Verifying Regularity Conditions for Logit-Normal GLMM

Yun Ju Sung Charles J. Geyer

January 10, 2006

In this note we verify the conditions of the theorems in Sung and Geyer (submitted) for the Logit-Normal GLMM model used for all the examples in that paper.

1 Model

Before we start on the conditions themselves, we make a few general comments about this model. First, due to the simplicity of the random effect, it is possible to use no importance weights (and this is what the `bernor` package for R does). The importance sampling distribution (h in the notation of the paper) from which the Monte Carlo simulations of the missing data are drawn is multivariate standard normal, which contains no unknown parameters. All of the unknown parameters are in the β and Δ in the linear predictor

$$\eta = X\beta + Z\Delta b \tag{1}$$

(here X and Z are known matrices, β is a vector, Δ a diagonal matrix, and b the multivariate standard normal vector of random effects).

The observed data for one individual is IIM (independent and identically modeled) Bernoulli with success probability vector having components $p(\eta_k)$, where we define

$$p(\eta_k) = \operatorname{logit}^{-1}(\eta_k) = \frac{1}{1 + \exp(-\eta_k)}$$

where η_k are the components of η. Then we have n IID individuals.

1.1 Differentiability of the Joint

Define $q(t) = 1 - p(t)$. Then we have

$$\begin{aligned} p'(t) &= p(t)q(t) \\ q'(t) &= -p(t)q(t) \end{aligned}$$

and

$$f_\theta(b, y) = h(b) \prod_{k \in K} p(\eta_k)^{y_k} q(\eta_k)^{1-y_k} \tag{2}$$

where the η_k are components of the vector η, which is a function of b and θ defined by (1). Then we have

$$\frac{\partial \log f_\theta(b, y)}{\partial \theta_l} = \sum_{k \in K} \bigl[y_k - p(\eta_k) \bigr] \frac{\partial \eta_k}{\partial \theta_l}$$

where

$$\begin{aligned} \frac{\partial \eta_k}{\partial \beta_l} &= x_{kl} \\ \frac{\partial \eta_k}{\partial \delta_l} &= z_{kl} b_l \end{aligned}$$

where x_{kl} are the components of X and similarly for z_{kl}, the b_l are the components of b, the β_l are the components of β, and the δ_l are the diagonal elements of Δ. Since these are not functions of the parameters, we have $\partial^2 \eta_k / \partial \theta_l \partial \theta_{l'} = 0$ and

$$\begin{aligned} \frac{\partial^2 \log f_\theta(b, y)}{\partial \theta_l \partial \theta_{l'}} &= - \sum_{k \in K} p(\eta_k) q(\eta_k) \frac{\partial \eta_k}{\partial \theta_l} \frac{\partial \eta_k}{\partial \theta_{l'}} \\ \frac{\partial^3 \log f_\theta(b, y)}{\partial \theta_l \partial \theta_{l'} \partial \theta_{l''}} &= \sum_{k \in K} \bigl[q(\eta_k) - p(\eta_k) \bigr] p(\eta_k) q(\eta_k) \frac{\partial \eta_k}{\partial \theta_l} \frac{\partial \eta_k}{\partial \theta_{l'}} \frac{\partial \eta_k}{\partial \theta_{l''}} \end{aligned}$$

Moreover, each $\partial \eta_k / \partial \theta_l$ is either constant, a constant times a component of b, or, in the case where several δ_l are constrained to be equal, a linear combination of several components of b. Thus, writing $\|b\|_1$ for the sum of the absolute values of the components of b, there are constants C_0 and C_1 such that

$$\left| \frac{\partial \eta_k}{\partial \theta_l} \right| \le C_0 + C_1 \|b\|_1 \tag{3}$$

for all k and l.

1.2 Differentiability of the Marginal

We now examine differentiability of

$$f_\theta(y) = \int f_\theta(b, y) \, d\mu(b) \tag{4}$$

where μ is Lebesgue measure. For differentiability under the integral sign we use the lemma on p. 124 in Ferguson (1996). What is required is that the partial derivatives of the integrand be dominated by an integrable function that does not contain the parameter on some neighborhood of the point where the derivative is taken. Now

$$\frac{\partial f_\theta(b, y)}{\partial \theta_l} = \frac{\partial \log f_\theta(b, y)}{\partial \theta_l} \cdot f_\theta(b, y) \tag{5}$$

and we see from the analysis in Section 1.1 that the right hand side is dominated by (3) times $h(b)$. Hence, because the normal distribution has first moments, we have differentiability under the integral sign

$$\nabla f_\theta(y) = \int \nabla f_\theta(b, y) \, d\mu(b)$$

and

$$\begin{aligned}
\nabla \log f_\theta(y) &= \frac{1}{f_\theta(y)} \int \nabla f_\theta(b, y) \, d\mu(b) \\
&= \frac{1}{f_\theta(y)} \int \bigl[\nabla \log f_\theta(b, y)\bigr] f_\theta(b, y) \, d\mu(b) \\
&= \int \bigl[\nabla \log f_\theta(b, y)\bigr] f_\theta(b \mid y) \, d\mu(b) \\
&= E_\theta\bigl\{\nabla \log f_\theta(X, Y) \bigm| Y = y\bigr\}
\end{aligned}$$

where in the last line we switched variable names from b to x because capital B doesn't look like a random variable. We will do that throughout this note.

Differentiating (5) again we get

$$\frac{\partial^2 f_\theta(b, y)}{\partial \theta_l \partial \theta_{l'}} = \left[\frac{\partial^2 \log f_\theta(b, y)}{\partial \theta_l \partial \theta_{l'}} + \frac{\partial \log f_\theta(b, y)}{\partial \theta_l} \cdot \frac{\partial \log f_\theta(b, y)}{\partial \theta_{l'}}\right] \cdot f_\theta(b, y)$$

and we see from the analysis in Section 1.1 that the right hand side is dominated by (3) squared times $h(b)$. Hence, because the normal distribution has second moments, we have differentiability under the integral sign

$$\nabla^2 f_\theta(y) = \int \nabla^2 f_\theta(b, y) \, d\mu(b)$$

and

$$\nabla^2 \log f_\theta(y) = \frac{1}{f_\theta(y)} \int \nabla^2 f_\theta(b, y)\, d\mu(b) - \bigl(\nabla \log f_\theta(y)\bigr)\bigl(\nabla \log f_\theta(y)\bigr)^T$$
$$= E_\theta\bigl\{\nabla^2 \log f_\theta(X, Y) \bigm| Y = y\bigr\} - \bigl(\nabla \log f_\theta(y)\bigr)\bigl(\nabla \log f_\theta(y)\bigr)^T$$

It should now be clear that $f_\theta(y)$ is infinitely differentiable, that derivatives of all orders can be passed under the integral sign in (4). The formulas for the derivatives get messier but not harder.

Comment It is clear that the only property of the normal distribution used here is having moments of all orders. Since we will not be interested in higher than third moments any other distribution with at least three moments would do as well.

2 Theorem 2.2

Now we start on the conditions of Theorem 2.2. Condition (1) follows from the parameter space being a subset of $\mathbb{R}^d$, hence a separable metric space when given the usual metric. Condition (2) follows from the differentiability discussed above. Condition (3) is made easy by the sample space $\mathcal{Y}$ being finite. From (2) we see that $f(b, y)/h(b)$ is bounded above by one. Hence

$$Q \log \left[P \sup_{\phi \in B_\theta} f_\phi(X, Y)/h(X)g(Y) \right] \le -Q \log g(Y)$$

(again using x interchangeably with b). So whatever misspecified model we have, so long as it puts positive probability at all points, that is, $g(y) > 0$ for all y, this condition is satisfied. In conditions (4) and (5), the index set $\mathcal{Y}$ being finite, verification of Glivenko-Cantelli amounts to no more than verifying that the elements of the family are $L^1(P)$. As we just noted,

$$\sup_{\phi \in B} f_\phi(b, y)/h(b)g(y) \le 1/g(y)$$

and hence is obviously integrable so long as $g(y) > 0$ for all y. Similarly,

$$\frac{f_\theta(b \mid y)}{h(b)} = \frac{f_\theta(b, y)}{h(b) f_\theta(y)} \le \frac{1}{f_\theta(y)}$$

and this is integrable so long as $f_\theta(y) > 0$ for all θ and y, which is true in the parameterization we are using. That finishes the verification for Theorem 2.2. The only condition we needed to impose was $g(y) > 0$ for $y \in \mathcal{Y}$.

3 Theorem 2.3

3.1 Differentiation Under the Integral Sign

The proof of condition (1) is similar to the proof in Section 1.2. We must verify that

$$\begin{aligned}
\frac{\partial f_\theta(b \mid y)}{\partial \theta_l} &= \frac{\partial}{\partial \theta_l} \frac{f_\theta(b, y)}{f_\theta(y)} \\
&= \frac{1}{f_\theta(y)} \frac{\partial f_\theta(b, y)}{\partial \theta_l} - \frac{f_\theta(b, y)}{f_\theta(y)^2} \frac{\partial f_\theta(y)}{\partial \theta_l} \\
&= \left[\frac{\partial \log f_\theta(b, y)}{\partial \theta_l} - \frac{\partial \log f_\theta(y)}{\partial \theta_l} \right] \frac{f_\theta(b, y)}{f_\theta(y)}
\end{aligned}$$

is dominated by an integrable function that does not contain θ when thought of as a function of b alone. We established in Section 1.1 that the first term in the square brackets is dominated by (5). We established the existence of the second term in the square brackets in Section 1.2, and it does not contain b. Then from (2) we see that the term outside the square brackets is dominated by $h(b)$. That establishes the first differentiation under the integral sign

$$\nabla \int f_\theta(b \mid y) \, d\mu(b) = \int \nabla f_\theta(b \mid y) \, d\mu(b).$$

To take another derivative under the integral sign, we must show that the second derivatives

$$\begin{aligned}
\frac{\partial^2 f_\theta(b \mid y)}{\partial \theta_l \partial \theta_{l'}} &= \left[\frac{\partial^2 \log f_\theta(b, y)}{\partial \theta_l \partial \theta_{l'}} - \frac{\partial^2 \log f_\theta(y)}{\partial \theta_l \partial \theta_{l'}} \right] \frac{f_\theta(b, y)}{f_\theta(y)} \\
&\quad + \left[\frac{\partial \log f_\theta(b, y)}{\partial \theta_l} - \frac{\partial \log f_\theta(y)}{\partial \theta_l} \right] \\
&\quad \times \left[\frac{\partial \log f_\theta(b, y)}{\partial \theta_{l'}} - \frac{\partial \log f_\theta(y)}{\partial \theta_{l'}} \right] \frac{f_\theta(b, y)}{f_\theta(y)}
\end{aligned} \tag{6}$$

is similarly dominated, and the argument is similar.

3.2 The Sandwich

This section is about condition (3), but before getting to that, we note that, since $\mathcal{Y}$ is finite, condition (2) is trivial.

For condition (3) consider the Kullback-Leibler information

$$K(\theta) = Q \log \frac{g(Y)}{f_\theta(Y)}$$

Because $\mathcal{Y}$ is finite and because of the assumption made in Section 2 that $g(y) > 0$ for all y, this is a finite sum of finite terms. Because of the form of the model, K is continuous, even infinitely differentiable (in Section 1.2 we established the existence of two derivatives, and claimed without detailed proof we could keep going in the same way to show more exist). Moreover, K is bounded below by zero (by a well known Jensen's inequality argument).

If (a big "if") we could prove that K had a local minimum θ^*, the differentiability under the expectation w. r. t. Q is again trivial because of the finiteness of $\mathcal{Y}$ and the differentiability of $f_\theta(y)$ established in Section 1.2 and would give

$$0 = \nabla K(\theta^*) = -Q \nabla \log f_\theta(Y).$$

The finiteness of the other expectations in condition (3) is again trivial, and

$$\nabla^2 K(\theta^*) = -Q \nabla^2 \log f_\theta(Y)$$

implies that the positive definiteness condition would follow from θ^* being a strong local minimum of K ("strong" meaning the left hand side is positive definite).

Unfortunately, we do not have an analysis that guarantees K has a local minimum, strong or otherwise, and it seems this sort of analysis is generally difficult. Even when the model is correctly specified and the true parameter value θ_0 is the unique global minimum of K (Ferguson, 1996, lemma on p. 113), the positive definiteness of the expected second derivative must be assumed rather than given by a general argument (Ferguson, 1996, Condition (4) of Theorem 18).

There are very simple Logit-Normal GLMM where the model is not even identifiable, take $\eta = \beta + \delta b$ with all of these objects scalar. Then the data are IID Bernoulli, which has one identifiable parameter, the mean $p(\eta)$, but we have two unknown parameters β and σ in the model, so this model must be non-identifiable. Hence, there could not be a general proof that the existence of a strong local minimum part of condition (3) holds for all Logit-Normal GLMM. Our exploration of the "flu" example in the paper, originally taken from Coull and Agresti (2000) gives us no reason for optimism that the log likelihood for that problem (even assuming the model correctly specified) has a strong local maximum (which would be a strong local minimum of K), although the model is clearly not non-identifiable.

3.3 Donsker

The Donsker condition, condition (5) of the theorem is fairly trivial, again because $\mathcal{Y}$ is finite. Verification of this condition comes down to verifying that each function in the class is $L^2(P)$. Now

$$
\begin{aligned}
\nabla \frac{f_\theta(b \mid y)}{h(b)} &= \nabla \frac{f_\theta(b, y)}{f_\theta(y) h(b)} \\
&= \nabla \frac{f_\theta(y \mid b)}{f_\theta(y)} \\
&= \frac{\nabla f_\theta(y \mid b)}{f_\theta(y)} - \frac{f_\theta(y \mid b)}{\bigl(f_\theta(y)\bigr)^2} \nabla f_\theta(y) \\
&= \bigl[\nabla \log f_\theta(y \mid b) - \nabla \log f_\theta(y)\bigr] \frac{f_\theta(y \mid b)}{f_\theta(y)}
\end{aligned}
$$

and we use the following facts, first, $f_\theta(y) > 0$ for all y and θ, second, $f_\theta(y \mid b) \le 1$ for all for all y, b, and θ, third, by the analysis in Section 1, $\nabla \log f_\theta(y \mid b)$ is bounded by (3). Thus we have $L^2(P)$ using only the fact that the normal distribution has second moments.

3.4 Glivenko-Cantelli

For condition (5), the finiteness of $\mathcal{Y}$ makes the class Glivenko-Cantelli if each member is $L^1(P)$, which is trivial because this just asserts that the conditional density $f_\theta(b \mid y)$ integrates.

For conditions (4) and (7), we use Theorem 2.7.11 in van der Vaart and Wellner (1996). Applied to condition (4) the theorem requires a function H such that if we write

$$
v_{ll'}(\theta, y) = \frac{\partial^2 f_\theta(y)}{\partial \theta_l \partial \theta_{l'}}
$$

then we have

$$
|v_{ll'}(\theta, y) - v_{ll'}(\theta', y)| \le \|\theta - \theta'\| H(y)
$$

with H being $L^1(Q)$. The finiteness of $\mathcal{Y}$ makes the integrability of any H we find trivial. Thus we only need to establish that $f_\theta(y)$ is twice differentiable with Lipschitz second derivative. This follows from the pattern we saw in Section 1.2, the third derivative of $f_\theta(y)$ is bounded by a constant (uniformly in θ), and provides the required Lipschitz condition. That the metric entropy for the index set is finite for each ϵ comes from the compactness of S_ρ.

For condition (7) we must find H such that if we write

$$
v_{ll'}(\theta, b, y) = \frac{1}{h(b)} \frac{\partial^2 f_\theta(b \mid y)}{\partial \theta_l \partial \theta_{l'}} \tag{7}
$$

then we have

$$
|v_{ll'}(\theta, b, y) - v_{ll'}(\theta', b, y')| \le \|\theta - \theta'\| H(b)
$$

with H being $L^1(P)$. The finiteness of $\mathcal{Y}$ makes it unnecessary to make the right hand side depend on the distance between y and y'. Note that the partial derivative in (7) is given by (6). If we go on to third derivatives, we get

$$
\begin{aligned}
\frac{\partial^3 f_\theta(b \mid y)}{\partial \theta_l \partial \theta_{l'} \partial \theta_{l''}} = & \left[\frac{\partial^3 \log f_\theta(b, y)}{\partial \theta_l \partial \theta_{l'} \partial \theta_{l''}} - \frac{\partial^3 \log f_\theta(y)}{\partial \theta_l \partial \theta_{l'} \partial \theta_{l''}} \right] \frac{f_\theta(b, y)}{f_\theta(y)} \\
& + \left[\frac{\partial^2 \log f_\theta(b, y)}{\partial \theta_l \partial \theta_{l'}} - \frac{\partial^2 \log f_\theta(y)}{\partial \theta_l \partial \theta_{l'}} \right] \\
& \quad \times \left[\frac{\partial \log f_\theta(b, y)}{\partial \theta_{l''}} - \frac{\partial \log f_\theta(y)}{\partial \theta_{l''}} \right] \frac{f_\theta(b, y)}{f_\theta(y)} \\
& + \text{two similar terms obtained by permuting } l, l' \text{ and } l'' \\
& + \left[\frac{\partial \log f_\theta(b, y)}{\partial \theta_l} - \frac{\partial \log f_\theta(y)}{\partial \theta_l} \right] \\
& \quad \times \left[\frac{\partial \log f_\theta(b, y)}{\partial \theta_{l'}} - \frac{\partial \log f_\theta(y)}{\partial \theta_{l'}} \right] \\
& \quad \times \left[\frac{\partial \log f_\theta(b, y)}{\partial \theta_{l''}} - \frac{\partial \log f_\theta(y)}{\partial \theta_{l''}} \right] \frac{f_\theta(b, y)}{f_\theta(y)}
\end{aligned}
\tag{8}
$$

Again we see that these are bounded uniformly in θ by the cube of (3) times $h(b)$, and this is integrable. This gives the desired Lipschitz property.

References

Coull, B. A. and Agresti, A. (2000). Random effects modeling of multiple binomial responses using the multivariate binomial logit-normal distribution. *Biometrics*, **56**, 73–80.

Ferguson, T. (1996). *A Course in Large Sample Theory.* Chapman & Hall.

Sung, Y. J. and Geyer, C. J. (submitted). Monte Carlo likelihood inference for missing data models. `http://www.stat.umn.edu/geyer/bernor/ms.pdf`.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer-Verlag.