Accurate Calculation and Maximization of Log-Likelihood for Mixed Logistic Regression

Eric V. Slud, Mathematics Dept.
University of Maryland, College, Park

May 2, 2000

Abstract. For the mixed-effect logistic regression model with a single vector of random effects per stratum, log-likelihood can be effectively and accurately calculated either by the well-known Laplace steepest-descent approach (Breslow and Lin 1995) to calculating likelihood integrals or by a series approximation of Crouch and Spiegelman (1990) based on residue integral expansions. The best accuracy available from such approximations is obtained by the Laplace method for large strata (of size in the hundreds or larger) and the Crouch-Spiegelman method for smaller strata. The Crouch-Spiegelman method is also particularly effective for numerical log-likelihood maximization because it simultaneously provides accurate derivatives of the log-likelihood with respect to parameters.

Keywords. *Generalized linear model, Laplace's method, nested error logistic regression, Newton-Raphson iteration, numerical integration, relative error, small-area estimation.*

1 Introduction

In recent years, several authors have attempted to extend the highly developed methodology for estimation and testing within mixed-effect linear

models to generalized linear models (GLM's) and in particular to logistic regression with random effects (Wong & Mason 1985, Zeger & Karim 1991, Breslow & Clayton 1993, Wolfinger & O'Connell 1993, Natarajan & Ghosh 1998, Booth & Hobert 1998). Several approximate computational strategies for estimation have been developed, notably those of Goldstein (1991, 1995, and related **MLn** software), Breslow & Clayton (1993), Wolfinger & O'Connell (1993, with a close relation to **SAS** *Proc Mixed* and *macro glmmix*, for which see Littell et al. 1996), and Breslow & Lin (1995). Although the algorithms and software development of these authors were designed to accommodate multi-level hierarchical models and general link functions, the most important application is to parameter estimation in two-level mixed-effect logistic regression models. In particular, the extension of small-area estimation methodology to GLM's is of interest primarily in this setting (Booth & Hobert 1998).

One drawback of the commercially available software for estimation in mixed-effect GLM's — not addressed in the otherwise very useful overview by Zhou et al. (1999) — is the limitation of the sizes (numbers of strata, dimension of covariates) of problems which they can handle. Specifically, since the algorithms they rely on (essentially those reviewed by Breslow & Clayton 1993 and Wolfinger & McConnell 1994) treat regression coefficients and stratum random effects iteratively as parameter elements of a large linear model, these software products cannot reliably handle even two-level models with several regressors and hundreds of strata. Yet such applications are of great interest in large national datasets, from which small-domain estimates must often be generated (Ghosh & Rao 1994). One such setting, involving decennial census data, was treated in Slud (1998).

Related strategies for mixed-effects nonlinear regression have been developed by Lindstrom & Bates (1990), and Pinheiro & Bates (1995b), and the **Splus** function *nlme* of Pinheiro & Bates (1995a) has been advocated for estimation within mixed-effect logistic regression with large cells, via a variance-stabilizing transformation, by Slud (1998).

The only direct numerical approximations generally studied for likelihood in mixed nonlinear models are based on Laplace's asymptotic method of steepest descents for integrals with sharply peaked integrands (Pinheiro & Bates 1995, Breslow & Lin 1995). Specifically for the case of random-intercept mixed logistic models, Breslow & Clayton (1993) also refer to a

series approximation of Crouch & Spiegelman (1990). The latter approximation does in fact apply very usefully to mixed logistic regression, as we shall demonstrate below, and potentially to other mixed GLM's as well.

The present paper first compares the best approximations available for direct computation of log-likelihood in two-level mixed logistic models — those arising from the Laplace method, the higher-order correction by Breslow & Lin (1995) of the Laplace method, and the application of the Crouch-Spiegelman (1990) series — with the computationally burdensome (but feasible) exact numerical integration. The main objective of the paper is then to show that in tandem the Breslow-Lin and Crouch-Spiegelman approximations enable extremely accurate and quick maximum-likelihood estimation of regression and variance-component parameters.

2 The Model and the Algorithms

Consider the following data structure, which is common in unit-level models for large survey and census databases. The response variables $\underline{Y} = (y_{ij},\ i = 1, \ldots, N,\ j = 1, \ldots, J_i)$ consist of counts of positive responses from the individuals within N strata of respective (positive) sampled sizes $(n_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, J_i)$, but we are interested in cases where many of the cell-counts n_{ij} may be small, although the overall sample size $\sum_{i=1}^{m} \sum_{j=1}^{J_i} n_{ij}$ is large. In addition, there is an $N \times p$ matrix $\mathbf{X}$ of covariates, the rows x_{ij} of which describe, but are constant within and are often used to define, the various strata. The n_{ij} individuals within the (i, j) stratum are modelled as independent of each other and of all individuals in other strata, *given the full complement of relevant covariates*, but some of the covariates (other than the p columns of $\mathbf{X}$ relevant either to the individual or the stratum) might be unobservable. Thus individuals sharing the same index i whose responses would be conditionally independent given the full set of covariates will be dependent through shared but unobservable *random effects* given only the observed covariates $\mathbf{X}$. The index-variable i is called the *cluster index*, and the shared random effects are assumed to enter as coefficients of observed sub-vectors z_i of dimension $q \leq p$ common to all $x_{ij},\ j = 1, \ldots, J_i,$ for $i = 1, \ldots, m.$ Mathematically, the mixed-effect

logistic-regression model which we consider has the form:

$$y_{ij} \sim Binom(n_{ij}, p_{ij}) \qquad \text{with} \qquad \log(\frac{p_{ij}}{1-p_{ij}}) = x_{ij}\beta + z_i u_i \qquad (1)$$

and the y_{ij} are conditionally independent given the random effects u_i, where the *i.i.d.* cluster random effects u_i are assumed $\mathcal{N}(0, \Omega)$, and the p-dimensional coefficients β and the $q \times q$ covariance matrix Ω parameterize the model. For simplicity in what follows, we denote

$$\eta_{ij} = x_{ij}\beta \quad , \qquad h(x) \equiv e^x/(1+e^x) \qquad (2)$$

so that in our model, $p_{ij} = h(\eta_{ij} + z_i u_i)$.

The model (1) arises in unit-level models for *small-area estimation* within large-scale surveys (*cf.* Ghosh & Rao 1994, where (1) is called the *nested error regression model*). It is essentially the same model considered by Zeger & Karim (1991) and Wong & Mason (1985). It is slightly more general than the 'single component of dispersion' models treated by Breslow & Lin (1995), although the results of that paper are applicable to it too, and fails being a fully general multilevel model because the variance-component covariates z_i are assumed to be constant over all strata (i,j) with the same cluster-index i. However, strata (i,j) within the same cluster i do share a common unobserved random effect, so within the hierarchy of multilevel models of Goldstein (1995), (1) is a fully general *two-level model*.

2.1 Laplace's Method for Loglikelihood

The log-likelihood at parameters (β, Ω) for the model (1) is

$$\begin{aligned} & \sum_i \log \int \prod_{j=1}^{J_i} \left\{ h(\eta_{ij} + z_i u_i)^{y_{ij}} (1 - h(\eta_{ij} + z_i u_i))^{n_{ij} - y_{ij}} \right\} f_{\mathbf{u}}(u_i)\, du_i \\ = & \sum_i \log \int \prod_{j=1}^{J_i} \left\{ h(\eta_{ij} + \sigma_i w)^{y_{ij}} (1 - h(\eta_{ij} + \sigma_i w))^{n_{ij} - y_{ij}} \right\} \phi(w)\, dw \qquad (3) \end{aligned}$$

where ϕ denotes the standard normal density and

$$\sigma_i^2 \equiv \sigma_i^2(\Omega) = \sqrt{z_i\, \Omega\, z_i^{tr}}$$

is assumed strictly positive, for all i. The idea of Laplace's Method of approximating the i^{th} integral in this expression is that almost all of the contribution to the integral occurs for values of w close (especially when all of the y_{ij} and $n_{ij} - y_{ij}$ are large) to the maximizer w_{ij}^* of the last integrand. The method consists of Taylor-expanding the logarithm of the integrand in w about w_{ij}^*. Following the usage of Breslow & Lin (1995) and previous authors, we will call *Laplace's method* the approximation based on Taylor expansion up to quadratic terms, and will refer to the method involving expansion to fourth order the *Breslow-Lin* approximation. These approximations can be summarized as follows (See Breslow & Lin 1995, where the development only seems to be less general because u_i is scalar and $z_i = 1$.). In what follows, since we give single-stratum formulas, for simplicity we drop the subscripts i. First, if w_j^* is the unique solution of

$$\frac{w_j^*}{\sigma} \;=\; \sum_{j=1}^{J} \Big(y_j \;-\; n_j \, h(\eta_j + \sigma w_j^*) \Big) \tag{4}$$

then the Laplace and Breslow-Lin approximations to the single-stratum likelihood

$$L(\underline{y}, \underline{n}, \underline{\eta}, \sigma) \;=\; \int \prod_{j=1}^{J} \Big\{ h(\eta_j + \sigma w)^{y_j} (1 - h(\eta_j + \sigma w))^{n_j - y_j} \Big\} \, \phi(w) \, dw \tag{5}$$

are respectively (Breslow & Lin 1995, p. 83)

$$L_{Lap} \;=\; \exp\left(- \sum_j \, l_j^* - \; \frac{1}{2} \, \log(1 + \sigma^2 \sum_j \, n_j \, p_j^* (1 - p_j^*)) \; - \; \frac{w_j^{*\,2}}{2\,\sigma^2} \right) \text{(Laplace)}$$

$$L_{Lap} \cdot \exp\left(- \frac{\sigma^4}{8} \cdot \frac{\sum_j \, n_j \, p_j^* (1 - p_j^*)(1 - 6 p_j^* (1 - p_j^*))}{(1 + \, \sigma^2 \, \sum_j \, n_j \, p_j^* \, (1 - p_j^*))^2} \right) \qquad \text{(Bres-Lin)}$$

where

$$p_j^* \;\equiv\; h(\eta_j \, + \, \sigma \, w_j^*) \quad , \qquad l_j^* \;=\; y_j \, \log p_j^* \, + \, (n_j - y_j) \, \log(1 - p_j^*)$$

Alternatives to Laplace's method exist, in which only the logarithm of the integrand divided by a fixed density factor is Taylor-expanded. Although such variant approximations may be comparably accurate to the standard Laplace method when the quantities y_j, n_j are large, they can also be dramatically less accurate when $\sum_j n_j$ is less than 10, and we do not consider them here.

2.2 Series Expansions of Crouch & Spiegelman

In Crouch & Spiegelman (1990), integrals such as the one defining the mixed logistic-normal likelihood $L(\underline{y}, \underline{n}, \underline{\eta}, \sigma)$ in (5) are approximated to high accuracy by series defined via residue calculus. (The reasoning leading to the series dates back to a 1943 paper of Alan Turing.) Several parameters can be chosen to modify the accuracy and types of terms, but in the notation of this paper, the most accurate approximation — and the one most amenable to numerical maximization — is the particular series defined by

$$\frac{\Delta}{\sqrt{\pi}} \sum_{k=-\infty}^{\infty} \exp\left(-(w_0 + k\Delta)^2\right) \prod_{j=1}^{J} \left\{ \left(h(\eta_j + \sigma\sqrt{2}(w_0 + k\Delta))\right)^{y_j} \cdot \left(1 - h(\eta_j + \sigma\sqrt{2}(w_0 + k\Delta))\right)^{n_j - y_j} \right\} \tag{6}$$

where w_0 is a constant intercept which can be chosen arbitrarily (and will be chosen for convenience equal to 0). In actual evaluations, the summations are truncated where the exponential terms fall below ϵ.

Unlike the Laplace-method approximations, the series (6) is guaranteed (Crouch & Spiegelman 1990, Sec. 3, p. 466) to yield absolute error at most ϵ whenever the step-size Δ is chosen, in terms of the constant

$$\alpha = \sqrt{2 \log(1/\epsilon)}$$

to satisfy:

$$\Delta = \begin{cases} \pi/\alpha & \text{if } 2\sqrt{2}\,\alpha \le \pi/\sigma \\ (\sigma\sqrt{2}\,\pi^2)/(\pi^2/4 + 2\sigma^2\alpha^2) & \text{otherwise} \end{cases}$$

This formula for step-size results in $\Delta \le \pi/\alpha$ which is a decreasing function of σ for $\sigma^2 > \pi^2/(8\sigma^2)$.

Although the integrand of (5) is slightly more complicated than the integrands f to which Crouch & Spiegelman (1990) apply their method, their results do apply without alteration to it. Crouch & Spiegelman (1990) also show that series involving additional residue-integral terms, with larger step-sizes, can be theoretically just as accurate. However, although the additional terms can be evaluated explicitly and exactly, they involve large magnitudes with alternating signs so that, after roundoff errors in double

precision computer arithmetic are taken into account, it becomes computationally burdensome, if not actually impossible, to maintain accuracies better than about 10^{-7}. Since we shall also see, in Section 4 below, that the series approximations (6) are particularly well suited to efficient computation of Newton-Raphson maximization iterations, we do not here consider any of the Crouch-Spiegelman series other than (6).

2.3 Exact Numerical Integrals for Likelihood

There are two obstacles to exact evaluation of the likelihood terms (5): the computational burden of numerically evaluating the required integrals, and the fact that multi-stratum problems will typically involve likelihood contributions which differ by many orders of magnitude, so that a uniformly high accuracy for all terms in a log-likelihood is not attainable if numerical integrations are done directly. However, as we shall see below, even when the likelihood terms L_{Lap} given by the Laplace and other approximations to (5) are not very accurate, they always give an answer correct within a factor a 2, so that numerical integration of the expression (5) with integrand divided by L_{Lap} can always be obtained to high accuracy. For example, using this idea with the relative tolerance setting of 10^{-6} of the **Splus** *integrate* function usually gives (relative) accuracy at least 10^{-7}, and it is in this sense that the 'exact' numerically integrated (log-) likelihoods are given for comparison in the following Section.

2.4 General Characteristics of the Approximations

One clear distinction can be drawn between the log-likelihood methods described above. The Crouch-Spiegelman approximation (6) and the computationally expensive method of direct numerical integration both give very accurate calculations of the likelihood integral (5), but the relative accuracy, which is the relevant measure for the resulting log-likelihood, is much less, becoming progressively less as the sample-size gets larger and the likelihood itself gets smaller. Relative accuracies of 10^{-6} can readily be maintained up to sample sizes of 80 or 90 with a single default setting of $\epsilon = 10^{-15}$. In problems with larger strata, successively smaller ϵ must be specified to ensure high *relative* accuracy of the approximation of Crouch and Spiegel-

man. We illustrate this in connection with Tables 5 and 7 below. By contrast, while the accuracies of the Laplace and Breslow-Lin approximate log-likelihoods are somewhat smaller (3 or 4 decimal places is typical, as illustrated in the following Section), they do have the virtue of applying without change to give high, but not increasing, relative or absolute accuracy when stratum sizes get large.

3 Performance of the Approximations

The key finding in applying the approximations to logistic mixed-effect log-likelihoods with typical combinations of fixed and random effect parameters is that the straightforward Laplace and Breslow-Lin approximations typically give relative accuracies respectively of order 10^{-2} to 10^{-3}, while the Crouch-Spiegelman series give accuracy no worse than 10^{-6} on strata of size $n \leq 50$. On the other hand, on very large strata (n from 100 to 200 and above, in our illustrations below), the Crouch-Spiegelman approximation gives *relative* accuracy for likelihood (or equivalently, absolute accuracy for log-likelihood) no better than the Breslow-Lin approximation.

Consider first the behavior of the approximations for a typical array of strata with sample sizes from 1 to 100. We first simulated data (X, U, Y) independently for 500 strata, five strata with each of the sample sizes $n = 1 \ldots, 100$, following the model

$$Y \sim Binom\Big(n,\, h(X \cdot \beta + \sigma U)\Big) \quad , \qquad X \sim \mathcal{N}(0,\, 1) \tag{7}$$

for the parameter-set $\sigma^2 = (0.75)^2 = 0.5625$, $\beta' = (-1.5, 0.6)$. Then we calculated, for each stratum, all of the approximations described above for the log-likelihood, along with the logarithm of the 'exact' numerically integrated log-likelihood (defined to achieve relative error no worse than 10^{-6}). The result of the approximations at the true parameter values are displayed, for every fifth stratum, in Table 1, and in summary form (for absolute errors, averaged over groups of similarly sized strata) in Table 2. These Tables, and other similar ones not shown, demonstrate clearly that the Breslow-Lin approximations to log-likelihood have a slight negative bias for small-to-moderate strata and are often no better than the Laplace approximations when the random-effect variance is large (e.g., > 0.5); that both of these

approximations have errors in the range 0.001 to 0.01 for small-to-moderate strata and correct fixed-effect parameter values; and that in this setting for a wide range of variance-component values the Laplace approximation is often closer to the correct log-likelihood than is the Breslow-Lin approximation.

TABLE 1. Display of errors in log-likelihood for correct fixed-effect parameters and over-large $\sigma^2 = 0.75$ with stratum-sizes 1–20 using Laplace, Breslow-Lin, or Crouch-Spiegelman approximations versus 'exact' integrated log-likelihood. Entry in 'Exact' column is the conservatively estimated absolute error from **Splus** function *integrate* with relative-error setting 1.e-6.

eta	rspct	cellct	logLik	Laplace	Bres-Lin	Cr-Sp	Exact
-2.460	0	1	-0.1071	0.00200	-0.00057	6.94e-17	1.13e-08
-1.481	0	2	-0.4570	0.00060	-0.00224	5.55e-17	4.77e-09
-0.823	0	3	-1.0322	-0.00299	-0.00284	1.11e-15	1.75e-08
-1.496	0	4	-0.8146	0.00097	-0.00465	8.88e-16	1.06e-08
-1.260	1	5	-2.7452	-0.00461	-0.00450	4.44e-16	1.14e-08
-1.933	0	6	-0.8204	0.00327	-0.00618	1.11e-15	1.22e-08
-1.503	1	7	-3.1622	-0.00312	-0.00706	-4.44e-16	7.45e-09
-1.892	1	8	-3.2710	-0.00104	-0.00842	8.88e-16	1.32e-08
-1.604	1	9	-3.4986	-0.00253	-0.00887	-8.88e-16	1.47e-08
-0.514	3	10	-6.6323	-0.00942	-0.00234	3.51e-13	1.54e-07
-1.546	2	11	-5.6076	-0.00564	-0.00839	-2.66e-15	6.64e-08
-1.440	3	12	-7.2620	-0.00755	-0.00641	2.04e-14	2.64e-07
-1.030	5	13	-9.3800	-0.00912	-0.00196	-2.13e-14	7.66e-07
-1.798	4	14	-9.2415	-0.00773	-0.00641	-3.91e-14	8.33e-07
-1.784	4	15	-9.4988	-0.00756	-0.00690	-1.85e-13	4.75e-07
-1.616	4	16	-9.6721	-0.00756	-0.00692	5.33e-14	6.27e-07
-1.547	3	17	-8.4391	-0.00659	-0.00935	8.88e-14	7.09e-08
-2.417	6	18	-13.5349	-0.00773	-0.00593	4.09e-14	3.25e-08
-1.767	2	19	-6.9036	-0.00449	-0.01265	-2.66e-15	3.95e-08
-0.852	3	20	-9.3939	-0.00704	-0.00856	1.78e-15	1.43e-07

Consider next the case of log-likelihood approximations for the same data as in the previous Tables, but evaluated for a smaller value of σ^2. Here again the Breslow-Lin approximation is biased slightly downward, especially in small strata, but one can see more clearly — a pattern that is repeated consistently — that the initial advantage for the former in small strata turns steadily with increasing stratum-size in favor of the latter in larger strata.

Again in this setting, for strata of sizes up to 100, the Crouch-Spiegelman approximation is virtually perfect — at least as good as the highly accurate numerical integrations.

TABLE 2. Summary of average absolute errors in log-likelihood for indicated stratum-size groups, for correct fixed-effect parameters and $\sigma^2 = 0.75$, for stratum-sizes 1–100 using Laplace, Breslow-Lin, or Crouch-Spiegelman approximations with respect to 'exact' integrated log-likelihood.

Stratum-size	# strata	Laplace	Bres-Lin	Cr-Sp	Exact	logLik
1 to 4	20	0.00301	0.00202	5.16e-16	1.29e-08	-0.8340
5 to 10	30	0.00501	0.00544	6.50e-14	7.76e-08	-3.5143
11 to 20	50	0.00620	0.00754	2.28e-13	2.58e-07	-7.6597
21 to 50	150	0.00547	0.00776	9.94e-11	2.45e-07	-17.7451
51 to 80	150	0.00451	0.00746	6.79e-08	2.27e-07	-30.1949
81 to 100	100	0.00385	0.00545	1.84e-06	2.14e-07	-44.0192

TABLE 3. Summary of average absolute errors in log-likelihood for indicated stratum-size groups, for correct fixed-effect parameters and $\sigma^2 = 0.25$, for stratum-sizes 1–100 using Laplace, Breslow-Lin, or Crouch-Spiegelman approximations with respect to 'exact' integrated log-likelihood. Average log-likelihood values over stratum-size groups is given in final column.

Stratum-size	# strata	Laplace	Bres-Lin	Cr-Sp	Exact	logLik
1 to 4	20	0.00056	0.00014	4.79e-16	1.04e-08	-0.8266
5 to 10	30	0.00108	0.00064	4.04e-12	4.77e-08	-3.4598
11 to 20	50	0.00157	0.00138	2.93e-15	8.00e-08	-7.7522
21 to 50	150	0.00202	0.00229	6.25e-11	2.21e-07	-17.8458
51 to 80	150	0.00196	0.00292	4.58e-08	3.07e-07	-30.2583
81 to 100	100	0.00205	0.00272	7.34e-07	3.33e-07	-44.0694

To make the comparison among approximations still clearer, we evaluate and approximate the log-likelihoods once more, for the same data again, but at the still smaller value $\sigma^2 = 0.09$. Now the Breslow-Lin approximation is improves over the standard Laplace method for strata of size up to 50 or so, but is clearly not better for larger strata.

TABLE 4. Summary of average absolute errors in log-likelihood for indicated stratum-size groups, for correct fixed-effect parameters and $\sigma^2 = 0.09$, for stratum-sizes 1–100 using Laplace, Breslow-Lin, or Crouch-Spiegelman approximations with respect to 'exact' integrated log-likelihood. Entry in 'Exact' column is the estimated absolute error in numerically integrated log-likelihood.

Stratum-size	# strata	Laplace	Bres-Lin	Cr-Sp	Exact	logLik
1 to 4	20	0.00009	0.00001	4.68e-16	1.03e-08	-0.8257
5 to 10	30	0.00020	0.00005	5.37e-16	1.24e-08	-3.4494
11 to 20	50	0.00032	0.00015	1.69e-12	7.05e-08	-7.8935
21 to 50	150	0.00054	0.00041	2.65e-11	1.43e-07	-18.1443
51 to 80	150	0.00065	0.00071	5.47e-09	1.60e-07	-30.6223
81 to 100	100	0.00073	0.00086	7.29e-08	2.11e-07	-44.5575

Finally, let us consider the approximation accuracies when the strata are large and the parameter values are only roughly correct. Now we simulate the model (7) with parameter values $\beta = (-1.3, 0.9)$, $\sigma^2 = 0.25$ with sample sizes from 120 to 960 (10 times each, in increments of 60, for a total of 150 strata) and repeat the calculations of log integrated likelihoods and their approximations at $\sigma^2 = 0.15$, first at correct fixed-effect values (in Table 5), and then with fixed-effect values η randomly simulated (independently of the data) in the range $[-1.5, -0.5]$ (in Table 6). The results are displayed, in the same format as before, in the following two Tables. All Crouch-Spiegelman calculations in the next two Tables are based on the same fixed value $\epsilon = 10^{-15}$. Adjustment of ϵ with larger stratum size will be discussed in connection with Table 7 below.

TABLE 5. Summary of average absolute errors in log-likelihood for indicated stratum-size groups, for incorrect fixed and random-effect parameters, for stratum-sizes 120—960 using Laplace, Breslow-Lin, or Crouch-Spiegelman approximations with respect to 'exact' integrated log-likelihood. All evaluations were done at the correct fixed-effect values with $\sigma^2 = 0.15$ (instead of the correct value 0.25).

Stratum-size	# strata	Laplace	Bres-Lin	Cr-Sp	Exact	log-Lik
120 to 240	30	0.00114	0.00115	0.00053	2.66e-07	-97.442
300 to 420	30	0.00082	0.00115	0.00852	1.20e-07	-178.773
480 to 600	30	0.00068	0.00096	0.02565	1.88e-07	-262.607
660 to 780	30	0.00056	0.00073	0.06514	2.49e-07	-361.094
840 to 960	30	0.00045	0.00059	0.14040	2.54e-07	-474.394

TABLE 6. Summary of average absolute errors in log-likelihood for indicated stratum-size groups, for incorrect fixed and random-effect parameters, for stratum-sizes 120—960 using Laplace, Breslow-Lin, or Crouch-Spiegelman approximations with respect to 'exact' integrated log-likelihood. All log-likelihoods and approximations were evaluated at $\sigma^2 = 0.15$ and at incorrect randomly re-simulated fixed-effect values.

Stratum-size	# strata	Laplace	Bres-Lin	Cr-Sp	Exact	log-Lik
120 to 240	30	0.00118	0.00103	0.00056	3.99e-07	-99.663
300 to 420	30	0.00084	0.00101	0.00968	1.76e-07	-181.605
480 to 600	30	0.00067	0.00086	0.02502	2.12e-07	-265.993
660 to 780	30	0.00054	0.00068	0.05060	3.10e-07	-363.539
840 to 960	30	0.00044	0.00055	0.12375	2.44e-07	-476.007

The message of Tables 5 and 6 is clear. For fixed absolute-accuracy parameter ϵ, the *relative* accuracy of the Crouch-Spiegelman series approximation to likelihood deteriorates sufficiently, when the stratum likelihood-contribution is very small (i.e., log-likelihood is 100 or more) that its corresponding log-likelihood approximation is much less reliable than the standard Laplace method approximation. The Breslow-Lin approximation is roughly comparable to the Laplace in these Tables because the random-effect variance-parameter at which evaluations are made is small, but in large strata (say, with count of 75 or more) there appears to be no reason ever to prefer Breslow-Lin to Laplace.

Next we address the issue of modifiying the absolute-accuracy parameter ϵ with larger stratum sizes in calculating the Crouch-Spiegelman approximation. We have seen in the previous few Tables that strata in the range of sizes 100 to 1000 result in very poor series aproximations for log-likelihood when the error parameter ϵ is fixed at the default value 10^{-15}. In Table 7, we now exhibit in the setting of Table 5 the increase in accuracy and computational work arising by specifing successively smaller values for ϵ. In the Table, the 'exponent' referred to in the row indexing is the negative logarithm to base 10 of the ϵ chosen. For the given range of exponents, the computational effort increased somewhat faster than linearly with the exponent, as measured by the number of terms in the series, shown in the final column. The Splus code for the function implementing the Crouch-Spiegelman approximation for log-likelihood is given in Appendix 7 below. So far, this code applies only to the case $J_i \equiv 1$ of model (1) where random-effect covariates Z_i are constant over each stratum.

TABLE 7. Summary of average absolute errors in Crouch-Spiegelman approximated log-likelihood for indicated stratum-size groups, for the same data and evaluation-point as for Table 5. Each row corresponds to re-calculated approximation with different $\epsilon = 10^k$, and the final column shows the number of summands in the resulting series approximating log-likelihood for each stratum.

Exponent	Size 120–240	Size 300–420	Size 480–600	Size 660–780	Size 840–960	Number of terms
15	9.953e-06	6.646e-04	3.656e-03	1.629e-02	0.03940	31
20	9.953e-06	6.646e-04	3.656e-03	1.629e-02	0.03940	43
25	1.327e-07	2.372e-05	3.092e-04	3.189e-03	0.01077	59
30	5.770e-10	1.244e-06	3.527e-05	4.646e-04	0.00195	75
35	2.426e-12	2.290e-08	1.854e-06	5.635e-05	0.00022	91

4 Accurate Maximum Likelihood Estimation

The paper of Breslow & Lin (1995) makes clear that in some problems the (higher-order) Laplace approximation to the log-likelihood of mixed logistic models is sufficiently accurate and well-behaved that it can be used to find approximate maximum likelihood estimators including estimators of variance

components, even in settings where the usual Laplace method cannot be guaranteed to have a local maximum in σ^2 near the true value. Although that paper primarily considered small strata (matched pairs binary-outcome logistic regression), it seems natural to try to obtain approximate maximum likelihood (ML) parameter estimators within the mixed-effect logistic model (1) by maximizing the Breslow-Lin approximate log-likelihood. The simplest way do that is to code the approximate log-likelihood as a function which is to be optimized using some standard numerical function maximizer, such as *nlminb* in **Splus**. This approach works fairly well (i.e. converges to roughly correct values) in several mixed logistic problems where we have tried it, but it can produce seriously biased estimators, as Breslow & Lin (1995) point out.

The Crouch-Spiegelman series (6) not only provides accurate approximations to the log-likelihood, but conveniently and simultaneously provides first and second derivatives which can be used in Newton-Raphson iterations to find the maximum-log-likelihood parameter estimates. To see clearly why this is so, restrict attention to the case where $J = 1$ in (5). (Analogous but messier calculations can be made in the more general case of (5) where $J > 1$.) In that case, with y replaced by k, we denote $L(k,n) = L(k,n,\eta,\sigma) = L(k,n,x\beta,\sigma)$. It is easy to check that (again with $\tau \equiv \sigma^2$)

$$\begin{aligned}
\nabla_\beta L(k,n,x\beta,\sigma) &= x^{tr}\left\{k\,L(k,n+1) - ,\, L(k+1,n+1)\right\} \\
\nabla_\tau L(k,n,x\beta,\sigma) &= \frac{1}{2}\left\{k^2\,L(k,n) \;-\; (2k+1)n\,L(k+1,n+1) \;+\right. \\
&\qquad \left. n(n+1)\,L(k+2,n+2)\right\}
\end{aligned}$$

Higher derivatives can easily be found by iterating these formulas.

Using the last displayed formulas, derivatives of log-likelihood with respect to parameters can be obtained from Crouch-Spiegelman-type series, evaluated at the same points as the series (6). The integrals against the normal density to which the Crouch-Spiegelman approximations apply in this way can be alternatively displayed in terms of the integrand $h(u) \equiv h(x\beta + u\sigma)$ (where x is a row vector) and parameter $\tau \equiv \sigma^2$, in the form

$$\nabla_\beta L(k,n,\,\beta\cdot\eta,\,\sigma) \;=\; x^{tr}\int (k-nh)\,h^k\,(1-h)^{n-k}\,\phi(u)\,du$$

$$\nabla_\beta \nabla_\beta^{tr}\, L(k,n,\, x\beta,\, \sigma) \;=\; 2\, x^{tr} x\, \frac{\partial}{\partial \tau}\, L(k,n,\, \beta\cdot\eta,\, \sigma) \;=$$

$$x^{tr} x \int \Big\{ (k-nh)^2 - nh(1-h) \Big\}\, h^k\, (1-h)^{n-k}\, \phi(u)\, du$$

$$\nabla_\beta \frac{\partial}{\partial \tau}\, L(k,n,\, x\beta,\, \sigma) \;=\; \frac{1}{2}\, x^{tr} \int \Big\{ (k-nh)^3 - 3nh(1-h)(k-nh)$$

$$-nh(1-h)(1-2h) \Big\}\, h^k\, (1-h)^{n-k}\, \phi(u)\, du$$

$$\frac{\partial^2}{\partial \tau^2}\, L(k,n,x\beta,\sigma) \;=\; \frac{1}{4} \int \Big\{ (k-nh)^4 - 6nh(1-h)(k-nh)^2 - 4nh(1-h)(1-2h)(k-nh)$$

$$+\, 3(nh(1-h))^2 + nh(1-h)(6h(1-h)-1) \Big\}\, h^k\, (1-h)^{n-k}\, \phi(u)\, du$$

These formulas have been used to implement in **Splus** (so far, only for the case $J = 1$) a Newton-Raphson iteration which very rapidly and efficiently calculates the maximum-likelihood estimators, even for problems with very many (eg, hundreds) of strata of sizes ranging from 1 into the hundreds. This is in stark contrast with the performance of commercial software based on iteration ideas involving linearized log-likelihoods and algorithms for mixed linear models. As an illustration of how well this Newton-Raphson algorithm (**Splus** code for which is given in Appendix 7) works, it was applied to the dataset of 500 strata described just after model-equation (7), which was used in generating Tables 1 and 2. Recall that the correct parameter values for that simulated dataset were $\beta = c(-1.5, 0.6)$, $\sigma^2 = 0.5625$. Starting from the not-very-good initial values $\beta = (-1, 0)$ and $\sigma^2 = 0.1$, the Newton-Raphson iteration took 8 seconds on a *Sun Sparc V* workstation, arriving in 9 steps at the MLE $(-1.4878, 0.5796)$ for β and 0.5399 for σ^2.

A more difficult example of approximate Maximum Likelihood estimation is given by the dataset of very large strata simulated in connection with Table 5. When the starting values $\beta = (1, 0)$ and $\sigma^2 = 0.1$ were used, with the accuracy parameter ϵ for Crouch-Spiegelman set at 10^{-15}, ten Newton-Raphson steps produced the following sequence of log-likelihood values:

```
-41754.33 -41289.38 -41273.55 -41266.35 -41323.88 -41384.16
-41297.40 -41290.09 -41292.56 -41295.34
```

Thus no convergence was to be expected with that coarse level of accuracy. (Recall that the per-stratum errors shown in Table 7 with $-\log_{10}(\epsilon) = 15$

were unacceptably large.) With the same initial values and $-\log_{10}(\epsilon) = 35$, the sequence of log-likelihood values upon 20 successive Newton-Raphson iterations was

```
-41754.02 -41406.87 -41286.39 -41249.87 -41231.52 -41225.01
-41224.97 -41224.93 -41224.89 -41224.85 -41224.81 -41224.76
-41224.70 -41224.48 -41224.49 -41224.50 -41224.52 -41224.55
-41224.69 -41224.65
```

and the final MLE's were $(-1.2502, 0.9454)$ for β, when the the correct values were $(-1.3, 0.9)$, and 0.275 for σ^2, which should be compared with the correct value of 0.25. Here also there was no ultimate convergence via Newton-Raphson, but restarting the iteration using the Splus function *nlminb* evaluated at the Laplace log-likelihood approximation resulted in rapid convergence with log-likelihood -41222.6 (which was evaluated as -41222.72 using the Laplace method approximation) at the values $\hat{\beta} = (-1.2432, 0.9439)$, $\hat{\sigma}^2 = 0.2119$.

For strata of sizes larger than 1000, it seems idle to pursue maximum log-likelihood calculations based upon either the Breslow-Lin or Crouch-Spiegelman approximations. Based upon Tables 5 through 7, the standard Laplace method appears quite serviceable and accurate for large strata. For parameter estimation based upon extremely large strata ($n > 1000$), an effective method justified and explored in Slud (1998) is based upon a variance-stabilizing transformation to yield a mixed-effect nonlinear regression model, within which parameter-estimates and analysis of deviance can be performed using existing software such as the **Splus** function *nlme*.

5 Summary & Discussion

The approximations and numerical computations of this paper establish that log-likelihoods for the mixed-effect logistic model (1) with a single vector of random effects per stratum can be effectively and accurately calculated either by the well-known Laplace steepest-descent approach (Breslow & Lin 1995) to calculating likelihood integrals or by a series approximation of Crouch & Spiegelman (1990) based on complex residue integrals. For most applications,

involving strata of size at most a few hundred, the most accurate and still easily computed approximations will be those of Crouch-Spiegelman, which are also ideally suited to Newton-Raphson maximum likelihood estimation. However, the Laplace-method approximation to log-likelihood (or the modified approximation of Breslow & Lin, for strata of size no bigger than 50) will also be adequate (accurate to two or three decimal place accuracy) for many purposes involving single evaluations of mixed logistic likelihood terms. One such purpose is the calculation of small-area estimators based on mixed-effect logistic models, as in Booth & Hobert (1998). There the quantity of interest is the conditional expected response probability given the response data for (a random sample taken from) the stratum. In the notation of this paper, restricted to the case $J_i = 1$ of (1), based on data from the entire stratum this conditional expectation is given by

$$E\Big(h(x_i\beta + zu_i) \,\Big|\, y_i\Big) \;=\; \frac{L(y_i + 1,\, n_i + 1,\, x_i\beta,\, \sigma)}{L(y_i,\, n_i,\, x_i\beta,\, \sigma)}$$

and the numerator and denominator of the displayed fraction can each be effectively calculated by either a Crouch-Spiegelman or Laplace method approximation.

For analysis of mixed logistic models with very large strata (say of size 1000 or more), the Crouch-Spiegelman approximate log-likelihoods are no longer reliable, but the standard Laplace method calculates log-likelihood satisfactorily, with relative error at worst one part per million, and numerical log-likelihood maximization based upon either the standard Laplace approximation or its Breslow-Lin refinement yields usable parameter estimators.

The recommended method of computing log-likelihoods within the class (1) of mixed logistic regression models is to use the Crouch-Spiegelman series (6) for strata of size up to several hundreds, and the Laplace approximation L_{Lap} for larger strata. The recommended method to maximize log-likelihood numerically is either to use a general numerical optimizer (such as *nlminb* in **Splus**) on this hybrid log-likelihood approximation, or, if no strata are larger than 1000, to apply a Newton-Raphson iteration based on Crouch-Spiegelman approximation of log-likelihood and its derivatives up to order two in the parameters. For parameter estimation when all strata are very large, a good alternative is provided by the mixed nonlinear-regression parameter estimates using **nlme** based on transformed data as in Slud (1998).

Acknowledgment. This research was partially supported by the Bureau of the Census and the American Statistical Association through an ASA/Census Fellowship.

6 Bibliography

Breslow, N. & Clayton, D. (1993) Approximate inference in generalized linear mixed models. *Jour. Amer. Statist. Assoc.* **88**, 9-25.

Breslow, N. & Lin, X. (1995) Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* **82**, 81-91.

Crouch, E. & Spiegelman, D. (1990) The evaluation of integrals of the form $\int f(t)e^{-t^2}dt$: Application to Logistic-normal models. *Jour. Amer. Statist. Assoc.* **85**, 464-469.

Goldstein, H. (1991) Nonlinear multilevel models, with an application to discrete response data. *Biometrika* **78**, 45-51.

Goldstein, H. (1995) **Multilevel Statistical Models**, 2nd ed. London: Edward Arnold.

Lindstrom, M. & Bates, D. (1990) Nonlinear mixed effects models for repeated measures data. *Biometrics* **46**, 673-87.

Littell, R., Milliken, G., Stroup, W. & Wolfinger, R. (1996) **SAS System for Mixed Models**, SAS Institute: Cary NC.

Pinheiro, J. & Bates, D. (1995a) lme and nmle: Mixed Effects, Models, Methods, and Classes for S and Splus, version 1.2. Documentation posted in StatLib, http://www.statlib.cmu.edu

Pinheiro, J. & Bates, D. (1995b) Approximations of the loglikelihood function in the nonlinear mixed effects model. *Jour. of Comp. & Graph. Statist.* **4**, 12-35.

Slud, E. (1998) Logistic regression with large cell-counts and multiple-level random effects. Preprint.

Wolfinger, R. & O'Connell, M. (1993) Generalized linear models: a pseudo-likelihood approach. *Jour. Statist. Comput. & Simul.* **48**, 233-43.

Zhou, X., Perkins, A., & Hui, Siu (1999) Comparisons of software packages for generalized linear multilevel models. *American Statistician* **53**, 282-290.

7 Appendix: Splus Code

The function *ResAppr* reproduced here calculates (for the case $J = 1$) both the Crouch-Spiegelman approximation to (log-) likelihood and (optionally and simultaneously) the gradients and Hessian with respect to parameters β and σ^2 in (1).

```
ResAppr
function(eta, sigsq, rspct, cellct, GrdHess = F, tolexp = 15)
{
# Required input parameters are:
#   vector of fixed-effects  eta (= x beta), 1  per stratum)
#   variance-component which may either be scalar or vector
#   vector of response counts, by stratum; and
#   vector of stratum-sizes.
# Optional parameters are:
#   boolean indicator GrdHess which if true signals
#       calculation of 1st & 2nd derivatives with respect
#       to beta and sigsq intercepts;
#   tolexp is  negative base-10 log of epsilon for all strata.
   nstrat <- length(eta)
   af <- sqrt(2 * sigsq)
   afmax <- max(af)
   if(length(sigsq) == 1)
      af <- rep(af, nstrat)
   alpha <- sqrt(log(2 * sqrt(pi)) + tolexp * log(10))
   if(afmax * alpha < pi/2) {
```

```
      Kap <- alpha
      dstep <- pi/Kap
   }
   else {
      Kap <- pi/(2 * afmax)
      dstep <- (2 * pi * Kap)/(Kap^2 + alpha^2)
   }
   krng <- c(-1, 1) * floor(0.999 + sqrt(tolexp *
      log(10))/dstep)
   klen <- krng[2] - krng[1] + 1
   assign("uaux", outer(rep(1, nstrat), dstep *
      (krng[1]:krng[2])), frame = 0)
   assign("paux", array(0, dim = c(nstrat, klen)), frame = 0)
   assign("Qaux", paux, frame = 0)
   assign("Yfr", rspct/cellct, frame = 0)
   if(GrdHess)
      assign("GrHaux", array(0, dim = c(nstrat, 4)), frame = 0)
# The 4 columns of GrHaux respectively concern the derivatives:
#   (1) 1st deriv with respect to beta coeff's;
#   (2) 1st deriv wrt sigma, = 2nd deriv for beta-intercept
#   (3) mixed beta & sigma derivative,   and
#   (4) second-order sigma derivative
   assign("paux", plogis(eta + af * uaux), frame = 0)
   assign("Qaux", (dstep/sqrt(pi)) * exp(rspct * log(paux) +
      (cellct - rspct) * log(1 - paux) - uaux^2), frame = 0)
   app1 <- c(Qaux %*% rep(1, klen))
   if(GrdHess) {
      kdev <- Yfr - paux
      pqovn <- (paux * (1 - paux))/cellct
      pmqon <- (1 - 2 * paux)/cellct
      assign("GrHaux", cbind((kdev * Qaux) %*% rep(1, klen),
 ((kdev^2 - pqovn) * Qaux) %*% rep(1, klen),
         (((kdev^3) - 3 * pqovn * kdev - pqovn * pmqon) *
         Qaux) %*% rep(1, klen), ((kdev^4 - 6 * pqovn *
         (kdev^2) - 4 * kdev * pqovn * pmqon + 3 * pqovn^2 +
         (pqovn * (6 * cellct * pqovn - 1))/cellct^2) *
         Qaux) %*% rep(1, klen)), frame = 0)
   }
```

```
   bestvec <- app1
   if(GrdHess) {
      gradvec <- cellct * (GrHaux[, 1]/bestvec)
      hessvecs <- (outer(cellct, (2:4), function(x, y)
      x^y) * GrHaux[, 2:4])/bestvec
   }
   c(list(Approx = log(app1), krange = krng, Delta = dstep),
      if(GrdHess) list(Gradvec = gradvec, Hessvecs = hessvecs)
      else NULL)
}
```

A key advantage of the Crouch and Spiegelman method of mixed logistic log-likelihood approximation is that the first and second derivatives with respect to model parameters β, σ^2 in the model (1) are readily calculated in parallel with the log-likelihood itself, enabling easy coding of Newton-Raphson iterations to find Maximum Likelihood parameter estimates. An Splus function to be used in tandem with the log-likelihood evaluation function *ResAppr* above is reproduced next. See Section 4 above for further discussion and an example of its use.

```
NRstpLgs
function(Dmat, beta.start, sigsq.start, rspct, cellct,
   minstp = 0.1, maxit = 10, tolexp = 15, toler = 1e-05)
{
   np <- length(beta.start)
   betasgsq.old <- rep( - Inf, np + 1)
   betasgsq <- c(beta.start, sigsq.start)
   nobs <- nrow(Dmat)
   loopind <- 0
   newLlik <- numeric(maxit)
   hessmat <- array(0, dim = rep(np + 1, 2))
   while(sum(abs(betasgsq - betasgsq.old)) > toler &
loopind < maxit) {
      sigsq <- betasgsq[np + 1]
      tmplist <- ResAppr(c(Dmat %*% betasgsq[1:np]), sigsq,
         rspct, cellct, GrdHess = T, tolexp = tolexp)
      loopind <- loopind + 1
```

```
      newLlik[loopind] <- sum(tmplist$Approx)
      gradllk <- cbind(tmplist$Gradvec * Dmat, 0.5 *
         tmplist$Hessvecs[, 1])
      hessmat[1:np, 1:np] <- t(Dmat) %*% (
         tmplist$Hessvecs[, 1] * Dmat)
      hessmat[1:np, np + 1] <- hessmat[np + 1, 1:np] <-
         0.5 * (t(Dmat) %*% tmplist$Hessvecs[, 2])
      hessmat[np + 1, np + 1] <- 0.25 *
         sum(tmplist$Hessvecs[, 3])
      hessmat <- hessmat - t(gradllk) %*% gradllk
      gradllk <- c(t(gradllk) %*% rep(1, nobs))
      betasgsq.old <- betasgsq
      Delta <- c(solve(hessmat, gradllk))
      if(Delta[np + 1] > betasgsq.old[np + 1]/2) {
         stp <- betasgsq.old[np + 1]/Delta[np + 1]
         minstp <- stp/3
      }
      else stp <- 2
      nextLlik <-  - Inf
      while(nextLlik < newLlik[loopind] & stp > minstp) {
         stp <- stp/2
         betasgsq <- betasgsq.old - stp * Delta
         nextLlik <- sum(ResAppr(c(Dmat %*% betasgsq[1:np]),
            betasgsq[1 + np], rspct, cellct)$
            Approx)
      }
   }
   list(neval = loopind, coeff = betasgsq, logLik =
      newLlik[1:loopind], lastGrad = gradllk, lastHess =
      hessmat, lastDelt = Delta, lastLlik = nextLlik)
}
```