Statistica Sinica **25** (2015), 1107-1132
doi:http://dx.doi.org/10.5705/ss.2013.137

FREQUENTIST INFERENCE ON RANDOM EFFECTS BASED ON SUMMARIZABILITY

Myunghee Cho Paik[1], Youngjo Lee[1] and Il Do Ha[2]

[1]*Seoul National University and* [2]*Pukyong National University*

Abstract: Although Bayesian methodologies have been successful in drawing inference about random effects, the frequentist literature has been limited. In this paper we consider inferences on random effects in hierarchical generalized linear models from a frequentist point of view using their summarizability. We show asymptotic distributional properties for the conditional and the marginal inference when the number of subunits is large. We conduct simulation studies when the number of subunits is small to moderate. A seizure study and an infertility study are used to illustrate the conditional and the marginal inference of random effects.

Key words and phrases: Conditional inference, confidence interval, marginal inference, nested generalized linear mixed models, random-effect inference.

1. Introduction

Random-effect models such as hierarchical generalized linear models (HGLMs), generalized linear mixed models, or multilevel generalized linear models are widely used for the analysis of clustered data. The random effects are unit-specific unobserved quantities arising independently from a common distribution representing heterogeneity among the independent units. Subunits within a unit are correlated by sharing the same random effects. Predicting random effects is of interest in applications including small area estimation, genetic evaluation of animals and quality management (Rao (2003)); Robinson (1991)). In the Bayesian approach, inference regarding random effects is based on the posterior distribution given the observed data (e.g., Carlin and Louis (2000)). In the empirical Bayesian approach, the empirical posterior distribution is used to predict random effects via posterior mean or mode evaluated at an estimated fixed parameter (e.g., Morris (1983); Maritz and Lwin (1989)).

Although Bayesian methodologies have been successful in drawing inference about random effects, the frequentist literature has been limited. The two intervals are fundamentally different in that Bayesian interval is a fixed interval around a random quantity and the frequentist's interval is a random interval around a fixed quantity. We call our approach frequentist since we state probabilistic statements about random intervals. In this paper we investigate drawing inference about random effects from a frequentist's stance. A hierarchical likelihood (h-likelihood) inference advocates estimating fixed parameters from the adjusted profile likelihood and estimating random effects as if they are fixed parameters and drawing inference using the variance obtained from the second derivative of negative log h-likelihood (Lee and Nelder (1996)). In the normal linear mixed models, treating random effects as if they are fixed provides the best linear unbiased predictor (BLUP) and the inverse of the Hessian matrix gives the variance of the residual of the mode from the random effect (Henderson (1975); Robinson (1991)). However, in models other than multivariate normal, inferential procedures lack rigorous theoretical justification. In this regard, Meng (2009, 2011) established Bartlett-like identities for h-likelihood: the score for the random effect has zero expectation and the variance of the score is the expected negative Hessian under easily verifiable conditions. However, the difference between the estimators of the random effects and the random effects themselves may not be quadratically summarizable, leaving the Bartlett identities meaningless. Meng (2009) also warned that even if summarizability is achieved, normality may not be claimed due to lack of independence among subunits in a unit.

There is some literature on drawing inference about random effects from a frequentist's stance. Ma, Krewski, and Burnett (2003) and Ma and Jørgensen (2007) presented early work on this, but their works focus on a special model. Inference on random effects differs from the usual inference on fixed effects. First, inference on random effects is regularized due to a distributional assumption on the random effect. As the number of subunits in the independent unit increases, the effect of this additional assumption diminishes, but plays an important role in the case of a small number of subunits. Therefore finite sample performance would not resemble that of fixed parameter inference, and summarizability can be achieved with a smaller number of subunits. Therefore empirical studies carry significance when the number of subunits is small, as in many applications. Second, unlike inference on fixed effects, two types of inference are possible, namely, conditional and marginal. Robinson (1991) made an important distinction between the realized value of random effect and the yet-to-be-realized value of random effect using an animal breeding example. Breeding value of an animal already born as the realized value of random effect can be *estimated* while breeding value of a mating between two potential parents can be *predicted*. However, the distinction was not made about interval estimation.

Our goal is to derive two types of inference on random effects. The aforementioned two types of random effects are first introduced by Robinson (1991), focusing on conceptual distinction and presenting the identical point estimation. We proceed from this and show separate inferential procedures, including interval estimation, about the two distinctive quantities, treating a realized value of random effect as fixed and yet-to-be-realized value as random. We propose to draw conditional inference given the realized value of random effect, which involves obtaining its estimator, and placing confidence intervals for the realized values via asymptotic distribution of the difference between its estimator and the fixed realized value. We propose to also draw marginal inference which involves obtaining its predictor and placing prediction intervals for the unrealized values via asymptotic distribution of the difference between the predictor and the random effect averaging over the joint distribution of outcomes and the random effect. We state a probabilistic statement about a random interval around random quantity, which is new in our knowledge. The distinction is not just conceptual but tangible reflected in setting the simulation studies. For the conditional inference, we generate single random effect for each individual throughout the replications. For the marginal inference we generate new random effect from replication to replication.

We showcase two examples in this paper to illustrate conditional and marginal inferences, respectively. The first example is a seizure study where repeated numbers of seizures are the outcome and the random effect is individual's seizure propensity. We can view that each individual possesses this quantity inherently and this quantity is already realized although we cannot observe, and draw conditional inference. The second example deals with repeated measures of prolactin levels of menstrual cycles from women. The repeated measures of prolactin levels are assumed to be a function of reproductive propensity. This reproductive propensity will be generated for every menstrual cycle. The reproductive propensity responsible for the observed prolactin values is realized but is not available for observation. However the reproductive propensity for a future menstrual cycle is yet to be realized. In this study the propensity of future cycle is more of interest than the current, since it characterizes pregnancy potential. We draw marginal inference for the second example.

Specifically we obtain an estimator or a predictor of random effect that maximizes h-likelihood and derive conditional and marginal inferences. For each case we derive asymptotic distribution of the difference between the esitmator/predictor and the random effect from a frequentist's stance when the number of subunits increases. Interestingly, we show that the asymptotic distribution of the difference between the predictor and the random effect is not normal even if the number of subunits increases. We evaluate its finite sample performance via simulation and show that behavior of coverage probabilities is markedly different between conditional and marginal inferences. Extra distributional assumption on random effects renders summarizability to be achieved with relatively small number of subunits and allows decent finite sample performance especially in marginal inference even when the number of subunit is small. All proofs are given in the Supplementary Material.

2. Settings

Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})^T$ be the response for the ith unit $(i = 1, \ldots, K)$ and ν_i be the corresponding unobserved random effect. We consider HGLMs and restrict our attention to nested hierarchical structure in that each outcome Y_{ij}, $j = 1, \ldots, n_i$, of $\mathbf{Y}_i$ is repeatedly measured within unit i. We assume that Y_{ij} is from an exponential family distribution given random effect ν_i, and follows a generalized linear model (GLM) with the density $f(Y_{ij}|\nu_i; \psi_{ij}, \phi)$, where

$$\log f(Y_{ij}|\nu_i; \psi_{ij}, \phi) = \frac{Y_{ij}\psi_{ij} - b(\psi_{ij})}{\phi} + d(Y_{ij}, \phi), \tag{2.1}$$

ψ_{ij} denotes the canonical parameter, and ϕ is the dispersion parameter. We write μ_{ij} for the conditional mean of Y_{ij} given ν_i, $\frac{d}{d\psi_{ij}} b(\psi_{ij}) = \mu_{ij}$, and $\eta_{ij} = q(\mu_{ij})$, with $q(\cdot)$ as the link function for the GLM relating μ_{ij} and η_{ij}. The linear predictor η_{ij} takes the form $\eta_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \nu_i$ with $\nu_i = \nu(u_i)$ for some strictly monotonic differentiable function of u_i, where $\mathbf{X}_{ij} = (1, x_{1ij}, \ldots, x_{pij})$ is a $1 \times (p+1)$ covariate vector corresponding to fixed effects $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$. We impose distributional assumption on u_i with density

$$f(u_i; \boldsymbol{\alpha}) = \exp\Big[\frac{\{a_1(\boldsymbol{\alpha})u_i - a_2(\boldsymbol{\alpha})\}}{\varphi} + c(u_i, \varphi)\Big].$$

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \phi, \boldsymbol{\alpha}^T)^T$, $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_K)^T$, $\mathbf{u} = (u_1, u_2, \ldots, u_K)^T$, and $\mathbf{Y} = (\mathbf{Y}_1^T, \mathbf{Y}_2^T, \ldots, \mathbf{Y}_K^T)^T$. The h-likelihood is defined as

$$H\{\boldsymbol{\theta}, \boldsymbol{\nu}(\mathbf{u}); \mathbf{Y}\} = \sum_{i=1}^{K} \frac{n_i}{K} h_i\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\}, \tag{2.2}$$

where

$$\begin{aligned} h_i\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\} &= \ell_{1i}\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\} + \ell_{2i}\{\boldsymbol{\theta}, \nu(u_i)\}, \qquad (2.3)\\ \ell_{1i}\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\} &= n_i^{-1} \sum_{j=1}^{n_i} \log f(Y_{ij}|\nu_i; \psi_{ij}, \phi)\\ &= n_i^{-1} \sum_{j=1}^{n_i} \Big[\frac{Y_{ij}\psi_{ij} - b(\psi_{ij})}{\phi} + d(Y_{ij}, \phi)\Big],\\ \ell_{2i}\{\boldsymbol{\theta}, \nu(u_i)\} &= n_i^{-1} \log f\{v(u_i); \boldsymbol{\alpha}\}\\ &= n_i^{-1}\Big\{\frac{a_1(\boldsymbol{\alpha})u_i - a_2(\boldsymbol{\alpha})}{\varphi} + c(u_i, \varphi) + \log(\frac{du_i}{d\nu_i})\Big\}. \end{aligned}$$

There is a subtle difference between the h-likelihood and the joint likelihood, or termed the extended likelihood by Bjørnstad (1996), in that the joint likelihood of $(\boldsymbol{\theta}, \boldsymbol{\nu}^*(\mathbf{u}))$, where $\boldsymbol{\nu}^*(\mathbf{u})$ represents a class of transformation of $\mathbf{u}$, is not invariant respect to the choice of function $\boldsymbol{\nu}^*(\mathbf{u})$ due to the Jacobian term for $\mathbf{u}$ (Lee and Nelder (2005)). The h-likelihood is the joint density of $\mathbf{Y}$ and the random effects $\boldsymbol{\nu} = \boldsymbol{\nu}(\mathbf{u})$, and therefore is a subclass of joint likelihood defined on a particular scale of $\mathbf{u}$, $\boldsymbol{\nu}(\mathbf{u})$, out of a class of scales $\boldsymbol{\nu}^*(\mathbf{u})$. The scale of the h-likelihood in which random effects enter linearly to the fixed effects is shown to provide the invariant inference about the random effects with respect to a linear transformation of the chosen scale (Lee and Nelder (2005, 2009)).

3. Inference about Random Effects

To focus on inference about $\boldsymbol{\nu}$, or $\mathbf{u}$, we first treat $\boldsymbol{\theta}$ as known. We consider each ν_i for $i = 1, \ldots, K$. Inference of u_i or ν_i only involves information from the i^{th} unit not from other units when fixed parameters are known. When fixed parameters are unknown, inference on u_i or v_i involves other units as shown in Section 5. The contribution of the log h-likelihood for the i^{th} unit becomes $h_i\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\}$ given in (2.3). Here ψ_{ij} is a function of $\nu(u_i)$ as well as $\boldsymbol{\theta}$. We denote by $a^{(p)}(u_i)$ the p^{th} derivative of $a(u_i)$ with respect to u_i. Let $\hat{u}_i$ be the solution of

$$h_i^{(1)}\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\} = l_{1i}^{(1)}\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\} + l_{2i}^{(1)}\{\boldsymbol{\theta}, \nu(u_i)\} = 0,$$

where

$$l_{1i}^{(1)}\{\boldsymbol{\theta}, \nu(u_i), \mathbf{Y}_i\} = n_i^{-1} \sum_{j=1}^{n_i} \frac{\partial \psi_{ij}}{\partial u_i} \frac{Y_{ij} - \frac{\partial}{\partial \psi_{ij}} b(\psi_{ij})}{\phi}$$

and

$$l_{2i}^{(1)}\{\boldsymbol{\theta}, \nu(u_i)\} = n_i^{-1} \Big[\frac{a_1(\boldsymbol{\alpha})}{\varphi} - c^{(1)}(u_i, \varphi) + \{\log(\frac{du_i}{d\nu_i})\}^{(1)} \Big].$$

Expanding $h_i^{(1)}\{\boldsymbol{\theta}, \nu(\hat{u}_i); \mathbf{Y}_i\} = 0$, we have

$$\begin{aligned} 0 = & h_i^{(1)}\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\} + (\hat{u}_i - u_i) h_i^{(2)}\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\} \\ & + \frac{1}{2}(\hat{u}_i - u_i)^2 h_i^{(3)}\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\} + o_p\{(\hat{u}_i - u_i)^2\}. \end{aligned}$$

As in inference for fixed parameters, we say that $\sqrt{n_i}(\hat{u}_i - u_i)$ is summarizable if expressible using the derivatives of an object function such as a h-likelihood and terms which vanish as n_i increases. There are two elements in each term in the expansion above, the polynomials in $(\hat{u}_i - u_i)$ and the derivatives of $h_i\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\}$. If u_i is a fixed parameter and $\hat{u}_i$ is a consistent estimator, $(\hat{u}_i - u_i)^2$ is small and we can summarize $(\hat{u}_i - u_i)$ using the first two terms. But u_i is not fixed, and the remainder terms are not guaranteed to vanish (Meng (2011)). In normal models, however, even if u_i is random, $h_i^{(3)}\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\} = 0$, and $(\hat{u}_i - u_i)$ is summarizable. This suggests that to achieve summarizability, we make either the polynomial terms in $(\hat{u}_i - u_i)$ or the derivative terms vanish.

3.1. Vanishing terms polynomial in $(\hat{u}_i - u_i)$

We can make the polynomials in $(\hat{u}_i - u_i)$ terms vanish by expanding around the realized value of u_i, say u_{0i}. In this case the conditional inference about $(\hat{u}_i - u_i)$ given $u_i = u_{0i}$ resembles the inference of fixed parameter, though not exactly since there is ℓ_{2i} term in (2.3). Assume that

$$I(\boldsymbol{\theta}, u_{0i}) = E\Big[- h_i^{(2)}\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\} | u_i = u_{0i}\Big]$$

exists and is positive.

Theorem 1. *Given u_{0i}, $\sqrt{n_i}(\hat{u}_i - u_{0i}) \to N(0, I(\boldsymbol{\theta}, u_{0i})^{-1})$ as $n_i \to \infty$.*

We can estimate $I(\boldsymbol{\theta}, u_{0i})^{-1}$ via $I(\boldsymbol{\theta}, \hat{u}_i)^{-1}$ or $-h_i^{(2)}\{\boldsymbol{\theta}, \nu(\hat{u}_i); \mathbf{Y}_i\}^{-1}$. Although this resembles the usual fixed parameter inference, a distinctive feature is contribution from l_{2i}. Usual frequentist's inference on u_i, as if u_i were a fixed parameter, utilizes l_{1i} alone. Our inference on random effects utilizes an additional assumption and l_{2i} plays an important role, especially when the number of subunits is small. We show in Section 6 that *average* coverage probabilities over K independent units can be close to the nominal value even when n_i is as small as 2. When the fixed parameters are unknown, the fact that random effects arise from a common distribution renders the parameters in l_{2i} estimable.

Remark 1. Conditionally, $\hat{u}_i$ is a biased estimator for u_{0i}, but the bias vanishes as n_i approaches infinity. The asymptotic conditional variance $I(\boldsymbol{\theta}, u_{0i})^{-1}$ can be improved by adding a higher-order term, $I(\boldsymbol{\theta}, u_{0i})^{-1}E\{l_{2i}^{(2)}(\boldsymbol{\theta}, \nu(u_i))|u_i = u_{0i}\}I(\boldsymbol{\theta}, u_{0i})^{-1}$ (see Supplementary Material for details). Since this term is negative, the asymptotic variance without the higher-order term is conservative.

We look next at the marginal inferences, when u_i is unrealized and the expectation is taken over the joint distribution of $\mathbf{Y}_i$ and u_i.

Theorem 2. *As $n_i \to \infty$, $\sqrt{n_i}(\hat{u}_i - u_i)$ has mean zero and variance $E_{u_i}[I(\boldsymbol{\theta}, u_i)^{-1}]$ and converges in distribution to a distribution whose moment generating function can be expressed as $E_{u_i}[\exp\{(1/2)t^2 I(\boldsymbol{\theta}, u_i)^{-1}\}]$, where $E_{u_i}[a(u_i)] = \int a(u_i) f(u_i; \boldsymbol{\alpha}) du_i$.*

Here the marginal variance of $\sqrt{n_i}(\hat{u}_i - u_i)$ is the expectation of the inverse of $I(\boldsymbol{\theta}, u_i)$, not the inverse of the expectation as in the case of the maximum likelihood estimator. Let $E_{\mathbf{Y}_i}\{a(\mathbf{Y}_i)\} = \int a(\mathbf{Y}_i) f(\mathbf{Y}_i;\boldsymbol{\theta}) d\mathbf{Y}_i$, where $f(\mathbf{Y}_i;\boldsymbol{\theta}) = \int \prod_{j=1}^{n_i}\{f(Y_{ij}|u_i;\psi_{ij}(\boldsymbol{\beta}),\phi)\} f(u_i;\boldsymbol{\alpha}) du_i$. Using the fact that

$$\lim_{n_i\to\infty}\left|E_{u_i}\left[I(\boldsymbol{\theta},u_i)^{-1}\right] - E_{\mathbf{Y}_i}\left[I(\boldsymbol{\theta},\hat{u}_i)^{-1}\right]\right| = 0,$$

we can use $E_{\mathbf{Y}_i}[I(\boldsymbol{\theta},\hat{u}_i)^{-1}]$, $I(\boldsymbol{\theta},\hat{u}_i)^{-1}$, or $-h_i^{(2)}\{\boldsymbol{\theta},\hat{u}_i;\mathbf{Y}_i\}^{-1}$ as an estimator of $E_{u_i}[I(\boldsymbol{\theta},u_i)^{-1}]$. We call $E_{\mathbf{Y}_i}[I(\boldsymbol{\theta},\hat{u}_i)^{-1}]$ and $I(\boldsymbol{\theta},\hat{u}_i)^{-1}$ the 'expected' version of the variance estimator while $-h_i^{(2)}\{\boldsymbol{\theta},\hat{u}_i;\mathbf{Y}_i\}^{-1}$ is the 'observed' version of the variance estimator, distinguishing whether the expectation is taken over $\mathbf{Y}_i$. Since the variance of the conditional mean of the score function is negligible, the estimator for the conditional variance from Theorem 1 can be used for the marginal variance from Theorem 2.

Remark 2. The variance of $(\hat{u}_i - u_i)$ vanishes as $n_i \to \infty$, since there is an accumulation of information about u_i in $\hat{u}_i$. As pointed out by Meng (2009), this information does not surface when predicting future independent observations.

Remark 3. Although the asymptotic marginal distribution of $\sqrt{n_i}(\hat{u}_i - u_i)$ is not normal, skewness is zero and kurtosis differs from that of normal distribution only by factor of $E_{u_i}\{I(\boldsymbol{\theta},u_i)^{-2}\}/\{E_{u_i} I(\boldsymbol{\theta},u_i)^{-1}\}^2$.

If $I(\boldsymbol{\theta},u_i)$ does not depend on u_i, $\sqrt{n_i}(\hat{u}_i - u_i)$ is marginally normal. In general, $I(\boldsymbol{\theta},u_i)$ depends on u_i. Specifically, when $\nu(u_i) = u_i$, $I(\boldsymbol{\theta},u_i) = n_i^{-1}\sum_{j=1}^{n_i}[\partial\mu_{ij}/\partial\eta_{ij}]^2 V_{ij}^{-1}$, where $V_{ij} \equiv Var(Y_{ij}|u_i)$. Since the term $[\partial\mu_{ij}/\partial\eta_{ij}]^2$ V_{ij}^{-1} is responsible for dependence on u_i, a link function that satisfies $[\partial\mu_{ij}/\partial\eta_{ij}]^2$ $= V_{ij}$, say a stabilizing link, can eliminate dependence on u_i.

Corollary. *Under stabilizing link functions, $\sqrt{n_i}(\hat{u}_i - u_i)$ is marginally normal with mean zero and variance $E_{u_i}\{I(\boldsymbol{\theta},u_i)^{-1}\} \equiv I(\boldsymbol{\theta})^{-1}$.*

Under the normal linear mixed models, the identity link is the stabilizing link, thus the corollary can be applied to the BLUP. The stabilizing function for Poisson is $\mu_{ij} = \eta_{ij}^2$: Poisson-normal with $E(Y_{ij}|u_i) = \mu_{ij} = \eta_{ij}^2$, $\eta_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + u_i$, where u_i is $N(0,\lambda)$ yields $I(\boldsymbol{\theta}) = n_i^{-1}(4n_i + 1/\lambda) = 4 + 1/(n_i\lambda)$. For binomial-normal models, a stabilizing link is $E(Y_{ij}|u_i) = \mu_{ij} = (1/2)\sin(\eta_{ij}) + 1/2$, $\eta_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + u_i$ where u_i is distributed as $N(0,\lambda)$ and $I(\boldsymbol{\theta}) = [1 + 1/(n_i\lambda)]$. For gamma-normal models with a stabilizing link, $E(Y_{ij}|u_i) = k\mu_{ij}$, $\eta_{ij} = \log\mu_{ij}$, $\eta_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + u_i$, and $u_i \sim N(0,\lambda)$, $I(\boldsymbol{\theta}) = n_i^{-1}(n_i k + 1/\lambda) = k + 1/(n_i\lambda)$. With stabilizing link functions for Poisson and binomial models, however, we need to restrict the range of these functions to allow one-to-one mappings for μ and η.

3.2. Vanishing derivative terms

In this section, we summarize $(\hat{u}_i - u_i)$ using different estimating functions and consider marginal inference. Although this type of summarizability is not generally applicable, it provides an alternative route to investigate properties of $(\hat{u}_i - u_i)$. Suppose that $h_i^{(1)}\{\boldsymbol{\theta}, \nu(u_i), \mathbf{Y}_i\}$ can be expressed as $h_i^{(1)}\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\} = g\{\boldsymbol{\theta}, \nu(u_i)\}U_i\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\}$, where $g\{\boldsymbol{\theta}, \nu(u_i)\} \neq 0$, such that the solution of $U_i\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\} = 0$ is the same as that of $h_i^{(1)}\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\} = 0$. Since solving for ν or u gives equivalent results, we use expressions $U_i(\boldsymbol{\theta}, u_i; \mathbf{Y}_i)$ and $g(\boldsymbol{\theta}, u_i)$ without loss of generality. We assume that the partition satisfies

(A1) $U_i^{(2)}(\boldsymbol{\theta}, u_i; \mathbf{Y}) = O_p(n_i^{-1/2})$ and $U_i^{(3)}(\boldsymbol{\theta}, u_i; \mathbf{Y}_i) = 0$, and

(A2) There exists κ_i with $\kappa_i \equiv \lim_{n_i \to \infty} U_i^{(1)}(\boldsymbol{\theta}, u_i; \mathbf{Y}_i)$.

Theorem 3. *If $\hat{u}_i$ is the solution of $U_i(\boldsymbol{\theta}, u_i; \mathbf{Y}_i) = 0$, then as $n_i \to \infty$, $\sqrt{n_i}(\hat{u}_i - u_i)$ converges in distribution to a distribution whose moment generating function is given by*

$$\begin{aligned} &E_{u_i} E_{\mathbf{Y}_i|u_i}\Big[\exp\{t\sqrt{n_i}\kappa^{-1}U_i(\boldsymbol{\theta}, u_i; \mathbf{Y}_i)\}|u_i\Big] \\ &\quad = E_{u_i}\Big[\exp\Big[\frac{1}{2}t^2 n_i \kappa_i^{-2} Var\{U_i(\boldsymbol{\theta}, u_i; \mathbf{Y}_i)\}\Big]\Big] + o(1), \end{aligned}$$

with mean 0 *and variance* $n_i\kappa_i^{-2}Var\{U_i(\boldsymbol{\theta}, u_i; \mathbf{Y}_i)\}$.

Remark 4. The bias of $\hat{u}_i$, $E(\hat{u}_i - u_i)$, is $\kappa_i^{-1}EU_i(\boldsymbol{\theta}, u_i; \mathbf{Y}_i)$ and is negligible for large n_i. In some cases, the Jacobian term due to the choice of scale $\nu(u_i)$ renders this bias term exactly zero, making $\hat{u}_i$ an unbiased predictor even in small sample cases. This fact also highlights the merit of summarizing $(\hat{u}_i - u_i)$ in terms of estimating function $U_i(\boldsymbol{\theta}, u_i; \mathbf{Y}_i)$.

Remark 5. Theorem 3 provides alternative asymptotic variance formula and alternative moment generating functions to those in Theorem 2. They can be shown to be asymptotically equivalent.

Example 1 (Normal-normal model)**.** Consider a normal-normal model with $E\{Y_{ij}|\nu(u_i)\} = \mu_{ij} + \nu(u_i)$, $\mu_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta}$, $\nu(u_i) = u_i$, and $f(u_i; D) = 1/\sqrt{2\pi D}$ $\exp(-u_i^2/2D)$. We have

$$h_i^{(1)}\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\} = n_i^{-1}\Big\{\sigma^{-2}\sum_{j=1}^{n_i}(Y_{ij} - \mu_{ij} - u_i) - \frac{u_i}{D}\Big\},$$

where $\sigma^2 = \text{Var}\{Y_{ij}|\nu(u_i)\}$. Since $h_i^{(1)}\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\}$ is already linear in u_i, we take $g(\boldsymbol{\theta}, u_i) = 1$, and the predictor for u_i is the solution of $h_i^{(1)}\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\} =$

$U_i(\boldsymbol{\theta}, u_i; \mathbf{Y}_i) = 0$. We can verify that

$$-U_i^{(1)}(\boldsymbol{\theta}, u_i; \mathbf{Y}_i) = -h_i^{(2)}\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\} = \sigma^{-2} + \frac{1}{n_i D},$$

and $U_i^{(2)}(\boldsymbol{\theta}, u_i; \mathbf{Y}_i) = 0$, satisfying (A1) and (A2). We can also verify that

$$Var(\hat{u}_i - u_i) = E\{-U_i^{(1)}(\boldsymbol{\theta}, u_i; \mathbf{Y}_i)\}^{-2} Var\{U_i(\boldsymbol{\theta}, u_i; \mathbf{Y}_i)\}.$$

Here conditional and marginal variances of $n_i Var(\hat{u} - u_i)$ are $I(\boldsymbol{\theta})^{-1} = \{\sigma^{-2} + 1/(n_i D)\}^{-1}$. We do not need to adjust the estimating function to make higher-order terms vanish and the identity link is the stabilizing link function.

Example 2 (Poisson-gamma model)**.** Consider a Poisson-gamma model with $E\{Y_{ij}|\nu(u_i)\} = \mu_{ij} = \exp(\eta_{ij})$, $\eta_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \nu(u_i)$, $\nu(u_i) = \log\ u_i$, $d\nu(u_i) = u_i^{-1} du_i$, and

$$f(u_i; \lambda, k) = \frac{u_i^{k-1}\exp(-u_i/\lambda)}{\Gamma(k)\lambda^k}.$$

In HGLMs, for identifiability we may put a constraint on the fixed β or random effects (e.g., Lee and Nelder (1996)). Here we set $E(u_i) = 1$, so $k = 1/\lambda$.

If $\mu_{ij} = \mu_{ij}^* u_i$, where $\mu_{ij}^* = \exp(\mathbf{X}_{ij}\boldsymbol{\beta})$, we have

$$h_i\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\} = n_i^{-1}\Big[\sum_{j=1}^{n_i}\Big(- u_i\mu_{ij}^* + Y_{ij}\log\ u_i + Y_{ij}\log\mu_{ij}^*\Big) + (k-1)\log\ u_i - \frac{u_i}{\lambda} + \log\ u_i\Big],$$

and $h_i^{(1)}\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\} = n_i^{-1}(-\mu_{i+}^* + Y_{i+}/u_i) + n_i^{-1}(k/u_i - k)$, where $\mu_{i+}^* = \sum_{j=1}^{n_i}\mu_{ij}^*$ and $Y_{i+} = \sum_{j=1}^{n_i} Y_{ij}$. Note that the solution of $h_i^{(1)}\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\} = 0$ is the same as that of

$$U_i(\boldsymbol{\theta}, u_i; \mathbf{Y}_i) = n_i^{-1}(Y_{i+} - u_i\mu_{i+}^*) + n_i^{-1}(k - ku_i).$$

Thus $h_i^{(1)}\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\} = g(\boldsymbol{\theta}, u_i)U_i(\boldsymbol{\theta}, u_i; \mathbf{Y}_i) = 0$, with $g(\boldsymbol{\theta}, u_i) = u_i^{-1}$, and it follows that $-U_i^{(1)}(\boldsymbol{\theta}, u_i; \mathbf{Y}_i) = n_i^{-1}(\mu_{i+}^* + k)$, free of u_i and $U_i^{(2)}\{\boldsymbol{\theta}, u_i; \mathbf{Y}_i\} = 0$. Since $E(u_i) = 1$, and $E(1/u_i) = k/(k-1)$, we can see that $E[h_i^{(1)}\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\}] = n_i^{-1}k\{E(1/u_i) - 1\} = n_i^{-1}k/(k-1)$. However, the expectation of $EU_i(\boldsymbol{\theta}, u_i; \mathbf{Y}_i)$ is zero. Also, the variance of $(\hat{u}_i - u_i)$ can be obtained by $E[-U_i^{(1)}\{\boldsymbol{\theta}, u_i; \mathbf{Y}_i\}]^{-2}$ $Var[U\{\boldsymbol{\theta}, u_i; \mathbf{Y}_i\}]$, where $E[-U_i^{(1)}\{\boldsymbol{\theta}, u_i; \mathbf{Y}_i\}] = n_i^{-1}(\mu_{i+}^* + k)$ and

$$\begin{aligned} n_i^2 Var[U\{\boldsymbol{\theta}, u_i; \mathbf{Y}_i\}] &= n_i^2 EVar[U\{\boldsymbol{\theta}, u_i; \mathbf{Y}_i)|u_i] + n_i^2 VarE[U\{\boldsymbol{\theta}, u_i; \mathbf{Y}_i\}|u_i] \\ &= Eu_i\mu_{i+} + Var(k - ku_i) = \mu_{i+}^* + k, \end{aligned}$$

yielding $E[-U_i^{(1)}\{\boldsymbol{\theta}, u_i; \mathbf{Y}_i\}]^{-2} Var[U\{\boldsymbol{\theta}, u_i; \mathbf{Y}_i\}] = 1/(\mu^*_{i+} + k)$. A direct evaluation of the variance results in the same variance as follows:

$$Var(\hat{u}_i - u_i) = \frac{Var(Y_{i+})}{(\mu^*_{i+} + \frac{1}{\lambda})^2} + Var(u_i) - 2Cov\Big(\hat{u}_i, u_i\Big)$$
$$= \frac{\lambda \mu^*_{i+}}{\mu^*_{i+} + k} + \lambda - 2\frac{\mu^*_{i+}\lambda}{\mu^*_{i+} + k} = \frac{1}{\mu^*_{i+} + k}.$$

On the other hand, there is no closed form for the variance from Theorem 2,

$$E[E[-h_i^{(2)}\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\}|u_i]^{-1}] = E\Big[\frac{1}{E[n_i^{-1}(Y_{i+} + k)/u_i^2]}\Big] = E\Big[n_i\Big\{\frac{u_i\mu^*_{i+} + k}{u_i^2}\Big\}^{-1}\Big].$$

Thus, in Poisson-gamma HGLMs the variance from Theorem 3 and the exact variance are the same while the one from Theorem 2 is hard to evaluate.

Example 3 (Binomial-beta model)**.** For binary outcome, the canonical link function gives $E\{Y_{ij}|\nu(u_i)\} = \mu_{ij} = \exp(\eta_{ij})/\{1 + \exp(\eta_{ij})\}$, so $\eta_{ij} = \log \mu_{ij}/(1 - \mu_{ij}) = \mathbf{X}_{ij}\boldsymbol{\beta} + \nu(u_i)$ and we set that $\nu(u_i) = \log u_i/(1 - u_i)$, which yields $d\nu(u_i) = \{u_i(1 - u_i)\}^{-1} du_i$ and $du_i/d\nu_i = u_i(1 - u_i)$. We assume that u_i has the beta density

$$f(u_i; \alpha_1, \alpha_2) = \frac{u_i^{\alpha_1 - 1}(1 - u_i)^{\alpha_2 - 1}}{B(\alpha_1, \alpha_2)},$$

where $B(\alpha_1, \alpha_2) = \Gamma(\alpha_1)\Gamma(\alpha_2)/\Gamma(\alpha_1 + \alpha_2)$. Here we set $\alpha_1 = \alpha_2 = \alpha$ to give $E(u_i) = 1/2$. A h-likelihood and its derivative have the forms:

$$h_i\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\} = n_i^{-1} \sum_{j=1}^{n_i} \Big[Y_{ij} \log \frac{\mu_{ij}}{1 - \mu_{ij}} + \log(1 - \mu_{ij})\Big]$$
$$+ n_i^{-1}\Big[\alpha \log u_i + \alpha \log(1 - u_i)\Big],$$
$$h_i^{(1)}\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\} = n_i^{-1} \sum_{j=1}^{n_i} \Big[\frac{Y_{ij} - \mu_{ij}}{u_i(1 - u_i)}\Big] + \frac{\alpha}{n_i u_i} - \frac{\alpha}{n_i(1 - u_i)}$$
$$= \frac{\{Y_{i+} - \mu_{i+} + \alpha\} - 2\alpha u_i}{n_i u_i(1 - u_i)}.$$

With $\mu^*_{ij} = \exp(\mathbf{X}_{ij}\boldsymbol{\beta})/\{1 + \exp(\mathbf{X}_{ij}\boldsymbol{\beta})\}$, we can see that $\mu_{ij} = u_i\mu^*_{ij}/[u_i\mu^*_{ij} + (1 - u_i)(1 - \mu^*_{ij})]$. Then, for $\mu_{ij} = \mu_i$ for all j, we can express $h_i^{(1)}\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\} = g(\boldsymbol{\theta}, u_i)U_i(\boldsymbol{\theta}, u_i; \mathbf{Y}_i)$ where

$$g(\boldsymbol{\theta}, u_i) = \Big[u_i(1 - u_i)\{u_i\mu^*_i + (1 - u_i)(1 - \mu^*_i)\}\Big]^{-1}.$$

Writing $\bar{Y}_i = Y_{i+}/n_i$,

$$U_i(\boldsymbol{\theta}, u_i; \mathbf{Y}_i) = \Big[\{u_i\mu_i^* + (1-u_i)(1-\mu_i^*)\}\bar{Y}_i - u_i\mu_i^*\Big] + n_i^{-1}\{\alpha(1-u_i) - \alpha u_i\}\{u_i\mu_i^* + (1-u_i)(1-\mu_i^*)\},$$

$$U_i^{(1)}(\boldsymbol{\theta}, u_i; \mathbf{Y}_i) = (2\mu_i^* - 1)\bar{Y}_i - \mu_i^* - n_i^{-1}(2\alpha)\{u_i\mu_i^* + (1-u_i)(1-\mu_i^*)\} + n_i^{-1}\alpha(2\mu_i^* - 1)(1-2u_i),$$

and

$$U_i^{(2)}(\boldsymbol{\theta}, u_i; \mathbf{Y}_i) = -4n_i^{-1}\alpha(2\mu_i^* - 1).$$

Therefore (A1) and (A2) are satisfied.

Example 4 (Gamma-inverse gamma model)**.** Let $E\{Y_{ij}|\nu(u_i)\} = k\mu_{ij}$, $\eta_{ij} = \log\mu_{ij}$, $\eta_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \nu(u_i)$, $\nu(u_i) = \log u_i$, and $d\nu(u_i) = (1/u_i)du_i$. Conditioning on $\nu(u_i)$, Y_{ij} has the density

$$f\{Y_{ij}|\nu(u_i);\boldsymbol{\beta},k\} = \frac{1}{\Gamma(k)}\Big(\frac{Y_{ij}}{\mu_{ij}}\Big)^k \exp(-\frac{Y_{ij}}{\mu_{ij}})\frac{1}{Y_{ij}}.$$

Suppose u_i has the inverse-gamma density

$$f(u_i;\alpha) = \frac{1}{\Gamma(\alpha+1)}\Big(\frac{\alpha}{u_i}\Big)^{\alpha+1}\exp(-\frac{\alpha}{u_i})\frac{1}{u_i}du_i,$$

with $E(u_i) = 1$. The contribution of the i^{th} unit to the h-likelihood is

$$h_i\{\boldsymbol{\theta},\nu(u_i);\mathbf{Y}_i\} = n_i^{-1}\Big[\sum_{j=1}^{n_i}[(k-1)\log Y_{ij} - \frac{Y_{ij}}{\{u_i\exp(\mathbf{X}_{ij}\boldsymbol{\beta})\}} - k\log\{u_i\exp(\mathbf{X}_{ij}\boldsymbol{\beta})\}] - (\alpha+1)\log u_i - \frac{\alpha}{u_i}\Big],$$

and $h_i^{(1)}\{\boldsymbol{\theta},\nu(u_i);\mathbf{Y}_i\} = n_i^{-1}[\sum_{j=1}^{n_i}\{Y_{ij}/(u_i^2\exp(\mathbf{X}_{ij}\boldsymbol{\beta})) - k/u_i\} - (\alpha+1)/u_i + \alpha/u_i^2]$. Setting $g(\boldsymbol{\theta},u_i) = 1/u_i^2$, we obtain

$$U_i(\boldsymbol{\theta},u_i;\mathbf{Y}_i) = n_i^{-1}\Big[\sum_{j=1}^{n_i}\Big\{\frac{Y_{ij}}{\exp(\mathbf{X}_{ij}\boldsymbol{\beta})} - ku_i\Big\} - (\alpha+1)u_i + \alpha\Big],$$

$U_i^{(1)}(\boldsymbol{\theta},u_i;\mathbf{Y}_i) = k-(\alpha+1)/n_i$, and $U_i^{(2)}(\boldsymbol{\theta},u_i;\mathbf{Y}_i) = 0$, satisfying (A1) and (A2). Furthermore, we can verify that $Var(\hat{u}_i - u_i)$ obtained by $\kappa_i^{-2}Var[U_i(\boldsymbol{\theta},u_i;\mathbf{Y}_i)]$, where $\kappa_i = n_i^{-1}(n_ik+\alpha+1)$, yields the directly evaluated exact variance.

4. Approximation to the Marginal Likelihood

In estimating the fixed parameter $\boldsymbol{\theta}$, we maximize

$$L(\boldsymbol{\theta};\mathbf{Y}) = \int\cdots\int e^{KH\{\boldsymbol{\theta},\boldsymbol{\nu}(\boldsymbol{u});\mathbf{Y}\}}d\nu_1(u_1)\cdots d\nu_K(u_K),$$

where $H\{\boldsymbol{\theta},\boldsymbol{\nu}(\boldsymbol{u})\}$ is given at (2.2). Standard likelihood inferential procedures can be applied to this marginal likelihood function, but a practical hurdle is computing the marginal likelihood. When the integral is hard to evaluate, one can use Laplace approximation. We show in this section how the partitions of the h-score shown in Section 3.2 can facilitate computing a higher-order correction term of Laplace approximation to the marginal likelihood.

Consider the accuracy of Laplace approximation from the i^{th} contribution which has the form

$$\int e^{n_i h_i\{\boldsymbol{\theta},\nu(u_i);\mathbf{Y}_i\}}d\nu(u_i) = e^{n_i h_i\{\boldsymbol{\theta},\nu(\hat{u}_i);\mathbf{Y}_i\}}\sqrt{2\pi}\sigma_i n_i^{-1/2}\Big[1-C_{n_i}\{\boldsymbol{\theta},\nu(\hat{u}_i)\}\Big]+O(n_i^{-2}),$$

where

$$\sigma_i^2 = -[h_i^{(2)}\{\boldsymbol{\theta},\nu_i(\hat{u}_i);\mathbf{Y}_i\}]^{-1},$$

$$C_{n_i}\{\boldsymbol{\theta},\nu(\hat{u}_i)\} = \frac{1}{8n_i}J_{1i}\{\boldsymbol{\theta},\nu(\hat{u}_i);\mathbf{Y}_i\} - \frac{5}{24n_i}J_{2i}\{\boldsymbol{\theta},\nu(\hat{u}_i);\mathbf{Y}_i\},$$

$$J_{1i}\{\boldsymbol{\theta},\nu(\hat{u}_i);\mathbf{Y}_i\} = -\frac{h_i^{(4)}\{\boldsymbol{\theta},\nu(\hat{u}_i);\mathbf{Y}_i\}}{[h_i^{(2)}\{\boldsymbol{\theta},\nu(\hat{u}_i);\mathbf{Y}_i\}]^2},$$

and

$$J_{2i}\{\boldsymbol{\theta},\nu(\hat{u}_i);\mathbf{Y}_i\} = -\frac{[h_i^{(3)}\{\boldsymbol{\theta},\nu(\hat{u}_i);\mathbf{Y}_i\}]^2}{[h_i^{(2)}\{\boldsymbol{\theta},\nu(\hat{u}_i);\mathbf{Y}_i\}]^3}.$$

An approximated marginal likelihood is

$$l(\boldsymbol{\theta};\mathbf{Y}) \approx \sum_{i=1}^{K}\Big[n_i h_i\{\boldsymbol{\theta},\nu(\hat{u}_i);\mathbf{Y}_i\} - \frac{1}{2}\log[-n_i h_i^{(2)}\{\boldsymbol{\theta},\nu(\hat{u}_i);\mathbf{Y}_i\}] + \log[1-C_{n_i}\{\boldsymbol{\theta},\nu(\hat{u}_i)\}]\Big]. \quad (4.1)$$

The first two terms are called the adjusted profile likelihood where the maximizer of u_i, $i = 1,\ldots,K$, is plugged in and the adjustment is made by subtracting the second term. The third term increases accuracy but can be computationally demanding. We illustrate how partitioning $h_i^{(1)}(.)$ into $g(.)$ and $U_i(.)$ simplifies computation of the higher-order correction term. Using the fact that $U_i(\boldsymbol{\theta},\hat{u}_i;\mathbf{Y}_i) = 0$, we have $h_i^{(2)}\{\boldsymbol{\theta},\nu(\hat{u}_i);\mathbf{Y}_i\} = g(\boldsymbol{\theta},\hat{u}_i)U_i^{(1)}(\boldsymbol{\theta},\hat{u}_i;\mathbf{Y}_i)$, $h_i^{(3)}\{\boldsymbol{\theta},\nu(\hat{u}_i);\mathbf{Y}_i\} = 2g^{(1)}(\boldsymbol{\theta},\hat{u}_i)U_i^{(1)}(\boldsymbol{\theta},\hat{u}_i;\mathbf{Y}_i) + g(\boldsymbol{\theta},\hat{u}_i)U_i^{(2)}(\boldsymbol{\theta},\hat{u}_i;\mathbf{Y}_i)$, and

$$h_i^{(4)}\{\boldsymbol{\theta},\nu(\hat{u}_i);\mathbf{Y}_i\} = 3g^{(2)}(\boldsymbol{\theta},\hat{u}_i)U_i^{(1)}(\boldsymbol{\theta},\hat{u}_i;\mathbf{Y}_i) + 3g^{(1)}(\boldsymbol{\theta},\hat{u}_i)U_i^{(2)}(\boldsymbol{\theta},\hat{u}_i;\mathbf{Y}_i).$$

When $U_i^{(2)}(\boldsymbol{\theta}, u_i; \mathbf{Y}_i) = 0$, further simplification is

$$J_{1i}\{\boldsymbol{\theta}, \nu(\hat{u}_i); \mathbf{Y}_i\} = -\frac{h_i^{(4)}\{\boldsymbol{\theta}, \nu(\hat{u}_i); \mathbf{Y}_i\}}{[h_i^{(2)}\{\boldsymbol{\theta}, \nu(\hat{u}_i); \mathbf{Y}_i\}]^2} = -\frac{3g^{(2)}(\boldsymbol{\theta}, \hat{u}_i)}{g^2(\boldsymbol{\theta}, \hat{u}_i)} U_i^{(1)}(\boldsymbol{\theta}, \hat{u}_i; \mathbf{Y}_i)^{-1},$$

$$J_{2i}\{\boldsymbol{\theta}, \nu(\hat{u}_i); \mathbf{Y}_i\} = -\frac{[h_i^{(3)}\{\boldsymbol{\theta}, \nu(\hat{u}_i); \mathbf{Y}_i\}]^2}{[h_i^{(2)}\{\boldsymbol{\theta}, \nu(\hat{u}_i); \mathbf{Y}_i\}]^3} = -\frac{4\{g^{(1)}(\boldsymbol{\theta}, \hat{u}_i)\}^2}{g^3(\boldsymbol{\theta}, \hat{u}_i)} U_i^{(1)}(\boldsymbol{\theta}, \hat{u}_i; \mathbf{Y}_i)^{-1}.$$

For $g(\boldsymbol{\theta}, u_i) = 1/u_i$, $C_{n_i}\{\boldsymbol{\theta}, \nu(\hat{u}_i)\} = \{12 n_i \hat{u}_i U_i^{(1)}(\boldsymbol{\theta}, \hat{u}_i; \mathbf{Y}_i)\}^{-1}$ and for $g(\boldsymbol{\theta}, u_i) = 1/u_i^2$, $C_{n_i}\{\boldsymbol{\theta}, \nu(\hat{u}_i)\} = 13/\{12 n_i U_i^{(1)}(\boldsymbol{\theta}, \hat{u}_i; \mathbf{Y}_i)\}$. For Poisson-gamma and gamma-inverse gamma models, $C_{n_i}\{\boldsymbol{\theta}, \nu(\hat{u}_i)\} = 1/[12 n_i \hat{u}_i\{\mu_{i+}/n_i - 1/(n_i\lambda)\}]$ and $C_{n_i}\{\boldsymbol{\theta}, \nu(\hat{u}_i)\} = 13/\{12 n_i (k - \alpha/n_i)\}$, respectively.

As the E- and M- steps resonate hot-deck style of 'fill in' then 'estimate', Laplace approximation also has intuitive appeal with E-step-like plugging-in mostly likely values and M-step-like estimating with penalty to avoid overfitting. Furthermore plugging-in missing data themselves, not a function of missing data as in the EM, offers simplicity in implementation. As for the accuracy of approximation the requirement is less stringent than in Bayesian application, since the approximation should be accurate enough to hold the argmax of the function, not the function itself.

5. Inference about Random Effects with Unknown Fixed Parameters

In this section we consider the case where $\boldsymbol{\theta}$ is unknown. Our Theorem 4 covers conditional inference on realized value u_{0i} and Theorem 5 covers marginal inference for random u_i.

Let $(\hat{\boldsymbol{\theta}}, \hat{\mathbf{u}})$ be the solution of

$$\begin{pmatrix} \frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}; \mathbf{Y}) \\ \mathbf{W}\{\boldsymbol{\theta}, \nu(\mathbf{u}); \mathbf{Y}\} \end{pmatrix} = 0,$$

where $\mathbf{W}\{\boldsymbol{\theta}, \nu(\mathbf{u}); \mathbf{Y}\} = \{W_1(\boldsymbol{\theta}, u_1; \mathbf{Y}_1), \ldots, W_K(\boldsymbol{\theta}, u_K; \mathbf{Y}_K)\}^T$ and $W_i(\boldsymbol{\theta}, u_i; \mathbf{Y}_i)$ is either $h_i^{(1)}\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\}$ or $U_i(\boldsymbol{\theta}, u_1; \mathbf{Y}_1)$. Similarly, $\mathbf{W}\{\boldsymbol{\theta}, \nu(\mathbf{u}); \mathbf{Y}\}$ is either

$$\mathbf{h}^{(1)}\{\boldsymbol{\theta}, \nu(\mathbf{u}); \mathbf{Y}\} = [h_1^{(1)}\{\boldsymbol{\theta}, \nu(u_1); \mathbf{Y}_1\}, \ldots, h_K^{(1)}\{\boldsymbol{\theta}, \nu(u_K); \mathbf{Y}_K\}]$$

or

$$\mathbf{U}(\boldsymbol{\theta}, \mathbf{u}; \mathbf{Y}) = \{U_1(\boldsymbol{\theta}, u_1; \mathbf{Y}_1), \ldots, U_K(\boldsymbol{\theta}, u_K; \mathbf{Y}_K)\}^T.$$

Consider an added condition.

(A3) $\|\frac{\partial}{\partial \boldsymbol{\theta}} W_i^{(1)}\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\}\| = O_p(1)$ for all i.

Suppose data arise as described in Section 2, (A1) and (A2) are satisfied if $W_i(\boldsymbol{\theta}, u_i; \mathbf{Y}_i) = U_i(\boldsymbol{\theta}, u_i; \mathbf{Y}_i)$, and (A3) is satisfied. Then we have results for conditional and marginal inference.

Theorem 4. *As* $n_i \to \infty$ *and* $n_i/K \to O(1)$, $\sqrt{n_i}(\hat{u}_i - u_{0i})$ *converges in distribution to normal with mean* 0 *and variance*

$$I(\boldsymbol{\theta}, u_{0i})^{-1} \quad + n_i I(\boldsymbol{\theta}, u_{0i})^{-1} B_{21i} A_{11}^{-1} Var\Big[\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}; \mathbf{Y}) | u_i = u_{0i}\Big] A_{11}^{-1} B_{21i}^T I(\boldsymbol{\theta}, u_{0i})^{-1}$$
$$-2 n_i I(\boldsymbol{\theta}, u_{0i})^{-1} B_{21i}^T A_{11}^{-1} Cov\Big[\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}; \mathbf{Y}), h_i^{(1)}\{\boldsymbol{\theta}, \nu(u_i)\mathbf{Y}_i\} | u_i = u_{0i}\Big],$$

where

$$A_{11} = E\left\{-\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} l(\boldsymbol{\theta}; \mathbf{Y})\right\}$$

and

$$B_{21i} = E\left\{\Big[\frac{\partial}{\partial \boldsymbol{\theta}^T} h_i^{(1)}\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}_i\}\Big] | u_i = u_{0i}\right\}.$$

Theorem 5. *As* $n_i \to \infty$, $\sqrt{n_i}(\hat{u}_i - u_i)$ *converges in distribution to a distribution with moment generating function*

$$E_{u_i}\left[\exp\Big[\frac{1}{2} t^2 n_i \kappa_i^{-2} Var\{U_i^*(\boldsymbol{\theta}, u_i; \mathbf{Y}_i)\}\Big]\right] + o(1),$$

where $U_i^*(\boldsymbol{\theta}, u_i; \mathbf{Y}_i) = U_i(\boldsymbol{\theta}, u_i; \mathbf{Y}_i) + [\frac{\partial}{\partial \boldsymbol{\theta}^T} U_i(\boldsymbol{\theta}, u_i; \mathbf{Y}_i)] A_{11}^{-1} [\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}; \mathbf{Y})]$.

6. Simulation Studies

We present results from simulation studies that evaluated finite sample performance and approximation to normality via the confidence intervals of realized but unobserved random effects and prediction intervals of unrealized random effects. Throughout, the number of replications was 500. In the first section, we consider the case when the fixed parameters were known and in the second section, the case when the fixed parameters were estimated. For conditional inference, $u_1, u_2 \ldots, u_K$ were generated and kept the same for all replications. For marginal inference, new random effects were generated for every replication. To place confidence intervals for realized random effects in the conditional inference, variance formulas in Theorem 1 or 4 were used depending whether fixed parameters are known or estimated. To place prediction intervals for unrealized random effects in the marginal inference, formulas in Theorem 2 or 5 were used. For every model, except for a Bernoulli-normal model, we used a binary covariate, x_{ij}, which is generated from a Bernoulli distribution with success probability 0.5. We considered the sample sizes $N = \sum_{i=1}^K n_i$ with $N = 100$, 200, 250, 500, and 1,000, and $(K, n_i) = (50,2), (100,2), (50,5), (100,5), (50,10)$ and $(50,20)$. The variance of random effects, λ in all presented models was $\lambda = 0.5$. All computations were conducted using SAS/IML.

Since we place K confidence intervals, we present coverage probabilities in two forms: in tables we report average coverage probabilities over K independent units; in figures we display individual coverage probabilities of K independent units obtained over 500 replications using Box-plots.

6.1. Coverage probability of random effects with known fixed parameters

In this section the confidence or prediction intervals of random effects in models with stabilizing link functions are examined. Here, under stabilizing link functions, $(\hat{u}_i - u_i)$ is normal when the number of subunits is large. We consider the Poisson-normal, Bernoulli-normal, and gamma-normal models where the random effect u_i is $N(0, \lambda)$.

For the Poisson-normal model, we set the conditional mean $E\{Y_{ij}|\nu(u_i)\} = \mu_{ij} = (\beta_0 + \beta_1 x_{ij} + u_i)^2$ with $\beta_0 = 2$, and $\beta_1 = 1$. With stabilizing link functions, the conditional and the marginal variance have the same asymptotic form. For the variance, we used two variance estimators, the 'observed' version, $-h_i^{(2)}\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}\} = n_i^{-1}\{2\sum_{j=1}^{n_i}(1+Y_{ij}/\mu_{ij})+1/\lambda\}$ and the 'expected' version, $I(\boldsymbol{\theta}) = E\Big[-h_i^{(2)}\{\boldsymbol{\theta}, \nu(u_i); \mathbf{Y}\}|u = u_{0i}\Big] = 4 + 1/(n_i\lambda)$. For a Bernoulli-normal model we set a stabilizing link $E\{Y_{ij}|\nu(u_i)\} = \mu_{ij} = (1/2)\sin(\eta_{ij}) + 1/2$ and $\eta_{ij} = \beta_0 + \beta_1 x_{ij} + u_i$, with $\beta_0 = -0.5$ and $\beta_1 = 1$. We generated x_{ij}'s from a uniform distribution on the interval [0,1]. We specified a gamma-normal model with $E\{Y_{ij}|\nu(u_i)\} = k\mu_{ij}$ and $\eta_{ij} = \log\mu_{ij} = \beta_0 + \beta_1 x_{ij} + u_i$, with $\beta_0 = \beta_1 = 1$. The shape parameter was $k = 2$.

For the specified models the means of K individual coverage probabilities from independent units for various combinations of (K, n_i) are shown in Table 1. In all models the average coverage probabilities using the two variance estimates maintained close to the nominal 95% level for both conditional and marginal inferences. In particular, they performed well even when $n_i \equiv n$ was as small as 2 with $K = 50$. However, the standard deviation of the empirical coverage probabilities was much smaller in the marginal inference than in the conditional difference. This point is visible in the figures.

Figures 1(a), 1(c), and 1(e) show Box-plots of K coverage probabilities of confidence intervals for the three models, and Figures 1(b), 1(d) and 1(f) show Box-plots of K coverage probabilities of prediction intervals for individual independent units. While most of individual coverage probabilities of confidence intervals in Figures 1(a), 1(c) and 1(e) show over 90% coverage, there are several individual coverage probabilities outside the lower inner fence. The width of fences of box plots of the coverage probabilities becomes narrower as the number of subunits in the unit (n_i) increases. Figure 3 shows bias and the coverage probabilities of ordered u_{0i}'s when $n_i = 2, 20$. Since the bias is $E(\hat{u}_i - u_{0i}) = \{(1-u_{0i})k\}/(\mu_{i+} + k)$ in the Poisson-gamma model, we find large bias and poor coverage probability associated with the values for extreme u_{0i}'s: the bias of $\hat{u}_{0i}$ is positive if $u_{0i} < 1$, negative for $u_{0i} > 1$, which implies that $\hat{u}_{0i}$ is conservative and corresponding confidence interval may miss u_{0i} by tilting toward the marginal mean. Figure 3

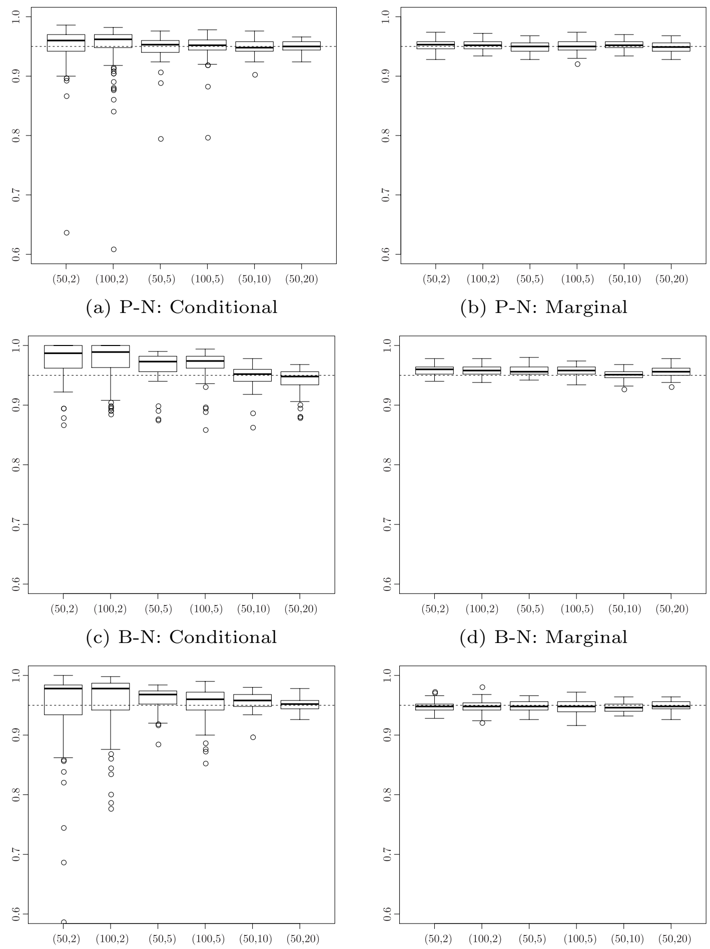
Table 1. Average coverage probabilities of the nominal 95% intervals over K independent units in Poisson-normal (P-N), Bernoulli-normal (B-N) and gamma-normal (G-N) models for known $\boldsymbol{\theta}$ with 500 replications.

				P-N	B-N	G-N
N	(K, n)	Type	Method	Mean (SD)	Mean (SD)	Mean (SD)
100	(50,2)	Conditional	expected	0.942 (0.047)	0.959 (0.047)	0.935 (0.076)
			observed	0.943 (0.053)	0.973 (0.035)	0.934 (0.102)
		Marginal	expected	0.945 (0.011)	0.943 (0.012)	0.938 (0.011)
			observed	0.949 (0.011)	0.959 (0.009)	0.945 (0.012)
200	(100,2)	Conditional	expected	0.948 (0.039)	0.965 (0.041)	0.950 (0.046)
			observed	0.949 (0.044)	0.976 (0.032)	0.959 (0.047)
		Marginal	expected	0.948 (0.009)	0.942 (0.011)	0.938 (0.010)
			observed	0.946 (0.009)	0.958 (0.008)	0.946 (0.010)
250	(50,5)	Conditional	expected	0.947 (0.025)	0.952 (0.030)	0.955 (0.022)
			observed	0.947 (0.028)	0.964 (0.027)	0.961 (0.021)
		Marginal	expected	0.948 (0.011)	0.937 (0.011)	0.944 (0.010)
			observed	0.949 (0.010)	0.957 (0.009)	0.946 (0.009)
500	(100,5)	Conditional	expected	0.949 (0.020)	0.956 (0.024)	0.949 (0.023)
			observed	0.949 (0.022)	0.968 (0.023)	0.953 (0.027)
		Marginal	expected	0.948 (0.010)	0.937 (0.010)	0.943 (0.012)
			observed	0.950 (0.010)	0.958 (0.008)	0.944 (0.011)
500	(50,10)	Conditional	expected	0.949 (0.013)	0.948 (0.020)	0.954 (0.014)
			observed	0.949 (0.014)	0.963 (0.011)	0.957 (0.014)
		Marginal	expected	0.947 (0.010)	0.933 (0.010)	0.947 (0.009)
			observed	0.948 (0.010)	0.951 (0.009)	0.947 (0.010)
1,000	(50,20)	Conditional	expected	0.950 (0.009)	0.943 (0.023)	0.951 (0.012)
			observed	0.950 (0.009)	0.955 (0.012)	0.951 (0.013)
		Marginal	expected	0.949 (0.009)	0.943 (0.010)	0.948 (0.009)
			observed	0.949 (0.010)	0.956 (0.009)	0.949 (0.010)

shows that the coverage probabilities are overstated for $u_{0i} < 1$ and understated for $u_{0i} > 1$. This trend becomes negligible when $n_i = 20$. In contrast to the conditional case, individual coverage probabilities are tight around the nominal value in the marginal case, as shown in Figures 1 (b), 1(d) and 1(f). Bias is small in the marginal case since it is averaged over the distribution of the random effect. Also the coverage indicator whether the interval includes the random effect is averaged over all possible random effects.

6.2. Coverage probability of random effects with estimated fixed parameters

In this section, we consider that the fixed parameters are estimated using two models. First we set a Poisson-gamma model with the conditional mean $E(Y_{ij}|u_i) = \mu_{ij} = \exp(\beta_0 + \beta_1 x_{ij} + \log u_i)$, with $\beta_0 = \beta_1 = 1$ and u_i distributed



(a) P-N: Conditional (b) P-N: Marginal

(c) B-N: Conditional (d) B-N: Marginal

(e) G-N: Conditional (f) G-N: Marginal

Figure 1. Individual coverage probabilities (y-label) of the nominal 95% (*dotted line*) confidence and prediction intervals from Poisson-normal (P-N), Bernoulli-normal (B-N), and gamma-normal (G-N) models for known $\boldsymbol{\theta}$. Here (50,2), (100,2), (50,5), (100,5), (50,10) and (50,20) indicate (K, n), where K is the number of clusters and n is the cluster size.

as gamma with mean 1 and variance $\lambda = 0.5$. We also consider a Poisson-lognormal model with conditional mean $\mu_{ij} = \exp(\eta_{ij})$ with $\eta_{ij} = \beta_0 + \beta_1 x_{ij} + u_i$, and $u_i \sim N(0, \lambda)$. We first assume that the fixed parameters are known, and then the fixed parameters are estimated by maximizing the second-order Laplace approximation given at (4.1).

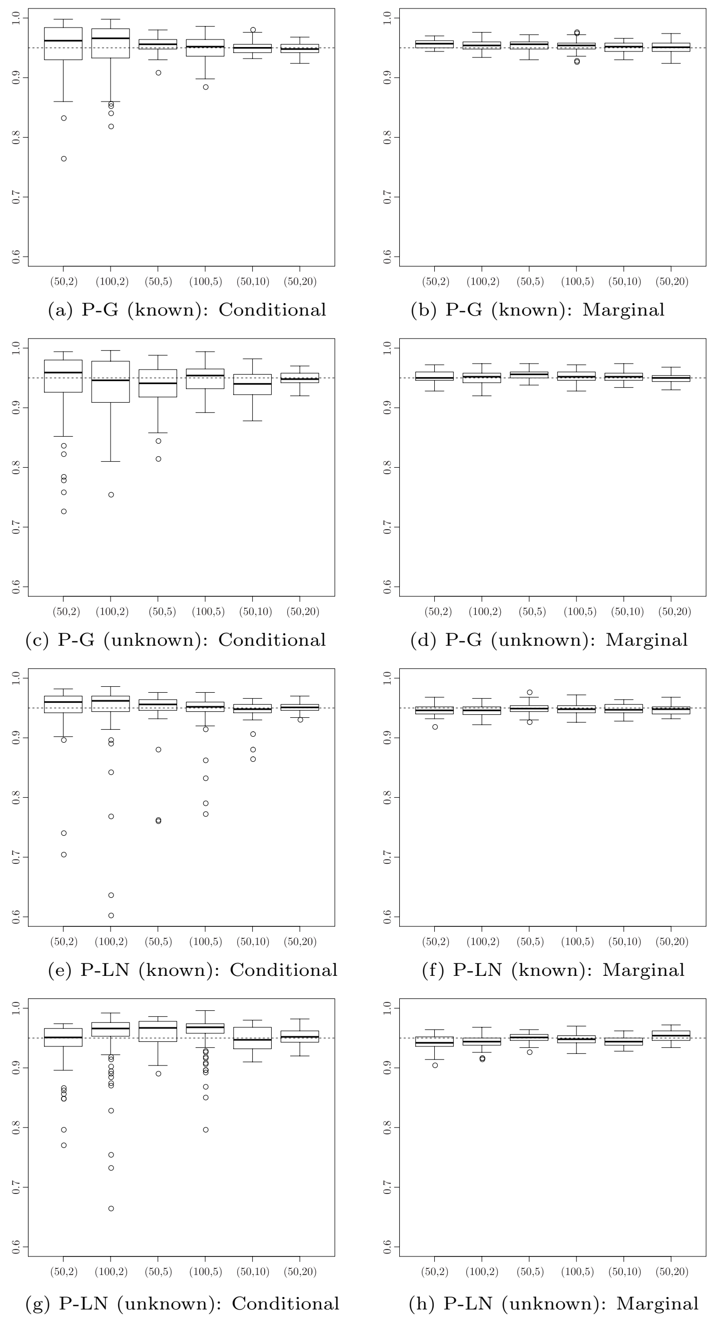
For these models, Table 2 shows that both conditional and marginal inferences provide good average coverage probabilities when the fixed parameters are known or unknown, even when n_i is as small as 2. As in Table 1, the standard

Table 2. Average coverage probabilities of the nominal 95% intervals over K independent units in Poisson-gamma (P-G) and Poissin-lognormal (P-LN) models with 500 replications.

			P-G		P-LN	
			$\boldsymbol{\theta}$: known	$\boldsymbol{\theta}$: unknown	$\boldsymbol{\theta}$: known	$\boldsymbol{\theta}$: unknown
N	(K, n)	Type	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
100	(50,2)	Conditional	0.963 (0.024)	0.977 (0.010)	0.946 (0.051)	0.936 (0.046)
		Marginal	0.956 (0.008)	0.954 (0.010)	0.946 (0.011)	0.944 (0.009)
200	(100,2)	Conditional	0.952 (0.040)	0.936 (0.052)	0.948 (0.055)	0.952 (0.050)
		Marginal	0.954 (0.009)	0.954 (0.011)	0.944 (0.010)	0.943 (0.010)
250	(50,5)	Conditional	0.956 (0.014)	0.935 (0.038)	0.947 (0.041)	0.960 (0.024)
		Marginal	0.953 (0.009)	0.953 (0.009)	0.947 (0.010)	0.946 (0.010)
500	(100,5)	Conditional	0.950 (0.020)	0.948 (0.024)	0.947 (0.031)	0.959 (0.030)
		Marginal	0.953 (0.009)	0.951 (0.010)	0.948 (0.010)	0.949 (0.010)
500	(50,10)	Conditional	0.951 (0.012)	0.940 (0.023)	0.945 (0.019)	0.946 (0.020)
		Marginal	0.950 (0.010)	0.952 (0.009)	0.951 (0.010)	0.945 (0.009)
1,000	(50,20)	Conditional	0.950 (0.012)	0.948 (0.012)	0.951 (0.009)	0.953 (0.015)
		Marginal	0.947 (0.011)	0.949 (0.008)	0.948 (0.009)	0.952 (0.009)

deviations of the empirical coverage probabilities are smaller for prediction intervals in the marginal inference than those in the conditional inference. Figures 2(a), 2(c), 2(e) and 2(g) show Box-plots of K individual coverage probabilities of the confidence intervals for the conditional inference in both models. Overall, interquartile ranges are visibly tighter in Poisson-lognormal model than in Poisson-gamma model. The marginal asymptotic distribution of $(\hat{u}_i - u_i)$ is not normal although skewness is zero. We display q-q plots of $\hat{u}_i - u_i$ where $i = 1, 2, 3$ for Poisson-lognormal and Poisson-gamma models when the fixed parameter is estimated. Figure S1 in Supplementary Material shows that tail behavior of Poisson-lognormal model is closer to that of normal distribution than that of Poisson-gamma, which explains tighter interquartile ranges in Poisson-lognormal model than in Poisson-gamma model. Figures 2(b), 2(d), 2(f) and 2(h) show Box-plots of individual coverage probabilities of prediction intervals in marginal inference. As in Figure 1, the coverage probabilities of prediction intervals are close to the nominal value possibly due to small bias.

We conducted simulations comparing the proposed method and Bayesian credible intervals for Poisson-gamma models with normal prior for β using WinBUGS 14. The results are reported in Table S1 of Supplementary Material. The two methods show similar coverage probabilities but the proposed method show faster computational time. Specifically, in terms of computing time for each replication using a Workstation with 2.3-GHz CPU and 64 GB RAM, Bayesian credible intervals took 59.1 times longer in CPU time on average than the proposed marginal approach.



(a) P-G (known): Conditional

(b) P-G (known): Marginal

(c) P-G (unknown): Conditional

(d) P-G (unknown): Marginal

(e) P-LN (known): Conditional

(f) P-LN (known): Marginal

(g) P-LN (unknown): Conditional

(h) P-LN (unknown): Marginal

Figure 2. Individual coverage probabilities (y-label) of the nominal 95% (*dotted line*) confidence and prediction intervals from Poisson-gamma (P-G) and Poisson-lognormal (P-LN) models for known & unknown $\boldsymbol{\theta}$. Here (50,2), (100,2), (50,5), (100,5), (50,10) and (50,20) indicate (K, n), where K is the number of clusters and n is the cluster size.

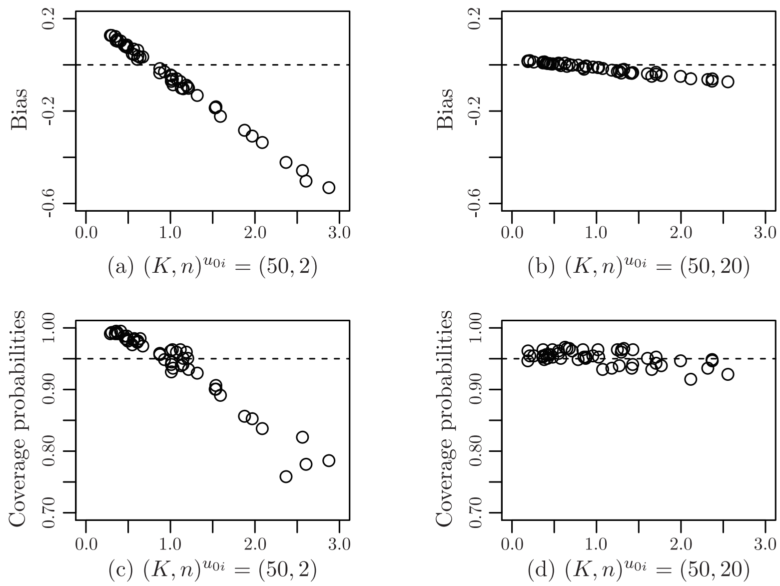
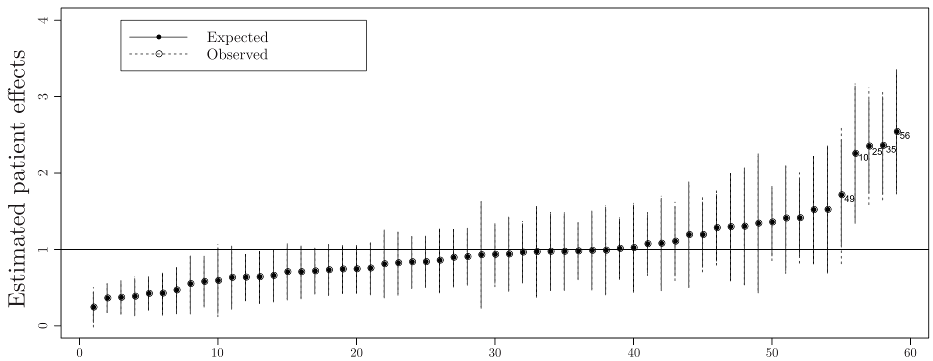


Figure 3. Bias and coverage probabilities against realized values u_{0i} in the conditional inference when $(K,n) = (50,2)$ and $(50,20)$ under the Poisson-gamma model.

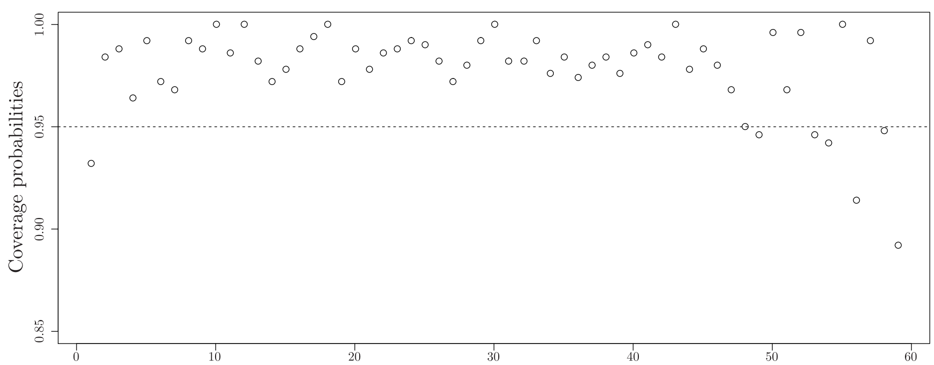
7. Applications

7.1. Conditional inference

We illustrate the proposed method using the epilepsy seizure count data from a clinical trial carried out by Leppik et al. (1985) and previously analyzed by Thall and Vail (1990). The data come from the randomized clinical trial conducted among patients suffering from simple or complex partial seizures to receive either the antiepileptic drug progabide or a placebo, as an adjuvant to standard chemotherapy. The primary outcome of interest (Y) is the number of seizures occurring over the previous 2 weeks measured at each of four successive postrandomization clinic visits. Thall and Vail (1990) took a quasi-likelihood approach and focused on comparing various types of overdispersion models. We assumed that extra variation was due to individual-specific seizure propensity and conducted a secondary analysis to quantify the seizure propensity. We formally identified patients with high seizure propensity using the inferential procedure described in Section 3. In this we assumed that inherent seizure propensity exists and is realized (subject was born with it) but cannot be observed. We would like to draw inference about the realized seizure propensity and apply the conditional inferential procedure described in Theorem 4. The data consist of four repeated measures ($n_i = 4$) of $K = 59$ epileptic patients, with covariates Constant, Base (x_1), Trt (x_2, placebo=0, progabide=1), Base.Trt (x_3), Age (x_4), and Visit ($x_5 =$



(a) Patients sorted in increasing order of estimated random effects



(b) Patients sorted in increasing order of estimated random effects

Figure 4. (a) Estimated random effects of 59 patients in the epileptic data and their 95% confidence intervals under the Poisson-gamma model; Expected and Observed, expected and observed variance estimates. (b) Coverage probabilities against estimated random effects.

$-0.3, -0.1, 0.1, 0.3$ for each visit). We assumed that the $Y_{ij}|u_i$ ($i = 1, \ldots, 59; j = 1, 2, 3, 4$) are Poisson with mean $\mu_{ij} = \exp(\eta_{ij})$; $\eta_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + \beta_5 x_{5ij} + \log u_i$ is the linear predictor and the random effect u_i is a gamma with mean 1 and variance λ. For the fixed parameters, we obtained the estimates by maximizing the second-order Laplace approximation given in (4.1). The estimates of fixed parameters and their standard errors (SEs) are $\hat{\beta}_0 = -1.32(SE = 1.25)$, $\hat{\beta}_1 = 0.88(SE = 0.13)$, $\hat{\beta}_2 = -0.90(SE = 0.40)$, $\hat{\beta}_3 = 0.35(SE = 0.20)$, $\hat{\beta}_4 = 0.50(SE = 0.37)$, $\hat{\beta}_5 = -0.29(SE = 0.10)$, and $\hat{\lambda} = 0.28(SE = 0.06)$, yielding significant difference between the two treatment groups.

We focused on potential heterogeneity between outcomes of patients and constructed 95% confidence intervals of realized values of random effects $u_{0i}(i = 1, \ldots, 59)$: $\{\hat{u}_i - 1.96SE(\hat{u}_i - u_{0i}), \hat{u}_i + 1.96SE(\hat{u}_i - u_{0i})\}$. Figure 4(a) gives the 95% confidence intervals for the realized but unobserved individual seizure propensity

($K = 59$). Confidence intervals obtained from both 'expected' and 'observed' versions of variance estimates show similar trends. Figure 4(a) demonstrates substantial variations in seizure propensity among patients. Especially, for four patients (patient id=10, 25, 35 and 56), the 95% confidence interval of u_{0i} does not contain 1, suggesting that the seizure propensity is significantly different from the norm. Patient id 49's interval excludes 1 using the variance estimate via expected information. These patients were identified as outliers via residual analysis by Thall and Vail (1990), Breslow and Clayton (1993), and Ma and Jørgensen (2007), but there were no formal inferential procedures. We also identify patients with low propensity significantly different from 1, which previous analyses did not.

We conducted a data-driven simulation using 500 replications based on the epilepsy data structures and the estimated coefficients to investigate the behavior of coverage probability against an increasing order of estimated random effects. The responses were generated from the Poisson-gamma model with the true conditional mean $\hat{\mu}_{ij} = \exp(\hat{\eta}_{ij})$; $\hat{\eta}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 x_{1ij} + \hat{\beta}_2 x_{2ij} + \hat{\beta}_3 x_{3ij} + \hat{\beta}_4 x_{4ij} + \hat{\beta}_5 x_{5ij} + \log \hat{u}_i$, where $\hat{\beta}_0, \ldots, \hat{\beta}_5$ and $\hat{u}_i$ are the estimated values from the data. Figure 4(b) shows that coverage probability for conditional inference tends to be low when the actual realized values were extreme.

7.2. Marginal inference

Another example is the infertility study of Archer (1987), previously analyzed by Paik (1992). Infertile women with normal serum prolactin levels have been known to establish a pregnancy after the use of bromocriptine, a dopamine agonist. Prolactin levels were measured four times repeatedly at 15-minute intervals after the injection of thyrotropin releasing hormone (TRH) in 30 subjects. Figure 5(a) displays prolactin measurements for the 30 subjects. An additional baseline prolactin level was measured before the injection of TRH. Subjects were divided into three groups depending on their fertility status: 6 were normal; 12 had anovulation and/or inphase endometrial biopsies; and 12 had histologic evidence of luteal phase deficiency. Paik (1992) showed that the patterns of responses differed in the three groups using an extended Generalized Estimating Equation approach (Liang and Zeger (1986)).

We conducted a secondary analysis to identify women among infertile group who have hypo- or hyper- responsiveness to TRH. We assumed that there is individual specific responsiveness of prolactin to TRH even after adjusting for the group effect. This responsiveness was assumed to arise randomly in each cycle. We are interested in the responsiveness of prolactin in a next cycle of TRH stimulation, which would be important for infertility, not the current responsiveness that is already realized (yet unobserved). This constitutes marginal inference.

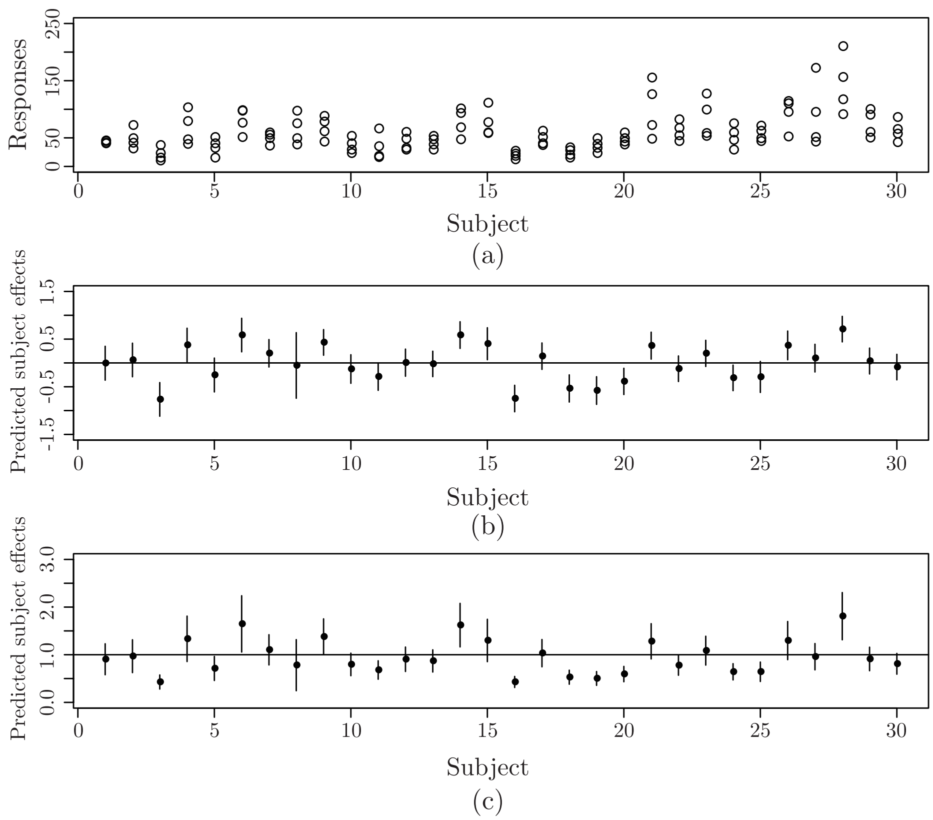


Figure 5. (a) Repeated-measure responses for 30 subjects in the prolactin data. (b) Predicted random effects and 95% prediction intervals, under the gamma-normal model. (c) Predicted random effects and 95% prediction intervals, under the gamma-inverse gamma model.

The prolactin responses $Y_{ij}(i = 1, \ldots, 30; j = 1, 2, 3, 4)$ were assumed to follow a gamma, as in Paik (1992). Specifically, the $Y_{ij}|u_i$ were gamma with shape parameter k and scale parameter μ_{ij}/k, so $E(Y_{ij}|u_i) = \mu_{ij}$ and $\text{var}(Y_{ij}|u_i) = \phi\mu_{ij}^2$ with $\mu_{ij} = \exp(\eta_{ij})$ and dispersion parameter $\phi = 1/k$. Covariates in the model were Constant, Indicator of Group 2 (x_1), Indicator of Group 3 (x_2), Time ($x_3 = 1, 2, 3, 4$) and Baseline prolactin level (x_4). The linear predictor was $\eta_{ij} = \beta_0+\beta_1 x_{1ij}+\beta_2 x_{2ij}+\beta_3 x_{3ij}+\beta_4 x_{4ij}+u_i$ and the random effect u_i was taken to be a normal with mean 0 and variance λ. The estimates of fixed parameters were $\hat{\beta}_0 = 4.368(SE = 0.179)$, $\hat{\beta}_1 = -0.055(SE = 0.215)$, $\hat{\beta}_2 = 0.391(SE = 0.215)$, $\hat{\beta}_3 = -0.267(SE = 0.012)$, $\hat{\beta}_4 = 0.006(SE = 0.005)$, $\hat{k} = 50.465(SE = 6.533)$, and $\hat{\lambda} = 0.170(SE = 0.052)$. The estimates of β are similar to those of the mean-model 5 by Paik (1992). The corresponding standard errors are slightly different due to different assumptions on the covariance structure.

We identified individuals with hypo- or hyper- responsiveness relative to their group means. Figure 5(b) displays the 95% prediction intervals for the random effects $u_i (i = 1, \ldots, 30)$ of individual subjects ($K = 30$) using the 'observed' versions of the variance estimate. The 95% prediction intervals show that patient number 3 has hypo-responsiveness while patient number 6 has hyper-responsiveness among the infertile group.

In addition, we fitted the gamma-inverse gamma model described in Section 3.2. The linear predictor was $\eta_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + \nu_i$ and the random effect $u_i = \exp(\nu_i)$ was assumed to be inverse gamma with mean 1 and variance $\lambda = 1/(\alpha - 1)$, $\alpha > 1$. The estimates of fixed parameters were very similar to those of the gamma-normal model: $\hat{\beta}_0 = 4.447(SE = 0.180)$, $\hat{\beta}_1 = -0.040(SE = 0.216)$, $\hat{\beta}_2 = 0.415(SE = 0.218)$, $\hat{\beta}_3 = -0.267(SE = 0.012)$, $\hat{\beta}_4 = 0.007(SE = 0.005)$, $\hat{\kappa} = 50.450(SE = 6.533)$, and $\hat{\alpha} = 4.836(SE = 1.293)$. Figure 5(c) shows similar inference as Figure 5(b), although different distributions were assumed for the random effects. This suggests that inference about individual responsiveness is robust against distributional specification of random effects here.

8. Discussion and Concluding Remarks

We have shown conditional and marginal inferences of random effects from a frequentist's stance. Conditionally, the estimators of realized but unobserved random effects are normally distributed as the number of subunits increases. Marginally, the predictors of unrealized and unobserved random effects are not necessarily normally distributed even if the number of subunits increases, but have zero skewness asymptotically. Simulations reveal that, for the models discussed here, coverage probabilities of the proposed inferential procedures for random effects are close to the nominal value even in small to moderate number of subunits. In conditional inference, some individual coverage probabilities fall short of claimed coverage, while in the marginal inference most coverage probabilities are tightly around the nominal value. Superior performance of coverage probabilities for the marginal inference is likely to be due to small bias.

We add cautionary remarks on applying results from finite sample performance shown via simulation in practice. When the number of subunits is as large as 20, asymptotic properties may hold reasonably well in both conditional and marginal cases. When n_i is small, the conditional inference requires greater caution in interpretation than the marginal inference, especially when predicted values are extreme, and one may accompany empirical analysis of coverage probability as shown in Section 7.1. Although interpretation is different, the marginal confidence intervals and Bayesian credible intervals seem to display similar coverage probabilities; the proposed marginal intervals take much shorter computational time. When random intervals for random quantities are needed, the proposed method offers a direct way to obtain them without resorting to frequentists' property of Bayesian credible intervals.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIP) (No. 2013R1A2A2A 01067262 and No. 2011-0030810) and Research Settlement Fund for the new faculty of SNU.

References

Archer, D. F. (1987). Prolactin response to thyrotropin releasing hormone in women with infertility or randomly elevate serum prolactin levels. *Fertility and Sterility* **47**, 559-564.

Bjørnstad, J. F. (1996). On the generalization of the likelihood function and likelihood principle. *J. Amer. Statist. Assoc.* **91**, 791-806.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed model. *J. Amer. Statist. Assoc.* **88**, 9-25.

Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis.* 2nd edition. Chapman and Hall, London.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**, 423-447.

Lee, Y and Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). *J. Roy. Statist. Soc. Ser. B* **58**, 619-678.

Lee, Y. and Nelder, J. A. (2005). Likelihood for random-effects (with discussion). *Statist. Operational Res. Trans.* **29**, 141-182.

Lee, Y. and Nelder, J. A. (2009). Likelihood inference for models with unobservables: another view (with discussion). *Statist. Sci.* **24**, 255-293.

Leppik, I. E., et al. (1985). A double-blind crossover evaluation of progabide in partial seizures. *Neurology* **35**, 285.

Liang, K. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

Ma, R. and Jørgensen, B. (2007). Nested generalized linear mixed models: Orthodox best linear unbiased predictor approach. *J. Roy. Statist. Soc. Ser. B* **69**, 625-641.

Ma, R., Krewski, D. and Burnett, R. T. (2003). Random effects Cox models: a Poisson modelling approach. *Biometrika* **90**, 157-169.

Maritz, J. S. and Lwin, T. (1989). *Empirical Bayes Methods*, 2nd edn. Chapman and Hall, London.

Meng, X.-L. (2009). Decoding the H-likelihood. *Statist. Sci.* **24**, 280-293.

Meng, X.-L. (2011). What's the H in H-likelihood: A holy grail or an achilles' heel? (with discussion). *Bayesian Statistics* **9**, 473-500.

Morris, C. N. (1983). Parametric empirical Bayes inference: theory and application. *J. Amer. Statist. Assoc.* **78**, 47-59.

Paik, M. C. (1992). Parametric variance function estimation for nonnormal repeated measurement data. *Biometrics* **48**, 19-30.

Rao, J. N. K. (2003). *Small Area Estimation*. Wiley, New York.

Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects (with discussion). *Statist. Sci.* **6**, 15-51.

Thall, P. F. and Vail, S. C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics* **46**, 657-671.

Department of Statistics, Seoul National University, Seoul 151-742, South Korea.

E-mail: myungheechopaik@snu.ac.kr

Department of Statistics, Seoul National University, Seoul 151-742, South Korea.

E-mail: youngjo@snu.ac.kr

Department of Statistics, Pukyong National University, Busan 608-737, South Korea.

E-mail: idha1353@pknu.ac.kr

(Received May 2013; accepted June 2014)