Published in final edited form as:
J R Stat Soc Ser A Stat Soc. 2011 July ; 174(3): . doi:10.1111/j.1467-985X.2011.00685.x.

Assessing variance components in multilevel linear models using approximate Bayes factors: A case study of ethnic disparities in birthweight

Benjamin R. Saville[1], **Amy H. Herring**[2], and **Jay S. Kaufman**[3]

[1]Department of Biostatistics, Vanderbilt University School of Medicine S-2323 Medical Center North, 1161 21st Avenue South Nashville, TN 37232-2158, b.saville@vanderbilt.edu [2]Department of Biostatistics, University of North Carolina at Chapel Hill [3]Department of Epidemiology, Biostatistics, and Occupational Health, McGill University

Abstract

Racial/ethnic disparities in birthweight are a large source of differential morbidity and mortality worldwide and have remained largely unexplained in epidemiologic models. We assess the impact of maternal ancestry and census tract residence on infant birth weights in New York City and the modifying effects of race and nativity by incorporating random effects in a multilevel linear model. Evaluating the significance of these predictors involves the test of whether the variances of the random effects are equal to zero. This is problematic because the null hypothesis lies on the boundary of the parameter space. We generalize an approach for assessing random effects in the two-level linear model to a broader class of multilevel linear models by scaling the random effects to the residual variance and introducing parameters that control the relative contribution of the random effects. After integrating over the random effects and variance components, the resulting integrals needed to calculate the Bayes factor can be efficiently approximated with Laplace's method.

Keywords

Bayes factors; Laplace approximation; hierarchical; multilevel linear model; variance components

1 Introduction

Many studies collect data that have hierarchical or clustered structures. Examples include randomized studies in which patients are clustered within practices, educational studies in which students are clustered in schools, or environmental studies in which individuals are clustered in homes clustered in counties. An analysis that ignores such clustering assumes all observations are independent, resulting in incorrect model-based standard errors that can lead to misleading scientific inferences. Multilevel models are used to account for the correlation of observations within a given group by incorporating group-specific random effects. These random effects can be nested (e.g. repeated observations of students nested in schools, with random effects at the student and school levels), cross-nested (e.g. repeated observations of students nested in high schools taking different courses, with random effects at the student, school, and course levels), or even non-nested (e.g. individuals clustered within job categories and states, with random effects at the job and state level). For an introduction to multilevel models, see Gelman and Hill (2007), Fitzmaurice et al. (2004),Sullivan et al. (1999), and Bryk and Raudenbush (1992).

1.1 Motivating data

Birth records were obtained for all live births in New York City in 2003 and linked to the hospital discharge data from the Statewide Planning and Research Cooperative System by the New York State Department of Health. These data include information on mother's demographic characteristics, previous births, smoking, weight gain rate during pregnancy, maternal birth outside the United States, and infant's gender, birth weight, and gestational age (Savitz et al. (2008)), all collected from the birth certificate. These data were also linked to U.S. Census data to obtain additional demographic information at the census tract level. Investigators are interested in identifying significant predictors of birth weight among term births adjusting for gestational age, with particular emphasis on exploring disparities related to race and ethnicity.

Research has shown a persistent racial disparity in birth outcomes in many countries (e.g., Osypuk and Acevedo-Garcia (2008); Kelly et al. (2009)). Although individual and community-level covariates have been shown to account for some of the racial disparity in low birth weight (Buka et al. (2003); Roberts (1997); Rauh et al. (2001); O'Campo et al. (1997)), much of this excess risk remains unexplained. Howard et al. (2006) found substantial variability in the risk of preterm birth and low birth weight among black race subgroups defined by 8 distinct maternal ancestries (African, American, Asian, Cuban, European, Puerto Rican, South and Central American, and West Indian and Brazilian). They also found nativity (U.S. or foreign born) to be a significant predictor that varied by ancestry. Additionally, 48 of 67 (72%) studies reviewed by Gagnon et al. (2009) found differences in birthweight outcomes between migrants and natives in western industrialized countries. These studies have been limited by coarse ethnic categorization that obscures substantial with-in group heterogeneity in behavioral, psychosocial and environmental exposures. Many data sets are also limited to the crude socioeconomic indicators on the birth certificate, such as mother's completed years of education.

To expand upon this research, investigators in the NYC birth study classified maternal ancestry into 62 country regions to determine whether birthweight variability in ancestries exists within smaller geographical regions, and whether potential ancestry effects are modified by the effects of race, maternal weight gain rate, and nativity. Such variability and potential patterns therein may help researchers further understand the factors associated with racial disparities in birth outcomes. More specifically, the association with race may depend on maternal ancestry (e.g. the association with black race may depend on whether the mother has Nigerian or Jamaican ancestry), the association with ancestry may depend on nativity (e.g. the association with Nigerian ancestry may depend on whether the mother lived primarily in or outside of the U.S.), and the association with maternal weight gain rate may depend on maternal ancestry (e.g. whether a mother has Nigerian or Jamaican ancestry). Additionally, it is common for individuals with similar demographic characteristics to live in close proximity, resulting in social as well as biological similarities between subjects. Hence, investigators are also interested in controlling for and assessing the impact of residential location as defined by census tract of residence. Neighborhood factors, such as the neighborhood deprivation index (NDI), a standardized score of various socioeconomic factors at the tract level (in which higher scores represent higher levels of deprivation), may explain some racial disparities in birth outcomes.

We fitted a multilevel linear model for infant's birth weight, predicted by infant gestational age, gender, maternal race, parity, smoking status, age, weight gain rate, nativity, and the NDI. Maternal weight gain rate is defined as total gestational weight gain (lbs.) divided by the length of each woman's pregnancy (weeks). We consider random effects that allow heterogeneity in birth weights across maternal ancestries and across census tract groups, as well as interactions between maternal ancestry and race, maternal weight gain rate, and nativity. To address the important question of whether heterogeneity exists within census tracts and maternal ancestries and whether potential heterogeneity across ancestries is affected by race, maternal weight gain rate, and nativity, one must be able to evaluate whether the variances of the random effects are different from zero.

1.2 Testing variance components

Testing whether the variance of a random coefficient is equal to zero is problematic because the null hypothesis lies on the boundary of the parameter space. Such issues are addressed extensively in the literature in the context of two level linear models, e.g., strategies that use a mixture of chi-square distributions (Self and Liang (1987); Stram and Lee (1994)), score tests (Lin (1997); Commenges and Jacqmin-Gadda (1997); Verbeke and Molenberghs (2003); Molenberghs and Verbeke (2007); Zhang and Lin (2008)), Wald tests (Molenberghs and Verbeke (2007); Silvapulle (1992)), and generalized likelihood ratio tests (Crainiceanu and Ruppert (2004)). These methods are only proposed in the context of the two-level linear model although some may be generalised to further cases. In this manuscript, we use the term “two-level linear model” to denote a class of linear models that accommodates two levels in the data hierarchy (e.g. repeated observations nested within subjects); A notable example of this model is the standard linear mixed model (see Laird and Ware (1982)) for repeated measures on subjects over time. The broader term “multilevel linear model” is used to denote a class of linear models that can have more than two levels in the data hierarchy or more than one level of clustering (e.g. repeated observations nested within subjects nested in schools). Such clusters can be nested, non-nested, or cross-nested with other clusters. The two-level linear model can then be viewed as a special case of the multilevel linear model.

Methods for testing variance components in the two-level linear model are useful to some extent in nested multilevel models for testing single variance components, but the null distributions are not easily obtained for testing multiple variance components, and random effects that are non-nested or cross-nested introduce additional complications. There is very little research specifically on testing variance components in multilevel models with more than two levels. Bryk and Raudenbush (1992) proposed a chi-square test of the residuals for evaluating variance components in multilevel models and incorporate this test in the multilevel modeling software package HLM. Other approaches for nested models include various versions of the likelihood ratio test (Snijders and Bosker (1999); Bliese (2002); Hox (2002)), e.g. using a one-tailed significance level or using a mixture of chi-square distributions. Berkhof and Snijders (2001) proposed three score tests for variance components in multilevel models and compared their methods via simulation to the likelihood ratio test, fixed F test, and Wald test. Their simulations only considered two level models and it is not clear whether generalizations to a larger number of levels are possible. Goldstein (1986) proposed a simple algorithm for fitting a multilevel linear mixed effects model for variance components near the boundary, but the manuscript did not provide a method for testing of whether or not a variance component is equal to zero. Fitzmaurice et al. (2007) proposed a permutation test for variance components in multilevel generalized linear mixed models. They applied their method to two-level generalized mixed models and suggested strategies for multilevel models with greater than two levels. Their strategy cannot be directly applied to multilevel models with crossed random effects and can only test one variance component at a time.

Bayes factors, or ratios of marginal likelihoods under equal prior probabilities, provide alternatives to frequentist hypothesis testing (see Kass and Raftery (1995)). In multilevel modeling settings, Bayes factors are ideal for comparing various types of models (e.g. multiple random effects, cross-nested or non-nested random effects), but the marginal likelihoods typically involve high dimensional integrals and are not available in closed form. Hence one must rely on approximations to the Bayes factor. The most widely used approximation to the Bayes factor is based on the Laplace approximation (Tierney and Kadane (1986)), resulting in the Bayesian information criterion (BIC) (Schwarz (1978)) under certain assumptions. These approximations suffer in performance from high-dimensionality (Kass and Raftery (1995)) and hence have limited applicability in multilevel models. The BIC and Laplace approximations are based on the assumption that the dimension of parameters is fixed as the sample size goes to infinity. This is problematic in multilevel models because the dimension of parameters increases as the sample size increases. Due to a violation of regularity conditions underlying the approximation, the Laplace method can fail when the parameter lies on the boundary of the parameter space (Pauler et al. (1999); Hsiao (1997); Erkanli (1994)).

Markov chain Monte Carlo (MCMC) methods provide alternatives for approximating Bayes factors. Many of these methods can fail for certain types of “default” priors on the variance components (Pauler et al. (1999)). Bayesian stochastic search variable selection methods using MCMC methods in the two-level linear model (e.g. Cai and Dunson (2006); Kinney and Dunson (2008)) may be generalizable to multilevel models, but these methods are generally computationally demanding and time consuming. Many other MCMC methods exist for model comparisons, e.g. the logarithm of the pseudo marginal likelihood (Gelfand (1996)), the Deviance Information Criterion (Spiegelhalter et al. (2002)), and other related methods for estimating marginal likelihoods (e.g., Chib and Jeliazkov (2001)). These methods generally require the fitting of each model being compared (i.e. MCMC samples from the posterior distribution for eachmodel) and are computationally demanding in high dimensional models. In addition, even though conceptually one can obtain a perfect estimate of the Bayes factor using MCMC methods run for infinitely-many iterations, in practice MCMC algorithms can only be run for a finite number of samples and the existing algorithms may require a very large number of iterations to obtain an accurate estimate. Hence, in practice MCMC-based estimates of Bayes factors are also approximate and it is not clear that such estimates will in general be closer to the truth (given chains of the length that are typically run for practical reasons) than faster analytic approximations. In an attempt to develop a more efficient approximation to the Bayes factor, Saville and Herring (2008) proposed a method for approximating Bayes factors in the two-level linear model via a relatively simple Laplace approximation to the marginal likelihood. Their method does not require the fitting of a model via MCMC methods but only applies to the simple case of a two-level multilevel linear model.

It is well known that Bayes factors can be sensitive to the choice of prior distributions (Kass and Raftery (1995)). This is challenging in model selection problems in which one has no prior information on the parameters. In these situations it is common to use default priors that do not require subjective inputs. One must choose these default priors with care, because as the prior variance increases, the Bayes factor will increasingly favor the null model (Bartlett (1957)). Berger and Pericchi (1996) discuss various procedures for default priors for model selection via Bayes factors. These include the authors’ proposed intrinsic Bayes factors, the Schwarz approximation (Schwarz (1978)), and the methods of Jeffreys (1961) and Smith and Spiegelhalter (1980). For improper non-informative priors, the Bayes factor involves an arbitrary constant, and hence is not well defined (Spiegelhalter and Smith (1982)). Gelman (2006) discusses various approaches to default priors specifically for variance components. Common approaches include the uniform prior (e.g. Gelman (2007)), the half-*t* family of prior distributions, and the inverse gamma distribution (Spiegelhalter et al. (2003)). These prior distributions can encounter difficulties when the variance components are close to 0. Other discussions of selecting default priors on variance components are presented by Natarajan and Kass (2000), Browne and Draper (2006), and Kass and Natarajan (2006). As an alternative to these approaches, Saville and Herring

(2008) scaled the random effects to the residual variance and introduced default priors that were shown to have good frequentist properties in the two-level linear model.

As noted previously, testing hypotheses on the boundary is problematic for certain classical (i.e. frequentist) approaches because traditional asymptotic results do not apply directly (e.g. it becomes more difficult to approximate the p-value for a likelihood ratio test). In the Bayesian case, there are no conceptual problems with testing null hypotheses on the boundary of the parameter space, but the Laplace approximation to the marginal likelihood under the alternative can be inaccurate when the parameter lies close to the boundary. We generalize the approach of Saville and Herring (2008) for testing variance components via Bayes factors to multilevel linear models with more than two levels in the data hierarchy (i.e. more than one level of clustering). The method does not require MCMC samples from the posterior distribution or the fitting of each model being compared; hence it is computationally more efficient than many of the current Bayesian methods available for multilevel linear models. The strategy is to scale the random effects to the residual variance and introduce parameters that control the relative contribution of the random effects. This scaling enables one to integrate over the random effects and variance components from the posterior in closed form, such that the resulting integrals needed to calculate the Bayes factor are of small dimension and can be efficiently approximated with Laplace's method. In addition, we have improved the accuracy of the Laplace approximation by transforming the scale parameter so that the boundary lies at negative infinity instead of zero. Our method is relatively fast to implement and may incorporate default prior distributions shown to have good frequentist properties in the two-level linear model (Saville and Herring (2008)). We present the Bayesian model selection problem in Section 2. We summarize our method for approximating the marginal likelihood in Section 3 and apply our method to the NYC birth data in Section 4. A discussion follows in Section 5.

2 Bayes factors and the multilevel linear model

2.1 Notation

We define the general multilevel linear model with q random factors as

$$\begin{aligned} Y_i = & \boldsymbol{x}_i' \boldsymbol{\beta} + \boldsymbol{z}_i' \mathbf{b}_{[i]} + \boldsymbol{\varepsilon}_i, \\ = & \boldsymbol{x}_i' \boldsymbol{\beta} + \sum_{h=1}^{q} \boldsymbol{z}_{ih}' \mathbf{b}_{h[i]} + \boldsymbol{\varepsilon}_i, \end{aligned} \qquad (1)$$

in which Y_i is the response for observation i, $i = 1, \dots, m$, $\boldsymbol{x}_i$ is a $p \times 1$ vector of predictors with corresponding fixed effects β. Defining

$\mathbf{b}_{[i]} = \left(\mathbf{b}_{1[i]}', \dots, \mathbf{b}_{q[i]}'\right)'$ and $\boldsymbol{z}_i = \left(\boldsymbol{z}_{i1}', \dots, \boldsymbol{z}_{iq}'\right)'$, we note $\boldsymbol{z}_{ih}$ is a $d_h \times 1$ vector of predictors with corresponding random effects $\boldsymbol{b}_{h[i]}$, in which $[i]$ indexes the group in factor h pertaining to the ith observation, and $\boldsymbol{b}_{h[i]} \sim N(\mathbf{0}, \Psi_h)$ is independent of $\varepsilon_i \sim N(0, \sigma^2)$, with $\boldsymbol{b}_{h[i]}$ independent of $\boldsymbol{b}_{h'[i]}$ for $h \neq h'$. From a Bayesian perspective, prior distributions are specified for β, Ψ_h, and σ^2 to reflect prior knowledge of the parameters. When one of the q random factors is nested within another random factor (e.g. maternal ancestry nested within geographical region), a hierarchical structure is created. A key feature of multilevel modeling is the incorporation of covariates $\boldsymbol{x}_i$ that can be measured at any level of the hierarchy. This allows one to address the effect of a given covariate, say at the ancestry level, while controlling for the effect of a higher level covariate, say at the geographical region level. One must interpret such regression parameters carefully because some covariates can operate at many different levels.

To illustrate, consider the NYC birth data for 2003, in which there are 104,710 observations within 62 ethnic ancestries and 2,128 census tracts. The aims of our analysis are to identify significant predictors of infant birth weight and to determine whether there is heterogeneity across ancestry groups and census tracts. To start, we will consider one predictor, maternal weight gain rate during pregnancy, which has been linked to infant birth weight. Because of social and biological characteristics shared by persons of the same ancestry, the effect of maternal weight gain rate may vary by country of origin. A non-nested multilevel linear model, with a random intercept and slope (for weight gain rate) at the ancestry level and a random intercept at the tract level, can evaluate this hypothesis. One model is

$$Y_i=\beta_0+b_{10[i]}+\left(\beta_1+b_{11[i]}\right)x_i+b_{20[i]}+\varepsilon_i, \quad (2)$$

in which Y_i is the weight of infant i, x_i is the weight gain rate of the ith mother, β_0 is the model intercept, β_1 is the parameter corresponding to weight gain rate, $b_{10[i]}$ is the random intercept and $b_{11[i]}$ the random slope corresponding to the ancestry of mother i, and $b_{20[i]}$ is the random intercept corresponding to the census tract of mother i. There are a total of $2 \times 62 = 124$ random effects at the ancestry level and 2,128 random effects at the census level. In order to test whether there is heterogeneity in birth weights across ancestries ($h = 1$) or census tracts ($h = 2$), one can conduct a test of whether the variance of the respective random effects is equal to 0. This corresponds to a test of $H_0 : \Psi_h = \mathbf{0}$, which lies on the boundary of the parameter space.

2.2 The Laplace approximation to the Bayes factor

From a Bayesian perspective, one can evaluate $H_0 : \Psi_h = \mathbf{0}$ by calculating the Bayes factor, or posterior odds of M_1 versus M_0 given equal prior odds, given by

$$B_{10}=\frac{p\left(\boldsymbol{Y}|M_1\right)}{p\left(\boldsymbol{Y}|M_0\right)}, \quad (3)$$

in which M_0 is model corresponding to the null hypothesis (variance components equal to 0) and M_1 is the model corresponding to the alternative hypothesis (variance components greater than 0). Calculating the Bayes factor requires the marginal likelihood

$$p\left(\boldsymbol{Y}|M_k\right)=\int p\left(\boldsymbol{Y}|\boldsymbol{\theta}_k, M_k\right)\pi\left(\boldsymbol{\theta}_k|M_k\right)d\boldsymbol{\theta}_k, \quad (4)$$

in which $p(\boldsymbol{Y}|\theta_k, M_k)$ is the data likelihood for model M_k, θ_k is the vector of model parameters, and $\pi(\theta_k|M_k)$ is the prior distribution of θ_k.

To approximate the marginal likelihood, we consider the Laplace approximation, which is based on a linear Taylor series approximation of $\tilde{l}(\theta_k) = \log\{p(\boldsymbol{Y}|\theta_k, M_k)\pi(\theta_k|M_k)\}$. The marginal likelihood $p(\boldsymbol{Y}|M_k)$ is estimated by

$$\widehat{p}\left(\boldsymbol{Y}|M_k\right)=(2\pi)^{d_k/2}\left|\tilde{\boldsymbol{\Sigma}}_k\right|^{1/2}p(\boldsymbol{Y}|\tilde{\boldsymbol{\theta}}_k, M_k)\pi(\tilde{\boldsymbol{\theta}}_k M_k), \quad (5)$$

in which $\tilde{\Sigma}_k$ is the inverse of the negative Hessian matrix of $l(\tilde{\theta}_k)$ evaluated at the posterior mode $\tilde{\theta}_k$, $p(\boldsymbol{Y}|\tilde{\theta}_k, M_k)$ is the marginal posterior evaluated at the posterior mode, $\pi(\tilde{\theta}_k|M_k)$ is the prior evaluated at the posterior mode, and d_k is the dimension of θ_k. Hence, in order to implement the Laplace approximation, one only needs the matrix of second partial derivatives and the posterior mode of $l(\tilde{\theta}_k)$, which for small dimensions is easily computed in standard statistical software packages. As noted previously, multilevel models are typically high-dimensional and may involve variance components near the boundary, meaning the Laplace approximation cannot be directly applied to the integral in (4).

3 Approximating the marginal likelihood

We outline the general strategy for the proposed methods and provide complete mathematical details in the Appendix. For computational convenience, we reparameterize the multilevel linear model given in (1) so that all random effects are contained in one vector $\boldsymbol{b}$. For example, in equation (2), there are 62 × 2 = 124 random effects corresponding to the random intercept and slope for maternal ancestry, and there are 2,128 random effects corresponding to the random intercept of census tract. These random effects are stacked into one vector $\boldsymbol{b}$ of dimension (2,252 × 1). A corresponding sparse design matrix $\boldsymbol{w}_i$ is created of the same dimension (i.e. a vector) that will contain mostly 0's, with non-zero elements corresponding to the appropriate random effects for observation i. Prior distributions specific to a given application are specified for β, σ^2 and Ψ_h, which is the covariance matrix of the random effects b_{hl} corresponding to factor h and classification l. We assume normality of the random effects $\boldsymbol{b}_{hl} \sim N(0, \Psi_h)$ which are independent of the residual error $\varepsilon_i \sim N(0, \sigma^2)$.

Extending the work of Saville and Herring (2008), we scale the random effects, now denoted as $\tilde{\boldsymbol{b}}$, to the residual variance such that $\tilde{\boldsymbol{b}}_{hl} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I})$ and introduce a parameter vector φ_h that controls the relative contribution of the scaled random effects. We also allow correlation between the respective random effects for a given factor through a parameter vector γ_h. For example, consider the cross-classified (non-nested) model given by equation (2). The re-parameterized model takes the form

$$Y_i=\beta_0+e^{\phi10}\tilde{b}_{10[i]}+\left(\beta_1+e^{\phi11}\tilde{b}^*_{11[i]}\right)x_i+e^{\phi20}\tilde{b}_{20[i]}+\varepsilon_i, \quad (6)$$

in which $\tilde{b}_{10[i]}$ is the scaled random intercept and $\tilde{b}_{11[i]}$ the scaled random slope corresponding to the ancestry of mother i, $\tilde{b}_{20[i]}$ is the scaled random intercept corresponding to the census tract of mother i, and $\tilde{b}^*_{11[i]}=\gamma_{1_{10}}\tilde{b}_{10[i]}+\tilde{b}_{11[i]}$, in which $\gamma 1_{10}$ allows correlation between the scaled random intercept and slope corresponding to ancestry (where the subscript h_{10} on γ denotes correlation between $\tilde{b}^*_{h1[i]}$ and $\tilde{b}_{h0[i]}$ for factor h). The random effects for this example correspond to a random intercept and slope at the ancestry level and a random intercept at the census level. Expression (6) is related to reparameterizations used to reduce autocorrelation in MCMC algorithms for multilevel models (Browne et al., 2009), though our focus and motivation are fundamentally different.

Let $\boldsymbol{Y} = (Y_i, \dots, Y_m)'$ and $\sigma^2 \sim \text{InvGam}(v, w)$. The primary reason for scaling the random coefficients to the residual variance is that it allows the integration of $\tilde{\boldsymbol{b}}$ and σ^2 from the posterior distribution in closed form, i.e. the marginal posterior $p(\boldsymbol{Y}|\beta, \varphi, \gamma)$ has a multivariate t-distribution. This enables one to obtain an accurate approximation of the marginal likelihood using Laplace's method. We assume the default prior $\varphi_{hk} \sim N(\log(0.3), 2)$ (corresponding to the kth random effect for factor h) suggested by Saville and Herring (2008) and use the Laplace method to integrate over (β, φ, γ) to obtain the marginal density $p(\boldsymbol{Y})$. This default prior was shown to have good frequentist properties in simulation studies in the two-level linear model. The prior distributions for β and γ as well as the values of v and w in the prior for σ^2 are set by the investigator based on the specific application. Following Gelman et al. (2006; 2008), we advocate weakly informative priors that are chosen by subject matter knowledge in an application area but with the prior variance modestly inflated relative to one's subjectively chosen prior variance to allow robustness. The elicitation process is illustrated through the motivating application in the following

Section. Because of the rescaling and subsequent integration, the dimension of the marginal posterior $p(\boldsymbol{Y}|\beta, \varphi, \gamma)$ is much smaller than that of the data likelihood $p(\boldsymbol{Y}|\beta, \varphi, \gamma, \tilde{\boldsymbol{b}}, \sigma^2)$ and lacks parameters with boundary constraints (i.e. variance components). For example, in the model given by (6), the density $p(\boldsymbol{Y}|\beta, \varphi, \gamma)$ only incorporates 2 parameters in β, 3 parameters in φ, and one parameter in γ. In addition, the number of parameters in the marginal posterior $p(\boldsymbol{Y}|\beta, \varphi, \gamma)$ is fixed as the sample size goes to infinity. Hence the Laplace approximation can be used to efficiently approximate the marginal likelihood $p(\boldsymbol{Y})$.

4 Application

We are interested in comparing various multilevel linear models for infant's birth weight, predicted by infant gestational age at delivery, gender, maternal race, parity, smoking status, age, weight gain rate, maternal nativity, and the neighborhood deprivation index, with random effects for census tracts and ethnic ancestries. We focus on singleton term births with a gestational age ≥ 37 weeks and a birth weight between 900 g and 5300 g. After exclusions, we have a total of 93,938 subjects with complete data available for the analysis.

We consider several competing models with various random coefficient structures (see Table 1). The first model we investigate allows a random intercept for ancestry (country of origin), defined as

$$M_1{:}Y_i{=}\boldsymbol{x}_i'\boldsymbol{\beta}{+}b_{1[i]}{+}\varepsilon_i, \quad (7)$$

with

$$\begin{aligned}\boldsymbol{x}_i'\boldsymbol{\beta}={}&\beta_0+\beta_1\mathrm{Black}_i+\beta_2\mathrm{Hisp}_i+\beta_3\mathrm{Asian}_i+\beta_4\mathrm{Other}_i+\beta_5\mathrm{Gest}_i\\&+\beta_6\mathrm{Gest}_i^2+\beta_7\mathrm{Pbirth}_i+\beta_8\mathrm{Female}_i+\beta_9\mathrm{Smoke}_i+\beta_{10}\mathrm{Foreign}_i\\&+\beta_{11}\mathrm{NDI}_i+\beta_{12}\mathrm{Age2}_i+\beta_{13}\mathrm{Age3}_i+\beta_{14}\mathrm{Age4}_i+\beta_{15}\mathrm{Age5}_i\\&+\beta_{16}\mathrm{Wtgain}_i+\beta_{17}\mathrm{Wtgain}_i^2+\beta_{18}\mathrm{Wtgain}_i^3.\end{aligned} \quad (8)$$

The explanatory variables Black_i, Hisp_i, Asian_i, and Other_i are indicator variables for race corresponding to black, Hispanic, Asian or Pacific Islander, and other (white is the reference group). Gest_i is the infant gestational age in weeks for subject i and Gest_i^2 is the corresponding quadratic variable. The variables Pbirth_i, Female_i, Smoke_i, and $\mathrm{Foreign}_i$ are indicator variables for any previous births, female infant gender, maternal smoking, and maternal birth outside of the United States, respectively. The variable NDI_i is the neighborhood deprivation index corresponding to the census tract of subject i (with higher values indicating more deprived living conditions). At the request of our epidemiologist collaborators, maternal age was categorized into the following groups: < 25yrs (reference group), 26–30 yrs (Age2i), 31–35 yrs(Age3i), 36–40 yrs (Age4i), and > 40 yrs ($\mathrm{Age5}_i$). The variable Wtgain_i is the difference in pounds in maternal pre-pregnancy weight and weight at delivery (deliver weight minus pre-pregnancy weight), and Wtgain_i^2 and Wtgain_i^3 are the corresponding quadratic and cubic variables, respectively. The continuous variables Gest_i, NDI_i, and Wtgain_i are centered and standardized by 2 standard deviations to place the regression coefficients on a similar scale as the binary indicators (Gelman (2008)). The quadratic and cubic versions of those variables are based on the standardized variables. The random intercept $b_{1[i]} \sim N(0, \psi_1)$ corresponds to the ancestry of subject i independent of $\varepsilon_i \sim N(0, \sigma^2)$.

We also consider a model with a random intercept for census tracts but without random effects for ancestries,

$$M_2{:}Y_i=\boldsymbol{x}_i^{\prime}\boldsymbol{\beta}+b_{2_{[i]}}+\varepsilon_i, \quad (9)$$

in which $b_{2[i]} \sim N(0, \psi_2)$ is the random intercept corresponding to the census tract of subject i. Incorporating random intercepts for both ancestries and census tracts, a two-factor cross-classified (non-nested) model takes the form

$$M_3{:}Y_i=\boldsymbol{x}_i^{\prime}\boldsymbol{\beta}+b_{1_{[i]}}+b_{2_{[i]}}+\varepsilon_i \quad (10)$$

in which $b_{1[i]} \sim N(0, \psi_1)$ is independent of $b_{2[i]} \sim N(0, \psi_2)$. As discussed previously, the effect of race may depend on maternal ancestry. Hence we consider a variation of M_3 with random intercepts for both ancestry and census tract, but we allow the effect of race to vary by ancestry. This model can be written as

$$M_4{:}Y_i=\boldsymbol{x}_i^{\prime}\boldsymbol{\beta}+b_{1p_{[i]}}+b_{2_{[i]}}+\varepsilon_i, \quad (11)$$

in which $b_{1p[i]} \sim N(0, \psi_{1p})$ is the random intercept corresponding to the ancestry (factor 1) of subject i within race p, independent of $b_{2[i]}$. This model assumes that two persons of the same ancestry with different races have different random intercepts. Similarly, it may also be the case that the effect of ancestry varies by nativity. Hence we consider

$$M_5{:}Y_i=\boldsymbol{x}_i^{\prime}\boldsymbol{\beta}+b_{1s_{[i]}}+b_{2_{[i]}}+\varepsilon_i, \quad (12)$$

in which $b_{1s[i]} \sim N(0, \psi_{1s})$ is the random intercept corresponding to the ancestry (factor 1) of subject i within nativity s, independent of $b_{2[i]}$. This model assumes that two persons of the same ancestry but different nativity (one foreign born and one not foreign born) have distinct random intercepts. It may also be the case that the effect of maternal weight gain rate on infant birth weight is affected by ancestry. This may result from either biological or social factors that are correlated with a given ancestry. We can model this heterogeneity by including a random slope for maternal weight gain rate for the ancestry factor. Adding this component to model M_3, we have

$$M_6{:}Y_i=\boldsymbol{x}_i^{\prime}\boldsymbol{\beta}+b_{10_{[i]}}+b_{2_{[i]}}+b_{11_{[i]}}\mathrm{Wtgain}_i+\varepsilon_i, \quad (13)$$

in which $b_{10[i]}$ is the random intercept and $b_{11[i]}$ is the random slope for weight gain rate corresponding to the ancestry of subject i, and $\boldsymbol{b}_{1[i]} = (b_{10[i]}, b_{11[i]})' \sim N(\mathbf{0}, \Psi_1)$ are independent of $b_{2[i]}$.

Previous research has shown heterogeneity of infant birth weights from women in different geographical regions (Howard et al. (2006)). Hence we also consider a model that includes random intercepts for the 15 geographical regions based on maternal ancestry in addition to random intercepts for maternal ancestry (country of origin) and census tract, given by

$$M_7{:}Y_i=\boldsymbol{x}_i^{\prime}\boldsymbol{\beta}+b_{1_{[i]}}+b_{2_{[i]}}+b_{3_{[i]}}+\varepsilon_i, \quad (14)$$

in which $b_{3[i]} \sim N(0, \psi_3)$ is the random intercept corresponding to the geographical region of subject i, independent of $b_{1[i]}$ and $b_{2[i]}$. Finally, we consider a model without random effects,

$$M_0{:}Y_i=\boldsymbol{x}_i^{\prime}\boldsymbol{\beta}+\varepsilon_i, \quad (15)$$

Our goal is to identify the preferred model and to proceed with inference using this chosen model.

The mean value for infant birth weight is 3,362 grams with a standard deviation of 460 g. Converting to kilograms for computational convenience, we use prior distributions $\beta_0 \sim N(3.36,1)$, $\beta \sim N(\mathbf{0}, \boldsymbol{I})$, and $\sigma^2 \sim \text{InvGam}(0.1, 0.1)$, which are weakly informative priors given the scale of the response and predictors. We found very strong evidence for heterogeneity in birth weights across census tracts and across ancestries (log $\hat{B}_{10} = 275$, log $\hat{B}_{20} = 32$, and log $\hat{B}_{30} = 283$, in which $\hat{B}_{kk'}$ denotes the estimated Bayes factor comparing M_k to $M_{k'}$ as given by equation 3), with birth weights tending to vary across maternal ancestries in greater magnitude than across census tracts. We found that the effects of race (log $\hat{B}_{43} = -6$), nativity (log $\hat{B}_{53} = -10$) and maternal weight gain rate (log $\hat{B}_{63} = -1$) do not vary by ancestry. Additionally, birth weights did not vary significantly by geographical region after accounting for maternal ancestry and census tract of residence (log$B_{73} = -2$).

We fitted the preferred model, M_3, using MCMC methods and based inference on 20,000 samples after discarding 5,000 as a burn-in. The posterior means and 95% credible intervals of the fixed effects are given in Table 2. Results are presented in grams for better interpretability. Predictors with 95% credible intervals greater than 0 include parity (99,111), maternal age 26–30 (45,60), maternal age 31–35 (64,80), maternal age 36–40 (74,93), maternal age >40 (60,92), and maternal foreign nativity (3,19). Hence, previous live births, greater maternal age, and maternal birth outside the U.S. are all associated with greater infant birth weights. Predictors with 95% credible intervals that are less than 0 include maternal Asian race (−92,−21), black race (−75,−5), infant female gender (−126,−115), maternal smoking (−186,−143), and higher neighborhood deprivation (95% CI=(−23,−9) for a 2 sd increase). Hence, Asian and black race (compared to white), female infants (compared to males), smokers (compared to nonsmokers), and greater NDI values are associated with lower infant birth weights. Both maternal weight gain rate and infant gestational age showed non-linear associations with infant birth weight. Figure 1 shows that in the range of 0.25–2 lbs./week, a greater maternal weight gain rate is associated with greater infant birth weights; in the range of less than 0.25 or greater than 2 lbs./week, a greater maternal weight gain rate is associated with smaller infant birth weights, although some caution should be exercised in the interpretation at the extremes of the data. Figure 1 shows greater gestational age is associated with greater infant birth weights, but this association flattens somewhat as gestational age nears the right tail of its distribution (44 weeks), perhaps due to inaccurate pregnancy dating. The variables with the largest effects on infant birth weight are smoking ($\hat{\beta}_9 = -165$), female infant gender ($\hat{\beta}_8 = -120$), maternal weight gain rate (non-linear), and infant gestational age (non-linear). Variables with weaker yet “significant” associations include a 2 sd increase in NDI ($\hat{\beta}_{11} = -16$), maternal foreign nativity ($\hat{\beta}_{10} = 11$), and black versus white race ($\hat{\beta}_1 = -40$). Although these smaller effects have little clinical relevance at the individual level, they are interesting findings for etiologic purposes, as a shift of the population distribution by a few grams can push many individuals beyond a critical point in the tail regions, potentially affecting perinatal mortality or other related outcomes at the population level. The 95% credible intervals for Hispanic ethnicity (−31, 57) and “other” race (−81, 75) contain zero, indicating non-significant associations with infant birthweight. The non-significant result for Hispanic race may be due to the nature in which the variable was constructed. Data were not initially collected for Hispanic race, and investigators therefore constructed a Hispanic indicator variable using the ethnic ancestry variable. Hence this predictor may lack the precision of the other race indicator variables. The “other” race group suffered from small sample size.

Figure 2 displays 95% credible intervals for the ancestry random intercepts. Ancestries with the greatest estimated infant birth weights include Peru, Morocco, and Nigeria, while ancestries with the lowest estimated infant birth weights include Guyana, Bangladesh, Gambia, and Ivory Coast. There were no notable trends across geographical regions.

5 Discussion

In these data with uniquely rich ancestry and geographic information, we found very strong evidence for heterogeneity in full-term infant birth weights across census tracts and across ancestries. Moreover, the variation in birth weight across maternal ancestries was greater in magnitude than across census tracts, and did not vary substantially by race, maternal weight gain rate, or nativity. We note that the tests of heterogeneity for birth weights across maternal ancestries by race or nativity may suffer from low power, due to the fact that many countries are comprised predominantly of one race and similar nativity (see Table 3). The finding of heterogeneity in birth weights across maternal ancestries is generally consistent with the findings of Howard et al. (2006), although those authors studied only black women in New York City and observed the effects of nativity to vary by maternal ancestry region. One limitation of their study was the grouping of West Indian and Brazilian ancestry, which was an artifact of the coding scheme used in data collection. Furthermore, Howard et al. (2006) and many previous papers focus on preterm birth (gestational age less than 37 weeks), whereas the current paper examines birthweight variability among full-term births only. The advantage of our outcome definition approach is to focus more clearly on variations in fetal growth, as small babies can arise from two mechanisms: shorter gestational age and intrauterine growth retardation, and the etiologies of these mechanisms may be entirely distinct (Wilcox and Skjaerven (1992)). While more etiologically focused on infant growth, however, this approach does restrict the distribution of birthweights included in our analyses, since much of the natural variability is contributed by gestational age. Therefore we are decomposing a subset of the true variability in birthweights, and our results apply specifically to mechanisms that operate through modifying intrauterine growth rate. The causal mechanisms for heterogeneity across ancestries may be due to any of a large number of unmeasured social and biological factors, including diet, physical activity, social support and maternal health conditions. Furthermore, it is important to note that coefficient estimates shown here are adjusted for measured covariates, but that in reality groups differ widely in mean values for these covariates. For example, the posterior means shown in Figure 2 hold constant all variables included as covariates in Model 3, but the reality is that these covariates are not constant across these groups in the population. Furthermore, the group means could differ even more dramatically if preterm births were also included.

The estimates here are useful for demonstrating how dramatically subpopulations can differ in outcomes, even when controlling for the important known determinants of birthweight. Groups in Figure 2 range over several hundred grams in their adjusted mean weights, a value which is large compared to known risk factors, such as maternal smoking. Furthermore, despite a great deal of literature on racial predisposition to adverse birth outcomes (Kistka et al. (2007)), the greatest variation observed in these data is at the level of national ancestry. For example, Nigeria and Gambia are both West African populations tied to the ancestral origins of the African-American population, and yet the former has an adjusted mean about 100g above the overall grand mean, whereas the latter has an adjusted mean about 100g below the overall grand mean. Unique patterns of selective migration from these countries are among a large number of possible explanations for such patterns, but they are less consistent with theories of racial predisposition. Ancestries with the greatest adjusted infant birth weights include Peru, Morocco, and Nigeria, which have no obvious connection. Nor do the ancestries with the lowest adjusted infant birth weights, such as Guyana, Bangladesh, Gambia, and Ivory Coast. As noted previously, there were no notable trends across broader geographical regions after accounting for country of origin. This contrasts somewhat with earlier work that found heterogeneity in adverse birth outcomes across large ancestry regions for black women (Howard et al. (2006)), though this work did not account for country-specific effects.

In summary, we have developed statistical methodology that has enabled the testing of random effects in the NYC birth weight study. Our approach avoids issues with testing on the boundary of the parameter space, uses low-dimensional approximations to the Bayes factor, and incorporates default priors for the variance components. Simulation studies (available from authors by request) suggest that these priors have good frequentist properties and large sample consistency. The methodology is applicable to designs with any number of random effects for any number of nested, non-nested, or crossnested factors, although computational limitations may exist for extremely high dimensional problems (see Appendix for discussion and proposed strategies). A major contribution of our method is the ability to test several variance components from multiple factors simultaneously, and to do so for nested, non-nested, or cross-nested multilevel designs.

Acknowledgments

We are grateful to Dr. David Savitz of the Mount Sinai School of Medicine for providing the NYC birth data and for his helpful comments on the manuscript. This study was supported by the National Institute of Child Health and Human Development (R21-HD050739) and National Institute of Environmental Health Sciences (T32-ES007018).

References

Bartlett MS. Comment on “A Statistical Paradox” by D. V. Lindley. Biometrika. 1957; 44:533–534.

Berger JO, Pericchi LR. The intrinsic Bayes factor for model selection and prediction. Journal of the American Statistical Association. 1996; 91:109–122.

Berkhof J, Snijders TA. Variance component testing in multilevel models. Journal of Educational and Behavioral Statistics. 2001; 26:133–152.

Bliese, PD. Multilevel random coefficient modeling in organizational research: Examples using SAS and S-PLUS. San Francisco: Jossey-Bass; 2002. p. 401-445.

Browne WJ, Draper D. A comparison of Bayesian and likelihood-based methods for fitting multilevel models (with discussion). Bayesian Analysis. 2006; 1:473–514.

Browne WJ, Steel F, Golalizadeh M, Green MJ. The use of simple reparameterizations to improve the efficiency of markov chain monte carlo estimation for multilevel models with applications to discrete time survival models. Journal of the Royal Statistical Society A - Statistics in Society. 2009; 172:579–598.

Bryk, AS.; Raudenbush, SW. Hierarchical Linear Models: Applications and Data Analysis Methods. Newbury Park, CA: Sage Publications; 1992.

Buka S, Brennan R, Rich-Edwards J, Raudenbush S, Earls F. Neighborhood support and the birth weight of urban infants. American Journal of Epidemiology. 2003; 157:1–8. [PubMed: 12505884]

Cai B, Dunson DB. Bayesian covariance selection in generalized mixed models. Biometrics. 2006; 62:446–457. [PubMed: 16918908]

Chib S, Jeliazkov I. Marginal likelihood from the metropolis-hastings output. Journal of the American Statistical Association. 2001; 96:270–281.

Commenges D, Jacqmin-Gadda H. Generalized score test of homogeneity based on correlated random effects models. Journal of the Royal Statistical Society, Series B. 1997; 59:157–171.

Crainiceanu CM, Ruppert D. Likelihood ratio tests in linear mixed models with one variance component. Journal of the Royal Statistical Society, Series B. 2004; 66:165–185.

Erkanli A. Laplace approximations for posterior expectations when the mode occurs at the boundary of the parameter space. Journal of the American Statistical Association. 1994; 89:250–258.

Fitzmaurice, GM.; Laird, NM.; Ware, JH. Applied Longitudinal Analysis. Hoboken, New Jersey: John Wiley & Sons, Inc; 2004.

Fitzmaurice GM, Lipsitz SR, Ibrahim JG. A note on permutation tests for variance components in multilevel generalized linear mixed models. Biometrics. 2007; 63:942–946. [PubMed: 17403100]

Gagnon A, Zimbeck M, Zeitlin J. The ROAM Collaboration. Migration to western industrialised countries and perinatal health: A systematic review. Social Science and Medicine. 2009; 69:934–946. [PubMed: 19664869]

Gelfand, AE. Model determination using sampling-based methods, pages 145–161. New York: Chapman & Hall; 1996.

Gelman A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Analysis. 2006; 1:515–533.

Gelman, A. Running WinBugs and OpenBugs from R. 2007. Available at www.stat.columbia.edu/_gelman/bugsR/. Accessed April 2008

Gelman A. Scaling regression inputs by dividing by two standard deviations. Statistics in Medicine. 2008; 27:2865–2873. [PubMed: 17960576]

Gelman, A.; Hill, J. Data Analysis using Regression and Multilevel/Hierarchical Models. New York, NY: Cambridge University Press; 2007.

Goldstein H. Multilevel mixed linear model analysis using iterative generalized least squares. Biometrika. 1986; 73:43–56.

Howard DL, Marshall SS, Kaufman JS, Savitz DA. Variations in low birth weight and preterm delivery among blacks in relation to ancestry and nativity: New York City, 19982002. Pediatrics. 2006; 118:E1399–E1405. [PubMed: 17079541]

Hox, J. Multilevel analysis: Techniques and applications. Mahwah, NJ: Lawrence Erlbaum; 2002.

Hsiao CK. Approximate Bayes factors when a mode occurs on the boundary. Journal of the American Statistical Association. 1997; 92:656–663.

Jeffreys, H. Theory of Probability. 3rd edition. Oxford, U.K.: Oxford University Press; 1961.

Kass RE, Natarajan R. A default conjugate prior for variance components in generalized linear mixed models (comment on article by Browne and Draper). Bayesian Analysis. 2006; 1:535–542.

Kass RE, Raftery AE. Bayes factors. Journal of the American Statistical Association. 1995; 90:773–795.

Kelly Y, Panico L, Bartley M, Marmot M, Nazroo J, Sacker A. Why does birthweight vary among ethnic groups in the UK? Findings from the Millennium Cohort Study. Journal of Public Health. 2009; 31:131–137. [PubMed: 18647751]

Kinney SK, Dunson DB. Fixed and random effects selection in linear and logistic models. Biometrics. 2008; 63:690–698. [PubMed: 17403104]

Kistka ZAF, Palomar L, Lee KA, Boslaugh SE, Wangler MF, Cole FS, DeBaun MR, Muglia LJ. Racial disparity in the frequency of recurrence of preterm birth. Am J Obstet Gynecol. 2007; 196:131.e1–131.e6. [PubMed: 17306652]

Laird N, Ware J. Random-effects models for longitudinal data. Biometrics. 1982; 38:963–974. [PubMed: 7168798]

Lin X. Variance components testing in generalised linear models with random effects. Biometrika. 1997; 84:309–326.

Molenberghs G, Verbeke G. Likelihood ratio, score, and Wald tests in a constrained parameter space. The American Statistician. 2007; 61:22–27.

Natarajan R, Kass RE. Reference Bayesian methods for generalized linear mixed models. Journal of the American Statistical Association. 2000; 95:227–237.

O'Campo P, Xue X, Wang M, Caughy M. Neighborhood risk factors for low birthweight in Baltimore: A multilevel analysis. American Journal of Public Health. 1997; 87:1113–1118. [PubMed: 9240099]

Osypuk TL, Acevedo-Garcia D. Are racial disparities in preterm birth larger in hyper-segregated areas? American Journal of Epidemiology. 2008; 167:1295–1304. [PubMed: 18367470]

Pauler DK, Wakefield JC, Kass RE. Bayes factors and approximations for variance component models. Journal of the American Statistical Association. 1999; 94:1242–1253.

Rauh V, Andrews H, Garfinkel R. The contribution of maternal age to racial disparities in birthweight: A multilevel perspective. American Journal of Public Health. 2001; 91:1815–1824. [PubMed: 11684610]

Roberts E. Neighborhood social environments and the distribution of low birthweights in Chicago. American Journal of Public Health. 1997; 87:597–603. [PubMed: 9146438]

Saville BR, Herring AH. Testing random effects in the linear mixed model using approximate Bayes factors. Biometrics. 2009; 65:369–376. [PubMed: 18759835]

Savitz DA, Janevic TM, Engel SM, Kaufman JS, Herring AH. Ethnicity and gestational diabetes in New York City, 1995–2003. British Journal of Obstetrics & Gynecology. 2008; 115:969–978.

Schwarz G. Estimating the dimension of a model. The Annals of Statistics. 1978; 6:461–464.

Self SG, Liang KY. Asymptotic properties of maximum likelihood estimators and the likelihood ratio tests under nonstandard conditions. Journal of the American Statistical Association. 1987; 82:605–610.

Silvapulle MJ. Robust Wald-type tests of one-sided hypotheses in the linear model. Journal of the American Statistical Association. 1992; 87:156–161.

Smith AFM, Spiegelhalter DJ. Bayes factors and choice criteria for linear models. Journal of the Royal Statistical Society, Series B. 1980; 42:213–220.

Snijders, TAB.; Bosker, RJ. Multilevel analysis: An introduction to basic and advanced multilevel modeling. Thousand Oaks, CA: Sage; 1999.

Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, Series B. 2002; 64:583–640.

Spiegelhalter DJ, Smith AFM. Bayes factors for linear and log-linear models with vague prior information. Journal of the Royal Statistical Society, Series B. 1982; 44:377–387.

Spiegelhalter, DJ.; Thomas, A.; Best, NG.; Gilks, WR.; Lunn, D. WinBUGS User Manual, Version 1.4. 2003. Available at www.mrc-bsu.cam.ac.uk/bugs

Stram DO, Lee JW. Variance components testing in the longitudinal mixed effects model. Biometrics. 1994; 50:1171–1177. [PubMed: 7786999]

Sullivan LM, Dukes KA, Losina E. Tutorial in Biostatistics: An introduction to hierarchical linear modelling. Statistics in Medicine. 1999; 18:855–888. [PubMed: 10327531]

Tierney L, Kadane JB. Accurate approximations for posterior moments and marginal densities. Journal of the American Statistician. 1986; 81:82–86.

Verbeke G, Molenberghs G. The use of score tests for inference on variance components. Biometrics. 2003; 59:254–262. [PubMed: 12926710]

Wilcox AJ, Skjaerven R. Birth weight and perinatal mortality: the effect of gestational age. Am J Public Health. 1992; 82:378–382. [PubMed: 1536353]

Zhang D, Lin X. Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics. Lecture Notes in Statistics. 2008; 192:19–36.

Appendix

A.1 Approximating the marginal likelihood

A.1.1 Reparameterization

For computational convenience, we reparameterize the multilevel linear model given in (1). Let

$$Y_i = \boldsymbol{x}'_i\boldsymbol{\beta} + \text{ðİŠŸ}'_i\mathbf{b} + \varepsilon_i = \boldsymbol{x}'_i\boldsymbol{\beta} + \sum_{h=1}^{q}\text{ðİŠŸ}'_{ih}\mathbf{b}_h + \varepsilon_i \quad \text{(A.1)}$$

in which $\text{ðİŠŸ}_i = \left(\text{ðİŠŸ}'_{i1}, \ldots, \text{ðİŠŸ}'_{iq}\right)'$, $\boldsymbol{w}_{ih}$ is an $(r_h \times 1)$ vector of predictors with corresponding random effects $\boldsymbol{b}_h$, and $r_h = d_h c_h$ is the total number of random effects for factor h, with d_h the number of random effects for one observation for factor h, and c_h the total number of classifications for factor h. For example, in equation (2), $d_1 = 2$ and $c_1 = 62$ corresponding to a random intercept and slope (two random effects for observation i) for 62 classifications of ethnicity, and $d_2 = 1$ and $c_2 = 2{,}128$ corresponding to a random intercept (one random coefficient for observation i) for 2,128 classifications of census tracts. Additionally, $\boldsymbol{w}_{ih} = [\delta_i \otimes \boldsymbol{z}_{ih}]$, in which δ_i is a $(c_h \times 1)$ vector of indicator variables (equals 1 if yes, 0 if no) for group membership of observation i in each of the c_h classifications, and $\otimes$ denotes the left Kronecker product. The basic idea of this reparameterization is that all random effects in the model are stacked into one large vector $\boldsymbol{b}$. The design matrix $\boldsymbol{w}_i$ will contain mostly 0's, with non-zero elements corresponding to the appropriate random effects for observation i, and has dimension ($r \times 1$), with $r=\sum_{h=1}^{g} r_h$ the total number of random effects in the model. Also, $\mathbf{b}=\left(\mathbf{b}_1^{'},\ldots,\mathbf{b}_q^{'}\right)^{'}$ in which $\mathbf{b}_h=\left(\mathbf{b}_{h1}^{'},\ldots,\mathbf{b}_{hc_h}^{'}\right)^{'}$ is the vector of all random effects for factor h. We assume $\boldsymbol{b}_{hl} \sim N_{dh}(\mathbf{0}_{dh}, \Psi_h)$ (corresponding to factor h and classification l) independent of $\varepsilon_i \sim N(0, \sigma^2)$. Prior distributions are specified for β, Ψ_h, and σ^2 that are appropriate for the application.

A.1.2 Rescaling the random effects

Extending the work of Saville and Herring (2008), we scale the random effects to the residual variance such that $\tilde{\boldsymbol{b}}_{hl} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$. We then express the model as

$$Y_i=\boldsymbol{x}_i^{'}\boldsymbol{\beta}+\text{ðİŠŸ}_i^{'}\boldsymbol{\Phi}\boldsymbol{\Gamma}\tilde{\mathbf{b}}+\boldsymbol{\varepsilon}_i, \quad \text{(A.2)}$$

in which $\tilde{\boldsymbol{b}}$ is the vector of scaled random effects and $\boldsymbol{\Phi}=\text{diag}\left(\exp\left(\phi_1^{'*},\ldots\phi_q^{'*}\right)\right)$ with $\phi_h^*=\left(\mathbf{1}_{c_h}\otimes\phi_h\right)$, and $\varphi_h = (\varphi_{h1}, \ldots, \varphi_{hd_h})'$ are parameters that control the relative contribution of the random effects. The role of $\boldsymbol{\Gamma}=\text{blockdiag}\left(\boldsymbol{\Gamma}_1^*,\cdots,\boldsymbol{\Gamma}_q^*\right)$ with $\boldsymbol{\Gamma}_h^*=\left(\text{ðİŚř}_{c_h}\otimes\boldsymbol{\Gamma}_h\right)$, in which $\boldsymbol{\Gamma}_h$ is a lower triangular matrix with $\mathbf{1}_{dh}$ along the diagonal and lower off-diagonal elements γ_h, is to induce correlation between the random effects within factor h. There are a total of $g=\sum_{h=1}^{q} d_h$ parameters in the matrix Φ, or one parameter for each "random effect" in the model.

We can stack all observations into one response vector $\boldsymbol{Y}$ and write the model as

$$\boldsymbol{Y}=\boldsymbol{x}\boldsymbol{\beta}+\text{ðİŠŸ}\boldsymbol{\Phi}\boldsymbol{\Gamma}\tilde{\mathbf{b}}+\boldsymbol{\varepsilon}, \quad \text{(A.3)}$$

in which $\boldsymbol{Y} = (Y_i, \ldots, Y_m)'$, $\mathbf{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m)'$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)'$, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_m)'$. Let $\sigma^2 \sim \text{InvGam}(v, w)$. By integrating out $\tilde{\boldsymbol{b}}$ and σ^2 from the posterior distribution, the marginal posterior $p(\boldsymbol{Y}|\beta,\varphi,\gamma)$ can be shown to have the multivariate t-distribution given by

$$p\left(\boldsymbol{Y}|\beta,\phi,\gamma\right)=\Gamma\left(\frac{2\upsilon+p}{2}\right)\frac{(\pi 2\upsilon)^{-p/2}|\Sigma|^{-1/2}}{\Gamma(2\upsilon/2)}\left\{1+\frac{1}{2\upsilon}(\boldsymbol{Y}-\boldsymbol{x}\beta)^{'}\Sigma^{-1}(\boldsymbol{Y}-\boldsymbol{x}\beta)\right\}^{-\frac{2\upsilon+p}{2}}, \quad \text{(A.4)}$$

in which $\Gamma()$ denotes the gamma function and $\Sigma=(\boldsymbol{W}\Phi\Gamma\Gamma'\Phi'\boldsymbol{W}' + \boldsymbol{I}_{\mathrm{m}})$.

A.2 Computational considerations

A.2.1 Product of likelihoods

Although the theory previously outlined can accommodate any number of random effects for any number of nested, non-nested, or cross-nested factors, there are computational limitations that should be considered. If the number of factors is extremely large (unrealistic for most settings), the Laplace approximation may eventually break down because the multivariate t-distribution may not be of sufficiently small dimension. Aside from this issue, for studies with large sample size m, the covariance matrix Σ in equation (A.4) may be too large to handle computationally. For example, in applying model (2) to the complete 2003 NYC data (m =104,710), the covariance matrix Σ is (104,710 × 104,710). We note that this matrix has the potential to be extremely sparse, and even with very large m may be computationally feasible using sparse matrix computations. When the matrix is large and not sufficiently sparse, it may be advantageous to work with the product of independent likelihoods (conditional on the random effects) as opposed to the likelihood of the vector of response variables. To illustrate, the marginal distribution can be written as

$$\begin{aligned} &p(Y|\beta,\phi,\gamma)\\ &\quad=\int p(\boldsymbol{Y}|\beta,\phi,\gamma,\tilde{\mathbf{b}},\sigma^2)\pi(\tilde{\mathbf{b}})\pi(\sigma^2)d\tilde{\mathbf{b}}d\sigma^2\\ &\quad=\int\left[\prod_{i=1}^{m}p(Y_i|\beta,\phi,\gamma,\tilde{\mathbf{b}},\sigma^2)\right]\pi(\tilde{\mathbf{b}})\pi(\sigma^2)d\tilde{\mathbf{b}}d\sigma^2\\ &\quad=\frac{\Gamma\left(\frac{2\upsilon+m}{2}\right)\left|\text{ðİŚí}\right|^{-1/2}}{(\pi 2\upsilon)^{m/2}\Gamma\left(2\upsilon/2\right)}\left\{1+\frac{1}{2\upsilon}\left(f(\boldsymbol{Y})-\text{ðİŚł}'\text{ðİŚí}^{-1}\text{ðİŚł}\right)\right\}^{-\frac{2\upsilon+m}{2}} \end{aligned} \qquad (A.5)$$

with

$$\text{ðİŚí}=\left\{\text{ðİŚř}_r+\boldsymbol{\Gamma}'\boldsymbol{\Phi}'\left(\sum_{i=1}^{m}\text{ðİŠŸ}_i\text{ðİŠŸ}_i'\right)\boldsymbol{\Phi\Gamma}\right\},\text{ðİŚł}=\boldsymbol{\Gamma}'\boldsymbol{\Phi}'\left\{\sum_{i=1}^{m}\text{ðİŠŸ}_i\left(Y_i-\boldsymbol{x}_i'\boldsymbol{\beta}\right)\right\},\ \text{and}\ f(\boldsymbol{Y})=\sum_{i=1}^{m}\left(Y_i-\boldsymbol{x}_i'\boldsymbol{\beta}\right)^2$$

in which $\boldsymbol{I}_r$ denotes the identity matrix with dimension (r × r).

Using this approach, it should be computationally possible to approximate the marginal likelihood regardless of the size of m. The computation is limited, however, by the total number of random effects r. If r is very large, it may not be feasible to compute the inverse and determinant of the (r × r) matrix $\boldsymbol{A}$ (or may be very computationally expensive). For example, in applying (2) to the NYC data, r = 2, 252. Although it may be possible to compute the inverse and determinant of $\boldsymbol{A}$ in this example, computations are likely to be very slow. Hence, an alternative computational approach is to write the data likelihood as products of marginal likelihoods for lower-dimensional response vectors or scalars.

A.2.2 Alternative for non-nested models (cross-classified)

Consider the NYC data in which there are two non-nested (cross-classified) factors, ancestry and census tracts. We denote the factor with fewer groups as h = 1 (ancestry) and the factor with a larger number of groups as h = 2 (census tracts). We can write the marginal likelihood as

$$\begin{aligned} p(\boldsymbol{Y}|\beta,\phi,\gamma)&=\int p(\boldsymbol{Y}|\beta,\phi,\tilde{\mathbf{b}}_2,\tilde{\mathbf{b}}_1,\sigma^2)\pi(\tilde{\mathbf{b}}_2)\pi(\tilde{\mathbf{b}}_1)\pi(\sigma^2)d\tilde{\mathbf{b}}_2d\tilde{\mathbf{b}}_1d\sigma^2,\\ &=\int\left\{\prod_{k=1}^{C_2}p(\boldsymbol{Y}_k|\beta,\phi,\tilde{\mathbf{b}}_{2k},\tilde{\mathbf{b}}_1,\sigma^2)\right\}\pi(\tilde{\mathbf{b}}_2)\pi(\tilde{\mathbf{b}}_1)\pi(\sigma^2)d\tilde{\mathbf{b}}_2d\tilde{\mathbf{b}}_1d\sigma^2,\\ &=\int\left\{\prod_{k=1}^{C_2}\int p(\boldsymbol{Y}_k|\beta,\phi,\tilde{\mathbf{b}}_{2k},\tilde{\mathbf{b}}_1,\sigma^2)\pi(\tilde{\mathbf{b}}_{2k})d\tilde{\mathbf{b}}_{2k}\right\}\pi(\tilde{\mathbf{b}}_1)\pi(\sigma^2)d\tilde{\mathbf{b}}_1d\sigma^2\\ &=\int\left\{\prod_{k=1}^{C_2}\int\left[\prod_{i=1}^{m_k}p(\boldsymbol{Y}_{ki}|\beta,\phi,\tilde{\mathbf{b}}_{2k},\tilde{\mathbf{b}}_1,\sigma^2)\right]\pi(\tilde{\mathbf{b}}_{2k})d\tilde{\mathbf{b}}_{2k}\right\}\pi(\tilde{\mathbf{b}}_1)\pi(\sigma^2)d\tilde{\mathbf{b}}_1d\sigma^2, \end{aligned} \qquad (A.6)$$

in which c_2 is the number of groups in factor 2, $\boldsymbol{Y}_k$ is the vector of responses for group k in factor 2, $\tilde{\boldsymbol{b}}_2$ are the random effects for factor 2, $\tilde{\boldsymbol{b}}_{2k}$ are the random effects corresponding to group k in factor 2, $\tilde{\boldsymbol{b}}_1$ are the random effects for factor 1, m_k is the number of subjects in group k of factor 2 and Y_{ki} is the response of the ith subject in group k of factor 2. This approach allows one to integrate out the random effects for factor 2 in smaller dimensions, as $\tilde{\boldsymbol{b}}_{2k}$ is only a (d_2 × 1) vector. For model (2) applied to the NYC data, $\tilde{\boldsymbol{b}}_{2k}$ is a scalar (representing a random intercept for census tract k) and results in matrices with smaller dimensions than those obtained from (A.5). These derivations are specific to a model with two cross-classified factors, but the general strategy could be applied to models with a larger number of cross-classified factors.

A.2.3 Alternative for nested models

Consider a 3-level nested design, such as subjects nested within maternal ancestry nested within geographical region. In such cases one can use the nested structure for easier computation. Let $h = 1$ denote the maternal ancestry factor and $h = 2$ denote the geographical region factor. Then

$$
\begin{aligned}
p(\boldsymbol{Y}\,|\beta,\phi,\gamma) &= \int p(\boldsymbol{Y}\,|\,\beta,\phi,\tilde{\mathbf{b}}_2,\tilde{\mathbf{b}}_1,\sigma^2)\pi(\tilde{\mathbf{b}}_2)\pi(\tilde{\mathbf{b}}_1)\pi(\sigma^2)d\tilde{\mathbf{b}}_2 d\tilde{\mathbf{b}}_1 d\sigma^2 \qquad \text{(A.7)}\\
&= \int \left\{\prod_{k=1}^{c_2} p(\boldsymbol{Y}_k|\beta,\phi,\tilde{\mathbf{b}}_{2k},\tilde{\mathbf{b}}_{1k},\sigma^2)\right\}\pi(\tilde{\mathbf{b}}_2)\pi(\tilde{\mathbf{b}}_1)\pi(\sigma^2)d\tilde{\mathbf{b}}_2 d\tilde{\mathbf{b}}_1 d\sigma^2\\
&= \int \left\{\prod_{k=1}^{c_2} \int p(\boldsymbol{Y}_k|\beta,\phi,\tilde{\mathbf{b}}_{2k},\tilde{\mathbf{b}}_{1k},\sigma^2)\pi(\tilde{\mathbf{b}}_{2k})\pi(\tilde{\mathbf{b}}_{1k})d\tilde{\mathbf{b}}_{2k}d\tilde{\mathbf{b}}_{1k}\right\}\pi(\sigma^2)d\sigma^2\\
&= \int \left\{\prod_{k=1}^{c_2} \int \left[\prod_{j=1}^{c_{1k}} p(\boldsymbol{Y}_{kj}|\beta,\phi,\tilde{\mathbf{b}}_{2k},\tilde{\mathbf{b}}_{1kj},\sigma^2)\right]\pi(\tilde{\mathbf{b}}_{2k})\pi(\tilde{\mathbf{b}}_{1k})d\tilde{\mathbf{b}}_{2k}d\tilde{\mathbf{b}}_{1k}\right\}\pi(\sigma^2)d\sigma^2\\
&= \int \left\{\prod_{k=1}^{c_2} \int \left[\prod_{j=1}^{c_{1k}} \int p(\boldsymbol{Y}_{kj}|\beta,\phi,\tilde{\mathbf{b}}_{2k},\tilde{\mathbf{b}}_{1kj},\sigma^2)\pi(\tilde{\mathbf{b}}_{1kj})d\tilde{\mathbf{b}}_{1kj}\right]\pi(\tilde{\mathbf{b}}_{2k})d\tilde{\mathbf{b}}_{2k}\right\}\pi(\sigma^2)d\sigma^2\\
&= \int \left\{\prod_{k=1}^{c_2} \int \left[\prod_{j=1}^{c_{1k}} \int \left(\prod_{i=1}^{m_{kj}} p(Y_{kji}|\beta,\phi,\tilde{\mathbf{b}}_{2k},\tilde{\mathbf{b}}_{1kj},\sigma^2)\right)\pi(\tilde{\mathbf{b}}_{1kj})d\tilde{\mathbf{b}}_{1kj}\right]\pi(\tilde{\mathbf{b}}_{2k})d\tilde{\mathbf{b}}_{2k}\right\}\pi(\sigma^2)d\sigma^2,
\end{aligned}
$$

in which c_{1k} is the number of groups for factor 1 within group k of factor 2, m_{kj} is the number of subjects in group j of factor 1 within group k of factor 2, $\boldsymbol{Y}_{kj}$ is the response vector for subjects in group j of factor 1 within group k of factor 2, Y_{kji} is the response of subject i within group j of factor 1 within group k of factor 2, $\tilde{\boldsymbol{b}}_{1k}$ are the random effects for factor 1 within group k of factor 2, and $\tilde{\boldsymbol{b}}_{1kj}$ are the random effects corresponding to group j of factor 1 within group k of factor 2. This approach allows one to integrate out the random effects $\tilde{\boldsymbol{b}}_{1kj}$ and $\tilde{\boldsymbol{b}}_{2k}$ which have smaller dimensions equal to (d_1 × 1) and (d_2 × 1), respectively. For the NYC data with a random intercept for maternal ancestry and geographical region, $\tilde{\boldsymbol{b}}_{1kj}$ and $\tilde{\boldsymbol{b}}_{2k}$ are both scalars. These derivations are specific to a 3-level nested design, but the general strategy could be applied to models with larger numbers of nested factors, or even combinations of nested and cross-nested factors. For example, such strategies could be used on the NYC data, which has both nested and cross-classified random effects via factors for census tracts and maternal ancestry nested within geographical region.

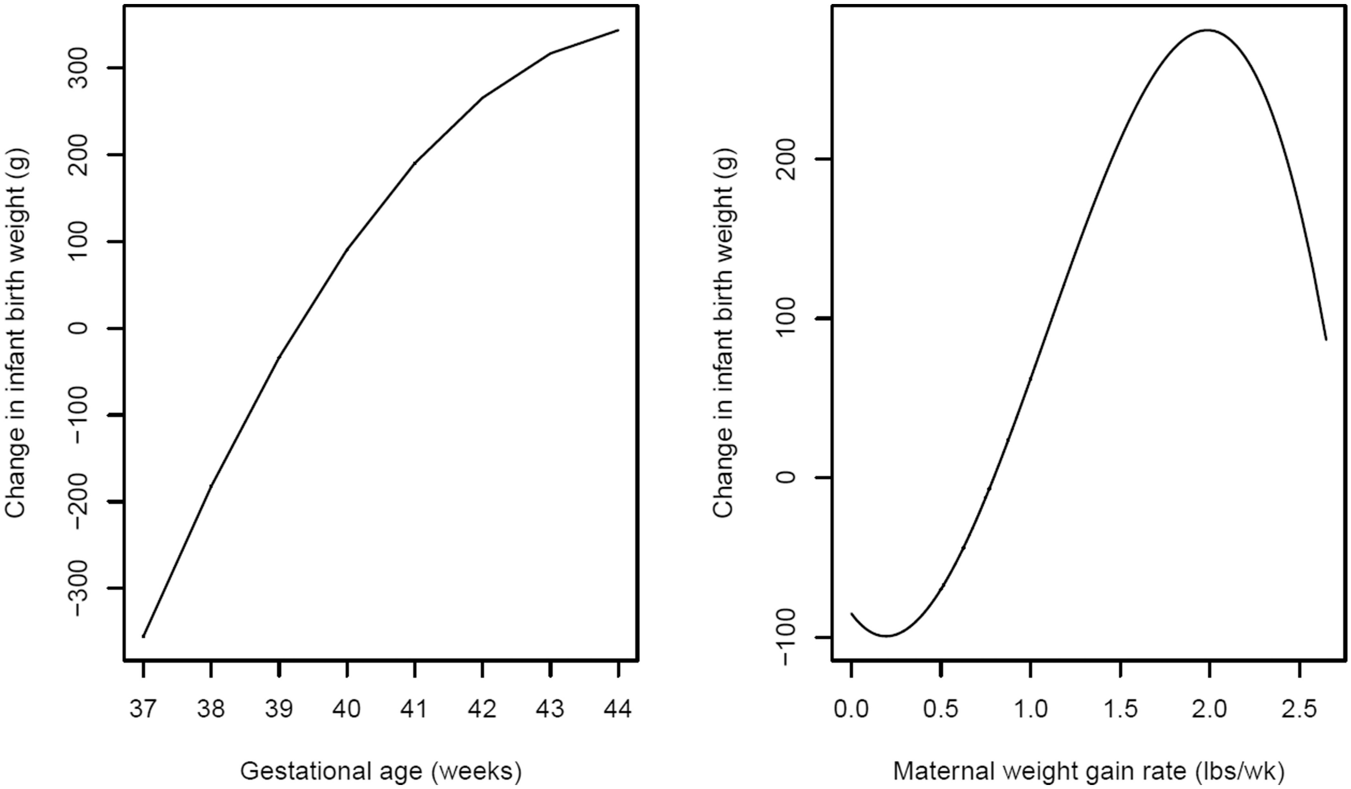

Figure 1.
Estimated change in infant birth weight by gestational age and maternal weight gain rate.

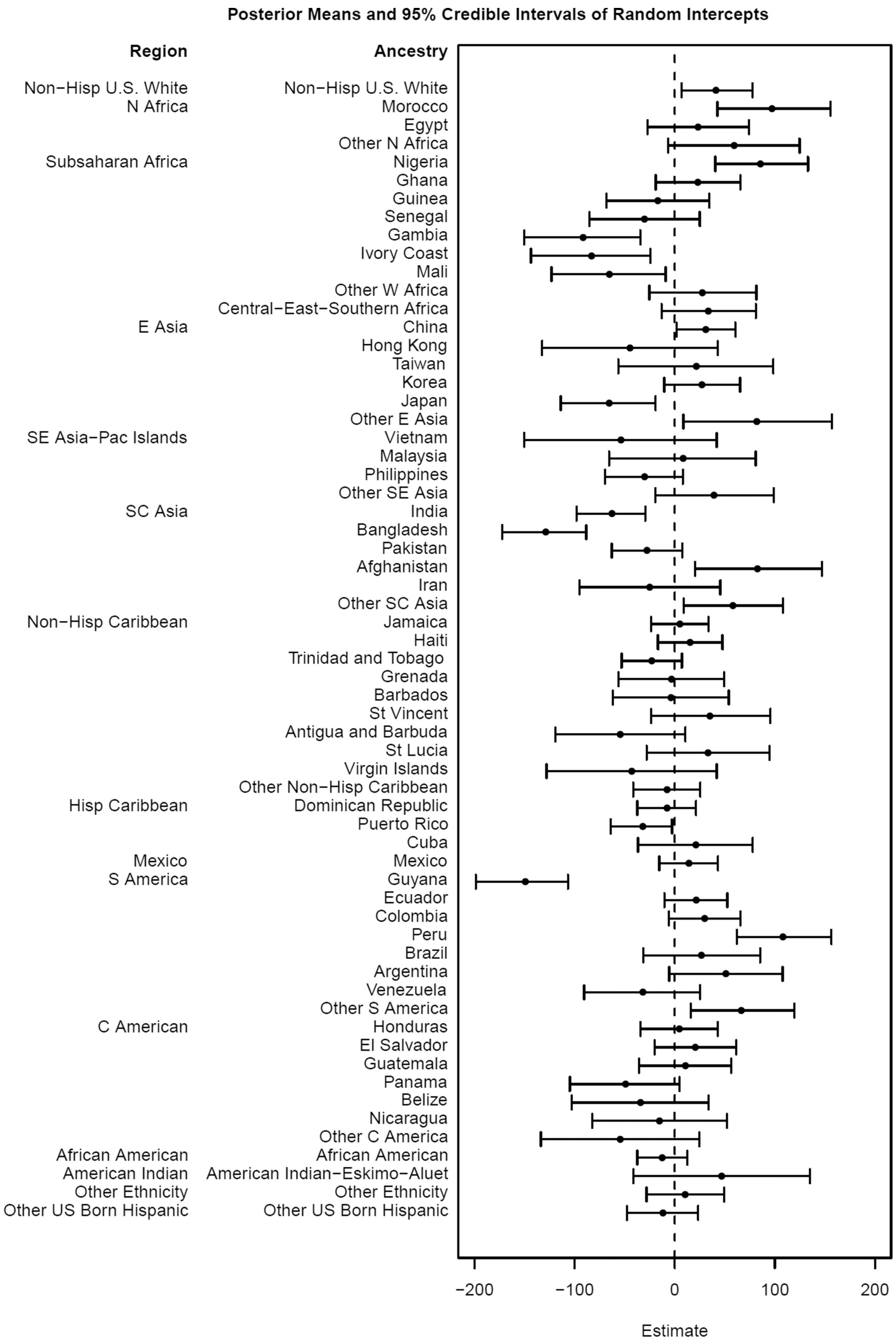

Figure 2.

Posterior means and 95% credible intervals of random intercepts.

Table 1

Random effects for models considered

Factor	Random coefficient	M_0	M_1	M_2	M_3	M_4	M_5	M_6	M_7
Ancestry (Country of origin)	Intercept		X		X			X	X
	Intercept: ancestry*race					X			
	Intercept: ancestry* nativity						X		
	Slope: maternal weight gain rate							X	
Census tract	Intercept			X	X	X	X	X	X
Geographical Region	Intercept								X

Table 2

Model posterior means and 95% credible intervals

Parameter	Posterior Mean	2.5%	97.5 %
β_0	3331	3295	3366
β_1 (Black)	−40	−75	−5
β_2 (Hisp)	13	−31	57
β_3 (Asian)	−57	−92	−21
β_4 (Other)	−4	−81	75
β_5 (Gest *)	296	290	301
β_6 (Gest *)^2	−63	−71	−54
β_7 (Pbirth)	105	99	111
β_8 (Female)	−120	−126	−115
β_9 (Smoke)	−165	−186	−143
β_{10} (Foreign)	11	3	19
β_{11} (NDP)	−16	−23	−9
β_{12} (Age 26–30)	52	45	60
β_{13} (Age 31–35)	72	64	80
β_{14} (Age 36–40)	84	74	93
β_{15} (Age > 40)	76	60	92
β_{16} (Wtgain *)	182	175	189
β_{17} (Wtgain *)^2	48	39	57
β_{18} (Wtgain *)^3	−35	−41	−29

*Estimates for a 2 sd increase

All estimates given in grams

Table 3

Frequency counts for ancestry by race

Region	Ancestry	White	Black	Hispanic	Asian	Other	Total
Non-Hisp U.S. White	Non-Hisp U.S. White	24749	0	0	0	0	24749
N Africa	Morocco	203	21	0	4	0	228
	Egypt	347	0	0	7	0	354
	Other N Africa	65	44	0	4	0	113
Subsaharan Africa	Nigeria	3	410	0	3	0	416
	Ghana	2	450	0	0	0	452
	Guinea	0	256	0	0	0	256
	Senegal	1	206	0	1	0	208
	Gambia	0	177	0	0	0	177
	Ivory Coast	0	161	0	0	0	161
	Mali	2	187	0	0	0	189
	Other W Africa	5	219	0	1	0	225
	Central-East-Southern Africa	38	283	0	4	0	325
E Asia	China	25	13	0	5506	0	5544
	Hong Kong	0	0	0	36	0	36
	Taiwan	1	0	0	65	0	66
	Korea	8	2	0	784	0	794
	Japan	9	3	0	352	0	364
	Other E Asia	19	3	0	51	0	73
SE Asia-Pac Islands	Vietnam	6	4	0	13	0	23
	Malaysia	0	0	0	78	2	80
	Philippines	22	9	0	646	0	677
	Other SE Asia	12	5	0	151	0	168
SC Asia	India	8	56	0	1374	7	1445
	Bangladesh	30	20	0	1190	0	1240
	Pakistan	40	10	0	960	0	1010
	Afghanistan	65	2	0	70	0	137
	Iran	96	0	0	2	0	98
	Other SC Asia	149	3	0	148	0	300

Region	Ancestry	White	Black	Hispanic	Asian	Other	Total
Non-Hisp Caribbean	Jamaica	5	2076	0	14	0	2095
	Haiti	6	1269	0	0	0	1275
	Trinidad and Tobago	12	1140	0	283	0	1435
	Grenada	0	220	0	3	0	223
	Barbados	0	175	0	0	0	175
	St Vincent	0	160	0	0	0	160
	Antigua and Barbuda	0	118	0	0	0	118
	St Lucia	1	142	0	1	0	144
	Virgin Islands	2	40	0	0	0	42
	Other Non-Hisp Caribbean	16	956	0	13	0	985
Hisp Caribbean	Dominican Republic	0	0	8426	0	1	8427
	Puerto Rico	0	0	7997	0	3	8000
	Cuba	0	0	192	0	0	192
Mexico	Mexico	0	0	6585	0	0	6585
S America	Guyana	0	0	1785	0	73	1858
	Ecuador	0	0	3053	0	0	3053
	Colombia	0	0	1239	0	1	1240
	Peru	0	0	521	0	0	521
	Brazil	0	0	178	0	0	178
	Argentina	0	0	198	0	0	198
	Venezuela	0	0	181	0	0	181
	Other S America	0	0	283	0	0	283
C American	Honduras	0	0	740	0	23	763
	El Salvador	0	0	640	0	0	640
	Guatemala	0	0	397	0	13	410
	Panama	0	0	226	0	0	226
	Belize	0	0	109	0	0	109
	Nicaragua	0	0	114	0	0	114
	Other C America	0	0	59	0	0	59
African American	African American	62	12323	0	12	6	12403
American Indian	American Indian-Eskimo-Aluet	5	18	0	0	12	35

Region	Ancestry	White	Black	Hispanic	Asian	Other	Total
Other Ethnicity	Other Ethnicity	59	344	0	137	7	547
Other US Born Hispanic	Other US Born Hispanic	0	0	1356	0	0	1356
Total		26073	21525	34279	11913	148	93938