Hierarchical Poisson Regression Modeling

Cindy L. Christiansen and Carl N. Morris

The Poisson model and analyses here feature nonexchangeable gamma distributions (although exchangeable following a scale transformation) for individual parameters, with standard deviations proportional to means. A relatively uninformative prior distribution for the shrinkage values eliminates the ill behavior of maximum likelihood estimators of the variance components. When tested in simulation studies, the resulting procedure provides better coverage probabilities and smaller risk than several other published rules, and thus works well from Bayesian and frequentist perspectives alike. The computations provide fast, accurate density approximations to individual parameters and to structural regression coefficients. The computer program is publicly available through Statlib.

KEY WORDS: Adjusted density method; Approximate Bayes methods; Empirical Bayes; Poisson regression interactive multilevel modeling; Restricted maximum likelihood.

1. INTRODUCTION

This article proposes a new procedure for Poisson data analyses and demonstrates its advantages over other procedures, including a popular sampling-based approach. The model allows varying exposures and individual shrinkage factors in a regression setting. The procedure for fitting the model calculates point estimates, variance estimates, and approximate distributions for all parameters. The available program is fast enough to be used interactively for model building and model fitting, and it also permits evaluation of the operating characteristics via repeated simulation. All tests made thus far of the internal validity and of external comparisons to other methods have been favorable.

Statistical applications and methods for hierarchical Poisson count and rate data are abundant. Home run data have been analyzed by Albert (1985, 1992) using various Poisson hierarchical methods. Poisson models have been used for literary analyses, as Efron and Thisted (1976) did with Shakespeare’s works and Mosteller and Wallace (1963) did for the Federalist papers. Many medical applications involve Poisson data, including the Rand Health Insurance Experiment work by Keeler and Rolph (1988) that estimated the effects of coinsurance on cost per treatment episode and the number of treatment episodes. Wolfe, Petroni, McLaughlin, and McMahon (1991) compared extra-Poisson variance models with Poisson error models for estimating diagnosis-specific hospital discharge rates. Disease incidence and mortality rates were analyzed by Bernardinelli and Montomoli (1992) using Gibbs methods and using Poisson empirical Bayes methods developed by Clayton and Kaldor (1987). Geographical variability in mortality rates is captured in the Poisson model used by Tsutakawa (1988). A “mixed-effects” Poisson regression model was offered by Vonesh (1990) to analyze risk factors associated with continuous ambulatory peritoneal dialysis.


We propose a new approach, Poisson regression interactive multilevel modeling (PRIMM), which has several advantages over the methods just referenced. First, PRIMM provides point and interval estimates for the individual and for the structural parameters, unlike some of the other methods. Second, in the cases we checked (and we believe, quite broadly), the nominal operating characteristics of PRIMM hold up better than the most commonly used alternative methods that we were able to consider (see Sec. 4). Third, although Monte Carlo–based Bayesian methods also can...
provide a full range of inferences, PRIMM’s computation speed (in our studies, approximately 1/30 of the time needed to run a Gibbs procedure through BUGS; cf. Sec. 4) permits its evaluation by repeated use in simulations, thus verifying that it has good frequentist properties. Fourth, good computational speed is due to using fitting procedures based on a standard linear model and on low-dimensional likelihood methods. That makes PRIMM an interactive tool for data analysis.

Because PRIMM provides inferences about individual and structural parameters, it is appropriate for many statistical applications. We now introduce basic terms and concepts for this model with a health services example. There are \( k = 11 \) renal transplantation hospitals in Massachusetts. Associated with each individual hospital \( i, i = 1, \ldots, k = 11 \), is a Poisson parameter \( \lambda_i \) that determines the 1-year mortality rate following renal transplantation at that hospital. Technically, we use the Poisson distribution here as a substitute for the sum of independent heterogeneous Bernoulli trials, with all probabilities small. The rates \( \lambda_1, \ldots, \lambda_k \) are called the “individual parameters.” The approximate distributions of these individual hospital parameters must be estimated to make inferences about mortality. Sometimes the individual parameters are predicted by \( r - 1 \) hospital level covariates (e.g., patient case mix) through a log-linear model involving regression coefficients, \( \beta_0, \ldots, \beta_{r-1} \). We call these regression coefficients the “structural parameters,” but sometimes others call them “population parameters.” The analysis of the structural parameters is appropriate when groups, clusters, correlated observations (Laird and Ware 1982; Zeger and Liang 1986), or overdispersion (equivalently, heterogeneity) exists. An additional structural parameter \( \zeta \) in our notation represents “between-individual” variability. The resulting multilevel (equivalently, hierarchical) analysis generally provides larger standard errors of these estimated structural regression coefficients, as is required to correctly reflect the separate levels of variation.

In Section 2 the model is described in two mathematically equivalent forms. At level 1, the “descriptive model” (Sec. 2.1) assumes that the observations follow Poisson distributions, conditionally on the individual rate parameters, possibly with different exposures. At level 2, it assumes that the \( k \) individual rates follow gamma distributions. The “inferential model” (Sec. 2.2), derived from the descriptive model, provides Poisson–gamma mixture (negative binomial) distributions to the observations, given the structural parameters. The \( k \) individual rates then follow independent gamma distributions, conditional on the data and on the structural parameters. In a third level, we allow a range of distributions for the structural parameters. Section 2 concludes by providing approximate distributions for the structural and the individual parameters, conditional only on the observed data. These approximations, some of which use a new theory, work well even for small \( k \).

Section 3 reports the PRIMM results for two datasets. Section 4 uses simulations to determine the operating characteristics in several examples for PRIMM and for six other methods described in Section 4.2. Section 4.3 gives the operating characteristics for PRIMM and for five methods that use a Poisson–gamma model (one an ideal Bayesian model where the hyperparameters are known), plus a method that assumes a Bayesian Poisson–lognormal model. Section 5 presents a summary and discussion. The Appendixes contain the proofs of results, information on the PRIMM computer program, and descriptions of several alternative Poisson methods tested here.

2. THE MODEL AND THE PRIMM METHOD

This section begins with a description of the model in Section 2.1 and concludes by determining the distributions, conditional on the data, of the structural parameters, \( \alpha = (\beta, \zeta) \), in Section 2.4, and of the individual level parameters \( \{\lambda_i\} \) in Section 2.5. The development of these procedures is provided in Sections 2.2–2.5, which specify the inferential model, the expressions for the log-likelihood function and for the posterior densities, the handling of nuisance parameters, and a new method for making beta distribution approximations to the shrinkage factors. Proofs of all theorems and lemmas appear in Appendix A.

2.1 The Descriptive Model

Level 1 of the descriptive model specifies the distributions of the observed data vector, \( z = (z_1, \ldots, z_k)' \), given the individual parameters \( \{\lambda_i\} \). Level 2 specifies gamma distributions for \( \{\lambda_i\} \), given the hyperparameters \( \alpha \equiv (\beta, \zeta) \).

Level 1: Individual Model. The observed counts \( z_1, \ldots, z_k \) have independent Poisson distributions with expected values \( E(z_i) = e_i \lambda_i, i = 1, \ldots, k \), for known exposures \( e_i > 0 \). Given \( \lambda = (\lambda_1, \ldots, \lambda_k)' \),

\[
z_i | \lambda_i \sim \text{Pois}(e_i \lambda_i) = \text{Pois}[e_i \lambda_i | e_i \lambda_i],
\]

\[
\text{independently } i = 1, \ldots, k. \tag{1}
\]

The observed rates, \( y_i \), with \( y_i \equiv z_i / e_i \) have expectations \( E(y_i) = \lambda_i \).

Notation. Throughout, as in (1), we use a square bracket convention with various distributions to indicate the mean and the variance of the random variable, whereas round parentheses indicate the standard parameter notation.

Level 2: Structural Model. The individual Poisson parameters \( \lambda_1, \ldots, \lambda_k \) follow conjugate gamma distributions for \( i = 1, \ldots, k \) independently, given the unknown hyperparameter vector \( \alpha = (\beta_0, \ldots, \beta_{r-1}, \zeta)' \), where \( r \) is the number of unknown regression coefficients. Thus

\[
\lambda_i(\alpha \sim \text{gam}(\zeta, \zeta / \mu_i)) = \text{gam}[\mu_i, \mu_i^2 / \zeta], \log(\mu_i) = x_i' \beta.
\]

The log link (i.e., the natural link for the Poisson distribution) is assumed for the structural means, so \( \log(\mu_i) = x_i' \beta \) for fixed covariates \( x_i = (x_{i0}, \ldots, x_{i(r-1)})' \), \( \beta = (\beta_0, \ldots, \beta_{r-1})' \), with the constant term represented by setting \( x_{i0} = 1 \). Here \( \zeta > 0 \) corresponds to an unobserved prior count, not necessarily an integer. The squared coefficient of variation of \( \lambda_i / \zeta \), is the same for all components in (2). The structural model (2) allows \( \mu_i \) to be known \( (r = 0) \) and
allows exchangeable (without covariates, \( r = 1 \)) and nonexchangeable (if \( r \geq 2 \)) distributions for \( \lambda_i \). In all cases standardized rates, \( \lambda_i / \mu_i \sim \text{Gamma}(\zeta, \zeta) \) are exchangeable, with unit expectation.

**Level 3: Distributions on the Structural Parameters.** The joint prior density for the \( r + 1 \)-dimensional hyperparameter vector is

\[
h(\beta, \zeta) = z_0 / (\zeta + z_0)^2
\]

with respect to \( d \zeta d \beta, (\beta', \zeta') \in \mathbb{R}^{r+1}. \) (3)

Equivalently,

\[
B_0 \equiv \zeta / (\zeta + z_0) \sim \text{Uniform}(0, 1).
\] (4)

This choice of an improper prior distribution for \( \beta \) is standard, made here to provide good repeated sampling properties to the resulting rules. We choose a proper prior distribution like (4) for \( \zeta \) because the maximum likelihood estimate of \( \zeta \) can occur at infinity (Aragon, Eberly, and Eberly 1992; Morris 1997). When that happens, maximum likelihood estimates of the shrinkages \( B_i = \zeta / (\zeta + e_i \mu_i) \) occur at the boundary point \( B_i = 1 \), interfering with further assessments and especially with evaluating precision. The uniform distribution for \( B_0 \) is relatively uninformative, giving both \( \zeta \) and \( 1 / \zeta \) infinite expectations. The constant \( z_0 \) is the median of \( \zeta \), so that small values of \( z_0 \) are less informative and encourage less shrinkage a priori. A conservative choice, one that mainly lets the data speak, would choose \( z_0 \) small enough so that \( z_0 < \zeta \). The mild conditions under which this model provides a proper posterior distribution are discussed in Section 2.4. Further justification of the choice of distribution (4) stems from the analogous hierarchical modeling problem with equal variance normal distributions. In that case the prior uniform distribution on a shrinkage factor provides minimax estimators of the \( k \) means whenever \( k \geq r + 6 \). The estimators are also admissible if \( r = 0 \) (Strawderman 1971).

Albert (1988) and Leonard and Novick (1986) provided inferences for a different two-level Poisson-gamma model, which we call “model 1” (Christiansen 1992), as opposed to “model 2” in (1)–(3). Our methods here account for unequal exposures \( e_i \) in (1), which were not treated in these two papers. Thus unlike theirs, our shrinkage factors \( B_i \), as defined in Section 2.2, must vary with \( e_i \). Almost all real Poisson data that we have encountered have had variable exposures. The Albert (1988) and Leonard and Novick (1986) model can be extended to include unequal \( e_i \) as in Albert and Pepple 1989 and Lu and Morris 1994. A second distinction between the two models lies in the parameterization of the gamma moments, with ours having a constant coefficient of variation. We chose this moment structure (2) because it holds also for the often-used level 2 lognormal specification. However, we know of no other reasons to prefer it. There is good reason to prefer the gamma to the lognormal specification at level 2, however (see secs. 2.2 and 4.3). Albert (1988) and Leonard and Novick (1986) used one of the hyperparameter distributions in (3), the one for the \( z_0 \) choice that arises naturally in their constant shrinkage formulation.

### 2.2 The Inferential Model

The marginal distributions of the observed data \( \{z_i\} \) are independent negative binomials, these being gamma mixtures of Poisson distributions. To fix notation, the negative binomial distribution, \( \text{NB}(\zeta, \mu) \equiv \text{NB}[\mu = \zeta p / (1 - p), \mu + \mu \zeta / \zeta] \) for \( \zeta > 0, 0 < p < 1 \) has density function

\[
\frac{\Gamma(\zeta + z)}{\Gamma(\zeta)!} p^z (1 - p)^{\zeta}, \quad z = 0, 1, 2 \ldots
\] (5)

As used here, (5) arises as the density of a mixture of Poisson(\( \Lambda \)) distributions on \( z \) if \( \Lambda \sim \text{Gam}(\zeta, 1 + (p / (1 - p))) \) is the mixing distribution. For fixed \( \mu \), large values of \( \zeta \) make the negative binomial approximately Poisson.

**Level 1: Marginal Model for the Observations.** The observed counts \( \{z_i\} \), given the hyperparameters \( \beta \) and \( \zeta \), have independent marginal distributions for \( i = 1, \ldots, k \)

\[
z_i | \beta, \zeta \sim \text{NB}(\zeta, 1 - B_i) = \text{NB}[e_i \mu_i, e_i \mu_i / \beta_i]
\] (6)

and

\[
B_i \equiv \zeta / (\zeta + e_i \mu_i)
\] (7)

with \( 0 < B_i < 1 \). For large \( \zeta \) or, equivalently, for “shrinkage factors” \( B_i \) near unity, the negative binomial distribution (6) is approximately the Poisson distribution. Smaller values of \( \zeta \) indicate extra variation beyond the Poisson variation in the true rate \( \lambda_i \).

**Level 2: Conditional Model for the Individual Parameters.** The conditional distributions (8) for the unobserved Poisson rate parameters, given \( (\beta, \zeta) \) and the data, are gamma distributions with means and variances that are linear in the data. (Throughout, “data” refers to \( \{z_i\} \) or, equivalently, to \( \{y_i\} \).)

\[
\lambda_i | \text{data, } \beta, \zeta \sim \text{Gam}(z_i + \zeta, e_i + \zeta / \mu_i) = \text{Gam}(\lambda^*_i, \sigma^*_i)^2
\] (8)

\( i = 1, \ldots, k \), independently, where the posterior mean \( \lambda^*_i \) and variances \( \sigma^*_i \) given the hyperparameters, are denoted by

\[
\lambda^*_i \equiv E(\lambda_i | \text{data, } \beta, \zeta) = (1 - B_i) y_i + B_i \mu_i
\] (9)

and

\[
\sigma^*_i \equiv \text{var}(\lambda_i | \text{data, } \beta, \zeta) = \lambda^*_i (1 - B_i) / e_i.
\] (10)

Linearity of the conditional mean (9) holds because the gamma distribution is conjugate to the Poisson likelihood. The gamma distribution choice for the structural model (2) is conservative (i.e., it provides minimax estimators for mean squared error) among all distributions with the same first two moments for this problem (Morris 1983a). Another popular choice of distribution for this problem, the lognormal distribution, does not provide linear conditional moments (and is not minimax). Thus the lognormal is not conservative, it does not produce shrinkage factors like \( B_i \).
naturally, and the posterior mean (9) is mathematically intractable. The gamma, however, leads to the simple negative binomial marginal form (8) with easily proven concave log-likelihood function for \( \beta \) (see Sec. 2.3).

Level 3 for the hyperparameters remains the same as in the descriptive model. The joint densities on \( (z, x, \beta, \zeta) \) are the same whether defined by the three components (1), (2), and (3) of the descriptive model or by the three components (6), (8), and (3) of the inferential model.

2.3 Likelihood Function for the Hyperparameters

The likelihood function of the hyperparameters, based on the marginal NB(\( \zeta, 1 - B_i \)) density (6), is

\[
L(\beta, \zeta) = \prod_{i=1}^{k} \frac{\Gamma(\zeta + z_i)}{\Gamma(\zeta)z_i!} (1 - B_i)^{z_i} B_i^\zeta. \tag{11}
\]

**Lemma 1.** Denote the log-likelihood function \( \mathcal{L}(\beta, \zeta) \equiv \log L(\beta, \zeta) \) as the logarithm of (11). Define \( M \equiv \max[z_1, \ldots, z_k], N_j \equiv \#\{i: z_i = j\} \) for \( j = 0, \ldots, M \), \( N_j \equiv \#\{i: z_i \geq j\} = n_j + \cdots + n_M \), \( \mathcal{Z} \equiv \sum z_i/k \), and \( X \equiv (x_1, \ldots, x_k) \) as the \( k \times r \) matrix of covariates. Then

\[
\mathcal{L}(\beta, \zeta) = \sum_{j=1}^{M} N_j \log(\zeta + j - 1) - k\mathcal{Z} \log(\zeta) + z'X\beta - \sum_{i=1}^{k} (\zeta + z_i) \log(1 + e_i \mu_i / \zeta). \tag{12}
\]

Fix \( \zeta > 0 \) and define \( \tau \equiv \log(\zeta) \) throughout. This log transformation is preferred partly to improve likelihood approximations. The log-likelihood of \( \beta \) alone, with \( \zeta \) fixed, involves only the last two terms of (12), these being precisely the log-likelihood function (although \( \zeta \) may not be integral) for a logistic regression model, with the ith case having \( z_i \) successes, \( \zeta \) failures, \( z_i + \zeta > 0 \) trials, offset \( \log(e_i / \zeta) \), and success probability \( 1 - B_i \). The score function, the \( r \)-dimensional gradient of (12) with respect to the vector \( \beta \), is

\[
\frac{\partial \mathcal{L}(\beta, \zeta)}{\partial \beta} = \sum_{i=1}^{k} (z_i - e_i \mu_i) B_i x_i. \tag{13}
\]

As is well known, any root \( \hat{\beta} \), (the subscript indicates dependence on \( \tau \)) of (13)—also the posterior mode in this plane—is unique, because (12) is concave in \( \beta \). It exists except in a few extreme cases on the \( \{z_i\} \). The negative Hessian of second derivatives with respect to \( \beta \), evaluated at \( \hat{\beta}_\tau \), is

\[
\hat{\mathbf{H}}_\tau \equiv -\frac{\partial^2 \mathcal{L}(\beta, \zeta)}{\partial(\beta^2)} \bigg|_{\hat{\beta}_\tau} = \mathbf{X}'\hat{\mathbf{D}}_\tau \mathbf{X} \tag{14}
\]

where \( \hat{\mathbf{D}}_\tau \) is the \( k \times k \) diagonal matrix having ith diagonal element \( e_i B_i \lambda_i^\tau > 0 \), with \( B_i \) in (7) and \( \lambda_i^\tau \) in (9) both evaluated at \( \mu_i = \exp(x_i' \hat{\beta}_\tau) \). Thus (14) is the observed information matrix with \( \tau \) known. Any logistic regression subroutine using the data \( \{z_i\} \), \( \zeta \), as described just before (13), quickly locates the root \( \hat{\beta}_\tau \) of (13) and provides the inverse of \( \hat{\mathbf{H}}_\tau \) as the approximate covariance matrix for \( \beta \).

The negative of the derivative of (13) with respect to \( \tau \), the cross-derivative, is

\[
-\frac{\partial^2 \mathcal{L}(\beta, \zeta)}{\partial \beta \partial \tau} = -\sum_{i=1}^{k} (z_i - e_i \mu_i) B_i (1 - B_i) x_i. \tag{15}
\]

Differentiating (13) with respect to \( \tau \), after first evaluating \( \beta_\tau \), gives \( 0 = \frac{\partial^2 \mathcal{L}(\beta, \zeta)}{\partial \beta \partial \tau} \). Differentiating \( \mathcal{L} \) and solving for \( \partial \hat{\beta}_\tau / \partial \tau \), a term needed in Section 2.4, gives

\[
\frac{\partial \hat{\beta}_\tau}{\partial \tau} = (\mathbf{X}'\hat{\mathbf{D}}_\tau \mathbf{X})^{-1} \sum_{i=1}^{k} (z_i - e_i \mu_i) B_i (1 - B_i) x_i, \tag{16}
\]

all terms in (16) being evaluated at \( \hat{\beta}_\tau \).

Because the expectation of (15) is 0, \( \beta \) and \( \tau \) become asymptotically independent, given the data, as \( k \to \infty \). This usually permits only minor statistical dependence between \( \beta \) and \( \tau \) for finite \( k \), and allows for only mild functional dependence of \( \hat{\beta}_\tau \) on \( \tau \). These mild dependencies simplify algorithms for maximizing over \( \beta \) and \( \tau \), improve the small-sample approximations developed in this section, and aid interpretations in applications to data.

The derivatives of the log-likelihood (12) with respect to \( \tau = \log(\zeta) \) are

\[
\frac{\partial \mathcal{L}(\beta, \zeta)}{\partial \tau} = \sum_{j=1}^{M} N_j \left( \frac{\zeta}{\zeta + j - 1} \right) + \zeta \sum_{i=1}^{k} \log(B_i) - \sum_{i=1}^{k} (z_i - e_i \mu_i) B_i \tag{17}
\]

\[
-\frac{\partial^2 \mathcal{L}(\beta, \zeta)}{\partial \tau^2} = \sum_{j=1}^{M} N_j \left( \frac{\zeta}{\zeta + j - 1} \right)^2 - \sum_{i=1}^{k} z_i B_i + \sum_{i=1}^{k} (z_i - e_i \mu_i) B_i (1 - B_i) - \frac{\partial \mathcal{L}(\beta, \zeta)}{\partial \tau}. \tag{18}
\]

Assuming that there is a constant term, \( x_1 = 1 \), the last term in (17) is \( \partial \mathcal{L}(\beta, \zeta)/\partial \beta_0 \), which vanishes at \( \hat{\beta}_\tau \).

2.4 Distributions of the Hyperparameters \( \beta \) and \( \tau = \log(\zeta) \)

The analysis here and in Section 2.5 is similar to that for maximum likelihood, but with several adjustments to improve performance for small and moderate \( k \). The first adjustment, introduced in Section 2.1, uses a proper prior distribution (3) on \( \zeta \), to ensure that the mode occurs at a finite value and also to ensure that \( \zeta \) has a proper posterior distribution. The posterior density \( p(\beta, \tau) \) may be written, for some constant \( c_0 \), as

\[
p(\beta, \tau) \equiv c_0 L(\beta, \zeta | \text{data}) \zeta/(\zeta + z_0)^2. \tag{19}
\]

This posterior density (19) can be shown to be proper (i.e., has a finite integral) provided that \( k - n_0 \), the
number of cases for which \( z_i > 0 \), is at least \( r \) and that the \((k - n_0) \times r\) submatrix of \( X \) for these cases is of full rank. This mild (and probably not necessary) condition is satisfied in almost all examples. The proof (not given due to space considerations and lack of central interest) shows that for each fixed \( \zeta \), the integral over \((\beta_1, \ldots, \beta_{r-1})\) of (11) has a finite upper bound under these conditions, with the bound independent of \( \zeta \). Thus the final integral over \( \zeta \) is finite, because (4) is a proper distribution on \( \zeta \).

A second adjustment, a restricted maximum likelihood (REML) type correction, takes into account using a modal estimate of \( \tau \) in the presence of the nuisance parameter \( \beta \). One computes the mode of the approximate marginal density,

\[
p_2(\tau) \equiv c_2 |\hat{\Theta}|^{-1/2} L(\hat{\beta}, \zeta | \text{data}) \zeta / (\zeta + z_0)^2. \tag{20}
\]

The usual REML approach maximizes (20) (without the final factor \((\zeta + z_0)^2\), of course) with respect to \( \tau \) (or \( \zeta \)). This produces a different and better estimate for \( \tau \) than that obtained by maximizing (19) (Harville 1977; Liang and Zeger 1995; Patterson and Thompson 1971; Reid 1995; Searle, Casella, and McCulloch 1992). Because the REML correction \(|\hat{\Theta}|^{-1/2}\) is complex and only approximate, we instead approximate \(|\hat{\Theta}|^{-1/2}\), the geometric mean of the \( \hat{\Theta}\) eigenvalues, by a constant multiple of the geometric mean of the \( k \) values \( \zeta / (\zeta + z_0) \), with \( m_0 \) being a constant chosen approximately equal to \( E(\sum z_i / \sum e_i) \). This simplifies differentiation of (20) substantially and makes almost the same adjustment as does the use of \(|\hat{\Theta}|^{-1/2}\). Thus we choose \( \beta \) and \( \bar{\tau} \) to maximize the following approximation to the log density, with respect to \( d\tau \), of \( (\beta, \tau)\):

\[
l_R(\beta, \tau) \equiv \log \left\{ L(\beta, \zeta | \text{data}) \frac{\zeta}{(\zeta + z_0)^2} \times \left\{ \zeta / (\zeta + e_i m_0) \right\}^{-r/2k} \right\} \\
= L(\beta, \zeta) + (1 - r/2) \tau - 2 \log(\exp(\tau) + z_0) \\
+ \frac{r}{2k} \sum_{i=1}^{k} \log(\exp(\tau) + e_i m_0), \tag{21}
\]

where \( L \) is the log-likelihood in (12). The subscript \( R \) is used on \( l_R(\beta, \tau) \) to remind us that the REML-type adjustment is made with this log density.

The first and second derivatives of \( l_R(\beta, \zeta) \) in (21) with respect to \( \beta \) and the cross-derivative \( -\partial^2 l_R(\beta, \tau) / \partial \beta \partial \tau \) are the same as for \( L(\beta, \zeta) \) and so are given by (13), (14), and (15). The derivatives of (21) with respect to \( \tau \) are

\[
- \frac{\partial^2 l_R(\beta, \tau)}{\partial \tau^2} = -\frac{\partial^2 L(\beta, \zeta)}{\partial \tau^2} + \frac{2z_0 \exp(\tau)}{(\exp(\tau) + z_0)^2} \\
- \frac{r}{2k} \sum_{i=1}^{k} \frac{e_i m_0 \exp(\tau)}{(\exp(\tau) + e_i m_0)^2}, \tag{22}
\]

with the leading terms in (22) and (23) provided by (17) and (18). Finding the root \( \bar{\beta} \) of (13) for each \( \tau \) has already been discussed as a logistic regression, and \( \bar{\tau} \), the root of (22), can be found easily using (22) and (23) in Newton’s method. At convergence, the inverse of the resulting Hessian given by (14), (15), and (23), all evaluated at \( \bar{\tau} \) and \( \bar{\beta} \equiv \bar{\beta}_R \), provides the approximate covariance matrix for \((\beta, \tau)\). Thus we have Result 1 for the hyperparameter distributions, which holds asymptotically as \( k \to \infty \). More important, Result 1 is designed to provide accurate approximations for relatively small \( k \).

**Result 1: Approximate hyperparameter distribution.** Assuming the model (1), (2), (3), the joint distribution of \( \beta \) and \( \tau \), given the data, is approximately

\[
\left( \begin{array}{c} \beta \\ \tau \end{array} \right) | \text{data} \sim \mathcal{N}_{r+1} \left( \left( \begin{array}{c} \bar{\beta} \\ \bar{\tau} \end{array} \right), \Sigma \right) \equiv \frac{\partial^2}{\partial \tau^2} \left( \frac{\exp(\tau \hat{\Theta}^T X)}{\hat{\psi}} \right)^{-1} \\
= \left( \begin{array}{c} \psi \end{array} \right) \right) \\
\times \left( \begin{array}{c} \hat{\psi} \end{array} \right)^{-1}, \tag{24}
\]

Here \( \psi \equiv \partial \bar{\beta} / \partial \tau \) (see (16)) and

\[
\hat{\psi} \equiv \text{var}(\tau | \text{data}) \\
\equiv \left( -\frac{\partial^2 l_R(\beta, \tau)}{\partial \tau^2} \right)^{-1} \frac{\partial^2 L(\beta, \zeta)}{\partial \beta \partial \tau} \left( \psi' \hat{\Theta}^T X \psi \right)^{-1} \tag{25}
\]

(see (23)), with all terms evaluated at \( \bar{\beta} \) and \( \bar{\tau} \). The right side of (25) also is \( -\partial^2 l_R(\bar{\beta}, \bar{\tau}) / \partial \tau^2 \), evaluated at \( \bar{\beta}, \bar{\tau} \).

### 2.5 Distributions of the Parameters \( \lambda_i \)

Having obtained the approximation (24), it remains to use it to calculate the distribution of each \( \lambda_i \). The means \( E\lambda_i \) and the variances \( \text{var}(\lambda_i) \) are obtained ultimately in Theorem 1, with the final distributional approximations in Result 2. All moments in this section assume (24) and are conditional on the data, although that often is omitted from the notation. To start, the mean and variance of \( \mu_i \equiv \exp(\chi_i' \beta) \) are

\[
\hat{\mu}_i \equiv E\mu_i | \text{data} = \exp(\chi_i' \hat{\beta} + \chi_i' \Sigma_{11} \chi_i / 2) \tag{26}
\]

and

\[
\hat{\sigma}^2_{\mu_i} \equiv \text{var}(\mu_i | \text{data}) = \hat{\mu}_i^2 \left( \exp(\chi_i' \Sigma_{11} \chi_i) - 1 \right), \tag{27}
\]

where \( \Sigma_{11} \) is the \( \times \) covariance matrix for \( \beta \), from (24). These results follow immediately from the moment-generating function of the normally distributed \( \chi_i' \beta \), per (24), given the data.

The mean of \( \lambda_i \) is \( E\lambda_i = E\lambda_i^* = (1 - EB_i) \gamma_i + EB_i \mu_i \) from (9), whereas \( \text{var}(\lambda_i) = E\lambda_i^* (1 - B_i) / e_i + \text{var}(\lambda_i^*) \)
from (10) and (9). These calculations require expressions for \(E \mu_s^t B_t^s\) for \(s, t = 0, 1, 2\). Lemma 2 reduces the two-dimensional expectations \(E \mu_s^t B_t^s\) to one-dimensional integrals. The notation \(E_{\mu_s^t B_t^s} \equiv (B_t^s(1 - B_t^s)) \mu_s^t B_t^s f_{0_t}(B_t^s) dB_t^s\) is used. Lemma 2 specifies \(f_{0_t}(B_t^s)\), the density of \(B_t^s\), given the data.

Lemma 2. Assume the multivariate normal distribution (24) for \((\beta', \tau')\) and define \(b_t = \text{cov}(x_t | \beta, \tau - x_t' \beta) / \text{var}(\tau - x_t' \beta)\). Then for \(t \geq 0\),

\[
E_{\mu_t^s B_t^s} \equiv \frac{E \mu_t^s B_t^s}{E \mu_t^s} = \frac{E B_t^s(1 - B_t^s) \mu_t^s B_t^s}{E(B_t^s(1 - B_t^s))},
\]

(28)\]

Even with the simplification provided by Lemma 2, there are too many integrals of the form (28) to evaluate quickly. We use the adjusted density method (ADM) to approximate the integrals in (28). The ADM, paralleling maximum likelihood estimation methods, uses first and second derivatives of the log density to fit an approximating Pearson family (see Morris 1988a). Maximum likelihood methods correspond to the normal distribution. In this application, because moments of \(B_t^s\) are needed and \(0 < B_t < 1\), the beta distribution is chosen as the approximating Pearson family.

The ADM works as follows for the beta distribution. Given \(0 < B < 1\), with a unimodal density function \(p(B)\), define \(l(B) = \log(p(B)B(1 - B))\), the extra factor \(B(1 - B)\) is the “adjustment” to \(p(B)\) required for the beta distribution. Let \(B\) maximize \(l(B)\), so that \(l' = 0\), and define \(i_0 = -l''(B)\) as the “Pearson information.” If \(p(B)\) is the beta(a1, a2) density, proportional to \(B^{a1-1}(1 - B)^{a2-1}\), then the operations yield \(B = a_1/(a_1 + a_2)\) and \(i_0 = (a_1 + a_2)/(B(1 - B))\). Thus \(a_1 = B^{a1-1}(1 - B)^{a2-1}\) and \(a_2 = (1 - B)^{a1-1}B^{a2-1}\), and so

\[
B \sim \text{Beta}(\hat{B}^2(1 - \hat{B})^2, \hat{B}(1 - \hat{B})^2)
\]

(29)

is the ADM approximation. Moment approximations are the beta distribution moments from (29). Lemma 3 shows how to evaluate (29), and then (28), when \(u = \log(B)\) has a normal distribution.

Lemma 3. Let \(u \sim N(\mu, \sigma^2)\) and \(B \equiv e^u/(1 + e^u)\). Then the ADM approximation to the distribution of \(B\) is

\[
B \sim \text{Beta}(a_1, a_2)
\]

\[
= \text{Beta} \left[ \frac{\hat{B}^2(1 - \hat{B})}{\hat{B}(1 - \hat{B})} \right],
\]

(30)

where \(\hat{B}^2 \equiv e^u/(1 + e^u), a_1 = 1/(\sigma^2(1 - \hat{B}^2))\), and \(a_2 = 1/(\sigma^2\hat{B}^2)\). Thus when \(t \geq 0\) and \(-a_1 < w < a_2\), ADM approximations to (28) follow from

\[
E(B_t^s(1 - B_t^s))^{w} \equiv \frac{(a_1 + w)^{w(t)}}{(a_1 + w)^{w(t)}},
\]

(31)

defining \(a + w)^{w(t)}\). It is good luck that for this problem the beta distribution approximation (31) to the complicated terms of \(B_t^s\) in (28) takes such a simple form. We now have all of the terms needed to approximate the moments of each \(\lambda_t\).

Theorem 1. Assume the distribution (24) for \(\beta, \tau\), and given the data. For \(i = 1, \ldots, k\), the ADM approximations to the first two moments of \(\lambda_t\), given the data, are

\[
\hat{\lambda}_t = E(\lambda_t | data) = (1 - \hat{B}_t)y_t + \hat{u}_t E_1(B_t)
\]

(32)

and

\[
\hat{\sigma}_t^2 = \text{var}(\lambda_t | data)
\]

\[
= \frac{1}{\hat{\epsilon}_t} y_t E_0(1 - B_t)^2
\]

\[
+ \frac{1}{\hat{\epsilon}_t} \hat{u}_t E_1 B_t(1 - B_t) + E_2 B_t^2
\]

\[
- 2y_t \hat{u}_t E(1 - B_t^2) + y_t^2 E_0 B_t^2
\]

(33)

Here, defining \(\omega_t^2 \equiv \text{var}(\tau - x_t' \beta)\),

\[
E_0 B_t = \hat{B}_t = e^\tau/(e^\tau + e^\beta x_t' \beta) = a_{11}/(a_{11} + a_{12}),
\]

\[
a_{11} = \omega_{11}^{-2}/(1 - \hat{B}_t),
\]

\[
a_{12} = \omega_{12}^{-2}/\hat{B}_t,
\]

\[
\hat{u}_t = \text{cov}(x_t' \beta, \tau - x_t' \beta)/\omega_t^2
\]

from (24),

\[
\hat{u}_t = \text{cov}(x_t' \beta, \tau - x_t' \beta)/\omega_t^2
\]

from (24),

\[
E_1(B_t) = (a_{11} + b_t)/(a_{11} + a_{12}),
\]

\[
E_2(B_t) = (a_{11} + sb_t)(a_{11} + sb_t + 1)
\]

\[
+ (a_{11} + a_{12})(a_{11} + a_{12} + 1).
\]

The proof of Theorem 1 is given in Appendix A.

The following result holds asymptotically as \(k \to \infty\) and quite accurately (see Sec. 4 for numerical verification) for relatively small values of \(k\), thereby providing reliable interval estimates for each \(\lambda_t\).

Result 2: Marginal distributions of \(\{\lambda_t\}\). Given the data, \(\lambda_t\) has an approximate gamma distribution

\[
\lambda_t | data \sim \text{Gam}(\hat{\lambda}_t, \hat{\sigma}_t^2)
\]

\[
= \frac{\hat{\sigma}_t^2}{\hat{\lambda}_t} \text{Gam} \left( \frac{\hat{\lambda}_t}{\hat{\sigma}_t^2}, 1 \right),
\]

(34)

with \(\hat{\lambda}_t\) and \(\hat{\sigma}_t\), defined in (32) and (33).

Justification. The gamma distribution holds approximately because \(\lambda_t\) is gamma distributed conditionally on the data, given \(\beta\) and \(\tau\) (8). This approximation improves as these hyperparameters are more accurately estimated. Theorem 1 provides the moment estimates (32) and (33), which are matched to the first two gamma moments in (34).
3. DATA EXAMPLES

3.1 Example 1: Pump failure rates

Gaver and O’Muircheartaigh (1987) analyzed pump failures at a pressurized water reactor nuclear power plant using empirical Bayes maximum likelihood techniques. These data were also analyzed by Carlin and Gelfand (1991), Gelfand and Smith (1990), George, Makov, and Smith (1993), and Johnson (1992). None of their models involves regression.

The approach of Section 2 allows for nonexchangeability and here we model pump type in a regression form. The four pumps with the largest exposures were operated continuously, $x_{11} = 1$, whereas the first six pumps ran intermittently, $x_{11} = 0$; see Table 1. The exposures $e_i$ are in units of 1,048 hours of operation—about 6 weeks. (The original exposures represent units of 1,000 hours. What remains of the original exposures by a constant (.954) has no effect on the final results, but doing so clarifies the exposure pattern in these data.) The vector $x$ represents pump failures. Therefore, $y_i$ is the $i$th pump failure rate in failures per 1,048 hours with the average rate of 1.14 (for the intermittently operated pumps 1–6) and .22 for the continuously operated pumps 7–10. The weighted average is $\frac{\sum z_i}{\sum e_i} = .225$ (used for $n_0$ in (21)). We have sorted the data by the exposures, and so have included the original observation numbers (“Orig. no.”) from Gaver and O’Muircheartaigh (1987).

The results of the PRIMM analysis are given in the last four columns of Table 1. The regression mean estimates are $\hat{\mu}_i = 1.27, i = 1, \ldots, 6$ or the intermittently operated pumps and $\hat{\mu}_i = 25, i = 6, \ldots, 10$ for the continuously operated pumps. Note that $\hat{\mu}_i > \exp(\mu_1/\beta) = \exp(-.54 - .67x_{11})$; see (26). The standard error of $\hat{\beta}_1$ is .29, with an estimate to standard error ratio of -2.29, which indicates nonexchangeability among the two pump types and thus the need for the regression. The average shrinkage $\bar{B} = .16$. Within pump type, shrinkage estimates $B_i$ decrease with increasing exposure. The pumps with the largest exposures have posterior mean estimates very close to the observed failure rate, because their estimated shrinkages are small. Note that $B_0 = \frac{\zeta}{\zeta + .225} \sim \text{Unif}(0, 1)$ from (4), reflecting a prior median choice of $z_0 = .225$ for $\zeta$, which is noticeably less than $\zeta = .96$ and hence is a conservative choice. No matter how small one picks $z_0, \zeta \geq .86$ is obtained, and the average shrinkages exceed .15. These limits are achieved as $z_0$ approaches 0 and affect the prior and posterior mean estimates only slightly. The estimated correlations between $r$ and $(\beta_0, \beta_1)'$ are $(-.032, -.015)$, which support the approximate independence discussed after (16).

The PRIMM analysis without regression produces estimates of the structural and individual parameters similar to those in previously cited work. A benefit of any deterministic method, like PRIMM, over sampling-based fitting methods is that it will produce identical estimates for observations 1 and 2, for which the data are identical. The sampling-based methods that have been reported have produced different answers (although the estimates computed exactly are the same). Randomness can make the use of sampling-based estimates awkward for some legal and public policy applications.

3.2 Example 2: Heart Transplant Data

A second sample involves the 30-day mortality rates at $k = 15$ U.S. heart transplant centers operating from October 1987–December 1989; see Table 2. This subset was chosen as a systematic sample from the 131 centers analyzed by Christiansen and Morris 1996. Fewer than 8% of the 3,646 patients died within 1 month of their transplant, justifying the Poisson approximation to the binomial distribution, with exposure $e_i \equiv$ number of treated patients at transplant center $i$. The severity index $x_{11}$, used as a covariate here, reflects the expected number of deaths at center $i$ and is based on seven patient-level demographic and health risk variables for each patient treated at the center. Larger values of the index $x_{11}$ mean that the hospital treated a more difficult than average patient case mix.

The estimates of $\tau = \log(\zeta)$ (24), the shrinkages (30), and the first two moments $\lambda_i$ and $\bar{\sigma}_i$, of the mortality rate distributions (32) and (33) appear in Table 2. The average shrinkage estimate is .48, indicating that differences in the hospital mortality rates are not fully explained by the severity index. The estimated regression coefficient for the index of $\beta_1 = 1.19$ with a standard error of 1.85, the latter from the $3 \times 3$ covariance matrix (24). Although this estimate is not statistically different from 0, the positive sign indicates that a more difficult case mix increases a hospital’s expected

| Table 1. Pump Failure Data and Analysis Results Using the PRIMM Method |
|-------------------|-------------------|-------------------|
| obs $i$           | Orig. no.         | $x_0 = .225$      |
|                   | $e_i$             | $z_i$             |
| 1                 | 7                 | 1                 |
| 2                 | 8                 | 1                 |
| 3                 | 9                 | 2                 |
| 4                 | 5                 | 3                 |
| 5                 | 10                | 2                 |
| 6                 | 2                 | 15                |
| 7                 | 6                 | 30                |
| 8                 | 3                 | 60                |
| 9                 | 1                 | 90                |
| 10                | 4                 | 120               |
|                   | $y_i = z_i/e_i$   | $x_{11}$          |
| 1.27              | .43               | 1.08              |
| 1.27              | .43               | 1.08              |
| 1.27              | .27               | 1.77              |
| 1.27              | .13               | .67               |
| 1.27              | .07               | 2.13              |
| 1.27              | .05               | 1.2               |
| .25               | .11               | .59               |
| .25               | .06               | .09               |
| .25               | .04               | .06               |
| .25               | .03               | .12               |

*Note:* The first six pumps ran intermittently; the last four ran continuously. Exposures, $e_i$, are in 1,048 hours of operation. The number of failures is $z_i$. The failures per 1,048 hours: $y$, average 1.14 for pumps 1–6 and .22 for pumps 7–10.
mortality rate. The convolution parameter ζ is estimated to be ζ = 3.1, so that the standardized individual parameters λi/μi, from Level 2 (2), are estimated to follow approximately $\chi^2_{0.2,6.2}$ distributions. Gamma distributions with moments $\hat{\lambda}_i$ and $\hat{\sigma}_{\lambda_i}$ (34) reflect uncertainty about the center’s true mortality rates for their given case mix of patients. Consider centers 5 and 7. Observation $y_5 > y_7$, but still $\lambda_5 < \lambda_7$, reflecting the less severe case mix of patients at center 5. After a standardization for case mix, the centers could be compared on an equal footing in a profile analysis based on probabilities calculated from this hierarchical modeling analysis (Morris and Christiansen 1996).

4. OPERATING CHARACTERISTICS AND COMPARISONS WITH OTHER METHODS

This section reports simulation results comparing the inferences from Result 2 with six alternative estimation methods. Three dataset sizes are used, with $k = 10, 15, 45$. All three scenarios are based on real data. Poisson-gamma (P-G) data and Poisson-lognormal (P-LN) data are generated, the latter to test the robustness of PRIMM.

Five alternative methods (one is “ideal”, the hyperparameters being known) are considered for P-G models, and a P-LN model is fitted using Markov chain Monte Carlo (MCMC) simulation for numerical integration. These methods are described in Section 4.2, with some additional details given in Appendix D. Section 4.3 begins with a report of the operating characteristics of the P-G methods when analyzing 100 P-G datasets for each size $k$. Next we give the results from PRIMM for several large simulation studies (10,000 datasets) of P-G and P-LN data. The section concludes by comparing results from a fully Bayesian procedure.

PRIMM improves substantially on the four estimation procedures that use P-G models, especially when $k$ is small. These results provide further evidence and warning that for small values of $k$, naive plug-in methods are dangerously nonconservative, especially when the interest is in the individual units (Morris 1983b). PRIMM also compares favorably with the P-G model where the hyperparameters are known. For P-G and for P-LN data, PRIMM achieves the intended operating characteristics for the three tested values of $k$ and the chosen hyperparameter values. In only $\frac{1}{10}$ of the computation time of a widely-used Gibbs procedure and model, the PRIMM method obtains better overall coverage and better coverage for each exposure size.

4.1 Simulated Data

The first study, based on the pump data (Sec. 3.1), has $k = 10$ observations and one covariate ($r = 2$). It uses the exposures $e_i$ and the covariate $x_1$ from the pump data (Table 1). Each of the 100 simulations generates 10 true rates, $\lambda_1 \ldots \lambda_{10}$ (2), and from these 10 observed counts ($y_i$) (1) using the true hyperparameters $\zeta = 1$ and $\beta' = (-.5, -.7)$, which are close to the estimated values for the pump data.

The second study has $k = 15$ observations and two covariates. This study is similar in size and in the range of exposures and shrinkages to the transplant data in Section 3.2, although here we include an additional regression parameter. It uses the exposures $e_i + i + 2$ and covariates $x_{11} = i - 8, x_{12} = (-1)^{i+1}$ for $i = 1, \ldots, 15$. The true values $\zeta = 10$ and $\beta' = (0.1, 2)$ give true shrinkage values, $B_i = 10/(10 + e_i \exp(0.1x_{11} + 0.2x_{12}))$, which range from .19 to .85 and average .52.

The third simulation study is based on traffic accident data from the North Carolina Department of Motor Vehicles (Cribbins, Arey, and Donaldson 1969). This example has $k = 45$ highway sections. Exposures are vehicle miles driven at each site; they range by a factor of 10. The actual application requires making an inference about the expected rate of injury accidents $\lambda$, at each of the $k = 45$ sites, using a measure $x_1$ of the total daily movements entering or leaving the site (see the data in Appendix C). In the original data, the observed accident rates average .67. We simulated the data with hyperparameter values $\zeta = 3$ and $(\beta_{01}, \beta_{11}) = ((-.5, -3)$, which are close to the PRIMM estimates for the Cribbins et al. data. The average true shrinkage is .40, and the true shrinkages $B_i$ range from .11 to .78.
To test the robustness of the PRIMM method when the data are not P–G, we compared PRIMM to the MCMC method on P–LN data. The Level 2 distribution is lognormal with its moments chosen to match the true gamma moments from above. Thus we generated the Level 2 parameters \( \{ \lambda_i \} \) using

\[
\lambda_i \sim \text{lognormal}(x_i' \beta - \beta \log(1 + z_i^-), \log(1 + z_i^-)) = \text{LNN}[\mu_i = \exp(x_i' \beta), \mu_i^2 / \zeta].
\]

We then generated the counts \( \{ z_i \} \) to follow Poisson distributions with means \( \{ e_i \lambda_i \} \).

### 4.2 Methods being compared

The GLM Pearson (a) and GLM-deviance (b) methods fit an overdispersed generalized linear model (GLM) to estimate the regression coefficients \( \beta \) and a scale parameter \( \phi \). Estimates of \( \phi \) are obtained from Pearson residuals (a) or from deviance residuals (b). Moment estimates of the individual rate parameters \( \{ \lambda_i \} \) use \( B_i = 1 / \phi \), with \( \phi \) constrained to the interval \([1, \infty)\) and constant in \( i \). These GLM models assume \( \lambda_i \) follows a gamma distribution using \( \phi \) and \( \beta \) naively as true values in (8), (9), and (10). More details on these two methods appear in Appendix D.

The plug-in empirical Bayes (EB) method (c) obtains maximum likelihood hyperparameter estimates from the negative binomial likelihood (11) and uses these as true values in the posterior distributions (8). When \( \hat{\tau} = \infty \), as can happen for the MLE, \( B_i = 1 \) for every \( i \) and the distribution (8) has point mass at \( \mu_i \). The fourth method, EB-REML, takes method (c) one step further by including the REML correction in the negative binomial log-likelihood equations. The REML likelihood function described in Section 2.4 is used, implementing (21)–(25), and its derivatives from (22)–(23) but with the prior term \( \zeta / (\zeta + z_0)^2 \) removed from (21). Then the resulting \( \beta \) and \( \hat{\tau} \) are used as true values in (8) even if \( \hat{\tau} = \infty \). The PRIMM method (e), described in Section 2, uses the value \( z_0 = e_{\min} \sum z_i / \sum e_i = e_{\min} \tilde{e} / \tilde{e} \) each time; \( e_{\min} \equiv \text{minimum} \{ e_i \} \). Method (f) “ideal” uses the Bayesian model (8) with the hyperparameter known; that is, equal to the stated true values (\( \beta, \zeta \)). This provides an unachievable standard for the other procedures.

We also compared PRIMM to a procedure implemented by the BUGS program (Gilks, Thomas, and Spiegelhalter 1994; Spiegelhalter, Thomas, Best, and Gilks 1994a; Thomas, Spiegelhalter, and Gilks 1992). At the time of this simulation, the model of Section 2 could not be handled by the BUGS program. BUGS would not allow fitting a gamma regression model (2) at level 2, or the uniform shrinkage prior, (3) and (4), at level 3. For the P–LN BUGS method, we used burn-ins of 5,000 samples with 5,000 values after burn-in for \( k = 10 \) and \( k = 15 \) and with 2,000 values after burn-in when \( k = 45 \). These values are much greater than those used in the BUGS example datasets (Spiegelhalter et al. 1994b). Details on the level 2 and 3 distributions for this method are given in Appendix D.

### 4.3 Operating characteristics: coverages; interval widths; and risk functions

The first four methods sometimes produce “full shrinkage” estimates of \( B_i = 1 \), equivalently \( \hat{\xi} = \infty \) in methods (c) and (d) or \( \phi \leq 1 \) in methods (a) and (b). This causes the gamma distributions in (8) to have point mass at \( \mu_i \) and hence provide 0 width intervals. Full shrinkage occurs in two of the 100 plug-in EB analyses and once in the EB-REML analyses when \( k = 10 \), and 8, 7, 22, and 15 times when \( k = 15 \) for the GLM-Pearson, GLM-deviance, plug-in EB, and EB-REML methods. The occurrence of an infinite mode for \( \zeta \) and the problem with overly narrow intervals are common concerns of the MLE and other “plug-in” empirical Bayes approaches as noted by Hill 1990 and Morris 1988b.

Figure 1 shows average probabilities that the interval estimates fail to cover the true values of \( \lambda_i \) in the \( k = 15 \) study, when \( \zeta = 10 \) and \( \beta' = (0, 1, 2) \). These are estimates of the noncoverage probabilities in repeated use. All methods are set to have nominal noncoverage probabilities of .05. The noncoverages reported are the averages of the 100 probabilities

\[
1 - \Pr(\hat{\lambda}_{i, 0.25} < \lambda_i < \hat{\lambda}_{i, 0.95} | \text{data}).
\]

Although only 100 trials are available for each \( i \), the simulation accuracy is improved substantially by Rao–Blackwellization; that is, by using the fact that the probability (36) can be calculated exactly with \( \lambda_i \sim \text{gamma}(10 + z_i, 1) / (10 / \exp(1x_{i1} + 2x_{i2}) + e_i) \) for every observation \( i \). The resulting simulation error for noncoverage at each of the 15 exposure levels is about \( \pm 1\% \) for 5% noncoverage, rising to about \( \pm 2\% \) for rules with the highest noncoverage.

The EB (maximum likelihood, plug-in to (10) for variances), and the GLM methods all result in serious under-coverage. This is partly due to the occasional cases with
### Table 3. Operating Characteristics: Coverages, Interval Widths, and Risk Functions

<table>
<thead>
<tr>
<th>Poisson–gamma data</th>
<th>Noncoverage probabilities (36)</th>
<th>Avg. interval width</th>
<th>Squared error risk (37)</th>
<th>( \hat{B} )</th>
<th>avg. ( \hat{\zeta} )</th>
<th>avg. ( \hat{\beta} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k = 10, r = 2 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a GLM-Pearson</td>
<td>.100</td>
<td>1.30</td>
<td>10.13</td>
<td>.18</td>
<td>NA</td>
<td>(-63, -69)</td>
</tr>
<tr>
<td>b GLM-deviance</td>
<td>.089</td>
<td>1.31</td>
<td>9.95</td>
<td>.17</td>
<td>NA</td>
<td>(-63, -69)</td>
</tr>
<tr>
<td>c plug-in EB</td>
<td>.123</td>
<td>1.17</td>
<td>9.52</td>
<td>.28</td>
<td>1.55*</td>
<td>(-59, -71)</td>
</tr>
<tr>
<td>d EB-REML</td>
<td>.097</td>
<td>1.22</td>
<td>9.53</td>
<td>.24</td>
<td>1.20*</td>
<td>(-59, -71)</td>
</tr>
<tr>
<td>e PRIMM</td>
<td>.058</td>
<td>1.33</td>
<td>9.05</td>
<td>.19</td>
<td>1.32</td>
<td>(-59, -71)</td>
</tr>
<tr>
<td>Ideal:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f known ( \zeta ) and ( \beta )</td>
<td>.050</td>
<td>1.28</td>
<td>8.24</td>
<td>.18</td>
<td>1.00</td>
<td>(-50, -70)</td>
</tr>
<tr>
<td>( k = 15, r = 3 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a GLM-Pearson</td>
<td>.193</td>
<td>.82</td>
<td>9.99</td>
<td>.55</td>
<td>NA</td>
<td>(01, 10, 20)</td>
</tr>
<tr>
<td>b GLM-deviance</td>
<td>.178</td>
<td>.84</td>
<td>8.99</td>
<td>.55</td>
<td>NA</td>
<td>(01, 10, 20)</td>
</tr>
<tr>
<td>c plug-in EB</td>
<td>.168</td>
<td>.60</td>
<td>10.00</td>
<td>.71</td>
<td>23.1*</td>
<td>(01, 10, 20)</td>
</tr>
<tr>
<td>d EB-REML</td>
<td>.258</td>
<td>.74</td>
<td>9.68</td>
<td>.61</td>
<td>13.1*</td>
<td>(01, 10, 20)</td>
</tr>
<tr>
<td>e PRIMM</td>
<td>.047</td>
<td>1.04</td>
<td>8.89</td>
<td>.52</td>
<td>11.7</td>
<td>(01, 10, 20)</td>
</tr>
<tr>
<td>Ideal:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f known ( \zeta ) and ( \beta )</td>
<td>.050</td>
<td>.89</td>
<td>6.68</td>
<td>.52</td>
<td>10.0</td>
<td>(00, 10, 20)</td>
</tr>
<tr>
<td>( k = 45, r = 2 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a GLM-Pearson</td>
<td>.072</td>
<td>.89</td>
<td>31.9</td>
<td>.37</td>
<td>NA</td>
<td>(-51, 30)</td>
</tr>
<tr>
<td>b GLM-deviance</td>
<td>.067</td>
<td>.90</td>
<td>31.6</td>
<td>.36</td>
<td>NA</td>
<td>(-51, 30)</td>
</tr>
<tr>
<td>c plug-in EB</td>
<td>.055</td>
<td>.83</td>
<td>29.0</td>
<td>.44</td>
<td>5.76</td>
<td>(-51, 31)</td>
</tr>
<tr>
<td>d EB-REML</td>
<td>.072</td>
<td>.84</td>
<td>29.0</td>
<td>.42</td>
<td>4.06</td>
<td>(-51, 31)</td>
</tr>
<tr>
<td>e PRIMM</td>
<td>.055</td>
<td>.88</td>
<td>28.8</td>
<td>.41</td>
<td>3.32</td>
<td>(-51, 31)</td>
</tr>
<tr>
<td>Ideal:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f known ( \zeta ) and ( \beta )</td>
<td>.050</td>
<td>.86</td>
<td>27.2</td>
<td>.41</td>
<td>3.00</td>
<td>(-50, 30)</td>
</tr>
</tbody>
</table>

*The average \( \hat{\zeta} \) is infinite. This value is the median.

**NOTE:** \( \hat{B} \) represents the average estimated shrinkage. When \( k = 10 \) and the number of datasets is 100, \( \hat{B} \) is the average of 1,000 \( \hat{B} \) values. NA indicates that \( \hat{\zeta} \) is not defined for these methods.

\[
\hat{B} = 1, \text{ when the variances (10) are estimated as 0. It happens mainly because the MLE and rules (a)-(d) do not account for uncertainty of the hyperparameter estimates in assessing the posterior variances. Even when average coverage achieves the nominal value, a good rule must have coverage probabilities unrelated to exposures. The increasing noncoverage pattern for the two GLM methods is caused by their (inappropriate) constant shrinkages. Their over shrinkage for large exposures makes the variances (10) underestimated, so they produce relatively shorter intervals for large exposures. The situation is reversed for small exposures with GLM. Undercoverage for the two EB methods has the opposite pattern, partly because the EB methods overestimate shrinkage for small exposures. In general, good rules will have coverage rates for each individual \( i \) that are unrelated to the individual exposures or to any other predictors.}

Table 3 reports additional results for the \( k = 10, k = 15 \), and \( k = 45 \) simulation studies. Accuracy in that table improves, relative to those of Figure 1, because the \( k \) cases are averaged, so that the coverage estimates for \( k = 10 \) and \( k = 15 \) are accurate to about \( \pm 0.004 \), and about \( \pm 0.002 \) for \( k = 45 \). PRIMM provides wider intervals (see Table 3) than the other four methods. These wider intervals, needed to account for hyperparameter uncertainty, are partly responsible for its better coverage. For example, in the \( k = 10 \) and \( k = 45 \) studies, PRIMM provides much better coverage than GLM, even though its average interval widths are approximately equal to those of the GLM methods.

The loss function used to compare estimators is the sum of the standardized squared error losses,

\[
\sum_{i=1}^{k} \frac{(e_i \hat{\lambda}_i - e_i \lambda_i)^2}{e_i \mu_i}.
\]

The optimum estimator for this loss function is the mean \( E(\lambda_i | \text{data}) \). Thus \( \lambda_i^* \) (9) is the best estimator when \( \zeta \) and \( \beta \) are known—method (f)—and gives a lower bound to what is realistically possible. The risks, estimated as the average of these losses, are given in column 3 of Table 3. The PRIMM estimate of \( \{\lambda_1, \ldots, \lambda_k\} \) has better risks than the other techniques in all three examples.

The PRIMM estimates of the regression parameters \( \beta \) (column 6) are relatively insensitive to the estimation of \( \zeta \) (column 5), because they are asymptotically independent, based on Fisher’s information. The two GLM cases have constant shrinkage, \( \phi^{-1} = B \), that is approximately equal to the average of the true shrinkages in these studies. This still could allow the GLM approach, despite its poor performance for estimating \( \{\lambda_i\} \), to work well when \( \phi \) is used only to correct the estimate of the regression coefficient covariance matrix, if overdispersion or heterogeneity exists. We are not evaluating this use of \( \phi \) in GLM here, but McNeeney and Petkau (1994) did so for Poisson data.

A more extensive evaluation of PRIMM on 10,000 P–G and on 10,000 P–LN datasets is reported for \( k = 15 \) and for \( k = 45 \) in Table 4. The Rao–Blackwellization method (36) is not helpful for lognormal data, because there is no exact analytic solution for P–LN mixtures, so coverage in column 2 of Table 4 must be scored as a binary outcome. Hence 10,000 samples were needed for accurate results.
Each PRIMM analysis requires 2–3 seconds of user time in S-PLUS (StatSci 1991) on a SPARC 20, making this many runs possible.

Columns 1, 3, and 5 of Table 4 for PRIMM (10,000 datasets) are in good agreement with the corresponding results of Table 3 (also reported as PRIMM 100 in Table 4), and further confirm PRIMM’s reliability in these examples. PRIMM does about as well (actually, slightly better) for lognormal $\lambda_i$; compare columns 2, 4, and 6 with columns 1, 3, and 5 for PRIMM 10,000. PRIMM’s coverages of each $\lambda_i$ continue to be uncorrelated with exposure $e_i$ for $k = 15$ (compare Fig. 1) and for $k = 45$. (These estimates were close enough to .05 so that any differences from .05 may be attributed to simulation error.) The case $k = 10$ is not reported here for these 10,000 datasets, but overall noncoverages for PRIMM are about 5.4%, the high value being consistent with the results of Table 3.

PRIMM’s excellent squared error risk reported in column 6 of Table 4 for lognormal $\lambda_i$ follows from the minimax property that the conjugate gamma distribution holds among all prior distributions with a fixed mean and variance (Morris 1983a). Thus PRIMM is robust to the prior specification (2), in this minimax sense.

Gibbs sampling, a MCMC method for numerical integration, is being used increasingly in data analyses (see, e.g., Gilks et al. 1993). Because of this, because the BUGS program (Spiegelhalter et al. 1994a) now makes Gibbs sampling widely available, and at the reviewers’ suggestion, we have chosen this method over other MCMC methods for our comparisons; for example, over the we have chosen this method, urged by the reviewers, for our comparisons over other MCMC methods; for example, the Metropolis algorithm (Gelman, Carlin, Stern, and Rubin 1995). To improve convergence, the start-up values that we used here were the true values of the hyperparameter.

When the data are P–G, PRIMM consistently covers better than the P–LN BUGS method (which uses a different level 3 hyperparameter distribution). Figure 1 for $k = 15$ also plots the P–LN BUGS average noncoverage probability against exposure for the 100 P–G simulated dataset. No P–LN BUGS estimate has a noncoverage probability less than .06. For $k = 15$, Table 4 shows that PRIMM beats the P–LN BUGS procedure, with an overall noncoverage rate of .047 versus .070 and with a risk of 8.9 versus 9.3. A similar dominance holds for PRIMM when $k = 45$.

The P–LN data permit a fairer evaluation of the P–LN BUGS method, although such simulations were too time-consuming for these evaluations to be performed on more than 100 datasets for each value of $k = 15$ and $k = 45$. ($k = 10$ does not appear, because we were unable in 40% of the datasets to get convergence.)

Figure 2 shows the fraction (out of 100) of P–LN BUGS and the fraction of PRIMM noncovering 95% intervals for the 15 lognormal cases $i = 1, \ldots, 15$. Both procedures have similar noncoverage patterns across the 15 exposure values (the jaggedness is due to the sizeable simulation error in this 100 dataset study), but PRIMM’s noncoverage is better than that of P–LN BUGS in all 15 cases, 6.0% versus 9.3% overall. We do not know for sure why BUGS does not do better, but we suspect that the “noninformative” prior used at level 3 is the culprit; see Appendix D.

### Table 4. Operating Characteristics for the P–LN BUGS and the PRIMM Methods

<table>
<thead>
<tr>
<th>Poisson–gamma and Poisson–lognormal data Method</th>
<th>No. of datasets</th>
<th>Noncoverage probabilities</th>
<th>Avg. interval width</th>
<th>Squared error risk (37)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k = 15, r = 3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P–LN BUGS</td>
<td>(100)</td>
<td>.070</td>
<td>.093</td>
<td>.96</td>
</tr>
<tr>
<td>PRIMM</td>
<td>(100)</td>
<td>.047</td>
<td>.060</td>
<td>1.04</td>
</tr>
<tr>
<td>PRIMM</td>
<td>(10,000)</td>
<td>.049</td>
<td>.047</td>
<td>1.03</td>
</tr>
<tr>
<td>P–G ideal</td>
<td>(10,000)</td>
<td>.050</td>
<td>.045</td>
<td>.88</td>
</tr>
<tr>
<td>$k = 45, r = 2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P–LN BUGS</td>
<td>(100)</td>
<td>.076</td>
<td>.050</td>
<td>.86</td>
</tr>
<tr>
<td>PRIMM</td>
<td>(100)</td>
<td>.055</td>
<td>.045</td>
<td>.88</td>
</tr>
<tr>
<td>PRIMM</td>
<td>(10,000)</td>
<td>.053</td>
<td>.049</td>
<td>.88</td>
</tr>
<tr>
<td>P–G ideal</td>
<td>(10,000)</td>
<td>.050</td>
<td>.041</td>
<td>.86</td>
</tr>
</tbody>
</table>

Figure 2. Average Fraction Missed, $k = 15$, 100 Samples of Poisson–Lognormal Data (---, PRIMM; ----, P–LN BUGS). PRIMM outperforms the P–LN BUGS method for every value of exposure, even for the P–LN data. The average fraction of these cases missed for PRIMM, 6.0%, is high due to simulation error.

### 5. SUMMARY AND DISCUSSION

The two-level Poisson model specified here and the anal-
yses developed for its inferences have several advantages. The modeling assumptions (1) and (2) enlarge the range of applications relative to many of its competitors by permitting varying exposures and a regression specification. Assumption (3) provides a hyperparameter distribution expected to produce, in conjunction with the gamma choice in (2), good frequency properties for the resulting rules when conservative (small) choices are made for $z_0$. The gamma specification helps make the procedure robust relative to other distributions such as the lognormal, for the level 2 parameters (2). The procedure developed here also allows for informative Bayesian inferences, if the prior median for $\zeta$, $z_0$, is set at an appropriate (i.e., larger and less conservative) value. Shrinkage factors, whose estimates provide valuable summary and diagnostic information about the model, are defined naturally for the gamma specification at level 2. The PRIMM procedure provides point and interval estimates for all of the individual parameters and all the structural parameters; see Results 1 and 2. We expected the procedure to perform well by both frequency and Bayesian standards for small $k$, as it has in all simulations thus far, partly because the fitting of this model accounts quite accurately for the problems of nuisance parameters (with REML), for nonlinearity and the bounded range of the shrinkage factors, by using the ADM approximation, and it provides accurate formulas for the posterior means and variances (Theorem 1).

We have seen from simulations based on several real datasets, with $k$ as small as 10, that the approach here, as implemented by the PRIMM program in S-PLUS, provides confidence intervals that have (approximately) their nominal coverages—not only overall coverages, but also for every exposure level. For making inferences about the $k$ individual Poisson rates, this procedure compares quite favorably, with respect to coverage and to mean squared error, to GLM and to naive MLE- and REML-based empirical Bayes methods.

PRIMM’s relatively fast computing time, a couple of seconds in an S-PLUS environment for the examples considered here, has permitted the simulations of Section 4 to evaluate its operating characteristics. This speed in empirical work permits data analysts to fit the hierarchical Poisson model interactively, freeing them to concentrate on model building and on checking the regression and other distribution specifications.

**APPENDIX A: PROOFS AND JUSTIFICATIONS**

**Proof of Lemma 1**

This follows straightforwardly from algebra applied to the logarithm of (11), with additive constants eliminated. The first term in (12) uses the relation $\Gamma(\zeta + z_1)/\Gamma(\zeta) = (\zeta + z_1 - 1)$. The remaining terms use $B_1 = \zeta/(\zeta + z_1\mu)$, as in (7).

**Justification of Result 1**

We need to show that $\Sigma$ in (24) is the inverse of the $r + 1 \times r + 1$ Hessian of $-\mathcal{L}(\beta, \zeta)$ with elements defined by (14), (15), and (23); that is, of

$$
\mathbf{H} = \begin{bmatrix}
H_{11} & H_{12} \\
H_{21} & H_{22}
\end{bmatrix} = \begin{bmatrix}
H_{11} & -\mathbf{H}_{11} \hat{\nu} \\
-\hat{\nu} \mathbf{H}_{11} & H_{22}
\end{bmatrix},
$$

where $H_{11} = X'X$, $H_{22} = \text{diag}(1, \ldots, 1)$, and $H_{12} = -\partial^2 \mathcal{L}(\beta, \zeta)/\partial \beta \partial \zeta$, with all terms evaluated at $\hat{\beta}, \hat{\zeta}$. Note that $H_{22} = -\mathbf{H}_{11} \hat{\nu}$, which follows by comparing (15) and (16). It is easily checked in this notation that $\Sigma \mathbf{H} = \mathbf{I}_{r+1}$, and so it is unnecessary to consider inverse relations here.

The final assertion uses (23), with $\hat{\beta}$ replacing $\beta$ there. Use of the $\mathbf{H}$ notation helps to obtain $\partial^2 \mathcal{L}(\beta, \zeta)/\partial \beta \partial \zeta = -\hat{\nu} \mathbf{H}_{11} \hat{\nu} - 2\hat{\nu} \mathbf{H}_{11} + \mathbf{H}_{22} - \partial^2 \mathcal{L}(\beta, \zeta)/\partial \zeta \partial \zeta (\partial \hat{\beta} + \partial \hat{\zeta} / \partial \zeta) = \hat{\beta} \mathbf{H}_{11} \hat{\nu} - \hat{\nu} \mathbf{H}_{11} \hat{\nu}$, at $\hat{\beta}, \hat{\zeta}$.

**Proof of Lemma 2**

Fix $i$. Denote $u_1 = \hat{x}_i \beta$ and $u_2 = \tau - \hat{x}_i \beta$, so that $\mu = \exp(u_1)$ and $B_i/(1 - B_i) = \exp(u_2) / e_i$. Then $u_1 - u_2$ is independent of $u_2$, and hence of $B_i$. Thus

$$
\frac{E_{\mu_i} B_i^2}{E \mu_i^2} = \frac{E_{\exp(u_1 - u_2)} e^{u_1} e^{u_2} B_i^2}{E_{\exp(u_1 - u_2)} e^{u_1} e^{u_2}} = \frac{E_{\exp(u_1)} e^{u_2} B_i^2}{E_{\exp(u_2)} e^{u_1}}
$$

by independence. The latter expression is (28), on substituting for $u_2$.

**Proof of Lemma 3**

With $u = \log(B)$, it follows that the density for $B$ is

$$
p(B) = \frac{1}{\sigma} \phi \left( \frac{u - \mu}{\sigma} \right) \frac{du}{dB} = \frac{1}{\sigma} \phi \left( \frac{u - \mu}{\sigma} \right) \left( B(1 - B) \right)^{-1/2}
$$

so $l'(B) = \log(B(1 - B)p(B)) = -(u - \mu)^2 / 2\sigma^2 - \log(2\pi\sigma^2) / 2$. Then $l'(B) = -(1/\sigma^2)(u - \mu)(du/dB) = 0$ at $\log(B) = \mu$; that is, $B = e^\mu / (1 + e^\mu)$. Thus $\mu_i = l''(B) = (1/\sigma^2)(du/dB)^2 = (1/\sigma^2)(u - \mu)(du/dB)^2 = 1/(\sigma^2 B^2 (1 - B)^2)$ at $B_i$. (30) now follows from (29). When $B$ has a beta($\alpha_1, \alpha_2$) distribution, (31) is simply the ratio of beta functions, $\beta(a_1 + 1 + w, a_2 - w) / \beta(a_1 + w, a_2 - w)$, from which (31) follows easily.

**Proof of Theorem 1**

From (9), $E_{\lambda_i} = E_{\lambda_i} = (1 - E_{\mu_i} y_i / E_{\mu_i} \mu_i)$. Because $\log(\beta_i) = u_2 = \tau - \hat{x}_i \beta - \log(e_i)$ has mean $\hat{\tau} = \hat{x}_i \beta - \log(e_i)$, the ADM gives $E_{\beta_i} = \hat{\beta}_i = e^\hat{\tau} / (e^{\hat{\tau}} + e^{\hat{\tau} \hat{x}_i \beta})$, from Lemma 3, which also provides the expressions for $\alpha_{i1}, \alpha_{i2}$. Lemma 2 gives $E_{\beta_i \mu_i} = (E_{\mu_i} / E_{\beta_i}) B_i$ with $E_{\mu_i} = \hat{\mu}_i$ given by (26) and $E_{\beta_i}(B_i)$ given by (31), with $t = 1$ and $w = b_i$. This establishes (32).

From (10) and (9), $\var(\lambda_i) = E(\gamma_i)^2 + \var(\lambda_i) / E = (1 - B_i)^2 / e_i + E_{\mu_i} B_i(1 - B_i) / e_i + (E(\mu_i - \mu_i)^2 / B_i^2) = (E(\mu_i - \mu_i)^2 / B_i^2)$. This again requires $E_{\mu_i} B_i = \mu_i E_{\mu_i} B_i$, and also $E_{\beta_i} = E_{\beta_i} B_i = \mu_i E_{\beta_i} B_i$, and $E_{\beta_i}(B_i) = (E_{\beta_i} B_i)^2$. These values $E_{\beta_i} B_i^2$ for $s = 0, 1, 2$ and $t = 1, 2$ are provided by Lemma 3, with $\alpha_1 = \alpha_{i1}, \alpha_2 = \alpha_{i2}$, and $w$ replaced by $s b_i$, as in (28).

**APPENDIX B: PRIMM SOFTWARE AND COMPUTING INFORMATION**

The PRIMM program implements the computations of Section 2 and is developed to cover a wide range of data and applications. It is available through Statlib as an S-PLUS (Becker, Chambers and Wilks 1988; Statistical Sciences, Inc. 1991) program. The program runs when $k$ is at least $r$, the dimension of the unknown regression coefficients. However, we recommend not using it unless $k \geq r + 3$, with improvements expected as $k - r$ increases further. The user provides data $e, z,$ and $x$ (or $x$ is optional) to the program. Key output provided includes information on the loglikelihood statistic and estimates, standard errors, and correlations for the hyperparameters ($\beta, \zeta$). For each of the individual cases, the program provides estimates $\hat{\mu}_i, \hat{B}_i, \hat{\lambda}_i$, and $\hat{\sigma}_i$, and 95% (default) interval estimates of each $\lambda_i$. Other quantities are available at the user’s option.
A computational identity for the log-likelihood in Section 2.3, needed to maintain computational speed with $M = \max(z_i)$ is large, uses

$$
\sum_{j=1}^{M} N_j \log(\zeta + j - 1) = \sum_{j=1}^{t} N_j \log(\zeta + j - 1) + \sum_{j=t+1}^{M} n_j \left( \frac{\Gamma(\zeta + j)}{\Gamma(\zeta + t)} \right)
$$

in (12) and the derivatives of these terms in (17) and (18). Accurate Sterling-type approximations are used in the last term. The value $t = \min(6, M)$ balances accuracy and computational speed (see Christiansen 1992 for details).

**APPENDIX C: DATA**

The exposure, $e$, values used in the $k = 45$ traffic injury accident data simulation study are as follows:

<table>
<thead>
<tr>
<th>Exposure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.80</td>
<td>2.24</td>
</tr>
<tr>
<td>5.44</td>
<td>5.52</td>
</tr>
<tr>
<td>10.00</td>
<td>10.36</td>
</tr>
<tr>
<td>12.32</td>
<td>13.68</td>
</tr>
<tr>
<td>4.18</td>
<td>4.59</td>
</tr>
<tr>
<td>7.56</td>
<td>7.60</td>
</tr>
<tr>
<td>11.33</td>
<td>11.55</td>
</tr>
<tr>
<td>16.00</td>
<td>16.80</td>
</tr>
</tbody>
</table>

The values of the covariate $x_i$ are as follows:

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.44</td>
<td>1.83</td>
</tr>
<tr>
<td>-0.66</td>
<td>-4.51</td>
</tr>
<tr>
<td>0.86</td>
<td>-1.54</td>
</tr>
<tr>
<td>-1.04</td>
<td>-0.72</td>
</tr>
<tr>
<td>0.18</td>
<td>1.50</td>
</tr>
<tr>
<td>-1.07</td>
<td>-3.22</td>
</tr>
<tr>
<td>1.93</td>
<td>-0.19</td>
</tr>
<tr>
<td>-1.29</td>
<td>1.71</td>
</tr>
</tbody>
</table>

**APPENDIX D: METHODS TESTED IN THE SIMULATION STUDIES**

(a) and (b) Overdispersed Poisson Model

Generalized linear models GLM’s (McCullagh and Nelder 1989; Nelder and Wedderburn 1972) are commonly used tools for the analysis of Poisson or overdispersed Poisson (or negative binomial) data. The Poisson GLM methods use asymptotic theory to estimate the scale parameter for the common problem of overdispersion in the data.

The GLM methods assume that $E(z_i) = e_i \mu_i$ and $\text{var}(z_i) = \phi e_i \mu_i$. Here $\phi \equiv$ scale parameter and $e_i \equiv$ offset; that is, the coefficient of $\log(e_i)$ is assumed to be 1. We consider rules based on two different estimators of $\phi$: one based on Pearson residuals, (a) $\hat{\phi}_p = \sum_i (y_i - \mu_i)^2 / \mu_i / (k-r)$; and one based on the deviance residuals, (b) $\hat{\phi}_d = 2 \sum_i (y_i \log(y_i / \mu_i) - (y_i - \mu_i)) / (k-r)$. The implicit assumption is that the observed counts follow negative binomial distributions

$$
z_i | n_i, \phi \sim \text{NBin} \left( \frac{e_i \mu_i}{\phi - 1}, \frac{\phi - 1}{\phi} \right) = \text{NBin}[e_i, \mu_i, \phi_{e_i} \mu_i];
$$

$\phi > 1$. (D.1)

Comparing (6) to (D.1), we define $\phi^{-1} \equiv B$, so $B$ is a constant shrinkage value. The posterior moments of $\lambda_i$ for the two GLM methods use $(\lambda_i | \text{data}) \sim (\lambda_i | \text{data}, \phi, \beta)$

$$
\lambda_i | \text{data} \sim \text{Gam} \left( \frac{\lambda_i + e_i \mu_i}{\phi - 1}, \frac{e_i \phi}{\phi - 1} \right)
$$

$$
= \text{Gam} \left[ y_i - \frac{y_i - \mu_i}{\phi} \equiv \frac{\lambda_i + \mu_i}{e_i}, \frac{1}{\phi} \right].
$$

P–LN BUGS Model

The model for the P–LN BUGS method used in the simulation studies has the Poisson counts distributed as in (1). The true rates follow lognormal distributions where

$$
\lambda_i \sim \text{lognormal}(\lambda_i | \beta, 1/\tau_0).
$$

A priori, the regression hyperparameters are given independent priors $N(0, 100^2)$, and $\tau_0 \equiv 1/\log(1 + 1/\zeta)$ is assumed to be distributed as 1,000 $\text{Gam}(0.01, 1)$. These a priori assumptions are suggested as “noninformative” (Spiegelhalter et al. 1994b). Burnins of 5,000 were used.

[Received April 1994. Revised September 1996.]

**REFERENCES**


Dept.


