Chapter

# Bayesian Multilevel Modeling in Dental Research

*Edilberta Tino-Salgado, Flaviano Godínez-Jaimes, Cruz Vargas-De-León, Norma Samanta Romero-Castro, Salvador Reyes-Fernández and Victor Othon Serna-Radilla*

## Abstract

Clinical designs in dentistry collect measurements of the teeth of each subject, forming complex data structures; however, standard statistical methods (Student's t-test, ANOVA, and regression models) do not treat the data as a grouped data type; that is, the measurements are treated as independent despite not being the case. A disadvantage of not considering the dependence on multilevel data is that if there is a significant correlation between the observations, it is ignored by the researcher and consequently finds statistically significant results when in fact they are not. Bayesian methods have the advantage of not assuming normality, unlike maximum likelihood estimation, and Bayesian methods are appropriate when you have small samples. We showed the minimum statistical theory for the use of multilevel models in dental research when the response variable is numerical. In this regard, it was proposed to carry out a Bayesian multilevel analysis to determine the clinical factors associated with the depth of periodontal probing. We adapted the bottom-up strategy to specify a multilevel model in the frequentist approach to the Bayesian approach. We checked the adequacy of the fit of the postulated model using posterior predictive density.

**Keywords:** periodontal probing depth, dental research, nested data structures, Bayesian multilevel modeling, bottom-up methodology

## 1. Introduction

The most widely used statistical methods in dental research are t-test, ANOVA (one, two and three factors), non-parametric tests, and regression models [1]. These methods assume that the observations of the studied variables are independent. Nested data structures are frequently found in dental research. An example is an experimental design in which multiple measurements are performed on the same individual. If, in addition to performing multiple measurements in an individual, we perform multiple measurements in each tooth, we will obtain a nested data structure. This nesting of the data results in grouped data. Typically, for clinical and dental data, contextual variables are measured in each individual (i.e., socioeconomic level,

educational level, etc.), and these characteristics can form another group of data. Considering the detection of bacterial plaque in each tooth of individuals who have a home with a high marginality index, two nested groups are distinguished, namely, teeth nested in individuals and individuals nested at group. The word "nested" can be understood as "within" or "contained in." It is to be expected that items from the same group may be more similar to each other than items from a different group; that is, measurements from one individual are expected to be more similar to each other in comparison with measurements from other individuals. This fact indicates that the assumption of independence does not apply to nested data. Multilevel models take into account the non-independence of the observations. One consequence of ignoring the dependence of observations is that the results of some tests may be statistically significant when, actually, they are not. Under the classical approach, the estimation of the parameters of a multilevel model is performed using maximum likelihood, which has optimal properties in many scenarios; however, problems such as non-compliance with model assumptions or lack of convergence of iterative methods can occur. The Bayesian approach has some advantages over the classical approach.

The purpose of this chapter is to show the minimum statistical theory for the use of multilevel models in dental research when the response variable is numerical. For this, we will remember the definitions of multilevel models and multilevel generalized linear models (MGLM), in addition to the main Bayesian concepts and their application to MGLM. We will use an adaptation of the bottom-up strategy to specify a multilevel model. Our adaptation proposal tries to use the Bayesian leave-one-out cross-validation (LOO-CV) between the different steps for the comparison of models. We will check the adequacy of the fit of the postulated model using posterior predictive density. Finally, we will provide an example of this model applied to a numerical response variable, such as periodontal probing.

## 2. Multilevel models

Multilevel models partition the variance of the dependent variable at different levels of data grouping. At least two types of variance are distinguished: *intra-group variance* $\sigma_w^2$, or individual-level variance (level one), and *between-group variance* $\sigma_b^2$, which defines the variation at the group level (level two).

The dependency of the observations in the same group is measured with the *intraclass correlation coefficient* (ICC). Shrout and Fleiss in 1979 defined the ICC as the ratio of the *between-group variance* and the *total variance* (the sum of the variances between groups and in intra-groups):

$$ICC = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2} \tag{1}$$

The ICC varies between 0 and 1, since the variance cannot be negative. Before using a multilevel model, it is necessary to determine whether the ICC is significant at each level of the data. To that end, using the null model (defined in the next section), we determine whether the variance of the residuals of each level is significant. If that occurs, the ICC is also significant, and this means that at the individual level, the observations are dependent, and therefore, it is necessary to use a multilevel model instead of an ordinary multiple regression model [2].

### 2.1 Two-level models

Let $y_{ij}$ be the dependent variable measured in the *i-th* individual in the *j-th* level-two unit (e.g., the j-th group); $i = 1, \ldots, N$, where $N$ is the total sample size, and $j = 1, \ldots, J$ for $J$ level-two units.

The simplest two-level model is the *null model (intercept-only model, unconditional means model, or one-way random-effects analysis of the variance)*. The model is defined by two equations:

$$\begin{aligned} y_{ij} &= \beta_{0j} + e_{ij} \\ \beta_{0j} &= \gamma_{00} + u_{0j} \end{aligned} \tag{2}$$

$\beta_{0j}$ is the mean of $y$ in the group $j$ that varies across groups; $e_{ij}$ is the individual variation around this mean; $\gamma_{00}$ is the overall intercept, that is, the grand mean of $y$; and $u_{0j}$ is the deviation of $\beta_{0j}$ with respect to $\gamma_{00}$.

Substitution of $\beta_{0j}$ in $y_{ij}$ produces the single-equation model:

$$y_{ij} = \gamma_{00} + u_{0j} + e_{ij} \tag{3}$$

Eq. (3) is composed of a fixed part, $\gamma_{00}$, and a random part corresponding to two random effects, $u_{0j}$ and $e_{ij}$. Assuming that $e_{ij} \sim N(0, \sigma^2)$ and $u_{0j} \sim N\left(0, \sigma_{u0}^2\right)$ and that $e_{ij}$ and $u_{0j}$ are independent, the variance of $y_{ij}$ in Eq. (3) is

$$\begin{aligned} var\left(y_{ij}\right) &= var\left(\gamma_{00} + u_{0j} + e_{ij}\right) \\ &= \sigma_{u0}^2 + \sigma^2 \end{aligned} \tag{4}$$

where $\sigma_{u0}^2 = \sigma_b^2$ is the variance between groups, and $\sigma^2 = \sigma_w^2$ is the variance within groups in Eq. (1) to calculate the ICC.

Now, let us consider a two-level model with two level-one independent variables, $x_1$ with a fixed effect $\alpha_1$, which does not vary between groups, and $x_2$ with a random effect $\beta_{2j}$, which does vary between groups.

This model is defined by

$$\begin{aligned} y_{ij} &= \beta_{0j} + \alpha_1 x_{1ij} + \beta_{2j} x_{2ij} + e_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} w_{1j} + u_{0j} \\ \beta_{2j} &= \gamma_{10} + \gamma_{11} w_{1j} + u_{1j} \end{aligned} \tag{5}$$

In the above equation, both $\beta_{0j}$ and $\beta_{2j}$ depend on a level-two independent variable; $w_1$ and $u_{qj}$ are the deviation of the effect of the variable $w_1$ on $y_{ij}$ in the group $j$ with respect to the average effect $\gamma_{q0}$.

Substitution of $\beta_{0j}$ and $\beta_{2j}$ in $y_{ij}$ produces the single-equation model:

$$y_{ij} = \gamma_{00} + \alpha_1 x_{1ij} + \gamma_{01} w_{1j} + \gamma_{10} x_{2ij} + \gamma_{11} w_{1j} x_{2ij} + \left(u_{0j} + u_{1j} x_{2ij} + e_{ij}\right) \tag{6}$$

Generalizing the above two-level model to the case where level one includes $P$ independent variables $x_p$ that have a fixed effect, $Q$, independent variables $x_q$ that have a random effect, and $M$ level-two independent variables $w_m$, which also have a fixed effect, we have:

$$y_{ij} = \gamma_{00} + \sum_{m=1}^{M} \gamma_{0m} w_{mj} + \sum_{p=1}^{P} \alpha_p x_{pij} + \sum_{q=P+1}^{P+Q} \gamma_{q0} x_{qij} + \sum_{q=P+1}^{P+Q} \sum_{m=1}^{M} \gamma_{qm} w_{mj} x_{qij} + \left( u_{0j} + \sum_{q=P+1}^{P+Q} x_{qij} u_{qj} + e_{ij} \right) \tag{7}$$

Model 7 is composed of *fixed effects* (the coefficients $\gamma$ and $\alpha$) and *random effects* (all terms in parentheses). Two-level models can also be expressed in a matrix form by

$$Y = X\beta + Wu + e \tag{8}$$

where $Y$ is the vector of measurements of the dependent variable, $X$ is the design matrix of the fixed effect parameter vector $\beta$ (containing the overall mean, main effects, and interactions), $W$ is the design matrix of the random effects given by the vector $U$, and $e$ is the vector of level-one residual errors.

## 2.2 Three-level models

Let $i$, $j$, and $k$ indicate the observation units of levels 1, 2, and 3, respectively. In addition, level 3 has $K$ units, each level-three unit has $J_k$ level-two units, and the $j$th level-two unit in the $k$th level-three unit has $n_{ijk}$ level-oneunits. The null model is

$$\begin{aligned} y_{ijk} &= \beta_{0jk} + e_{ijk} \\ \beta_{0jk} &= \gamma_{00k} + u_{0jk} \\ \gamma_{00k} &= \xi_{000} + \nu_{00k} \end{aligned} \tag{9}$$

In the first equation, $\beta_{0jk}$ is the level-one random intercept that varies between the groups of level two, and $e_{ijk}$ is the residual variance at level one with respect to $\beta_{0jk}$. In the second equation, $\gamma_{00k}$ is the level-two random intercept that varies between the level-three units, and $u_{0jk}$ is the residual variation of the group $j$ with respect to $\gamma_{00k}$. In the third equation, $\xi_{000}$ is the general intercept, that is, the grand mean of $y$, and $\nu_{00k}$ is the variation between the means of the level-three groups (i.e., the deviation of the mean of group $k$ with respect to the grand mean).

Substituting $\gamma_{00k}$ in $\beta_{0jk}$ and then $\beta_{0jk}$ in $y_{ijk}$ yields

$$y_{ijk} = \xi_{000} + \nu_{00k} + u_{0jk} + e_{ijk} \tag{10}$$

Assuming that $e_{ijk} \sim N(0, \sigma^2)$, $u_{0jk} \sim N\left(0, \sigma^2_{u_0}\right)$, and $\nu_{00k} \sim N\left(0, \sigma^2_{\nu_0}\right)$ and that $e_{ijk}$, $u_{0jk}$, and $\nu_{00k}$ are independent, the variance of $y_{ij}$ in Eq. (10) is

$$var\left(y_{ijk}\right) = \sigma^2_{\nu_0} + \sigma^2_{u0} + \sigma^2 \tag{11}$$

One way to define the ICC at levels two and three, attributed to Davis and Scott [3], is

$$\rho_{\text{level 3}} = \frac{\sigma^2_{\nu_0}}{\sigma^2_{\nu_0} + \sigma^2_{u_0} + \sigma^2} \tag{12}$$

$$\rho_{\text{level 2}} = \frac{\sigma^2_{u_0}}{\sigma^2_{\nu_0} + \sigma^2_{u_0} + \sigma^2}, \tag{13}$$

Now, we consider a three-level multilevel model with two level-one independent variables, $x_1$ and $x_2$; the former has a fixed effect and the latter a random effect:

$$y_{ijk} = \beta_{0jk} + \alpha_1 x_{1ijk} + \beta_{2jk} x_{2ijk} + e_{ijk} \tag{14}$$

Let suppose that the random coefficients $\beta_{0jk}$ and $\beta_{2jk}$ are explained by a second-level variable, $w_1$, by the relationships

$$\begin{aligned} \beta_{0jk} &= \gamma_{00k} + \gamma_{01k} w_{1jk} + u_{0jk} \\ \beta_{2jk} &= \gamma_{10k} + \gamma_{11k} w_{1jk} + u_{1jk} \end{aligned} \tag{15}$$

And the random coefficients in Eq. (15) are explained by a third-level variable, $z_1$, by the equations

$$\begin{aligned} \gamma_{00k} &= \xi_{000} + \xi_{001} z_{1k} + \nu_{00k} \\ \gamma_{01k} &= \xi_{010} + \xi_{011} z_{1k} + \nu_{01k} \\ \gamma_{10k} &= \xi_{100} + \xi_{101} z_{1k} + \nu_{10k} \\ \gamma_{11k} &= \xi_{110} + \xi_{111} z_{1k} + \nu_{11k} \end{aligned} \tag{16}$$

Substituting Eqs. (16) in (15) and then in Eq. (14), we have the three-level multilevel model:

$$\begin{aligned} y_{ijk} = {} & \xi_{000} + \alpha_1 x_{1ijk} + \xi_{001} z_{1k} + \xi_{010} w_{1jk} + \xi_{100} x_{2ijk} + \xi_{011} z_{1k} w_{1jk} \\ & + \xi_{101} z_{1k} x_{2ijk} + \xi_{110} w_{1jk} x_{2ijk} + \xi_{111} z_{1k} w_{1jk} x_{2ijk} \\ & + \left(u_{0jk} + \nu_{00k} + \nu_{10k} x_{2ijk} + u_{1jk} x_{2ijk} + \nu_{01k} w_{1jk} + \nu_{11k} w_{1jk} x_{2ijk} + e_{ijk}\right) \end{aligned} \tag{17}$$

where the regression coefficients $\xi$ and $\alpha$ are the fixed part of the model, and the residual terms of each level contained in parentheses are the random part.

We can generalize the three-level model of Eq. (17). Suppose level one contains $P$ independent variables $x_p$ that have a fixed effect, $\alpha_p$, and $Q$ independent variables $x_q\, q = P+1, \ldots, P+Q$. Level two contains $M$ variables $w_m\, m = 1, \ldots, M$. Level three contains $L$ independent variables $z_l\, l = 1, \ldots, L$.

$$\begin{aligned} y_{ijk} = {} & \xi_{000} + \sum_{p=1}^{P} \alpha_p x_{pijk} + \sum_{q=P+1}^{P+Q} \xi_{q00} x_{qijk} + \sum_{m=1}^{M} \xi_{0m0} w_{mjk} + \sum_{l=1}^{L} \xi_{00l} z_{lk} \\ & + \sum_{m=1}^{M} \sum_{l=1}^{L} \xi_{0ml} z_{lk} w_{mjk} + \sum_{q=P+1}^{P+Q} \sum_{l=1}^{L} \xi_{q0l} z_{lk} x_{qijk} + \sum_{q=P+1}^{P+Q} \sum_{m=1}^{M} \xi_{qm0} w_{mjk} x_{qijk} \\ & + \sum_{q=P+1}^{P+Q} \sum_{m=1}^{M} \sum_{l=1}^{L} \xi_{qml} z_{lk} w_{mjk} x_{qijk} + \left( \nu_{00k} + u_{0jk} + \sum_{q=P+1}^{P+Q} \nu_{10k} x_{qijk} \right. \\ & \left. + \sum_{q=P+1}^{P+Q} u_{qjk} x_{qijk} + \sum_{m=1}^{M} \nu_{0mk} w_{mjk} + \sum_{q=P+1}^{P+Q} \sum_{m=1}^{M} \nu_{qmk} w_{mjk} x_{qijk} + e_{ijk} \right). \end{aligned} \tag{18}$$

When the three-level model includes a random slope of level one and a random slope of level two, the model easily includes many parameters (interaction and a residual effect by each random slope coefficient) that easily cause convergence

problems, except for sufficiently large data sets. Therefore, most three-level models have few random slope coefficients.

The equivalent matrix model for a three-level model is

$$Y = X\beta + Wu + Z\nu + e \tag{19}$$

Again, $X$ is the design matrix of the fixed effect parameter vector $\beta$ (containing the overall mean, main effects, and interactions), $W$ is the design matrix of the random effects given by the vector $u$, $Z$ is the design matrix of the random effects given by the vector $\nu$, and $e$ is the vector of level-one residual errors.

### 2.3 Assumptions of multilevel models

Statistical assumptions such as normal distribution, variance at each level, and independence between errors at different levels have been mentioned in the definition of the null multilevel model. They are explicitly defined in this section.

The dimension of the vector $u$ depends on the number of random coefficients in the level-one equation; for example, in Eq. (14), the dimension is two. Similarly, the dimension of the vector $\nu$ depends on the number of random coefficients in the level-two equation; for example, in Eq. (15), the dimension is four. Let $e = (e_{111112}\cdots)^T$, $u = \left(u_{0jk}\, u_{1jk}\cdots u_{sjk}\right)^T$, and $\nu = (\nu_{00k}\cdots\nu_{0tk}\cdots\nu_{t0k}\cdots\nu_{ttk})^T$ with dimensions $N$, $s$, and $t^2$, respectively.

Multilevel models' assumptions are:

$$\begin{pmatrix} \nu \\ u \\ e \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \quad \begin{pmatrix} D & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & G & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & R \end{pmatrix}\right) \tag{20}$$

where $\mathbf{0}$ is the vector of zeros with the appropriate dimension and

$$D = \begin{pmatrix} \sigma^2_{\nu 0} & \sigma_{\nu 01} & \cdots & \sigma_{\nu 0t} \\ \sigma_{\nu 01} & \sigma^2_{\nu 1} & \cdots & \sigma_{\nu 1t} \\ & \cdots & & \\ \sigma_{\nu 0t} & \sigma_{\nu 1t} & \cdots & \sigma^2_{\nu t} \end{pmatrix}, \quad G = \begin{pmatrix} \sigma^2_{u0} & \sigma_{u01} & \cdots & \sigma_{u0s} \\ \sigma_{u01} & \sigma^2_{u1} & \cdots & \sigma_{u1s} \\ & \cdots & & \\ \sigma_{u0s} & \sigma_{u1s} & \cdots & \sigma^2_{us} \end{pmatrix}, \quad R = \sigma^2 I \tag{21}$$

Eq. (20) says:

1. Level-one errors, $e$, are independent, identically normal distributed with mean zero and variance $\sigma^2$.

2. Level-two errors, $u$, follow a multivariate normal distribution with mean $\mathbf{0}$ and covariance $G$.

3. Level-three errors, $\nu$, follow a multivariate normal distribution with mean $\mathbf{0}$ and covariance $D$.

4. Level-one and level-two errors are independent $Cov(e, u) = \mathbf{0}$.

5. Level-one and level-three errors are independent $Cov(e, \nu) = \mathbf{0}$.

6. Level-two and level-three errors are independent $Cov(u, \nu) = \mathbf{0}$.

### 2.4 Multilevel model estimation

The estimation of a multilevel model is complex because, in addition to the residuals at the individual level in the model, there are more residual terms of random intercepts and/or slopes of higher levels. Simultaneously, three types of parameters need to be estimated: the fixed effects, the random effects, and the residual variance/covariance components in matrices $D$, $G$, and $R$. Statistical theory and estimation algorithms for multilevel modeling are beyond the scope of this chapter, but some ideas are given.

When matrices $D$, $G$, and $R$ are known, they can be used to estimate the combined model using generalized least square (GLS). The variance of $y$, given that the matrices $D$ and $G$ are known, is

$$\hat{V} = WDW' + ZGZ' + R \tag{22}$$

The inverse of the $\hat{V}$ matrix can be used as a weight; the regression coefficients of the model can be estimated using GLS. However, the matrices $D$ and $G$ are unknown.

The maximum likelihood estimation method is the most used for estimating multilevel models. It consists of maximizing the likelihood function that generally involves an iterative process that takes the parameter estimates as the initial parameter values for the next iteration of parameter estimation. This process is repeated until the parameter estimates have stabilized from one iteration to the next. The default *tolerance number*, which is sometimes defined by the users, is usually a sufficiently small number, for example, $10^{-8}$. The model converges if the tolerance number is reached between two consecutive iterations. However, sometimes this does not happen. If the limit of specified iterations is reached and the tolerance number between two consecutive iterations has not been reached, the method is said to not converge, and this fact may indicate model specification problems or a small sample size.

Other estimation methods used in multilevel models are *generalized estimating equations*, *bootstrap methods,* and *Bayesian methods* [3]. When the assumptions of the multilevel models (Section 2.3) are not met, these methods are adequate.

### 2.5 Multilevel generalized linear models

Multilevel generalized linear models (MGLM) are an extension of generalized linear models. What makes both models different is that the former assumes dependence in the observations of the dependent variable and the latter assumes independence in the observations.

A three-level MGLM of the dependent variable $Y$ conditioned in the random effects $\nu$ and $u$ is

$$g[E(Y|\nu,\ u)] = \eta = X\beta + Wu + Z\nu \tag{23}$$

where $g(\cdot)$ is the link function, which is a known monotonic, differentiable function, and $\eta$ is the linear predictor. As in multilevel models, the random effects are

assumed to have a normal distribution with zero mean vector and variance/covariance matrixes $D$ and $G$, respectively. The multilevel models described in Sections 2.1 and 2.2 are a particular case of MGLM with $g$, the identity function.

## 3. Bayesian inference

Bayesian inference is a more attractive alternative to frequentist maximum likelihood estimation when: (1) we have information about the parameters in the model, (2) the frequentist estimation method does not converge, (3) the sample size is small at the highest level of the data, or (4) nonlinear functions of the parameters are to be estimated. With this motivation, let us define some concepts.

The heart of the Bayesian inference is the posterior distribution of $\theta$, $p(\theta|y)$, which is defined as the joint probability distribution of the observed data $y$ and the parameter $\theta$, $p(y, \theta) = p(y|\theta)p(\theta)$, conditioned on the known value of $y$, $p(y) = \int p(\theta)p(y|\theta)d\theta)$. Using Bayes' theorem, we obtain

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)} \tag{24}$$

where $p(y|\theta)$ is the *likelihood* of the data $y$, and $p(\theta)$ is the *prior distribution* of *theta*. $p(y) = \int p(\theta)p(y|\theta)d\theta$ with fixed $y$ is a normalization constant not depending on $\theta$. So, an equivalent equation to (24) is

$$p(\theta|y) \propto p(\theta)p(y|\theta) \tag{25}$$

Prior distributions can be *informative* or *non-informative*. When the researcher has a high degree of certainty about $\theta$, the prior distribution will have a small variance and so will also be informative. If this fact does not happen, that is, the researcher has low degree of certainty about $\theta$, the prior distribution will have a large variance and so will be non-informative. Since the prior distribution is a factor in the posterior distribution, when the prior is informative, it will have a great impact on the posterior, so the researcher must be careful when an informative prior distribution is used.

Bayesian estimators are only mean or median vector of the posterior distribution, that is, $\hat{\theta} = \int \theta \frac{p(\theta)p(y|\theta)}{p(y)} d\theta$. However, if $\theta$ has high dimension, this implies to obtain multiple integrals that usually do not have a closed solution. Sometimes, $\theta = (\theta_a, \theta_b)$ and $\theta_b$ are nuisance parameters that must be ignored. The solution is to integrate the posterior distribution with respect to the nuisance parameters, but again, this multiple integral may have no closed solution.

The most widely used method is Markov chain Monte Carlo to obtain means, medians, and quantiles of the posterior distribution.

### 3.1 Markov chain Monte Carlo

#### *3.1.1 Markov chain*

A discrete-time *Markov chain* is a sequence of random variables, $X_n$,$n \geq 1$, that take values in a finite or countable $\Omega$ set that satisfies

$$p(X_{n+1} = j | X_0 = i_0, \ \ldots, X_n = i_n) = p(X_{n+1} = j | X_n = i_n) \tag{26}$$

for all $n$ and any states $i_0, \ldots, i_n, j$ in Ω. Under regularity conditions, the chain will gradually forget its initial state $i_0$, and starting from a state $t$, $p^t(\cdot | X_0 = i_0)$ will converge to a unique stationary distribution $\phi(\cdot)$ (invariant) that does not depend on $t$ or $i_0$.

As the number of sampled points $\{X_t\}$ increases, they will look more like dependent samples from $\phi(\cdot)$. The *burn-in* of an MCMC is the number of iterations, $m$, to eliminate so that the rest show a behavior of dependent samples from the stationary distribution $\phi(\cdot)$ [4]. When the number of burn-in samples is $m$, an estimator of the expectation of $f(X)$ is

$$E[\hat{f(X)}] = \frac{1}{n-m} \sum_{t=m+1}^{n} f(X_t) \tag{27}$$

### *3.1.2 Hamiltonian Monte Carlo*

The Gibbs sampling and the random walk Metropolis are methods whose distributions converge to the target distributions; however, complex models with a large number of parameters may require an unacceptably long time to converge to the target distribution. This problem is largely caused by inefficient random walks that estimate the parameters' space.

The Hamiltonian Monte Carlo (HCM) algorithm or hybrid Monte Carlo algorithm eliminates random walks using momentum variables that transform the target distribution sampling problem into the Hamiltonian dynamics simulation problem. The Störmer–Verlet "leapfrog" (jump steps) integrator is used to simulate the time evolution of this system. Given a sample $m$, a step size $\varepsilon$, and a number of steps $L$, the HMC algorithm consists of resampling the momentum variables $r_d$ from a standard multivariate normal distribution (it can be considered a Gibbs sampling update) and then applying $L$ "leapfrog" updates to the position and momentum variables ($\theta$ and $r$) to generate a pair of proposed position and momentum variables ($\tilde{\theta}$, $\tilde{r}$), which are defined as $\theta^m = \tilde{\theta}$ and $r^m = \tilde{r}$, and will be accepted or rejected according to the Metropolis algorithm. For more details, see [5]. In general, specifying the step size ($\varepsilon$) and number of steps ($L$) is quite difficult when the path is too short, too long, or too straight.

This method for generating MCMC is implemented in the brms package [6] to perform Bayesian estimation in multilevel models.

### *3.1.3 MCMC diagnostics*

After a large enough number of iterations, the MCMC eventually converges to the posterior distribution. A diagnostic statistic is needed to determine whether the MCMC has already converged to the stationary distribution or more iterations are needed. Several diagnostic statistics have been proposed, but we will use the Gelman and Rubin and graphical diagnostics.

***Gelman and Rubin diagnostic*** *(GR)* [7]. This diagnostic uses several chains, $\{X_{i0}, \ \ldots, X_{in-1}\}, \ i = 1, \ldots, m$, drawn from an overdispersed density with respect to the target density $\pi(\cdot)$. In 1992, Gelman and Rubin defined two estimators of the variance of $X$ when $X \sim \pi(\theta)$:

1. The *within-chain variance:* $W = \sum_{i=1}^{m}\sum_{j=0}^{n-1}\left(X_{ij} - \overline{X}_{i\cdot}\right)^2/(m(n-1))$ and

2. The *pooled variance:* $\hat{V} = ((n-1)/n)W + B/n$.

where $B/n = \sum_{i=1}^{m}\left(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot}\right)^2/(m-1)$ is the *between-chain variance* estimate, $\overline{X}_{i\cdot}$ is the mean of the chain $i$, $i = 1, \ldots, m$, and $\overline{X}_{\cdot\cdot}$ is the overall mean. The potential scale reduction factor (PSRF) or Rhat is defined by:

$$\hat{R} = \frac{\hat{V}}{W} \tag{28}$$

The variance in the numerator of $\hat{R}$ overestimates the target variance, while the variance in the denominator underestimates it. This fact produces $\hat{R}$ greater than 1. One criterion for stopping the MCMC simulation is that $\hat{R} \approx 1$ or $\hat{R} < 1.1$. The GR and ESS diagnostics are implemented in the *coda* package [8].

***Graphical diagnostics***. MCMC trace plots are the most widely used diagnostic plots to determine convergence. They are a time series that shows the behavior of the Markov chains around their state space and their achievements at each iteration. When the visible trends show changes in the dispersion of the chain trace, the MCMC has not reached a stationary state. In contrast, when good mixing is observed, the MCMC sampling is said to converge to the target distribution.

## 3.2 Model checking and model comparison

Any Bayesian analysis should include a check of the adequacy of the fit of the postulated model to the data. The adequacy of the fit of a model is measured by how well the distribution of the proposed model approximates the distribution of the data; the better the fit of the postulated model to the data, the better the model. But if the fit is poor, it does not mean that the model is bad, but rather that it contains deficiencies that can be improved. This section explains a model assessment method based on the posterior predictive distribution.

Let us define the replicated data $y^{rep}$ as one that could be observed tomorrow if the experiment that produced the current data $y$ were replicated tomorrow with the same model and the same values of $\theta$ that produced $y$. The distribution of $y^{rep}$ given the current data $y$ is called *posterior predictive distribution* and defined as [9].

$$p(y^{rep}|y) = \int p(y^{rep}|\theta)p(\theta|y)d\theta \tag{29}$$

If the model is accurate, that is, it has a reasonably good fit, the replicated data should be similar to the observed data.

### *3.2.1 Log pointwise predictive density*

The performance of the fitted model can be measured by the quality of its predictions in the new data $y^f$. Pointwise predictions are predictions of each element $y_i^f$ in $y^f$ that are summarized using an appropriate statistic.

Access to $y^f$ is not always easy and sometimes impossible. Instead, performance of the fitted model can be done using the current data $y$. This method for calculating predictive accuracy and to compare models is known as *within-sample predictive accuracy*.

The *log pointwise predictive density* (lppd) of the fitted model to the observed data and unknown parameter $\theta$ is defined as

$$lppd = \log \prod_{i=1}^{n} p(\theta|y_i) = \sum_{i=1}^{n} \log \int p(y_i|\theta) p(\theta|y_i) d\theta \tag{30}$$

In general, the expected predictive accuracy of a model fitted to new data is poorer than the expected predictive accuracy of the same model with the observed data. With the *computed lppd* (clppd), we can evaluate the expression using draws from $p(\theta|y)$ obtained with MCMC, $\theta^s$, $s = 1, \ldots, S$ using sufficient draws:

$$clppd = \sum_{i=1}^{n} \log \left( \frac{1}{S} \sum_{s=1}^{S} p(y_i|\theta^s) \right) \tag{31}$$

The clppd of the observed data $y$ is an overestimate of the clppd for future data.

A second method to assess posterior predictive expectation is the *adjusted within-sample predictive accuracy* that consists of a bias correction of the lppd estimated using information criteria such as Akaike information criterion, deviance information criterion, or Watanabe–Akaike information criterion.

A third method to assess posterior predictive expectation is the *cross-validation*, which captures the out-of-sample predictive error by fitting the model to the training data and assessing the predictive fit in the holdout data [9]. In model comparison, the best model is the one with the lowest predictive error. Let us explain this method in detail:

*Leave-one-out cross-validation* (LOO-CV) works with $n$ partitions in which each holdout set has only one observation, which generates $n$ different inferences, $p_{post(-i)}$, obtained through $S$ posterior simulations, $\theta^{is}$.

The Bayesian LOO-CV estimate of the predictive fit out of the sample is

$$lppd_{loo-cv} = \sum_{i=1}^{n} \log p_{post(-i)}(y_i) \approx \sum_{i=1}^{n} \log \left( \frac{1}{S} \sum_{s=1}^{S} p(y_i|\theta^{is}) \right) \tag{32}$$

Each prediction is conditioned in $n - 1$ data points, which underestimates the predictive fit. For large $n$, the difference is insignificant; however, for small $n$, a first-order bias correction $b = lppd - \overline{lppd}_{-i}$ can be used, where

$$\overline{lppd}_{-i} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \log p_{post(-i)}(y_j) \approx \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \log \left( \frac{1}{S} \sum_{s=1}^{S} p(y_j|\theta^{is}) \right) \tag{33}$$

The bias-corrected Bayesian LOO-CV is

$$lppd_{cloo-cv} = lppd_{loo-cv} + b \tag{34}$$

An estimation of the effective number of parameters is

$$p_{1oo-cv} = ldpp - ldpp_{loo-cv} \tag{35}$$

When comparing two fitted models, we can estimate the difference in their expected predictive accuracy by the difference in $elppd_{loo-cv}$. The standard error of the difference can be computed using a paired estimate to take advantage of the fact that the same set of $n$ data points is used to fit both models.

Suppose we are comparing models I and II, with corresponding fit measures $elpd^{I}_{loo-cv}$ and $elpd^{II}_{loo-cv}$; then difference and its standard error are

$$\begin{aligned} elpd\_diff &= elpd^{I}_{loo-cv} - elpd^{II}_{loo-cv} \\ se\_diff &= se\left(elpd^{I}_{loo-cv} - elpd^{II}_{loo-cv}\right) = \sqrt{nV^{n}_{i}\left(elpd^{I}_{loo,i} - elpd^{II}_{loo,i}\right)} \end{aligned} \tag{36}$$

When two models are compared using the LOO-CV statistic, the one with the lowest value of this statistic is declared the best model. If elpd_diff is used with the loo_compare function of the brms library [6], the value of the difference is reported in the best model accompanied by its se_diff. When comparing two models, the value of the difference is reported in the column of the best model. There is more evidence of the superiority of one model over another when the elpd_diff is larger than the se_diff.

## 4. Multilevel model methodology

To propose a multilevel model, it is necessary to determine which variables will be in the fixed part, which in the random part, and the cross-level interactions. This task can be complex, so we need a strategy to build the model.

### 4.1 Multilevel model building strategy

In this section, we show an adaptation of the bottom-up strategy to specify a three-level multilevel model. The bottom-up methodology is used in the frequentist approach [3]. Our adaptation proposal tries to use the Bayesian LOO-CV from Step 2 to Step 7 for model comparison.

*Step 1.* Fitting the intercept-only model:

$$y_{ijk} = \xi_{000} + \left(\nu_{00k} + u_{0jk} + e_{ijk}\right) \tag{37}$$

This model gives a basal line to compare with the next models.

*Step 2.* Add all the level-one independent variables fixed:

$$y_{ijk} = \xi_{000} + \sum_{p=1}^{P} \alpha_p x_{pijk} + \sum_{q=P+1}^{P+Q} \xi_{q00} x_{qijk} + \left(\nu_{00k} + u_{0jk} + e_{ijk}\right) \tag{38}$$

It must be determined which level-one variable has a significant effect on $y$. We will assume that all the $P$ level-one variables are statistically significant. Models 38 and 37 must be compared.

*Step 3.* Add the level-two independent variables fixed:

$$y_{ijk} = \xi_{000} + \sum_{p=1}^{P} \alpha_p x_{pijk} + \sum_{q=P+1}^{P+Q} \xi_{q00} x_{qijk} + \sum_{m=1}^{M} \xi_{0m0} w_{mjk} + \left(\nu_{00k} + u_{0jk} + e_{ijk}\right) \quad (39)$$

It must be determined which level-two variable has a significant effect on $y$. If the variables $w_m$ explain the variability of $y$, Model 39 should be superior to Model 38. Again, we assume that all the $M$ level-two independent variables are statistically significant.

*Step 4.* Add the level-three independent variables fixed:

$$\begin{aligned} y_{ijk} = {} & \xi_{000} + \sum_{p=1}^{P} \alpha_p x_{pijk} + \sum_{q=P+1}^{P+Q} \xi_{q00} x_{qijk} + \sum_{m=1}^{M} \xi_{0m0} w_{mjk} + \sum_{l=1}^{L} \xi_{00l} z_{lk} \\ & + \left(\nu_{00k} + u_{0jk} + e_{ijk}\right) \end{aligned} \quad (40)$$

It must be determined which level-three variable has a significant effect on $y$. If the variables $z_l$ explain the variability of $y$, Model 40 should be superior to Model 39.

Steps 1–3 consider the specification of the fixed part of the three-level multilevel model. Now we will specify the random part of the model.

*Step 5.* Assessing whether any of the slopes of the independent variables at level one has a significative variance component between groups at level two or level three.

$$\begin{aligned} y_{ijk} = {} & \xi_{000} + \sum_{p=1}^{P} \alpha_p x_{pijk} + \sum_{q=P+1}^{P+Q} \xi_{q00} x_{qijk} + \sum_{m=1}^{M} \xi_{0m0} w_{mjk} + \sum_{l=1}^{L} \xi_{00l} z_{lk} \\ & + \left( \nu_{00k} + u_{0jk} + \sum_{q=P+1}^{P+Q} \nu_{q0k} x_{qijk} + \sum_{q=P+1}^{P+Q} u_{qjk} x_{qijk} + e_{ijk} \right) \end{aligned} \quad (41)$$

where $u_{qjk}$ are the level-two residuals of the slopes of the level-one independent variable $x_q$, and $\nu$'s are the level-three residuals of the slopes of the level-two independent variable $w_m$.

Level-one independent variables that do not have a significant slope may have a significant random slope. This step and the next should be carefully performed, because the model can easily become overparameterized and/or have problems such as non-convergence or extremely slow calculations. It is advisable to assess significance of the slopes variable by variable. Next, the model is formulated with all the variables with significant random slopes. If Model 41 is not better than Model 40, the procedure for specifying a three-level multilevel model stops.

*Step 6.* Assessing whether any of the slopes of the level-two independent variable has a significant variance component among level-three groups.

$$\begin{aligned} y_{ijk} = {} & \xi_{000} + \sum_{p=1}^{P} \alpha_p x_{pijk} + \sum_{q=P+1}^{P+Q} \xi_{q00} x_{qijk} + \sum_{m=1}^{M} \xi_{0m0} w_{mjk} + \sum_{l=1}^{L} \xi_{00l} z_{lk} \\ & + \left( \nu_{00k} + u_{0jk} + \sum_{q=P+1}^{P+Q} \nu_{q0k} x_{qijk} + \sum_{q=P+1}^{P+Q} u_{qjk} x_{qijk} + \sum_{m=1}^{M} \nu_{0mk} w_{mjk} + e_{ijk} \right) \end{aligned} \quad (42)$$

where the $\nu$'s are the level-three residuals of the slopes of the level-two independent variable $w_m$.

The assessment of random slopes of the level-two variables should be performed variable by variable, and then, all these variables should be included into a model to assess the improvement of the model with respect to Model 41.

*Step 7.* Adding interactions between level-three independent variables and the level-one and level-two independent variables that have a significant slope variance in Steps 5 and 6. This produces the full model:

$$\begin{aligned} y_{ijk} = {} & \xi_{000} + \sum_{p=1}^{P} \alpha_p x_{pijk} + \sum_{q=P+1}^{P+Q} \xi_{q00} x_{qijk} + \sum_{m=1}^{M} \xi_{0m0} w_{mjk} + \sum_{l=1}^{L} \xi_{00l} z_{lk} \\ & + \sum_{m=1}^{M} \sum_{l=1}^{L} \xi_{0ml} z_{lk} w_{mjk} + \sum_{q=P+1}^{P+Q} \sum_{l=1}^{L} \xi_{q0l} z_{lk} x_{qijk} + \sum_{q=P+1}^{P+Q} \sum_{m=1}^{M} \xi_{qm0} w_{mjk} x_{qijk} \\ & + \sum_{q=P+1}^{P+Q} \sum_{m=1}^{M} \sum_{l=1}^{L} \xi_{qml} z_{lk} w_{mjk} x_{qijk} + \left( \nu_{00k} + u_{0jk} + \sum_{q=P+1}^{P+Q} \nu_{q0k} x_{qijk} \right. \\ & \left. + \sum_{q=P+1}^{P+Q} u_{qjk} x_{qijk} + \sum_{m=1}^{M} \nu_{0mk} w_{mjk} + \sum_{q=P+1}^{P+Q} \sum_{m=1}^{M} \nu_{qmk} w_{mjk} x_{qijk} + e_{ijk} \right). \end{aligned} \tag{43}$$

When explaining the variances of the random slopes in terms of contextual variables, the model automatically includes interaction terms between levels that compose the fixed part of the model. It is recommended to add variables that explain the variance of the random slope coefficients one by one and not as shown in this step (this was done here to avoid specifying more equations).

When it comes to an MGLM, the methodology changes slightly; that is, instead of defining models in terms of $y$, models are defined in terms of $g\left(\mu_{ijk}\right) = g\left(E\left[Y_{ijk}|\nu,\ u\right]\right)$, and the residual errors at the individual level are no longer specified. An example of this methodology for an MGLM is illustrated below.

## 5. Application: periodontal probing depth

In this section, an example is given in which a multilevel generalized linear model is used for data from a cross-sectional study conducted by Romero-Castro et al. [10]. This study was carried out among adults who reside in the state of Guerrero, Mexico, and who went to the external dental clinical service of the Dental School of the Autonomous University of Guerrero (UAGro) in search of treatment, during the period from August 2015 to February 2016. The protocol was approved (registration no. CB005/2015) by the ethic committee at UAGro.

The goal of this multilevel analysis was to determine the clinical factors associated with the depth of periodontal probing.

Thirty-two teeth were examined in each of the 116 patients. Probing pocket depth was recorded at six sites in each tooth, that is, *mesiobuccal, mid buccal, distobuccal, mesiolingual, mid lingual,* and *distolingual* locations of each tooth. Pocket depth was recorded by use of Florida probe in the six sites. The response variable was *probing*

*depth* measured in millimeters; that is, probing depth is a continuous variable and greater than zero ($\geq 0$). The data set consisted of 18,358 observations.

The independent variables, except the age, were all dichotomous: *bleeding, mobility, plaque, calculus, insulin resistance (fasting plasma glucose > 100 mg/dL), smoking, root remnants, and mismatched restorations*, where 0 indicated absence and 1 presence.

**Figure 1** and **Table 1** show the three levels of the data and the variables at each level. The first level corresponded to the probing sites where the independent variables *bleeding* and *furcation* and the response variable *probing depth* were measured. Level two corresponded to the dental piece, that is, teeth that only had the independent variable *mobility*, and level three corresponded to the patients, measuring the independent variables *age, plaque, calculus, insulin resistance, smoking, root remnants*, and *mismatched restorations*.

A first data analysis was done using a three-level multilevel model assuming a normal distribution for the probing depth. The frequentist fit had two problems: the residuals did not have a normal distribution and the numerical method to obtain the estimates did not converge.

The minimum of probing depth was 0.2 mm, Q1 was 0.8 mm, Q2 was 1.2 mm, Q3 was 1.8 and the maximum was 9 mm. In addition, its distribution was asymmetric to the right (skewness = 1.6 and kurtosis = 8.0). Therefore, it was assumed that probing depth had gamma distribution with mean $\mu$ and variance $\mu^2/\alpha$:

$$f(y) = \frac{(\alpha/\mu)^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} \exp\left(-\frac{\alpha y}{\mu}\right) \tag{44}$$

It is well known that gamma regression belongs to the generalized linear model family. But as the data studied is of hierarchical nature, the appropriate model is the multilevel

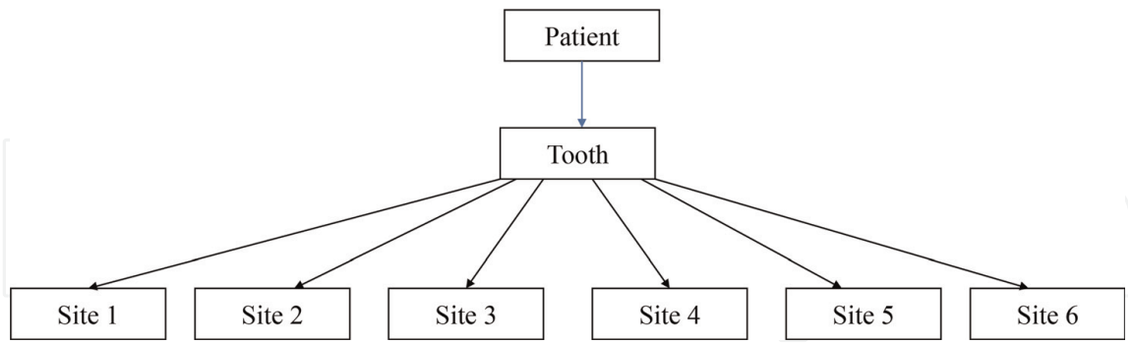


**Figure 1.**
*Multilevel structure of the probing depth of 1 tooth out of 32 teeth for each patient.*

| Levels | Variables |
|---|---|
| Level 3: Patient | Age, plaque, calculus, insulin resistance, smoke, root remnants, and mismatched restorations |
| Level 2: Tooth | Mobility |
| Level 1: Probing site | Bleeding |

**Table 1.**
*Independent variables on levels.*

generalized linear model. Given that the response variable is non-negative, the link function used was the natural logarithm to get expected probing depth greater than zero.

### 5.1 Bayesian estimation

**Likelihood:** It was assumed that probing depth follows a gamma distribution.

**Prior distribution:** It was defined as a product of marginal prior distributions for each component of $\beta$ in Model 23. $\beta$ is composed by the overall mean, main effects, and interactions: $\xi_{000}$,$\xi_{q00}$,$\xi_{0m0}$,$\xi_{00l}$,$\xi_{0ml}$,$\xi_{q0l}$,$\xi_{qm0}$,$\xi_{qml}$, and all of them had $N\left(0,\ 10^2\right)$ prior. *brms* function uses a special parameterization for matrices $D$ and $G$ in Eq. (21). This parameterization is $G = F(\sigma_k)\Omega_k F(\sigma_k)$, where $F(\sigma_k)$ is a diagonal matrix with diagonal elements $\sigma_k$ ([6]). Priors for $D$ and $G$ needed only to specify priors for $\sigma_k$ and $\Omega_k$, which were $\sigma_k \sim HalfCauchy(10)$ and $\Omega_k \sim CorrLKJ(1)$. Finally, the shape hyper-parameter was $shape \sim Gamma(0.01,0.01)$.

The analysis of this model was performed with the *brms* library ([6, 11]) that uses the probabilistic programming language *Stan* ([12]) in the environment of *R Software 4.0.5*.

**Simulation:** All the MCMC had four chains; the number of iterations and burn-in was not the same for the models studied, but all used a final sample of 4000.

The MCMC of the models (37, 38 and 39), that is, null, with level-one, and with level-two variables, were obtained using 4000 iterations and a burn-in of 3000. Model 40 used 5000 iterations and a burn-in of 4000, Model 41 used 7000 iterations and a burn-in of 6000, and Model 43 used 8000 iterations and a burn-in of 7000.

**Bayesian estimators:** The mean of the posterior distribution was used as the Bayesian estimator; this is related with minimizing the squared loss function.

**Models studied:** We studied a three-level multilevel generalized linear model, where $i$ represented the level-one units, $j$ the level-two units, and $k$ the level-three units. Although the values of the Rhats are not shown, all the MCMC of the studied models converged since all the Rhats were at most 1.01.

**Step 1.** The null model is

$$log\left(\mu_{ijk}\right) = \xi_{000} + \left(\nu_{00k} + u_{0jk}\right) \tag{45}$$

Columns 2 and 3 of **Table 2** show the Bayesian estimations of the null model. The credible intervals did not contain zero, so that the variances at the tooth level and at the patient level were significant. This supports the use of MGLM.

**Step 2.** The model with level-one variable, *bleeding*, is

$$log\left(\mu_{ijk}\right) = \xi_{000} + \alpha_1 \text{bleeding}_{1ijk} + \left(\nu_{00k} + u_{0jk}\right) \tag{46}$$

The Bayesian estimations of the model showed that the bleeding coefficient was significant (columns 4 and 5 of **Table 2**). The comparison of Models 45 and 46, using LOO-CV, indicates that the model including the level-one variables was better (before the last row and column 5 in **Table 2**).

**Step 3.** The model with level-two variable, *mobility*, is

$$log\left(\mu_{ijk}\right) = \xi_{000} + \alpha_1 \text{bleeding}_{1ijk} + \xi_{010}\text{mobility}_{1jk} + \left(\nu_{00k} + u_{0jk}\right) \tag{47}$$

| | Model 45 | | Model 46 | | Model 47 | | Model 48 | | Model 49 | | Model 50 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Coef** | **(95%CrI)** | **Coef** | **(95%CrI)** | **Coef** | **(95%CrI)** | **Coef** | **(95%CrI)** | **Coef** | **(95%CrI)** | **Coef** | **(95%CrI)** |
| **Group-level effects:** | | | | | | | | | | | | |
| Patient (116 levels) | | | | | | | | | | | | |
| sd(Intercept) | 0.20 | (0.17, 0.23) | 0.20 | (0.17, 0.23) | 0.20 | (0.17, 0.23) | 0.19 | (0.17, 0.22) | 0.19 | (0.17, 0.22) | 0.19 | (0.17, 0.22) |
| sd(Bleeding) | | | | | | | | | 0.12 | (0.02, 0.21) | 0.13 | (0.02, 0.23) |
| Patient:Tooth (3131 levels) | | | | | | | | | | | | |
| sd(Intercept) | 0.23 | (0.22, 0.24) | 0.22 | (0.21, 0.23) | 0.22 | (0.21, 0.23) | 0.22 | (0.21, 0.23) | 0.22 | (0.21, 0.23) | 0.22 | (0.21, 0.23) |
| sd(Bleeding) | | | | | | | | | 0.25 | (0.16, 0.33) | 0.25 | (0.16, 0.34) |
| **Population-level effects:** | | | | | | | | | | | | |
| Intercept | 0.34 | (0.30, 0.37) | 0.33 | (0.29, 0.37) | 0.33 | (0.29, 0.36) | 0.29 | (0.24, 0.34) | 0.29 | (0.25, 0.34) | 0.29 | (0.25, 0.34) |
| Bleeding | | | 0.15 | (0.11, 0.19) | 0.15 | (0.10, 0.19) | 0.15 | (0.10, 0.19) | 0.13 | (0.07, 0.20) | 0.11 | (0.02, 0.19) |
| Mobility | | | | | 0.04 | (−0.00, 0.08) | 0.03 | (−0.01, 0.07) | 0.03 | (−0.01, 0.07) | 0.03 | (−0.01, 0.08) |
| Calculus | | | | | | | 0.10 | (0.03, 0.18) | 0.10 | (0.03, 0.18) | 0.10 | (0.03, 0.18) |
| Smoking | | | | | | | −0.02 | (−0.14, 0.09) | −0.02 | (−0.14, 0.10) | −0.02 | (−0.13, 0.09) |
| Bleeding:Calculus | | | | | | | | | | | 0.06 | (−0.07, 0.19) |
| **Specific parameters:** | | | | | | | | | | | | |
| Shape | 4.50 | (4.41, 4.60) | 4.51 | (4.41, 4.61) | 4.51 | (4.41.4.61) | 4.51 | (4.42, 4.61) | 4.55 | (4.45, 4.64) | 4.54 | (4.45, 4.64) |
| elpd_diff (se_diff) | | | | −16.2 (8.2)† | | −2.3 (2.0)‡ | | −3.4 (1.5)◆ | | −10.7 (7.9)# | | −1.3 (1.8)* |

*Coef: Coefficient.*
**95%CrI: 95% Credible Interval. Comparisons.†Model 45 vs. 46.‡Model 46 vs. 47. Model 47 vs. 48.#Model 48 vs. 49.*Model 49 vs. 50.*

**Table 2.**
*Bayesian estimates of Models 45–50.*

Mobility fixed effect was not significant (columns 6 and 7 of **Table 2**); however, it was retained in the model because it was the only level-two variable and to get estimations of the effect of the third-level independent variables adjusted for the effect of level-two variable. The LOO-CV criterion indicated that this model was slightly better (before the last row and column 7 in **Table 2**).

There are seven level-three contextual variables (**Table 1**); before specifying the model containing only the significant level-three variables, a forward selection of variables was performed to avoid having an overparameterized model. **Table 3** shows the variable selection procedure where each model contains the level-one variable, *bleeding*, and the level-two variable, *mobility*. The LOO-CV model comparison indicates that the model that includes *calculus* and *smoking* variables is the best model.

**Step 4.** The model with level-three variables, *calculus* and *smoking*, is

$$\begin{aligned} log\left(\mu_{ij|k}\right) &= \xi_{000} + \alpha_1 \text{bleeding}_{1ijk} + \xi_{010}\text{mobility}_{1jk} \\ &\quad + \xi_{001}\text{calculus}_{1k} + \xi_{002}\text{smoking}_{2k} + \left(\nu_{00k} + u_{0jk}\right) \end{aligned} \tag{48}$$

Columns 8 and 9 in **Table 2** show that the variable *smoking* was not significant; however, the model that contains smoking is better than the others. Model 48 was better than Model 47 (before the last row and column 9 in **Table 2**).

**Step 5.** The model with a random slope for the variable bleeding.

In Eq. (49), a random slope for the variable bleeding is added that varies at patient and teeth levels; that is, the relationship between probing depth and bleeding varied between patients and between teeth.

$$\begin{aligned} log\left(\mu_{ij|k}\right) &= \xi_{000} + \alpha_1 \text{bleeding}_{1ijk} + \xi_{010}\text{mobility}_{1jk} \\ &\quad + \xi_{001}\text{calculus}_{1k} + \xi_{002}\text{smoking}_{2k} \\ &\quad + \nu_{10k}\text{bleeding}_{1ijk} + u_{1jk}\text{bleeding}_{1ijk} + \left(\nu_{00k} + u_{0jk}\right) \end{aligned} \tag{49}$$

Finally, in the next model interaction, terms were added based on signs that occur in periodontal disease.

Columns 10 and 11 of **Table 2** show that the random slope of bleeding was significant at patient and teeth levels. Again, this model was compared with Model 48 using the LOO-CV criterion, and the best model was Model 48, which contained random slopes (before the last row and column 11 in **Table 2**).

**Step 7.** The model with cross-level interactions is

$$\begin{aligned} log\left(\mu_{ij|k}\right) &= \xi_{000} + \alpha_1 \text{bleeding}_{1ijk} + \xi_{010}\text{mobility}_{1jk} \\ &\quad + \xi_{001}\text{calculus}_{1k} + \xi_{002}\text{smoking}_{2k} + \xi_{101}\text{bleeding}_{1ijk}\text{calculus}_{1k} \\ &\quad + \nu_{10k}\text{bleeding}_{1ijk} + u_{1jk}\text{bleeding}_{1ijk} + \left(\nu_{00k} + u_{0jk}\right) \end{aligned} \tag{50}$$

Eq. (50) has an interaction between the level-three variable calculus with the level-one variable bleeding. Columns 12 and 13 of **Table 2** show that the interaction was not significant (its credible interval contained zero). Finally, the comparison of models indicated that the best model was Model 49 corresponding to the bleeding random slope model (the last row and column 13 in **Table 2**). So, this model is interpreted.

**Figure 2** shows the posterior predictive fit of Model 49 to the data. The replicated data are plotted in a light color, and the observed data are plotted in black. As both

| Model | elpd_diff | se_diff |
|---|---|---|
| Comparison of models with one level-three independent variable. | | |
| $\xi_{001}\text{calculus}_{1k}$ | 0.0 | 0.0 |
| $\xi_{003}\text{insulin resistance}_{3k}$ | −0.1 | 1.6 |
| $\xi_{004}\text{root remnants}_{4k}$ | −0.8 | 1.5 |
| $\xi_{005}\text{plaque}_{5k}$ | −1.5 | 1.6 |
| $\xi_{002}\text{smoking}_{2k}$ | −1.7 | 1.6 |
| $\xi_{006}\text{age}_{6k}$ | −1.8 | 1.5 |
| $\xi_{007}\text{mismatched restorations}_{7k}$ | −2.5 | 1.5 |
| Comparison of models with two level-three independent variables. | | |
| $\xi_{001}\text{calculus}_{1k} + \xi_{002}\text{smoking}_{2k}$ | 0.0 | 0.0 |
| $\xi_{001}\text{calculus}_{1k} + \xi_{003}\text{insulin resistance}_{3k}$ | −2.1 | 1.5 |
| $\xi_{001}\text{calculus}_{1k} + \xi_{006}\text{age}_{6k}$ | −2.6 | 1.5 |
| $\xi_{001}\text{calculus}_{1k} + \xi_{005}\text{plaque}_{5k}$ | −2.9 | 1.5 |
| $\xi_{001}\text{calculus}_{1k} + \xi_{007}\text{mismatched restorations}_{7k}$ | −4.4 | 1.5 |
| $\xi_{001}\text{calculus}_{1k} + \xi_{004}\text{root remnants}_{4k}$ | −5.2 | 1.5 |
| Comparison of the best models with one and two level-three independent variables | | |
| $\xi_{001}\text{calculus}_{1k} + \xi_{002}\text{smoking}_{2k}$ | 0.0 | 0.0 |
| $\xi_{001}\text{calculus}_{1k}$ | −2.6 | 1.5 |
| Comparison of models with three level-three independent variables | | |
| $\xi_{001}\text{calculus}_{1k} + \xi_{002}\text{smoking}_{2k} + \xi_{004}\text{rootremnants}_{4k}$ | 0.0 | 0.0 |
| $\xi_{001}\text{calculus}_{1k} + \xi_{002}\text{smoking}_{2k} + \xi_{006}\text{age}_{6k}$ | −0.2 | 1.6 |
| $\xi_{001}\text{calculus}_{1k} + \xi_{002}\text{smoking}_{2k} + \xi_{003}\text{insulin resistance}_{3k}$ | −0.6 | 1.5 |
| $\xi_{001}\text{calculus}_{1k} + \xi_{002}\text{smoking}_{2k} + \xi_{005}\text{plaque}_{5k}$ | −1.2 | 1.6 |
| $\xi_{001}\text{calculus}_{1k} + \xi_{002}\text{smoking}_{2k} + \xi_{007}\text{mismatchedrestorations}_{7k}$ | −1.6 | 1.5 |
| Comparison of the best models with two and three level-three independent variables | | |
| $\xi_{001}\text{calculus}_{1k} + \xi_{002}\text{smoking}_{2k}$ | 0.0 | 0.0 |
| $\xi_{001}\text{calculus}_{1k} + \xi_{002}\text{smoking}_{2k} + \xi_{004}\text{root remnants}_{4k}$ | −2.9 | 1.5 |

**All the models have the structure:* $log\left(\mu_{ij|k}\right) = \xi_{000} + \alpha_1\text{bleeding}_{1ijk} + \xi_{010}\text{mobility}_{1jk} + \text{var1} + \text{var2} + \text{var3} + \left(\nu_{00k} + u_{0jk}\right)$, *where* var1 *is the independent variable that produces the best fit among all the seven models with one independent variable. Similarly,* var2 *is the second independent variable that produces the best fit among all the six models, having* var1 *in common, with two independent variables, and so on for* var3.

**Table 3.**
*Forward variable selection for the level-three variables*.*

curves agree very well, the posterior predictive density fits very well with the distribution of the probing depth. Both distributions are clearly not symmetric, and they seem to follow a gamma distribution. Definitely, normal distribution was not an appropriate assumption for probing depth. In conclusion, the random slope model (49) had a good fit.

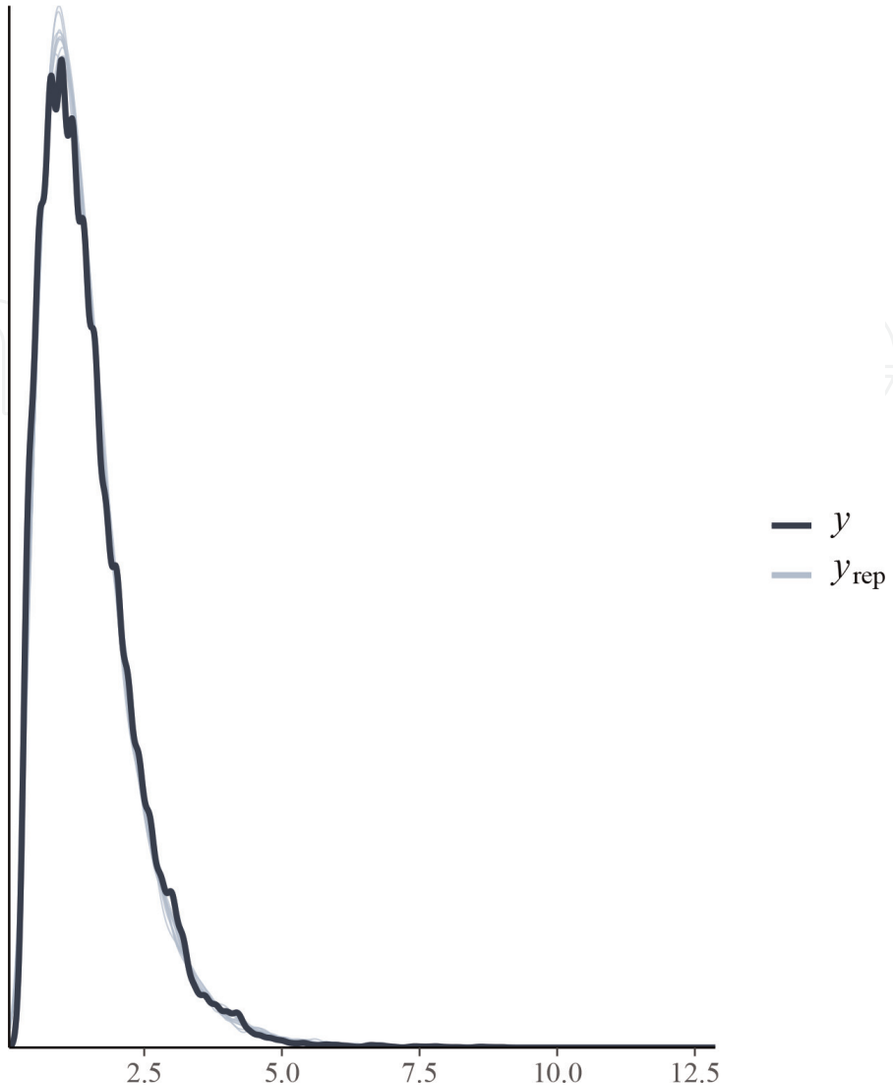


**Figure 2.**
*Posterior predictive density of Model 49 with a random slope for bleeding between teeth and patient.*

**Figure 3** shows the histograms of the empirical posterior distributions of the parameters. Finally, the MCMC of Model 49 converged since all the Rhats were at most 1.01, and the trace plots of **Figure 3** show that the chains mix well.

## 5.2 Discussion

In this example of probing depth, the variance at the tooth level (1.59) and the variance at the patient level (1.49) were significant (**Table 2**); that is, the mean of the dependent variable varied between teeth nested in patients, and the ICC at the tooth level (0.45) was higher than that at the patient level (0.42); that is, there was greater dependence between the measurements of the probing sites of different teeth than between measurements of the probing sites of different patients. This finding probes that using a multilevel model for these probing depth data was better than using a single-level model, and the former produced more accurate estimates and credible intervals. In addition, the random slope of bleeding was significant between teeth; that is, there was a positive relationship between probing depth and bleeding that varied between teeth in the patients (probing depth between teeth increased by an average of 1.28 mm if the site was bleeding). On the other hand,

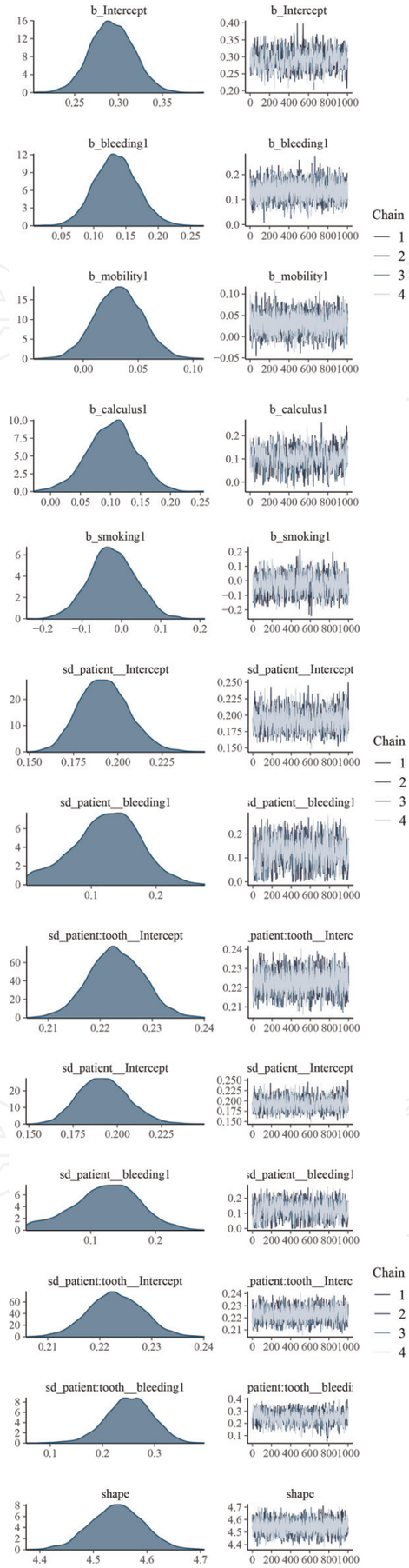


**Figure 3.**
*Posterior distribution of parameters of Model 49 and time series plots showing the MCMC output.*

calculus is a form of hardened dental plaque. In the random slope model (Eq. (49)), bleeding and calculus were significant parameters that estimated that, on average, the depth of bleeding probing sites was 1.14 mm greater than the sites that did not exhibit bleeding. On average, the probing depth of patients who had calculus on any of the teeth was 1.11 mm greater than of patients who did not have calculus. The plausible intervals for bleeding and calculus were (1.07,1.22) and (1.03,1.20), respectively.

In the random slope model, mobility and smoking were not significantly associated with probing depth, but if we decide to give them an interpretation, we can say that, on average, the probing depth of patients who presented dental mobility was 1.03 mm greater than the probing depth of patients who did not have dental mobility. Similarly, smoking patients had, on average, a probing depth 0.98 mm greater than that in non-smoking patients. Different results and interpretations could be obtained from measuring the independent variables at levels other than those given in this example. Specifically, the variable calculus could have been measured at the tooth level. Before fitting the Bayesian multilevel model, we tried to estimate the multilevel model using restricted maximum likelihood; however, the numerical method did not converge. More practical examples using the *R Software* can be found at [13].

## 6. Conclusions

In certain clinical research designs, the data have a nested structure (in other words, a hierarchical structure). The data that make up a nested structure are modeled using multilevel models because they simultaneously estimate the effects of the variables at the individual level and the effects of the contextual variables or variables at the group level. A significant ICC determines whether it is necessary to use a multilevel model. If the ICC is not significant, an ordinary regression model is sufficient to model the nested data. A disadvantage of multilevel models is that they easily contain a large number of parameters to be estimated. On the other hand, modeling the data levels separately incurs a large type 1 error even when the ICC is small. This fact causes the inferences to be incorrect. The maximum likelihood estimation of the parameters of a multilevel model requires that the assumptions of the distribution are satisfied. More general methods such as Bayesian estimation make it possible to estimate the parameters without requiring that the assumptions of the multilevel models be satisfied. In addition, the Bayesian estimation is robust to a small sample size, a situation that is more likely to occur in higher level observations, and in general, it is able to deal with technical problems such as multicollinearity of the data.

In this chapter, we adapted the bottom-up strategy to specify a multilevel model in the frequentist approach to the Bayesian approach. Our proposal was to use the Bayesian LOO-CV between the different steps for the comparison of models. Deviance information criterion (DIC) could also be used instead of Bayesian LOO-CV.

Two factors had a significant association with probing depth. Bleeding (site-level covariate) and dental calculus (patient-level covariate). At the tooth level, a factor associated with the probing depth was not found.

The methodology set out in this chapter can be applied to other areas of the health sciences with data with a hierarchical structure and numerical response variable.

## Acknowledgements

Tino-Salgado is indebted to CONACYT for fellowship that enabled him to pursue graduate studies for the degree of Maestría en Matemáticas Aplicadas. The authors thank the sixth semester periodontics students of the 2013–2018 batch for their assistance and logistical support during the patient recruitment stage.

## Conflict of interest

The authors declare no conflict of interest.

## Abbreviations

| | |
|---|---|
| DIC | Deviance information criterion |
| HMC | Hamiltonian Monte Carlo |
| ESS | Effective sample size |
| GEE | Generalized estimating eqs. |
| GR | Gelman and Rubin diagnostic |
| ICC | Intraclass correlation coefficient |
| LOO-CV | Leave-one-out cross-validation |
| lppd | Log pointwise predictive density |
| MCMC | Markov chain Monte Carlo |
| MGLM | Multilevel generalized linear models |
| NUTS | No-U-turn sampling |
| PSRF | Potential scale reduction factor |

## Author details

Edilberta Tino-Salgado[1†], Flaviano Godínez-Jaimes[1†], Cruz Vargas-De-León[1,2*], Norma Samanta Romero-Castro[3], Salvador Reyes-Fernández[3] and Victor Othon Serna-Radilla[3]

1 Facultad de Matemáticas, Maestría en Matemáticas Aplicadas, Universidad Autónoma de Guerrero, Chilpancingo de los Bravo, Mexico

2 División de Investigación, Hospital Juárez de México, Ciudad de México, Mexico

3 Facultad de Odontología, Especialidad en Implantología y Rehabilitación Bucal, Universidad Autónoma de Guerrero, Acapulco, Mexico

*Address all correspondence to: leoncruz82@yahoo.com.mx

† These authors contributed equally.

## References

[1] Kim JS, Kim D-K, Hong SJ. Assessment of errors and misused statistics in dental research. International Dental Journal. 2011;**61**(3): 163-167

[2] Wang J, Xie H, Fisher JF. Multilevel Models, Applications Using SAS. Berlin, Germany: de Gruyter; 2011. DOI: 10.1515/9783110267709

[3] Joop J Hox, Mirjam Moerbeek y Rens Van de Schoot. Multilevel Analysis: Techniques and Applications. New York, United States: Routledge; 2017

[4] Gilks WR, Richardson S, Spiegelhalter D. Markov Chain Monte Carlo in Practice. Florida, United States: CRC Press; 1995

[5] Matthew D Hoffman, Andrew Gelman y col. The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. The Journal of Machine Learning Research 2014;**15**(1): 1593–1623

[6] Bürkner P-C. brms: An R package for Bayesian multilevel models using stan. Journal of Statistical Software. 2017;**80**(1):1-28. DOI: 10.18637/jss.v080.i01

[7] Roy V. Convergence diagnostics for markov chain Monte Carlo. Annual Review of Statistics and Its Application. 2020;**7**:387-412

[8] Plummer M, Best N, Cowles K, Vines K. CODA: Convergence diagnosis and output analysis for MCMC. R News. 2006;**6**(1):7-11

[9] Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis. Florida, United States: CRC Press; 2013

[10] Romero-Castro NS, Castro-Alarcon N, Reyes-Fernández S, Flores-Alfaro E, Parra-Rojas I. Periodontal disease distribution, risk factors, and importance of primary healthcare in the clinical parameters improvement. International Journal of Odonto Stomatology. 2020;**14**(2):183-190. DOI: 10.4067/S0718-381X2020000200183

[11] Bürkner P-C. Advanced Bayesian multilevel modeling with the R package brms. The R Journal. 2018;**10**(1):395-411. DOI: 10.32614/RJ-2018-017

[12] Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: A probabilistic programming language. Journal of Statistical Software. 2017;**76**(1):1-32. DOI: 10.18637/jss.v076.i01

[13] Holmes Finch W, Bolin JE, Kelley K. Multilevel Modeling Using R. CRC Press; 2019