon^2\right) \text{ for } i = 1, \ldots, n_j, j = 1, \ldots, J, \\
\boldsymbol{\beta}_{Aj} \mid \boldsymbol{\Gamma}_A, \boldsymbol{\tau}, \mathbf{v} \sim N\left(\boldsymbol{\Gamma}_A \mathbf{v}'_{Aj}, \boldsymbol{\tau}\right) \text{ for } j = 1, \ldots, J, \\
\mathbf{x}_{i(j)\text{mis}} \mid \boldsymbol{\mu}_{xj}, \boldsymbol{\Sigma}_{xj}, \mathbf{x}_{i(j)\text{obs}} \sim N\Big(\boldsymbol{\mu}_{x_{i(j)\text{mis}}} + \boldsymbol{\Sigma}_{x_{i(j)\text{mis}}x_{i(j)\text{obs}}}\boldsymbol{\Sigma}^{-1}_{x_{i(j)\text{obs}}}\left(\mathbf{x}_{i(j)\text{obs}} - \boldsymbol{\mu}_{x_{i(j)\text{obs}}}\right), \\
\boldsymbol{\Sigma}_{x_{i(j)\text{mis}}} - \boldsymbol{\Sigma}_{x_{i(j)\text{mis}}x_{i(j)\text{obs}}}\boldsymbol{\Sigma}^{-1}_{x_{i(j)\text{obs}}}\boldsymbol{\Sigma}_{x_{i(j)\text{obs}}x_{i(j)\text{mis}}}\Big) \\
\text{for } i = 1, \ldots, n_j, \ j = 1, \ldots, J, \\
\mathbf{x}_{i(j)\text{obs}} \mid \boldsymbol{\mu}_{xj}, \boldsymbol{\Sigma}_{xj} \sim N\left(\boldsymbol{\mu}_{x_{i(j)\text{obs}}}, \boldsymbol{\Sigma}_{x_{i(j)\text{obs}}}\right) \text{ for } i = 1, \ldots, n_j, j = 1, \ldots, J, \\
\boldsymbol{\Sigma}_{xj} \sim \text{Inv-Wishart}(\boldsymbol{\Sigma}_{x0}, d_x) \text{ for } j = 1, \ldots, J, \\
\mathbf{v}_{j\text{mis}} \mid \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v, \mathbf{v}_{j\text{obs}} \sim N\Big(\boldsymbol{\mu}_{v_{j\text{mis}}} + \boldsymbol{\Sigma}_{v_{j\text{obs}}v_{j\text{mis}}}\boldsymbol{\Sigma}^{-1}_{v_{j\text{obs}}}\left(\mathbf{v}_{j\text{obs}} - \boldsymbol{\mu}_{v_{j\text{obs}}}\right), \\
\boldsymbol{\Sigma}_{v_{j\text{mis}}} - \boldsymbol{\Sigma}_{v_{j\text{mis}}v_{j\text{obs}}}\boldsymbol{\Sigma}^{-1}_{v_{j\text{obs}}}\boldsymbol{\Sigma}_{v_{j\text{obs}}v_{j\text{mis}}}\Big) \\
\text{for } j = 1, \ldots, J, \\
\mathbf{v}_{j\text{obs}} \mid \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v \sim N\left(\boldsymbol{\mu}_{v_{j\text{obs}}}, \boldsymbol{\Sigma}_{v_{j\text{obs}}}\right) \text{ for } j = 1, \ldots, J, \\
\boldsymbol{\mu}_v \sim N\left(\boldsymbol{\mu}_{\boldsymbol{\mu}_v}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}_v}\right), \\
\boldsymbol{\Sigma}_v \sim \text{Inv-Wishart}(\boldsymbol{\Sigma}_{v0}, d_v), \\
\sigma_\varepsilon^2 \sim \text{Inv-Gamma}\left(\nu_0/2, \nu_0\sigma^2_{\varepsilon_0}/2\right), \\
\boldsymbol{\gamma} = vec(\boldsymbol{\Gamma}_A) \sim N\left(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\right), \\
\text{and } \boldsymbol{\tau} \sim \text{Inv-Wishart}(\boldsymbol{\tau}_0, d_\tau).
\end{gathered}$$

Several aspects are worth noting. First, we do not have a separate entry for the latent means, as they are modeled as missing level-2 predictors. Second, different software packages have different capabilities of specifying such a model. We defer further discussion until after presenting the full conditional distributions.

5.5. Full conditional distributions

Full conditional for the level-1 coefficients

For each group-specific set of level-1 coefficients the full conditional can be expressed as

$$p\left(\boldsymbol{\beta}_{Aj} \mid \sigma_{\varepsilon}^2, \boldsymbol{\Gamma}_A, \boldsymbol{\tau}, \mathbf{y}_j, \mathbf{x}_j, \mathbf{v}_j\right) \propto p\left(\mathbf{y}_j \mid \boldsymbol{\beta}_{Aj}, \sigma_{\varepsilon}^2, \mathbf{x}_j\right) p\left(\boldsymbol{\beta}_{Aj} \mid \boldsymbol{\Gamma}_A, \boldsymbol{\tau}, \mathbf{v}_j\right).$$

The first term on the right-hand side is the normal distribution of the outcomes in (12). The second term is the normal distribution in (13). Here again we have the potential to apply the results of the standard Bayesian theory for normal models. The resulting full conditional distribution for $\boldsymbol{\beta}_{Aj}$ is

$$\boldsymbol{\beta}_{Aj} \mid \sigma_{\varepsilon}^2, \boldsymbol{\Gamma}_A, \boldsymbol{\tau}, \mathbf{y}_j, \mathbf{x}_j, \mathbf{v}_j \sim N\left(\boldsymbol{\mu}_{\boldsymbol{\beta}_{Aj} \mid \ldots}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}_{Aj} \mid \ldots}\right), \tag{21}$$

where

$$\boldsymbol{\mu}_{\boldsymbol{\beta}_{Aj} \mid \ldots} = \left(\boldsymbol{\tau}^{-1} + \mathbf{x}_{Aj}'\left(\sigma_{\varepsilon}^2\mathbf{I}\right)^{-1}\mathbf{x}_{Aj}\right)^{-1}\left(\boldsymbol{\tau}^{-1}\boldsymbol{\Gamma}_A\mathbf{v}_{Aj} + \mathbf{x}_{Aj}'\left(\sigma_{\varepsilon}^2\mathbf{I}\right)^{-1}\mathbf{y}_j\right)$$

$$\text{and } \boldsymbol{\Sigma}_{\boldsymbol{\beta}_{Aj} \mid \ldots} = \left(\boldsymbol{\tau}^{-1} + \mathbf{x}_{Aj}'\left(\sigma_{\varepsilon}^2\mathbf{I}\right)^{-1}\mathbf{x}_{Aj}\right)^{-1}.$$

5.5.1. Full conditional for the level-1 error variance

Turning to σ_{ε}^2, the full conditional is analogous to that in the single-level regression, only now we must compute the sums of squares relative to a group-specific model and then aggregate them

$$\sigma_{\varepsilon}^2 \mid \boldsymbol{\beta}_A, \nu_0, \sigma_{\varepsilon_0}^2, \mathbf{y}, \mathbf{x}^{(1)} \sim \text{Inv-Gamma}\left(\frac{\nu_0 + n}{2}, \frac{\nu_0\sigma_{\varepsilon_0}^2 + SS(\mathbf{E})}{2}\right) \tag{22}$$

where

$$SS(\mathbf{E}) = \sum_j \left(\mathbf{y}_j - \mathbf{x}_{Aj}\boldsymbol{\beta}_{Aj}\right)'\left(\mathbf{y}_j - \mathbf{x}_{Aj}\boldsymbol{\beta}_{Aj}\right).$$

5.5.2. Full conditional for the level-2 coefficients

Turning to the level-2 coefficients γ, the full conditional can be expressed as

$$\begin{aligned} p\left(\boldsymbol{\gamma} \mid \boldsymbol{\beta}_A, \boldsymbol{\tau}, \mathbf{v}, \boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\right) &\propto p\left(\boldsymbol{\beta}_A \mid \boldsymbol{\gamma}, \boldsymbol{\tau}, \mathbf{v}\right) p\left(\boldsymbol{\gamma} \mid \boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\right) \\ &= \prod_j p\left(\boldsymbol{\beta}_{Aj} \mid \boldsymbol{\gamma}, \boldsymbol{\tau}, \mathbf{v}_j\right) p\left(\boldsymbol{\gamma} \mid \boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\right). \end{aligned}$$

The first term on the right-hand side is the multivariate normal distribution of the level-1 coefficients, and the second term on the right-hand side is the multivariate normal prior for the level-2 coefficients. We can recognize that we have the potential to apply the results of the standard Bayesian theory for normal models. Thus, the full conditional distribution for $\boldsymbol{\gamma}$ is

$$\boldsymbol{\gamma} \mid \boldsymbol{\beta}_A, \boldsymbol{\tau}, \mathbf{v}, \boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}} \sim N\left(\boldsymbol{\mu}_{\boldsymbol{\gamma} \mid \ldots}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma} \mid \ldots}\right), \tag{23}$$

where

$$\boldsymbol{\mu}_{\boldsymbol{\gamma} \mid \ldots} = \left[\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}{}^{-1} + \mathbf{v}_A'\mathbf{v}_A \otimes \boldsymbol{\tau}^{-1}\right]^{-1}\left[\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}{}^{-1}\boldsymbol{\mu}_{\boldsymbol{\gamma}} + (\mathbf{v}_A'\mathbf{v}_A \otimes \boldsymbol{\tau}^{-1})\hat{\boldsymbol{\gamma}}\right]$$

$$\text{with } \hat{\boldsymbol{\gamma}} = vec\left(\boldsymbol{\beta}_A'\mathbf{v}_A(\mathbf{v}_A'\mathbf{v}_A)^{-1}\right) \text{ and } \boldsymbol{\Sigma}_{\boldsymbol{\gamma} \mid \ldots} = \left[\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}{}^{-1} + \mathbf{v}_A'\mathbf{v}_A \otimes \boldsymbol{\tau}^{-1}\right]^{-1}.$$

5.5.3. Full conditional for the level-2 error covariance

Turning to the level-2 error covariance matrix $\boldsymbol{\tau}$, the full conditional can be expressed as

$$p(\boldsymbol{\tau} \mid \boldsymbol{\beta}_A, \boldsymbol{\Gamma}_A, \mathbf{v}, \boldsymbol{\tau}_0, d_\tau) \propto p(\boldsymbol{\beta}_A \mid \boldsymbol{\Gamma}_A, \boldsymbol{\tau}, \mathbf{v}) p(\boldsymbol{\tau} \mid \boldsymbol{\tau}_0, d_\tau) = \prod_j p\left(\boldsymbol{\beta}_{Aj} \mid \boldsymbol{\Gamma}_A, \boldsymbol{\tau}, \mathbf{v}\right) p(\boldsymbol{\tau} \mid \boldsymbol{\tau}_0, d_\tau).$$

The first term on the right-hand side is the multivariate normal for the level-1 coefficients, and the second term on the right-hand side is the inverse-Wishart prior for the level-2 error covariance matrix. We can recognize that we have the potential to apply the results of the standard Bayesian theory for normal models. The full conditional for $\boldsymbol{\tau}$ is then

$$\boldsymbol{\tau} \mid \boldsymbol{\beta}_A, \boldsymbol{\Gamma}_A, \mathbf{x}, \mathbf{v}, \boldsymbol{\tau}_0, d_\tau \sim \text{Inv-Wishart}\left(d_\tau \boldsymbol{\tau}_0 + J\mathbf{S}_{\boldsymbol{\beta}_A}, d_\tau + J\right) \tag{24}$$

where

$$\mathbf{S}_{\boldsymbol{\beta}_A} = \frac{1}{J} \sum_j \left(\boldsymbol{\beta}_{Aj} - \boldsymbol{\Gamma}_A \mathbf{v}_j\right)' \left(\boldsymbol{\beta}_{Aj} - \boldsymbol{\Gamma}_A \mathbf{v}_j\right).$$

5.5.4. Full conditional for the missing values of the outcome

Assuming exchangeability of subjects within groups, the full conditional distribution for $\mathrm{y}_{\mathrm{mis}}$ is given by

$$p\left(\mathbf{y}_{\mathrm{mis}} \mid \boldsymbol{\beta}_A, \sigma_\varepsilon^2, \boldsymbol{\Gamma}_A, \boldsymbol{\tau}, \boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2, \mathbf{y}_{\mathrm{obs}}, \mathbf{x}, \mathbf{v}\right) = \prod_{j=1}^{J} \prod_{i=1}^{n_{j\mathrm{mis}}} p\left(y_{\mathrm{i}(j),\,\mathrm{mis}} \mid \boldsymbol{\beta}_{Aj}, \sigma_\varepsilon^2, \mathbf{x}_{i(j)}\right),$$

where $y_{i(j),\mathrm{mis}}$ is the value for the outcome for subject i in group j, which is missing, and $n_{j\mathrm{mis}}$ are the number of subjects with missing values on y in group j. Following the regression model, the full conditional distribution for each subject is

$$y_{i(j),\,mis} \mid \boldsymbol{\beta}_{Aj}, \sigma_\varepsilon^2, \mathbf{x}_{i(j)} \sim N\left(\mathbf{x}'_{Ai(j)} \boldsymbol{\beta}_{Aj}, \sigma_\varepsilon^2\right). \tag{25}$$

5.5.5. Full conditional for the missing values of the level-2 predictors

All of the latent means employed as level-2 predictors may be viewed as missing values. We may also have missing values on the additional level-2 predictors. We treat these two classes of missing values simultaneously; let $\mathrm{v}_{\mathrm{mis}}$ denote the missing values along all of these level-2 predictors.

Assuming exchangeability of groups, the full conditional distribution for $\mathbf{v}_{\mathrm{mis}}$ may be factored as

$$\begin{aligned} p(\mathbf{v}_{\mathrm{mis}} \mid \boldsymbol{\Omega}_2, \mathbf{v}_{\mathrm{obs}}, \boldsymbol{\beta}_A, \boldsymbol{\Gamma}_A, \boldsymbol{\tau}) &\propto p(\boldsymbol{\beta}_A \mid \boldsymbol{\Gamma}_A, \boldsymbol{\tau}, \mathbf{v}_{\mathrm{obs}}, \mathbf{v}_{\mathrm{mis}}) p(\mathbf{v}_{\mathrm{mis}} \mid \boldsymbol{\Omega}_2, \mathbf{v}_{\mathrm{obs}}) \\ &= \prod_{j=1}^{J} p\left(\boldsymbol{\beta}_{Aj} \mid \boldsymbol{\Gamma}_A, \boldsymbol{\tau}, \mathbf{v}_{j\mathrm{obs}}, \mathbf{v}_{j\mathrm{mis}}\right) p\left(\mathbf{v}_{j\mathrm{mis}} \mid \boldsymbol{\Omega}_2, \mathbf{v}_{j\mathrm{obs}}\right). \end{aligned}$$

The implication is that we have a full conditional distribution for each group.

We describe the full conditional for the missing values in a univariate fashion, that is, in terms of each missing value separately. Extensions to multivariate full conditionals are possible, notably for cases with the same pattern of missing data, but the univariate approach facilitates later comparisons with alternative univariate and conditional specifications for the distributions of the observed variables. Let v_{jq} denote the value for group j on predictor q, which is missing, and let $\mathbf{v}_{j(-q)}$ denote the values for group j on the remaining predictors.

The full conditional for the missing value is

$$p\left(v_{jq} \mid \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v, \boldsymbol{\beta}_{Aj}, \boldsymbol{\gamma}, \boldsymbol{\tau}, \mathbf{v}_{j(-q)}\right) \propto p\left(\boldsymbol{\beta}_{Aj} \mid \boldsymbol{\Gamma}_A, \boldsymbol{\tau}, v_{jq}, \mathbf{v}_{j(-q)}\right) p\left(v_{jq} \mid \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v, \mathbf{v}_{j(-q)}\right). \tag{26}$$

The first term on the right-hand side is the conditional distribution of the level-1 coefficients given in (13). As we did in the single-level regression, we can rewrite this isolating the role of the missing value as

$$\boldsymbol{\beta}_{Aj} \mid \boldsymbol{\Gamma}_A, \boldsymbol{\tau}, v_{jq}, \mathbf{v}_{j(-q)} \sim N\left(\tilde{\boldsymbol{\Gamma}}_{A(-q)} \mathbf{v}_{Aj(-q)} + \tilde{\boldsymbol{\gamma}}_q v_{jq}, \boldsymbol{\tau}\right),$$

where $\tilde{\boldsymbol{\gamma}}_q$ refers to the column of *effective coefficients* obtained by factoring v_q out of $\boldsymbol{\Gamma}_A \mathbf{v}'_{Aj}$ and $\tilde{\boldsymbol{\Gamma}}_{A(-q)}$ refers to the remaining coefficients in the model. Recognizing that $\tilde{\boldsymbol{\Gamma}}_{A(-q)} \mathbf{v}_{Aj(-q)}$ is a constant in the full conditional for v_{jq}, we can rewrite the conditional distribution of $\boldsymbol{\beta}_{Aj}$ as

$$\left[\boldsymbol{\beta}_{Aj} - \tilde{\boldsymbol{\Gamma}}_{A(-q)} \mathbf{v}_{Aj(-q)}\right] \Big| \boldsymbol{\Gamma}_A, \boldsymbol{\tau}, v_{jq}, \mathbf{v}_{j(-q)} \sim N\left(\tilde{\boldsymbol{\gamma}}_q v_{jq}, \boldsymbol{\tau}\right).$$

We can recognize this form as one that conforms to that laid out above as the standard Bayesian analysis for normal distributions, where the mean of the normal distribution is defined by having the unknown quantity (here, v_{jq}) multiplied by a known quantity (here, $\tilde{\boldsymbol{\gamma}}_q$).

The second term on the right-hand side of (26) is the distribution for the missing value conditional on the other predictors, which is a univariate version of that given in (19)

$$v_{jq} \sim N\left(\mu_{v_q} + \boldsymbol{\Sigma}_{v_q v_{(-q)}} \boldsymbol{\Sigma}^{-1}_{v_{(-q)}} \left(\mathbf{v}_{j(-q)} - \boldsymbol{\mu}_{v_{(-q)}}\right), \sigma^2_{v_q} - \boldsymbol{\Sigma}_{v_q v_{(-q)}} \boldsymbol{\Sigma}^{-1}_{v_{(-q)}} \boldsymbol{\Sigma}_{v_{(-q)} v_q}\right)$$

where

$$\boldsymbol{\mu}_v = \begin{bmatrix} \mu_{v_q} \\ \boldsymbol{\mu}_{v_{(-q)}} \end{bmatrix} \text{ with sizes } \begin{bmatrix} 1 \times 1 \\ (Q-1) \times 1 \end{bmatrix};$$

$$\boldsymbol{\Sigma}_v = \begin{bmatrix} \sigma^2_{v_q} & \boldsymbol{\Sigma}_{v_q v_{(-q)}} \\ \boldsymbol{\Sigma}_{v_{(-q)} v_q} & \boldsymbol{\Sigma}_{v_{(-q)}} \end{bmatrix} \text{ with sizes } \begin{bmatrix} 1 \times 1 & 1 \times (Q-1) \\ (Q-1) \times 1 & (Q-1) \times (Q-1) \end{bmatrix}.$$

In essence, this prior distribution is what emerges from regressing the predictor with missingness on the other predictors.

Following the standard Bayesian analysis for normal distributions, the full conditional distribution for the missing value on the predictor is

$$v_{jq} \mid \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v, \boldsymbol{\beta}_{Aj}, \boldsymbol{\gamma}, \boldsymbol{\tau}, \mathbf{v}_{j(-q)} \sim N\left(\mu_{v_{jq}|\ldots}, \sigma^2_{v_{jq}|\ldots}\right), \tag{27}$$

where

$$\mu_{v_{jq}|\ldots} = \left(\frac{1}{\sigma^2_{v_q} - \boldsymbol{\Sigma}_{v_q v_{(-q)}} \boldsymbol{\Sigma}^{-1}_{v_{(-q)}} \boldsymbol{\Sigma}_{v_{(-q)} v_q}} + \tilde{\boldsymbol{\gamma}}'_q \boldsymbol{\tau}^{-1} \tilde{\boldsymbol{\gamma}}_q\right)^{-1}$$
$$\left(\frac{\mu_{v_q} + \boldsymbol{\Sigma}_{v_q v_{(-q)}} \boldsymbol{\Sigma}^{-1}_{v_{(-q)}} \left(\mathbf{v}_{j(-q)} - \boldsymbol{\mu}_{v_{(-q)}}\right)}{\sigma^2_{v_q} - \boldsymbol{\Sigma}_{v_q v_{(-q)}} \boldsymbol{\Sigma}^{-1}_{v_{(-q)}} \boldsymbol{\Sigma}_{v_{(-q)} v_q}} + \tilde{\boldsymbol{\gamma}}'_q \boldsymbol{\tau}^{-1} \left(\boldsymbol{\beta}_{Aj} - \tilde{\boldsymbol{\Gamma}}_{A(-q)} \mathbf{v}_{Aj(-q)}\right)\right)$$
$$\text{and } \sigma^2_{v_{jq}|\ldots} = \left(\frac{1}{\sigma^2_{v_q} - \boldsymbol{\Sigma}_{v_q v_{(-q)}} \boldsymbol{\Sigma}^{-1}_{v_{(-q)}} \boldsymbol{\Sigma}_{v_{(-q)} v_q}} + \tilde{\boldsymbol{\gamma}}'_q \boldsymbol{\tau}^{-1} \tilde{\boldsymbol{\gamma}}_q\right)^{-1}.$$

5.5.6. Full conditional for the missing values of the level-1 predictors

Assuming exchangeability of groups and conditional exchangeability within groups, the full conditional distribution can be factored as

$$p\left(\mathbf{x}_{\text{mis}} \mid \mathbf{\Omega}_1, \mathbf{x}_{\text{obs}}, \mathbf{y}, \boldsymbol{\beta}_A, \sigma_\varepsilon^2\right) = \prod_{j=1}^{J} \prod_{i=1}^{n_{j\text{mis}}} p\left(\mathbf{x}_{i(j)\text{mis}} \mid \mathbf{\Omega}_{1j}, \mathbf{x}_{i(j)\text{obs}}, y_{i(j)}, \boldsymbol{\beta}_{Aj}, \sigma_\varepsilon^2\right),$$

where $n_{j\text{mis}}$ is the number of subjects with missing data on the predictors in group j. The implication is we have a full conditional distribution for each subject.

We describe the full conditional for the missing values in a univariate fashion, that is, in terms of each missing value separately. Extensions to multivariate full conditionals are possible, but the univariate approach facilitates later comparisons with alternative univariate and conditional specifications for the distributions of the observed variables. Let $x_{i(j)s}$ denote the value for subject i (in group j) on predictor s, which is missing, and let $\mathbf{x}_{i(j)(-s)}$ denote the values for subject i on the remaining predictors. The full conditional for the missing value is

$$\begin{aligned} p\left(x_{i(j)s} \mid \mathbf{\Omega}_{1j}, y_{i(j)}, \mathbf{x}_{i(j)(-s)}, \boldsymbol{\beta}_{Aj}, \sigma_\varepsilon^2\right) &\propto p\left(y_{i(j)} \mid \boldsymbol{\beta}_{Aj}, \sigma_\varepsilon^2, x_{i(j)s}, \mathbf{x}_{i(j)(-s)}\right) \\ &\quad \times p\left(x_{i(j)s} \mid \mathbf{\Omega}_{1j}, \mathbf{x}_{i(j)(-s)}\right). \end{aligned} \tag{28}$$

The first term on the right-hand side of is the conditional distribution of the outcome. Again, we can rewrite this expression to isolate the missing value,

$$y_{i(j)} \mid \boldsymbol{\beta}_{Aj}, \sigma_\varepsilon^2, x_{i(j)s}, \mathbf{x}_{i(j)(-s)} \sim N\left(\mathbf{x}'_{Ai(j)(-s)}\tilde{\boldsymbol{\beta}}_{Aj(-s)} + \tilde{\beta}_{sj}x_{i(j)s}, \sigma_\varepsilon^2\right),$$

where $\tilde{\beta}_{sj}$ is the *effective coefficient* for x_s in group j, obtained by factoring x_s out of $\mathbf{x}'_{Ai(j)}\boldsymbol{\beta}_{Aj}$, $\tilde{\boldsymbol{\beta}}_{Aj(-s)}$ are the remaining coefficients for group j, and $\mathbf{x}'_{Ai(j)(-s)}$ are the remaining augmented predictors. Recall that the model here accommodates interaction terms in a substantive model, as they appear in the matrix of augmented predictors $\mathbf{x}_A$ (see the Single-Level Regression, Missingness on a Predictor section for an example). We can rewrite the conditional distribution of the outcome as

$$\left[y_{i(j)} - \mathbf{x}'_{Ai(j)(-s)}\tilde{\boldsymbol{\beta}}_{Aj(-s)}\right] \Big| \boldsymbol{\beta}_{Aj}, \sigma_\varepsilon^2, x_{i(j)s}, \mathbf{x}_{i(j)(-s)} \sim N\left(\tilde{\beta}_{sj}x_{i(j)s}, \sigma_\varepsilon^2\right).$$

We can recognize this form as one that conforms to that laid out above as the standard Bayesian analysis for normal distributions, where the mean of the normal distribution is defined by having the unknown quantity (here, $x_{i(j)s}$) multiplied by a known quantity (here, $\tilde{\beta}_{sj}$).

The second term on the right-hand side of (28) is the prior distribution for the missing value conditional on the other predictors, which is univariate version of that given in (18):

$$x_{i(j)s} \sim N(\mu_{x_{js}} + \mathbf{\Sigma}_{x_{js}x_{j(-s)}}\mathbf{\Sigma}^{-1}_{x_{j(-s)}}(\mathbf{x}_{i(j)(-s)} - \boldsymbol{\mu}_{x_{j(-s)}}), \sigma^2_{x_{js}} - \mathbf{\Sigma}_{x_{js}x_{j(-s)}}\mathbf{\Sigma}^{-1}_{x_{j(-s)}}\mathbf{\Sigma}_{x_{j(-s)}x_{js}}) \tag{29}$$

where

$$\boldsymbol{\mu}_{xj} = \begin{bmatrix} \mu_{x_{js}} \\ \boldsymbol{\mu}_{x_{j(-s)}} \end{bmatrix} \text{ with sizes } \begin{bmatrix} 1 \times 1 \\ (P-1) \times 1 \end{bmatrix};$$

$$\mathbf{\Sigma}_{xj} = \begin{bmatrix} \sigma^2_{x_{js}} & \mathbf{\Sigma}_{x_{js}x_{j(-s)}} \\ \mathbf{\Sigma}_{x_{j(-s)}x_{js}} & \mathbf{\Sigma}_{x_{j(-s)}} \end{bmatrix} \text{ with sizes } \begin{bmatrix} 1 \times 1 & 1 \times (P-1) \\ (P-1) \times 1 & (P-1) \times (P-1) \end{bmatrix}.$$

In essence, this prior distribution is what emerges from regressing the predictor with missingness on the other predictors.

Appling the standard Bayesian analyses for normal distributions, the full conditional distribution for the missing value for the predictor is

$$p\left(x_{i(j)s} \mid \mathbf{\Omega}_{1j}, y_{i(j)}, \mathbf{x}_{i(j)(-s)}, \boldsymbol{\beta}_{Aj}, \sigma_\varepsilon^2\right) \sim N\left(\mu_{x_{i(j)s} \mid \ldots}, \sigma^2_{x_{i(j)s} \mid \ldots}\right), \tag{30}$$

where

$$\mu_{x_{i(j)s}\,|\,\cdots} = \left(\frac{1}{\sigma^2_{x_{js}} - \boldsymbol{\Sigma}_{x_{js}x_{j(-s)}}\boldsymbol{\Sigma}^{-1}_{x_{j(-s)}}\boldsymbol{\Sigma}_{x_{j(-s)}x_{js}}} + \frac{\tilde{\beta}^2_{sj}}{\sigma^2_{\varepsilon}}\right)^{-1}$$
$$\left(\frac{\mu_{x_{js}} + \boldsymbol{\Sigma}_{x_{js}x_{j(-s)}}\boldsymbol{\Sigma}^{-1}_{x_{j(-s)}}\left(\mathbf{x}_{i(j)(-s)} - \boldsymbol{\mu}_{x_{j(-s)}}\right)}{\sigma^2_{x_{js}} - \boldsymbol{\Sigma}_{x_{js}x_{j(-s)}}\boldsymbol{\Sigma}^{-1}_{x_{j(-s)}}\boldsymbol{\Sigma}_{x_{j(-s)}x_{js}}} + \frac{\tilde{\beta}_{sj}\left[y_{i(j)} - \mathbf{x}'_{Ai(j)(-s)}\tilde{\boldsymbol{\beta}}_{Aj(-s)}\right]}{\sigma^2_{\varepsilon}}\right)$$

$$\text{and } \sigma^2_{x_{i(j)s}\,|\,\cdots} = \left(\frac{1}{\sigma^2_{x_{js}} - \boldsymbol{\Sigma}_{x_{js}x_{j(-s)}}\boldsymbol{\Sigma}^{-1}_{x_{j(-s)}}\boldsymbol{\Sigma}_{x_{j(-s)}x_{js}}} + \frac{\tilde{\beta}^2_{sj}}{\sigma^2_{\varepsilon}}\right)^{-1}.$$

5.5.7. Full conditional for the parameters that govern the distribution of the level-2 predictors

Turning to the parameters that govern the distribution of the level-2 predictors, the full conditional for $\boldsymbol{\Omega}_2$ may be expressed as

$$p(\boldsymbol{\Omega}_2 \,|\, \mathbf{v}_{\text{mis}}, \mathbf{v}_{\text{obs}}) \propto p(\mathbf{v}_{\text{mis}}, \mathbf{v}_{\text{obs}} \,|\, \boldsymbol{\Omega}_2)p(\boldsymbol{\Omega}_2) = p\left(\mathbf{v}_{\text{mis}}, \mathbf{v}_{\text{obs}} \,|\, \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v\right)p(\boldsymbol{\mu}_v)p(\boldsymbol{\Sigma}_v),$$

and we may proceed with the means and the covariance matrix separately.

The full conditional for the means is

$$p\left(\boldsymbol{\mu}_v \,|\, \mathbf{v}, \boldsymbol{\Sigma}_v, \boldsymbol{\mu}_{\boldsymbol{\mu}_v}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}_v}\right) \propto p\left(\mathbf{v} \,|\, \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v\right)p\left(\boldsymbol{\mu}_v \,|\, \boldsymbol{\mu}_{\boldsymbol{\mu}_v}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}_v}\right).$$

The first term is the normal distribution of the predictors. The second term is the normal prior distribution for the means of the predictors. Following the standard theory, the full conditional distribution is

$$\boldsymbol{\mu}_v \,|\, \mathbf{v}, \boldsymbol{\Sigma}_v, \boldsymbol{\mu}_{\boldsymbol{\mu}_v}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}_v} \sim N\left(\boldsymbol{\mu}_{\boldsymbol{\mu}_v\,|\,\cdots}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}_v\,|\,\cdots}\right),$$

where

$$\boldsymbol{\mu}_{\boldsymbol{\mu}_v\,|\,\cdots} = \left(\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\mu}_v} + J\boldsymbol{\Sigma}^{-1}_v\right)^{-1}\left(\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\mu}_v}\boldsymbol{\mu}_{\boldsymbol{\mu}_v} + J\boldsymbol{\Sigma}^{-1}_v\bar{\mathbf{v}}\right),\ \ \boldsymbol{\Sigma}_{\boldsymbol{\mu}_v\,|\,\cdots} = \left(\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\mu}_v} + J\boldsymbol{\Sigma}^{-1}_v\right)^{-1},$$

and $\bar{\mathbf{v}}$ is the $(Q \times 1)$ vector of means of the treated-as-known predictors over the J groups.

The full conditional distribution of the covariance matrix of the predictors is

$$p\left(\boldsymbol{\Sigma}_v \,|\, \mathbf{v}, \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_{v0}, d_v\right) \propto p\left(\mathbf{v} \,|\, \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v\right)p(\boldsymbol{\Sigma}_v \,|\, \boldsymbol{\Sigma}_{v0}, d_v).$$

The first term is the normal distribution of the predictors. The second term is the inverse-Wishart prior distribution. By the standard Bayesian analyses, the full conditional distribution is

$$\boldsymbol{\Sigma}_v \,|\, \mathbf{x}, \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_{v0}, d_v \sim \text{Inv-Wishart}(d_v\boldsymbol{\Sigma}_{v0} + J\mathbf{S}_v, d_v + J),$$

where

$$\mathbf{S}_v = \frac{1}{J}\sum_j (\mathbf{v}_j - \boldsymbol{\mu}_v)\left(\mathbf{v}_j - \boldsymbol{\mu}_v\right)'.$$

5.5.8. Full conditional for the parameters that govern the distribution of the level-1 predictors

Turning to the parameters that govern the distribution of the level-1 predictors, the full conditional for $\boldsymbol{\Omega}_1$ may be expressed as

$$p(\boldsymbol{\Omega}_1 \,|\, \mathbf{x}_{\text{mis}}, \mathbf{x}_{\text{obs}}) \propto p(\mathbf{x}_{\text{mis}}, \mathbf{x}_{\text{obs}} \,|\, \boldsymbol{\Omega}_1)p(\boldsymbol{\Omega}_1) = p\left(\mathbf{x}_{j\text{mis}}, \mathbf{x}_{j\text{obs}} \,|\, \boldsymbol{\mu}_{xj}, \boldsymbol{\Sigma}_{xj}\right)p(\boldsymbol{\mu}_{xj})p\left(\boldsymbol{\Sigma}_{xj}\right)$$

and we may proceed with the means and the covariance matrix separately, for each group.

The full conditional for the means in each group has already been specified, as the means are part of the missing level-2 predictors. The relevant full conditional is given in (27). The full conditional distribution of the covariance matrix of the predictors in group j is

$$p\left(\boldsymbol{\Sigma}_{xj} \mid \mathbf{x}, \boldsymbol{\mu}_{xj}, \boldsymbol{\Sigma}_{x0}, d_x\right) \propto p\left(\mathbf{x} \mid \boldsymbol{\mu}_{xj}, \boldsymbol{\Sigma}_{xj}\right) p\left(\boldsymbol{\Sigma}_{xj} \mid \boldsymbol{\Sigma}_{x0}, d_x\right).$$

The first term is the normal distribution of the predictors. The second term is the inverse-Wishart prior distribution. By the standard Bayesian analyses, the full conditional distribution is

$$\boldsymbol{\Sigma}_{xj} \mid \mathbf{x}, \boldsymbol{\mu}_{xj}, \boldsymbol{\Sigma}_{x0}, d_x \sim \text{Inv-Wishart}\left(d_x \boldsymbol{\Sigma}_{x0} + n_j \mathbf{S}_{xj}, d_x + n_j\right),$$

where

$$\mathbf{S}_{xj} = \frac{1}{n_j} \sum_i (\mathbf{x}_{i(j)} - \boldsymbol{\mu}_{xj})(\mathbf{x}_{i(j)} - \boldsymbol{\mu}_{xj})'.$$

6. Discussion

In this paper we have derived analytical forms for the full conditional distributions for the parameters and missing data in multilevel models for normally distributed outcomes and predictors. We started with single-level models with no missing data and built up to two-level models with varying intercepts and slopes, with possibly missing values for the outcomes, level-1 predictors, and level-2 predictors. In doing so, we have sought to provide a coherent, complete account of the full conditional distributions that form the basis of a fully Bayesian analysis using Gibbs sampling, where the analyst seeks inference regarding both the missing data and the parameters of the model, as well as how to generate the imputations in the first phase of a multiple imputation approach. In the Bayesian analysis focusing on inference for model parameters, missing-data imputation is a means to the end of summarizing the posterior distribution of the substantive model parameters (e.g., the distributions of samples drawn from (21)-(24)). Meanwhile, invoking a multiple imputation perspective, estimating the model components provides a tool for generating imputations, which are then used in complete-data analyses to average over uncertainty about the missing values. Our work supports both a fully Bayesian approach and a multiple imputation approach, and it does so in coherent fashion for a broader set of models than has appeared in the literature to date. Furthermore, it clarifies that these full conditional distributions have known analytical forms, which obviates the need for more computationally expensive sampling strategies that previous research has suggested and current software employs (Goldstein, Carpenter, and Browne 2014; Enders, Du, and Keller 2020).

This work has employed multivariate normal distributions for the predictors. A special case occurs when using univariate normal distributions for the predictors. For the level-1 predictors, this could be accomplished in the current framework by specifying the $\boldsymbol{\Sigma}_x$s to be diagonal. Similarly, this could be accomplished for the level-2 predictors by specifying $\boldsymbol{\Sigma}_v$ to be diagonal. Several useful consequences emerge from this framework. First, the full conditional distributions for parameters that govern the predictors simplify. Instead of having an inverse-Wishart full conditional for the covariance matrix (in each case), we have a series of univariate inverse-gamma full conditionals for the variances.

More importantly, the full conditionals for the missing values simplify, such that they no longer involve the other predictors at this level. This is easiest to see in the simplest context of a single-level model with missingness on a predictor. If the prior distribution for the predictor is univariate normal, the full conditional becomes

$$x_{is} \mid \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x, y_i, \mathbf{x}_{i(-s)}, \boldsymbol{\beta}_A, \sigma^2_\varepsilon \sim N\left(\mu_{x_{is} \mid \ldots}, \sigma^2_{x_{is} \mid \ldots}\right),$$

where

$$\mu_{x_{is}\,|\,\ldots} = \left(\frac{1}{\sigma^2_{x_s}} + \frac{\tilde{\beta}^2_s}{\sigma^2_{\varepsilon}}\right)^{-1}\left(\frac{\mu_{x_s}}{\sigma^2_{x_s}} + \frac{\tilde{\beta}_s\left(y_i - \mathbf{x}'_{Ai(-s)}\tilde{\boldsymbol{\beta}}_{A(-s)}\right)}{\sigma^2_{\varepsilon}}\right) \text{ and } \sigma^2_{x_{is}\,|\,\ldots} = \left(\frac{1}{\sigma^2_{x_s}} + \frac{\tilde{\beta}^2_s}{\sigma^2_{\varepsilon}}\right)^{-1}.$$

The upshot is that, if the analyst wishes to allow the draws for the missing values of a predictor to involve the other predictors, specifying the missing predictor to have a univariate prior distribution will not suffice. In multiple imputation strategies, it is common to include all the predictors in generating an imputation for a missing value. The current work suggests this could be accomplished by specifying a fully Bayesian model with a multivariate prior for the predictors, but not by specifying a univariate prior.

Returning to the context of a multivariate prior, we noted that the full conditional for the missing value may been as combining a likelihood expression with a prior distribution, where the prior for the missing value may be seen as resulting from regressing the missing predictor on the remaining predictors. This could be specified directly by the analyst, by specifying a distribution for the missing value as conditional on the other predictors through a regression structure. For example, consider the distribution for the missing values for the level-2 predictors for group *j*. Instead of the multivariate normal specifications that induce the expression in (19), we could directly model the distribution via regression structures as

$$\mathbf{v}_{j\text{mis}}\,|\,\boldsymbol{\Omega}_2, \mathbf{v}_{j\text{obs}} \sim N\left(\mathbf{v}_{j\text{obs}}\boldsymbol{\eta}_A, \boldsymbol{\Sigma}_{\mathbf{v}_{j\text{mis}}\,|\,\mathbf{v}_{j\text{obs}}}\right), \tag{31}$$

where $\boldsymbol{\eta}_A$ is a set of augmented regression coefficients and $\boldsymbol{\Sigma}_{\mathbf{v}_{j\text{mis}}\,|\,\mathbf{v}_{j\text{obs}}}$ is the error covariance matrix from such a regression. This would more closely mimic that which is typically done in multiple imputation, where, for each missing predictor, a regression model is built regressing the predictor on the remaining predictors. We note in passing that that multiple imputation also typically uses the values of the outcome as another predictor in the regression model for a missing predictor. In the fully Bayesian framework adopted here, that is not necessary, as the relationship between the missing predictor and outcome variable is already captured via the likelihood expression involved in the full conditional.

The multivariate approach described here can be implemented in BUGS (Spiegelhalter et al. 2007) and Blimp (Keller and Enders 2019), but not JAGS (Plummer 2017, p. 15). The conditional specification may be appealing in situations where a multivariate specification is not available. To illustrate the so-called "sequential" decomposition, let us consider the simplest case of a single-level model with three predictors. The model for the predictors could be built with a sequence of univariate distributions following the partition from Ibrahim, Chen, and Lipsitz (2002). For subject *i*, one such partition orders the distributions as follows:

$$p(x_{i1}, x_{i2}, x_{i3}) = p(x_{i1}\,|\,x_{i2}, x_{i3})p(x_{i2}\,|\,x_{i3})p(x_{i3}). \tag{32}$$

Under the normality specifications here, each component of the composite density on the right follows from an additive linear regression model (with parameters that need prior specifications) with homoscedastic errors. Although beyond the scope of this work, it is important to note that the sequential parameterization can accommodate certain non-linearities among the predictors (e.g., x_{i2} could be a quadratic function of x_{i3}). Recent work by Erler and colleagues (Erler et al. 2016, 2019) discuss the sequential approach with simple illustrations in JAGS; sequential approaches are also available in 'mdmb' R package (Lüdtke, Robitzsch, and West 2020) and Blimp (Keller and Enders 2019). The current work is limited in that it assumes normality for the predictors. We believe this is a reasonable place to start because normality is commonly employed in applied settings, is often an adequate working model for empirical data, and may be seen as a conservative choice under mild assumptions (McElreath 2020, Ch. 10). Nevertheless, departures from normality are common (Micceri 1989), and the extent to which normality assumptions are

robust to departures from normality in this context is deserving of future research, as are extensions to this work that do not involve specifying normality.

As noted previously, another advantage of the multivariate approach is that it readily accommodates discrete variables by assuming an underlying normal latent variable distribution (Chib and Greenberg 1998; Goldstein et al. 2009; Carpenter and Kenward 2013; Enders, Keller, and Levy 2018; Enders, Du, and Keller 2020). For example, the Blimp application (Keller and Enders 2019) uses cumulative and multinomial probit models to accommodate ordinal and nominal variables, respectively). The estimation steps described herein apply perfectly to the underlying latent variables, but a Metropolis sampling step (Hastings 1970) is used to sample latent imputations that account for the discretized predictor in the substantive analysis model.

Extending the previous points, if multiple variables have missingess, the covariate models can be constructed through a similar factoring through recursive regression structures (i.e., regressing a predictor with missingness on a predictor with complete data, and then using both of those as predictors for yet another predictor with missingness). Again, this algorithmic approach is equivalent to our predictor if the sequence of models is additive linear regressions (with parameters that need prior specifications) with homoscedastic errors. In the context of imputation, this method is closely aligned with the substantive model-compatible approach in the literature (Goldstein, Carpenter, and Browne 2014; Bartlett et al. 2015). A fruitful direction for future research is to fully explicate the similarities and differences among the various parameterizations of the covariate distributions.

Finally, we note that multiple imputation strategies often involve auxiliary variables that are believed to be useful for imputing missing values of predictors, but not otherwise related to the outcome (Collins, Schafer, and Kam 2001). The results of this work suggest two possibilities for incorporating such auxiliary variables. First, they may be folded into the model as covariates, on par with the other predictors. As has been discussed above, by using the multivariate (or conditional) prior specification, the full conditionals for the predictors with missingness would then involve those auxiliary variables. The drawback to this approach is a loss of statistical efficiency in that it would model these auxiliary variables as related to the outcome variable, which may not concur with the researcher's theory. A second approach that avoids this limitation would be to adopt a conditional specification for the predictors with missingness by regressing them on auxiliary variables, which are *not* also included in the regression model for the outcome variable. We could view that component of the model as yet another regression; the full conditional distributions for that component would then follow what has been derived here. Further research might then extend what is done here to other approaches that include auxiliary variables in ways that do not change the theoretical model (Graham 2003).

Funding

Research reported here was supported by Award Number R305D150056, administered by the Institute of Education Sciences, U.S. Department of Education. The opinions expressed are those of the authors and do not necessarily reflect the positions or policies of the Institute of Education Sciences, or the U.S. Department of Education.

ORCID

Roy Levy http://orcid.org/0000-0001-7737-9176

References

Bartlett, J. W., S. R. Seaman, I. R. White, and J. R. Carpenter. 2015. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research* 24 (4): 462–87. doi:10.1177/0962280214521348.

Besag, J. 1974. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society Series B* 36:192–236.

Carpenter, J. R., H. Goldstein, and M. G. Kenward. 2011. REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. *Journal of Statistical Software* 45 (5):1–14. doi:10.18637/jss.v045.i05.

Carpenter, J. R., and M. G. Kenward. 2013. *Multiple imputation and its application*. West Sussex, UK: Wiley.

Chen, S.-H., and E. H. Ip. 2015. Behaviour of the Gibbs sampler when conditional distributions are potentially incompatible. *Journal of Statistical Computation and Simulation* 85 (16):3266–75. doi:10.1080/00949655.2014.968159.

Chib, S., and E. Greenberg. 1998. Analysis of multivariate probit models. *Biometrika* 85 (2):347–61. doi:10.1093/biomet/85.2.347.

Collins, L. M., J. L. Schafer, and C.-M. Kam. 2001. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods* 6 (4):330–51. doi:10.1037/1082-989X.6.4.330.

Enders, C. K. 2010. *Applied missing data analysis*. New York: The Guilford Press.

Enders, C. K., H. Du, and B. T. Keller. 2020. A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and nonlinear terms. *Psychological Methods* 25:88–112. doi:10.1037/met0000228.

Enders, C. K., B. T. Keller, and R. Levy. 2018. A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychological Methods* 23 (2):298–317. doi:10.1037/met0000148.

Erler, N. S., D. Rizopoulos, V. W. Jaddoe, O. H. Franco, and E. M. Lesaffre. 2019. Bayesian imputation of time-varying covariates in linear mixed models. *Statistical Methods in Medical Research* 28 (2):555–68. doi:10.1177/0962280217730851.

Erler, N. S., D. Rizopoulos, J. v Rosmalen, V. W. V. Jaddoe, O. H. Franco, and E. M. E. H. Lesaffre. 2016. Dealing with missing covariates in epidemiologic studies: A comparison between multiple imputation and a full Bayesian approach. *Statistics in Medicine* 35 (17):2955–74. doi:10.1002/sim.6944.

Gelfand, A. E., and A. F. M. Smith. 1990. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* 85 (410):398–409. doi:10.2307/2289776.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2013. *Bayesian data analysis*. 3rd ed. Boca Raton, FL: Chapman and Hall/CRC.

Gelman, A., and J. Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Goldstein, H., J. R. Carpenter, and W. J. Browne. 2014. Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 177 (2):553–64. doi:10.1111/rssa.12022.

Goldstein, H., J. Carpenter, M. G. Kenward, and K. A. Levin. 2009. Multilevel models with multivariate mixed response types. *Statistical Modelling* 9 (3):173–97. doi:10.1177/1471082X0800900301.

Graham, J. W. 2003. Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal* 10 (1):80–100. doi:10.1207/S15328007SEM1001_4.

Grund, S., O. Lüdtke, and A. Robitzsch. 2016. Multiple imputation of missing covariate values in multilevel models with random slopes: A cautionary note. *Behavior Research Methods* 48 (2):640–49. doi:10.3758/s13428-015-0590-3.

Grund, S., O. Lüdtke, and A. Robitzsch. 2018. Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods* 21 (1):111–49. doi:10.1177/1094428117703686.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57 (1):97–109. doi:10.2307/2334940.

Ibrahim, J. G., M.-H. Chen, and S. R. Lipsitz. 2002. Bayesian methods for generalized linear models with covariates missing at random. *Canadian Journal of Statistics* 30 (1):55–78. doi:10.2307/3315865.

Jackman, S. 2009. *Bayesian analysis for the social sciences*. Chichester, UK: Wiley.

Keller, B. T., and C. K. Enders. 2019. *Blimp user's guide (version 2)*.

Kim, S., C. A. Sugar, and T. R. Belin. 2015. Evaluating model-based imputation methods for missing covariates in regression models with interactions. *Statistics in Medicine* 34 (11):1876–88. doi:10.1002/sim.6435.

Lindley, D. V., and A. F. M. Smith. 1972. Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B* 34:1–41.

Liu, J., A. Gelman, J. Hill, Y.-S. Su, and J. Kropko. 2014. On the stationary distribution of iterative imputations. *Biometrika* 101 (1):155–73. doi:10.1093/biomet/ast044.

Lüdtke, O., H. W. Marsh, A. Robitzsch, U. Trautwein, T. Asparouhov, and B. Muthén. 2008. The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods* 13 (3):203–29. doi:10.1037/a0012869.

Lüdtke, O., A. Robitzsch, and S. Grund. 2017. Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods* 22 (1):141–65. doi:10.1037/met0000096.

Lüdtke, O., A. Robitzsch, H. W. Marsh, and U. Trautwein. 2011. A 2 × 2 taxonomy of multilevel latent contextual models: Accuracy–bias trade-offs in full and partial error correction models. *Psychological Methods* 16 (4): 444–67. doi:10.1037/a0024376.

Lüdtke, O., A. Robitzsch, and S. G. West. 2020. Regression models involving nonlinear effects with missing data: A sequential modeling approach using Bayesian estimation. *Psychological Methods* 25 (2):157–81. doi:10.1037/met0000233.

McElreath, R. 2020. *Statistical rethinking: A Bayesian course with examples in R and Stan*. 2nd ed. Boca Raton, FL: CRC Press.

Micceri, T. 1989. The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin* 105 (1): 156–66. doi:10.1037/0033-2909.105.1.156.

Plummer, M. 2017. *JAGS version 4.3.0 user manual*. https://mcmc-jags.sourceforge.io/

Quartagno, M., and J. Carpenter. 2018. Package 'jomo.' cran.r-project.org/web/packages/jomo/

Robert, C. P., and G. Casella. 2004. *Monte Carlo statistical methods*. 2nd ed. New York: Springer.

Robitzsch, A., and O. Lüdtke. 2021. Package 'mdmb.' (R package version 1.5-8). cran.r-project.org/web/packages/mdmb/mdmb.pdf

Rowe, D. B. 2003. *Multivariate Bayesian statistics: Models for source separation and signal unmixing*. Boca Raton, FL: Chapman and Hall/CRC.

Schafer, J. L., and R. M. Yucel. 2002. Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics* 11 (2):437–57. doi:10.1198/106186002760180608.

Shin, Y., and S. W. Raudenbush. 2010. A latent cluster-mean approach to the contextual effects model with missing data. *Journal of Educational and Behavioral Statistics* 35 (1):26–53. doi:10.3102/1076998609345252.

Speidel, M., J. Drechsler, and J. W. Sakshaug. 2018. Biases in multilevel analyses caused by cluster-specific fixed-effects imputation. *Behavior Research Methods* 50 (5):1824–40. doi:10.3758/s13428-017-0951-1.

Spiegelhalter, D. J., A. Thomas, A. G. Best, and D. Lunn. 2007. *WinBUGS user manual: Version 1.4.3*. Cambridge: MRC Biostatistics Unit.

van Buuren, S. 2011. Multiple imputation of multilevel data. In *Handbook of advanced multilevel analysis*, ed. by J. J. Hox and J. K. Roberts, 173–96. New York, NY: Routledge. doi:10.4324/9780203848852.ch10.

Yucel, R. M. 2008. Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 366 (1874):2389–403. doi:10.1098/rsta.2008.0038.

Zhang, Q., and L. Wang. 2017. Moderation analysis with missing data in the predictors. *Psychological Methods* 22 (4):649–66. doi:10.1037/met0000104.