Biostatistics (2012), **13**, 1, *pp.* 179–192
doi:10.1093/biostatistics/kxr030
Advance Access publication on September 13, 2011

Relative risk regression: reliable and flexible methods for log-binomial models

IAN C. MARSCHNER*
Department of Statistics, Macquarie University, NSW 2109, Australia and National Health and Medical Research Council Clinical Trials Centre, University of Sydney, NSW 2006, Australia
ian.marschner@mq.edu.au

ALEXANDRA C. GILLETT
Department of Statistics, Macquarie University, NSW 2109, Australia

SUMMARY

Relative risks (RRs) are generally considered preferable to odds ratios in prospective studies. However, unlike logistic regression for odds ratios, the standard log-binomial model for RR regression does not respect the natural parameter constraints and is therefore often subject to numerical instability. In this paper, we develop a reliable and flexible method for fitting log-binomial models. We use an Expectation–Maximization (EM) algorithm where the multiplicative event probability is viewed as the joint probability for a collection of latent binary outcomes. This gives a simple iterative scheme that provides stable convergence to the maximum likelihood estimate. In addition to reliability, the method offers some flexible generalizations, including models with unspecified isotonic regression functions. We examine the method's performance using simulations and data analyses of the age-specific RR of mortality following heart attack. These analyses demonstrate the potential for numerical instability in RR regression and show how this can be overcome using the proposed approach. Source code to implement the method in R is provided as supplementary material available at *Biostatistics* online.

Keywords: Binomial regression; EM algorithm; Generalized linear models; Isotonic regression; Relative risk.

1. INTRODUCTION

Relative risks (RRs) are easier to understand than odds ratios, and there have been many recommendations for their use in prospective studies (McNutt *and others*, 2003; Speigelman and Hertzmark, 2005; Lumley *and others*, 2006). Nonetheless, risk factor modeling commonly uses logistic regression, which measures associations using odds ratios. This is primarily due to convenience since regression models for RRs require the fitting of a binomial generalized linear model (GLM) with a log link function. This model, which we refer to as the log-binomial model, is well known to experience numerical instability because the log link function allows probabilities greater than 1. This means standard methods for fitting GLMs, such as iteratively reweighted least squares (IRLS), may fail to converge to the maximum likelihood

*To whom correspondence should be addressed.

estimate (MLE). Such instability is common (Carter *and others*, 2005; Blizzard and Hosmer, 2006) and can occur even when the MLE is well inside the parameter space.

Over the years, there have been many proposals for circumventing problems inherent in the log-binomial model. Some involve ignoring the parameter constraints and using a non-MLE approximation. For example, McNutt *and others* (2003), Zou (2004), Speigelman and Hertzmark (2005), and Carter *and others* (2005) have all discussed the log-linear Poisson model as an approximate way of fitting the log-binomial model. Similarly, proportional hazards regression and nonlinear least squares can both be used if one is prepared to ignore the parameter constraints (Lee, 1994; Lumley *and others*, 2006). However, such approaches have the obvious drawback that they do not necessarily produce models with risks in the range [0, 1], which may lead to difficulties interpreting and applying the models. An alternative approach is to modify IRLS by preventing it from exiting the parameter space. This was suggested as far back as Wacholder (1986) and many software packages supplement standard IRLS with modifications such as step halving. However, such modifications are still subject to convergence problems, as illustrated in Section 5. Some authors have suggested using a generic constrained optimization routine since fitting the log-binomial model is a constrained optimization problem (Lumley *and others*, 2006; Yu and Wang, 2008). This can be a useful approach but is often infeasible for flexible models with high dimensionality, such as the 139 parameter model we consider in Section 5. Finally, approximations such as the COPY method of Deddens and Petersen (2004) make modifications to the data that cause the constraints to be satisfied. This can be effective in some contexts, but in others, it need not be close to the MLE, as discussed by Lumley *and others* (2006).

The above discussion shows that there remains a need for reliable methods to fit the log-binomial model. This paper presents a new approach yielding the genuine MLE based on the constrained binomial likelihood function. The Expectation–Maximization (EM) algorithm is used, which provides stable constrained maximization of the likelihood function even in high dimensions. It also allows flexible generalizations, including models with unspecified isotonic regression functions. We demonstrate the usefulness of the new method with analyses of a large cardiovascular clinical trial and simulations.

2. BASIC METHOD

We begin by describing the basic method for categorical covariates. Our strategy uses the fact that the log-binomial model is a product probability model. Each of the terms in this product can itself be viewed as a probability, and we can then specify a collection of underlying latent binary outcomes with event probabilities corresponding to the terms in the product. The event probability for the observed outcome can then be viewed as the joint probability that every latent outcome is an event. Since estimation based on the unobserved latent outcomes is simple, the EM algorithm can be used.

2.1 *Model with categorical covariates*

Consider n independent binomial outcomes $Y_i \sim \text{Bin}(N_i, p_i)$, $i = 1, 2, \ldots, n$, each with A categorical covariates $x_{ij} \in \{1, 2, \ldots, k_j\}$, for $j = 1, 2, \ldots, A$. To accommodate an intercept in the model below, we also introduce a constant covariate $x_{i0} = 1$, with $k_0 = 1$. The covariate vector for outcome i is $\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{iA})$ and the covariate space $\mathcal{X}$ is the Cartesian product over j of the discrete sets of covariate values $\{1, 2, \ldots, k_j\}$. Note that these covariates can correspond to either main effects or interactions.

The log-binomial model is specified by the log link GLM

$$p_i = P(\mathbf{x}_i; \boldsymbol{\theta}) = \exp\left\{\sum_{j=0}^{A} \alpha_j(x_{ij})\right\} = \prod_{j=0}^{A} \theta_j(x_{ij}), \tag{2.1}$$

where $\theta_j(x) = \exp\{\alpha_j(x)\}$, $\boldsymbol{\theta}_j = (\theta_j(0), \theta_j(1), \ldots, \theta_j(k_j))$, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_A)$. In RR regression, the probability $P(\mathbf{x}; \boldsymbol{\theta})$ is the risk of an event with covariate pattern $\mathbf{x}$, while the ratio $P(\mathbf{x}_1; \boldsymbol{\theta}) \big/ P(\mathbf{x}_2; \boldsymbol{\theta})$ is the RR for covariate pattern $\mathbf{x}_1$ compared to $\mathbf{x}_2$. To interpret the parameters in (2.1) as RRs, we choose a reference level for each covariate, $r_j \in \{1, 2, \ldots, k_j\}$, and impose identifiability restrictions $\theta_j(r_j) = 1$, for $j = 1, 2, \ldots, A$. The parameter $\theta_j(x)$ is then the RR for level x of covariate j, compared to the reference level r_j, adjusting for the other covariates. The choice of reference vector $\mathbf{r} = (r_0, r_1, \ldots, r_A)$ is arbitrary.

The parameter space Θ is defined by the natural constraints on the event probabilities

$$\Theta = \{\boldsymbol{\theta} : 0 \leqslant P(\mathbf{x}; \boldsymbol{\theta}) \leqslant 1 \quad \text{for all } \mathbf{x} \in \mathcal{X}\}. \tag{2.2}$$

The MLE is then determined by maximization over Θ of the log likelihood

$$L(\boldsymbol{\theta} | \{Y_i\}) = \sum_{i=1}^{n} Y_i \log\{P(\mathbf{x}_i; \boldsymbol{\theta})\} + (N_i - Y_i) \log\{1 - P(\mathbf{x}_i; \boldsymbol{\theta})\}. \tag{2.3}$$

2.2 *Latent outcome model*

In order to construct an EM algorithm to maximize L over Θ, we postulate a collection of unobserved latent outcomes underlying each observed outcome. As a binomial outcome variable, Y_i is the sum of independent binary outcomes

$$Y_i = \sum_{k=1}^{N_i} Y_i^{(k)} \quad \text{where} \quad \Pr\left(Y_i^{(k)} = z\right) = p_i^z (1 - p_i)^{1-z} \quad \text{for} \quad z = 0, 1.$$

Using (2.1), $Y_i^{(k)}$ can be viewed as the product of independent latent binary outcomes

$$Y_i^{(k)} = \prod_{j=0}^{A} Z_{ij}^{(k)} \quad \text{where} \quad \Pr\left(Z_{ij}^{(k)} = z\right) = \theta_j(x_{ij})^z \{1 - \theta_j(x_{ij})\}^{1-z} \quad \text{for} \quad z = 0, 1.$$

This correspondence arises because the event probability associated with the observed outcomes is the joint probability associated with the latent outcomes

$$\Pr\left(Y_i^{(k)} = 1\right) = \Pr\left(Z_{i0}^{(k)} = 1, Z_{i1}^{(k)} = 1, \ldots, Z_{iA}^{(k)} = 1\right),$$

and this joint probability is identical to the product probability in (2.1). Under the latent outcome model, estimation of $\boldsymbol{\theta}$ would be straightforward if $\{Z_{ij}^{(k)}\}$ were observable, and we will exploit this to develop an EM algorithm in Section 2.3.

While the latent outcome model has the same parameter vector as the observed data model, the parameter space is different. Since $\theta_j(x)$ is a probability in the latent outcome model, $0 \leqslant \theta_j(x) \leqslant 1$ for all j and x. This is in contrast to the model for the observed data, which requires only the overall probabilities $P(\mathbf{x}; \boldsymbol{\theta})$ to be in $[0, 1]$ and not the individual terms within the product (2.1). Thus, the latent outcome model is defined over a restricted parameter space, in which the risk associated with any covariate pattern $\mathbf{x} \in \mathcal{X}$ does not exceed the risk associated with the reference covariate pattern $\mathbf{r}$:

$$\Theta(\mathbf{r}) = \{\boldsymbol{\theta} : 0 \leqslant P(\mathbf{x}; \boldsymbol{\theta}) \leqslant P(\mathbf{r}; \boldsymbol{\theta}) \leqslant 1 \quad \text{for all } \mathbf{x} \in \mathcal{X}\} \subset \Theta. \tag{2.4}$$

The latent outcome log likelihood is then defined using the sufficient statistics

$$Z_{ij} = \sum_{k=1}^{N_i} Z_{ij}^{(k)} \sim \text{Bin}(N_i, \theta_j(x_{ij})) \quad \text{independently,}$$

which leads to the log likelihood for $\boldsymbol{\theta} \in \Theta(\mathbf{r})$,

$$\ell(\boldsymbol{\theta}|\{Z_{ij}\}) = \sum_{i=1}^{n} \sum_{j=0}^{A} Z_{ij} \log\{\theta_j(x_{ij})\} + (N_i - Z_{ij}) \log\{1 - \theta_j(x_{ij})\}.$$

2.3 *EM algorithm*

Making use of the above latent outcome model, we first describe an EM algorithm to maximize L over $\Theta(\mathbf{r})$ and then discuss an extension that maximizes L over Θ.

Beginning with an initial parameter value $\hat{\boldsymbol{\theta}}_{(0)} \in \Theta(\mathbf{r})$, iteration $c+1$ of the EM algorithm updates the current iterate $\hat{\boldsymbol{\theta}}_{(c)}$ to a new iterate $\hat{\boldsymbol{\theta}}_{(c+1)}$ in 2 steps (McLachlan and Krishnan, 2008). In the present context, the E-step consists of calculating

$$Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}_{(c)}) = E\left\{\ell(\boldsymbol{\theta}|\{Z_{ij}\})\big|\{Y_i\}; \hat{\boldsymbol{\theta}}_{(c)}\right\} = \ell(\boldsymbol{\theta}|\{\hat{Z}_{ij(c)}\}),$$

where

$$\hat{Z}_{ij(c)} = E(Z_{ij}|Y_i; \hat{\boldsymbol{\theta}}_{(c)}) = Y_i + (N_i - Y_i) e_{ij}(\hat{\boldsymbol{\theta}}_{(c)})$$

and

$$e_{ij}(\boldsymbol{\theta}) = E\left(Z_{ij}^{(k)}|Y_i^{(k)} = 0\right) = \frac{\theta_j(x_{ij}) - P(\mathbf{x}_i; \boldsymbol{\theta})}{1 - P(\mathbf{x}_i; \boldsymbol{\theta})}.$$

The M-step then consists of maximizing $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}_{(c)})$ over $\Theta(\mathbf{r})$ to obtain $\hat{\boldsymbol{\theta}}_{(c+1)}$. Unlike the model for the observed outcomes, the latent outcome model allows separation of the individual parameter components in the log likelihood function, which reduces the M-step maximization to a collection of simple one-parameter estimation problems. Thus, letting $I_{jx} = \{i : x_{ij} = x\}$ be the set of observations that have covariate j equal to x, the components of $\hat{\boldsymbol{\theta}}_{(c+1)}$, for $j = 0, \ldots, A$ and $x = 1, \ldots, k_j$, are

$$\hat{\theta}_j^{(c+1)}(x) = \sum_{i \in I_{jx}} \hat{Z}_{ij(c)} \Big/ \sum_{i \in I_{jx}} N_i. \tag{2.5}$$

As long as $\hat{\boldsymbol{\theta}}_{(0)}$ belongs to $\Theta(\mathbf{r})$ then so too will each subsequent iterate. The EM algorithm therefore increases L within $\Theta(\mathbf{r})$, ultimately leading to maximization over $\Theta(\mathbf{r})$.

To fit (2.1), we need to maximize L over the full parameter space Θ, not the restricted parameter space $\Theta(\mathbf{r}) \subset \Theta$. However, at least one of the restricted parameter spaces must contain the maximum of L over the full parameter space because

$$\Theta = \bigcup_{\mathbf{r} \in \mathcal{X}} \Theta(\mathbf{r}).$$

This means that if we cycle through all $K = \prod_j k_j$ possible choices of $\mathbf{r} \in \mathcal{X}$ and implement the above EM algorithm for each choice, then at least one of the K limit points will provide the maximum of L over the full parameter space. Thus, maximization of L over Θ involves running the above EM algorithm K times, once for each possible choice of $\mathbf{r}$, and inspecting the limit points to determine which one achieves

the greatest value of L. This limit point is then the MLE $\hat{\boldsymbol{\theta}}$, being the constrained maximum of L over Θ. As discussed in Section 6, in practice, this search can terminate if one of the limit points is found to be stationary. The technique of cycling through restricted parameter spaces is analogous to that used by Marschner (2010) for identity link Poisson models, albeit with a different EM algorithm and different parameter constraints.

3. Linear covariates

A latent outcome model analogous to Section 2.2 can be used to develop an EM algorithm accommodating covariates that are linear on the log link scale. The justification is more complicated in this context, although the idea is the same.

Suppose that Y_i has B covariates that enter the model linearly on the log link scale, along with a constant covariate for the intercept term. For succinctness, we assume these are the only covariates; however, it is straightforward to use the methods of Section 2 in parallel to those presented in this section. The covariate vector is $\mathbf{x}_i = (x_{i0}, \ldots, x_{iB})$ with $x_{i0} = 1$. The covariate space is taken to be the Cartesian product of the covariate ranges $\mathcal{X} = \prod_j \left[x_j^{(0)}, x_j^{(1)}\right]$, where $x_j^{(0)} = \min_i\{x_{ij}\}$ and $x_j^{(1)} = \max_i\{x_{ij}\}$. Note that the covariates can be polynomial or other transformations of the observed explanatory variables, allowing for parametrically specified nonlinearities.

The log-binomial model is specified by the log link GLM

$$p_i = P(\mathbf{x}_i; \boldsymbol{\theta}) = \exp\left\{\sum_{j=0}^{B} \alpha_j x_{ij}\right\} = \prod_{j=0}^{B} \theta_j^{x_{ij}}, \tag{3.1}$$

where $\theta_j = \exp(\alpha_j)$ and $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_B)$. For applications of the log-binomial model to RR regression, θ_j is interpreted as the adjusted RR associated with a one-unit increase in covariate $j = 1, \ldots, B$. With $P(\mathbf{x}_i; \boldsymbol{\theta})$ defined by (3.1), the parameter space Θ and log likelihood function L are again given by (2.2) and (2.3).

While (3.1) is the natural way to express the model, a reparameterization is needed so that a latent outcome model can be used to interpret the parameters as probabilities, as in Section 2.2. First, for each covariate, we choose a reference value r_j that is either the minimum or maximum of the observed values of the covariate; that is, $r_j = x_j^{(s_j)}$ for $s_j \in \{0, 1\}$. Any of the 2^B elements of the Cartesian product $\mathcal{P} = \prod_j \left\{x_j^{(0)}, x_k^{(1)}\right\}$ may be chosen for the reference vector $\mathbf{r} = (r_0, \ldots, r_B)$, and different choices will lead to different forms for the reparameterization. Next, for $j = 1, \ldots, B$ define:

$$u_{i0} = 1; \quad u_{ij} = (-1)^{s_j}(x_{ij} - r_j); \quad \beta_0 = \alpha_0 + \sum_{j=1}^{B} \alpha_j r_j; \quad \beta_j = (-1)^{s_j}\alpha_j.$$

This leads to an equivalent parameterization of the model (3.1),

$$P(\mathbf{x}_i; \boldsymbol{\theta}) = P(\mathbf{u}_i; \boldsymbol{\lambda}) = \exp\left\{\sum_{j=0}^{B} \beta_j u_{ij}\right\} = \prod_{j=0}^{B} \lambda_j^{u_{ij}}, \tag{3.2}$$

where $\lambda_j = \exp(\beta_j)$, $\boldsymbol{\lambda} = (\lambda_0, \ldots, \lambda_B)$, and $\mathbf{u}_i = (u_{i0}, \ldots, u_{iB})$. The advantage of the parameterization (3.2) is that, unlike x_{ij} in (3.1), $u_{ij} \geqslant 0$ and this allows the components of $\boldsymbol{\lambda}$ to be viewed as probabilities in a latent outcome model. In particular, without loss of generality, assume that u_{ij} is an integer. This is

possible because any continuous covariate must be measured to a finite number of decimal places and is therefore able to be rescaled to integer values. Then, similarly to Section 2.2, the N_i binary outcomes underlying Y_i can be viewed as the product of independent latent binary outcomes

$$Y_i^{(k)} = \prod_{j=0}^{B} \prod_{u=1}^{u_{ij}} Z_{iju}^{(k)} \quad \text{where} \quad \Pr\left(Z_{iju}^{(k)} = z\right) = \lambda_j^z (1-\lambda_j)^{1-z} \quad \text{for} \quad z = 0, 1.$$

This correspondence arises because the event probability associated with the observed outcomes is the joint probability associated with the latent outcomes

$$\Pr\left(Y_i^{(k)} = 1\right) = \Pr\left(\bigcap_{j=0}^{B} \bigcap_{u=1}^{u_{ij}} \left\{Z_{iju}^{(k)} = 1\right\}\right),$$

and this joint probability is identical to the product probability in (3.2).

Similarly to Section 2.2, λ_j is a probability in the latent outcome model and is therefore restricted to the interval $[0, 1]$. Thus, since $u_{ij} \geqslant 0$ with equality when $x_{ij} = r_j$, (3.2) implies that under the latent outcome model, $\boldsymbol{\lambda}$ (or equivalently $\boldsymbol{\theta}$) belongs to a restricted parameter space, in which the risk associated with any covariate pattern $\mathbf{x} \in \mathcal{X}$ does not exceed the risk associated with the reference pattern $\mathbf{r} \in \mathcal{P}$. This restricted parameter space $\Theta(\mathbf{r})$ can again be written in the form (2.4). For the latent outcome model defined over $\Theta(\mathbf{r})$, the sufficient statistics are

$$Z_{ij} = \sum_{k=1}^{N_i} \sum_{u=1}^{u_{ij}} Z_{iju}^{(k)} \sim \text{Bin}(N_i u_{ij}, \lambda_j) \quad \text{independently.}$$

It follows that iteration $c+1$ of the EM algorithm involves updating the current iterate $\hat{\boldsymbol{\lambda}}_{(c)} = \left(\hat{\lambda}_0^{(c)}, \ldots, \hat{\lambda}_C^{(c)}\right)$ using

$$\hat{\lambda}_j^{(c+1)} = \sum_{i=1}^{n} u_{ij} \hat{Z}_{ij(c)} \Big/ \sum_{i=1}^{n} u_{ij} N_i,$$

where

$$\hat{Z}_{ij(c)} = Y_i + (N_i - Y_i) \frac{\hat{\lambda}_j^{(c)} - P(\mathbf{u}_i; \hat{\boldsymbol{\lambda}}_{(c)})}{1 - P(\mathbf{u}_i; \hat{\boldsymbol{\lambda}}_{(c)})}.$$

This EM algorithm maximizes the observed outcome log likelihood L over $\Theta(\mathbf{r})$. Using a similar justification as Section 2.3, maximization of L over Θ is achieved by repeatedly invoking the above EM algorithm for each of the 2^B possible choices of $\mathbf{r} \in \mathcal{P}$ and then choosing the restricted maximum that obtains the greatest log likelihood value.

4. FLEXIBLE REGRESSION

For unspecified functions $f_1, \ldots, f_B$, a flexible generalization of (3.1) is

$$p_i = P(\mathbf{x}_i; \boldsymbol{\theta}) = \exp\left\{\alpha_0 + \sum_{j=1}^{B} f_j(x_{ij})\right\}. \tag{4.1}$$

Since the method in Section 2 is highly stable, f_j in (4.1) can be estimated at the observed covariate values using a categorical specification. Let $\{z_j(x); x = 1, \ldots, m_j + 1\}$ be the $m_j + 1$ ordered unique

values from $\{x_{ij}; i = 1, \ldots, n\}$, so that $z_j(x) < z_j(y)$ for $x < y$. Then the methodology of Section 2 can be applied with B categorical covariates using the notation $\alpha_j(x) = f_j(z_j(x))$, or equivalently $\theta_j(x) = \exp\{f_j(z_j(x))\}$, for $j = 1, \ldots, B$ and $x = 1, \ldots, k_j$ where $k_j = m_j + 1$.

In practice, the above approach may lead to regression functions that are implausibly erratic if the covariates have a large number of unique values. The EM algorithm is particularly suited to address this through regularization of the unspecified regression functions. In this section, we discuss such regularization via a monotonicity requirement. In the supplementary material available at *Biostatistics* online, we provide additional details on flexible regression models, including smoothing via penalized likelihood estimation.

If the unspecified regression functions $f_1, \ldots, f_B$ are assumed to be nondecreasing, then (4.1) becomes a flexible isotonic regression model. Such models are often appropriate in applications of RR regression because some risk factors may be known to have a nondecreasing relationship with risk. However, existing methods for isotonic GLMs do not cover the log-binomial model because they are limited to canonical models (Bacchetti, 1989; Moreton-Jones *and others*, 2000).

Since the regression functions are unspecified, they are only estimable at the observed values of the covariates; so for an isotonic model, we can assume f_j is a step function with nonnegative increments at $\{z_j(x)\}$. An identifiability restriction is required on each function f_j, which we take as $f_j(z_j(m_j+1)) = 0$ for $j = 1, \ldots, B$. Then, for $x = 1, \ldots, m_j$, the step function f_j evaluated at its jump points can be written as

$$f_j(z_j(x)) = \sum_{r=x}^{m_j} \beta_j(r) \quad \text{where} \quad \beta_j(r) \leqslant 0. \tag{4.2}$$

Next define $B^* = \sum_{j=1}^{B} m_j$ dummy binary covariates $\{w_{il}; l = 1, \ldots, B^*\}$. These dummy variables take the values 1 or 2 using $w_{il} = 1\{x_{ij} > z_j(x)\} + 1$, where $1\{\cdot\}$ is an indicator function and l ranges over $1, \ldots, B^*$ according to

$$l = \sum_{q=1}^{j-1} m_q + x \quad \text{for} \quad j = 1, \ldots, B \quad \text{and} \quad x = 1, \ldots, m_j. \tag{4.3}$$

The methodology presented in Section 2 can then be applied with B^* binary categorical covariates $\{w_{il}; l = 1, \ldots, B^*\}$. In each case, $k_l = 2$ and the computations from Section 2.3 can be applied after using (4.3) to define $\alpha_l(1) = \beta_j(x)$ and $\alpha_l(2) = 0$, or equivalently $\theta_l(1) = \exp(\beta_j(x))$ and $\theta_l(2) = 1$. The EM algorithm will then automatically restrict $\hat{\theta}_l(1)$ to be in $[0, 1]$, or equivalently $\hat{\alpha}_l(1) \leqslant 0$, which automatically preserves the monotonicity of the fitted regression functions through (4.2). This is what makes the EM algorithm a particularly stable and attractive method for isotonic log-binomial regression. Note that, in contrast with the general categorical covariates discussed in Section 2, monotonicity restrictions require that the reference levels for the dummy binary covariates are always $r_l = k_l = 2$, so there is no cycling through the other possible reference level combinations in the manner described in Section 2.3.

5. Applications and simulations

5.1 *Analysis of heart attack data*

We illustrate the proposed method with an analysis of the age-specific RR of mortality within 30 days of a heart attack. The data come from the ASSENT-2 study, which was a clinical trial comparing 2 agents, tenecteplase and alteplase, in 16 949 individuals treated within 6 h of the onset of heart attack symptoms (ASSENT-2 Investigators, 1999). For our illustrative analyses, we consider 3 covariates in addition to age: the severity of heart failure as measured using the standard Killip classification scheme, the time delay

from symptom onset to the receipt of treatment, and the geographical region where the patient was treated. The type of treatment was not considered as a predictor since the mortality risk was virtually identical for both agents.

The proposed method is implemented in an R function named `logbin`, which is given in the supplementary material available at *Biostatistics* online. It is compared here with the standard R function `glm`, which uses IRLS enhanced by step halving (R Development Core Team, 2010). We begin with a basic analysis that demonstrates the potential for numerical difficulties and illustrates the use of the proposed method for overcoming these difficulties. A model with 9 parameters is used where all 4 predictors are categorical variables at 3 levels—age: <65, 65–75, or >75 years; severity: Killip class I, II, or III/IV; treatment delay: <2, 2–4, or >4 h; and region: Western countries, Latin America, or Eastern Europe. Of 81 possible covariate combinations, $n = 74$ were observed, each providing one binomial observation for a residual degrees of freedom of $74-9 = 65$.

The observed mortality risk overall was $1045/16949 = 0.062$, but this varied considerably across the various covariate combinations. The log-binomial model fitted by `glm` failed to converge, beginning with initial log RRs of zero for all covariates and an initial intercept of log(0.062). The supplementary material available at *Biostatistics* online include the iterates from `glm`, which were attracted to a suboptimal repeating sequence of period 8, with deviance cycling in the range 149.52–165.31. Also included in the supplementary material available at *Biostatistics* online is an example of aperiodic nonconvergence from `glm`. In contrast, the EM algorithm implemented in `logbin` converged to a stationary MLE with a model deviance of 149.32. Table 1 displays the fitted RRs obtained from `logbin` and 95% confidence intervals based on 1000 bootstrap replications sampled with replacement from the observed data. The EM algorithm converged for all 1000 replications, whereas `glm` converged in only 25.9% of replications. Also presented are the ranges of fitted mortality risks for each covariate level. Thus, for example, the mortality risk in the youngest age group ranges from 0.018 to 0.14, depending on the values of the other covariates. With estimated mortality risks ranging between 0.018 and 0.93, it can be seen that the MLE is in the interior of the parameter space.

Also displayed in Table 1 is a comparison of the MLE with one of the popular alternatives discussed in Section 1, the log-linear Poisson approximation. While some RRs are similar for the 2 methods, differences in excess of 20% are observed within some of the higher risk covariate levels. This observation is consistent with our simulation studies in Section 5.2 and the supplementary material available at *Biostatistics* online, where the greatest differences arise in higher risk settings. The differences between the methods in Table 1 cannot be explained by the fact that the MLE is constrained since all MLE risks lie in the interior of $[0, 1]$. In this respect, the MLE would seem preferable to the unconstrained Poisson method, which has some risks that substantially exceed 1.

Next, we consider 3 further analyses with the same covariates except that age is now modeled in years within the range 40 to 85, providing $n = 752$ observed covariate combinations. The 3 analyses involve incorporating age through: a one-parameter linear function, a 45-parameter categorical specification, and an unspecified isotonic regression function. Additional analyses presented in the supplementary material available at *Biostatistics* online also illustrate a smoothed version of the isotonic analysis. The high-dimensional categorical model is included for illustrative purposes to demonstrate the stability of the proposed method, whereas the linear and isotonic specifications provide realistic models given that mortality risk is known to increase with age. The 2 models that `glm` can accommodate, those with linear and categorical age specifications, again failed to converge. In contrast, `logbin` converged for both models, as well as the isotonic analysis. Figure 1 presents the fitted age-specific mortality risks for the 3 analyses and shows that the linear model provides a good approximation to the more flexible isotonic model.

Further analyses revealed an interaction between age and severity and are presented here as examples of higher dimensional analyses. The first of these uses 3 flexible isotonic regression functions for age, one for each severity level, and involves a total of 139 parameters. This is compared to an analogous

Table 1. *RR of mortality for 16 949 heart attack patients, using the log-binomial MLE and the log-linear Poisson approximation. For each level of the covariates, the risk range is the range of fitted mortality probabilities across all combinations of the other covariates. Age is in years, severity is Killip class, treatment delay is in hours, and regions are Western countries (West), Latin America (LA), and Eastern Europe (EE). The 95% CIs are based on 1000 bootstrap replications, for which both methods always converged*

	MLE			Poisson		
	RR	CI	Risk range	RR	CI	Risk range
Age						
<65	1	—	0.018–0.14	1	—	0.017–0.20
65–75	3.02	1.65–4.32	0.041–0.41	3.00	1.76–4.21	0.051–0.61
>75	6.87	3.35–9.55	0.12–0.93	7.26	3.74–9.99	0.12–1.48
Severity						
I	1	—	0.018–0.24	1	—	0.017–0.30
II	2.02	1.56–2.88	0.035–0.48	2.06	1.62–2.95	0.035–0.62
III/IV	3.96	3.09–5.87	0.071–0.93	4.88	3.64–7.58	0.083–1.48
Delay						
<2	1	—	0.018–0.79	1	—	0.017–1.19
2–4	1.06	0.76–1.45	0.020–0.83	1.07	0.76–1.45	0.018–1.27
>4	1.19	0.85–1.58	0.021–0.93	1.25	0.84–1.63	0.020–1.48
Region						
West	1	—	0.018–0.58	1	—	0.017–0.75
LA	1.08	0.69–1.56	0.019–0.62	1.05	0.68–1.53	0.018–0.79
EE	1.62	1.22–1.98	0.029–0.93	1.98	1.45–2.54	0.034–1.48
Constant	0.018	0.013–0.034	0.018–0.93	0.017	0.012–0.031	0.017–1.48

CI, confidence interval.

model having 3 linear functions for age, involving just 10 parameters. Unlike the main effects models discussed above, `glm` did converge for the linear interaction model, yielding the same parameter estimates as `logbin` and a model deviance of 716.99 on 742 degrees of freedom. Table 2 displays the results of the linear interaction analysis, showing that the RR of mortality increases less steeply with age when the heart attack is more severe. Figure 2 compares the linear model with the more flexible isotonic model obtained from `logbin`, confirming that linear age-dependence suffices. This illustrates the use of our flexible regression methods for investigating an appropriate parametric form for the regression functions.

5.2 *Simulation studies*

Carter *and others* (2005) conducted simulations of the log-linear Poisson method for estimating age-specific RR. We conducted simulations using the same assumptions to assess the MLE using the EM algorithm and to compare it with the log-linear Poisson method. Age in years was distributed $N(50.07, 10.55^2)$, with sample sizes from $n = 100$ to $n = 700$. Six parameter combinations were simulated 2000 times, with RRs per year from 0.966 to 0.996. The absolute risk at the average age ranged from 0.15 to 0.93. The $n = 300$ results are discussed here and are similar to the other sample sizes reported in the supplementary material available at *Biostatistics* online.

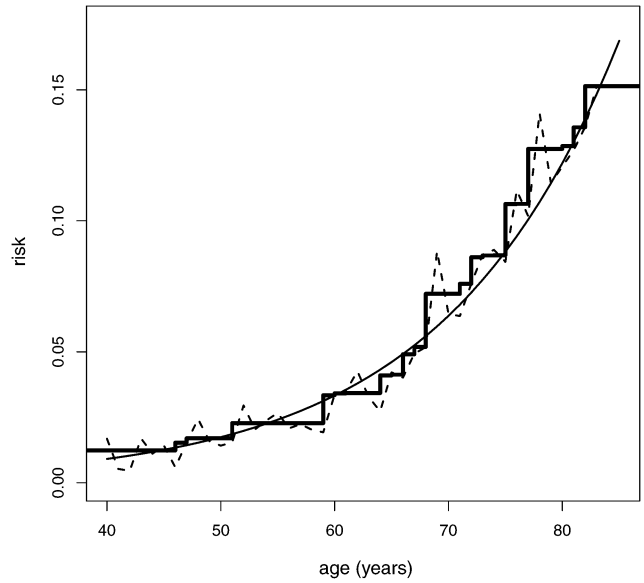

Fig. 1. Age-specific heart attack mortality from RR regression models with either unrestricted (dotted line), isotonic (thick line), or linear (thin line) dependence on the log link scale. Adjusted risk estimates are displayed for individuals from the Western region with low severity event and treatment delay <2 h.

Table 2. *RR from a log-binomial model for heart attack mortality with an interaction between severity and the linear age effect. CIs are calculated using the log-binomial information matrix. Risk ranges are defined as in Table 1*

Severity	Class I	Class II	Class III or IV
RR at age 85 years (95% CI)	1	1.15 (0.92–1.44)	1.56 (1.13–2.16)
RR per year younger (95% CI)	0.92 (0.91–0.93)	0.96 (0.95–0.97)	0.99 (0.98–1.01)
Risk range	0.005–0.47	0.04–0.55	0.24–0.74
Treatment delay	<2 h	2–4 h	>4 h
RR (95% CI)	1	1.03 (0.90–1.18)	1.19 (1.02–1.39)
Risk range	0.005–0.62	0.005–0.64	0.006–0.74
Region	Western	Latin America	Eastern Europe
RR (95% CI)	1	1.12 (0.79–1.58)	1.85 (1.51–2.27)
Risk range	0.005–0.40	0.006–0.45	0.01–0.74

We first discuss the 5 simulation combinations that had risk levels in the range 0.15 to 0.72 at the average age. The EM algorithm converged to the MLE in 100% of the simulations. While the 2 methods behaved very similarly at the lower risk levels, there was a general deterioration in the performance of the log-linear Poisson method as the risk level increased. This is depicted in Figure 3, where the relative efficiency of the log-linear Poisson method decreases with risk, in tandem with an increasing rate of

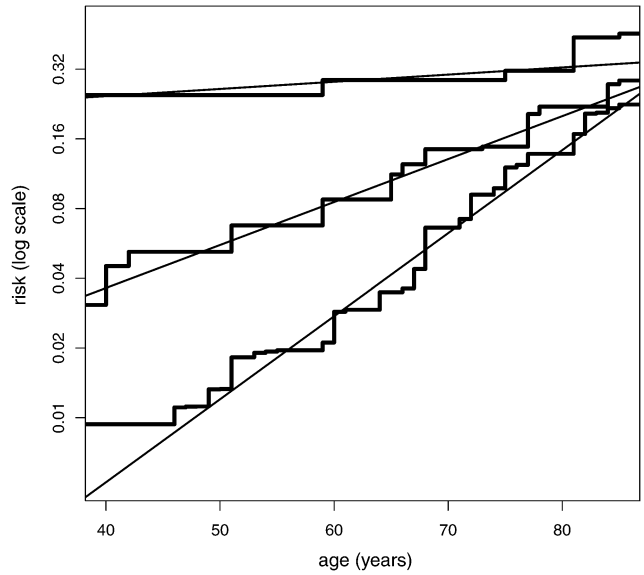

Fig. 2. RR regression models with either linear or isotonic age dependence and an interaction between age and severity. Risk estimates are for individuals from the Western region with <2 h treatment delay and are displayed separately by event severity (lower lines = Class I, upper lines = Class III/IV).

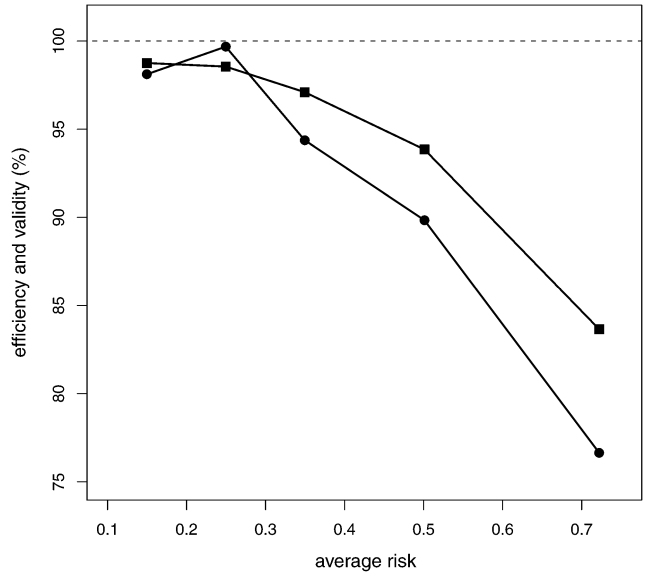

Fig. 3. Comparison of log-binomial and log-linear Poisson estimation of RR, for 5 sets of 2000 simulations with $n = 300$. Average risk means the risk at the average age in the population. Circles denote the relative efficiency of the Poisson method compared to the binomial method, using a ratio of mean squared errors. Squares denote the percent of simulations in which the Poisson method gave a valid fit (estimated probabilities not exceeding 1). The log-binomial fit was valid in all simulations and both methods always converged.

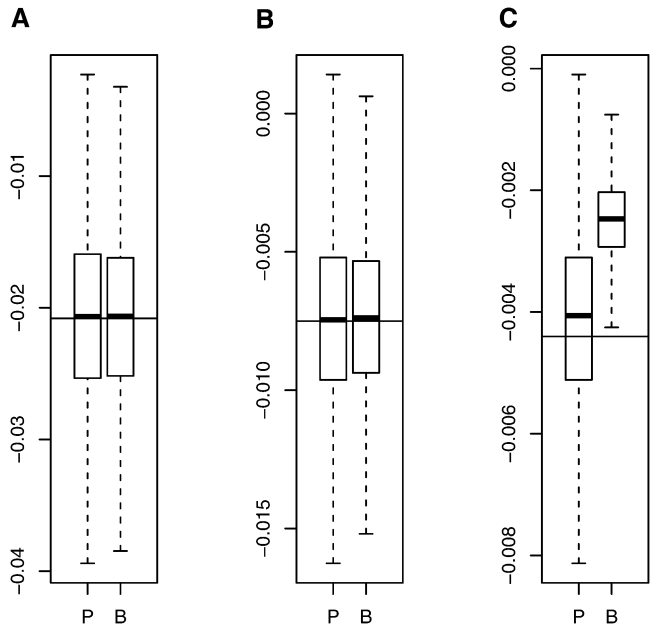

Fig. 4. Boxplots of log-binomial (B) and log-linear Poisson (P) estimates of log RR, for 2000 simulations with $n = 300$ and average risk levels of (A) 0.35, (B) 0.72, and (C) 0.93. Horizontal lines denote the true log RR.

parameter space violations. For both methods, the bias was acceptable across all 5 combinations, for all sample sizes, as illustrated in Figure 4 panels A and B.

The sixth parameter combination had an extremely high risk of 0.93 at the average age, with results displayed in Figure 4 panel C. The EM algorithm again achieved 100% convergence to the MLE, and there was lower variation in the MLE compared to the log-linear Poisson estimate. However, at these high risk levels, there was bias in the MLE since the imposition of an upper bound caused attenuation of the gradient between risk and age. The bias was much lower for the unconstrained Poisson method, although in that case only 10.9% of simulations had valid fitted values.

In interpreting these results, it is important to distinguish between the behavior of the MLE and the behavior the EM algorithm as a method for computing the MLE. The simulations confirmed that the EM method is highly reliable for computing the MLE. However, they have also shown that the imposition of the parameter constraints can lead to bias in the MLE when the risk levels are extremely high. For risk levels at more realistic levels, the MLE was superior to the log-linear Poisson method.

6. DISCUSSION

This paper has presented a new method for fitting log-binomial models and has illustrated its use in RR regression. Our approach uses the stability of the EM algorithm in constrained estimation contexts to overcome problems with IRLS-based methods. It offers some flexible generalizations and remains reliable even in high dimensions.

When IRLS does converge, it will be faster than the EM algorithm. Our method will therefore be most useful when IRLS fails. In such cases, convergence rate is a secondary issue and the priority is to have a method that works. Nonetheless, convergence acceleration may be a worthwhile improvement if it can be achieved without compromising stability. Computational time can be reduced by starting the

EM algorithm in the parameter space subset (2.4) that contains the MLE. If a stationary point is found, then the other restricted parameter spaces do not need to be checked. For this reason, the R code in the supplementary material available at *Biostatistics* online allows an initial guess of the highest risk covariate pattern.

Our approach produces the MLE over the constrained parameter space. Lumley *and others* (2006) argued that unconstrained non-MLE methods are more appropriate because restricting fitted probabilities to the interval [0, 1] can mean some observations exert undue leverage on the fit. This was evident in our simulations, but it only led to bias in the MLE when the risks were extremely high. For more realistic risks, our simulations demonstrated superior performance of the MLE compared to the unconstrained log-linear Poisson approximation. Accordingly, we would generally recommend the MLE; however, in high-risk contexts, it would seem prudent also to examine an unconstrained estimate or to use the more flexible constrained estimates presented here.

Although we only illustrated numerical instability in R, Carter *and others* (2005) and Blizzard and Hosmer (2006) reported high nonconvergence rates using SAS and Stata, and we have also observed nonconvergence in S-PLUS and SPSS.

Finally, we emphasize that even when standard methods do converge to the MLE, numerical instability remains a barrier to using the log-binomial model. Prospective studies, particularly clinical trials, often require prespecified analysis plans, which make it problematic to commit to the log-binomial model before one knows it will work. Furthermore, auxiliary analyses such as bootstrapping, cross-validation, and multiple imputation may require many log-binomial models to be fitted, all of which must converge for a valid analysis. The reliability of our method may therefore provide greater confidence in using RR regression as a primary analysis method in prospective studies.

Supplementary materials

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

We thank 2 reviewers and an associate editor for their comments and the ASSENT-2 investigators for providing data to the National Health and Medical Research Council Clinical Trials Centre. *Conflict of Interest:* None declared.

Funding

The Australian Research Council (DP110101254).

References

ASSENT-2 Investigators (1999). Single-bolus tenecteplase compared with front-loaded alteplase in acute myocardial infarction: the ASSENT-2 double-blind randomised trial. *Lancet* **354**, 716–722.

Bacchetti, P. (1989). Additive isotonic models. *Journal of the American Statistical Association* **84**, 289–294.

Blizzard, L. and Hosmer, D. W. (2006). Parameter estimation and goodness-of-fit in log binomial regression. *Biometrical Journal* **48**, 5–22.

Carter, R. E., Lipsitz, S. R. and Tilley, B. C. (2005). Quasi-likelihood estimation for relative risk regression models. *Biostatistics* **6**, 39–44.

DEDDENS, J. A. AND PETERSEN, M. R. (2004). Re: "Estimating the relative risk in cohort studies and clinical trials of common outcomes". *American Journal of Epidemiology* **159**, 213–214.

LEE, J. (1994). Odds ratio or relative risk for cross-sectional data? *International Journal of Epidemiology* **23**, 201–203.

LUMLEY, T., KRONMAL, R. AND MA, S. (2006). Relative risk regression in medical research: models, contrasts, estimators and algorithms. *University of Washington Biostatistics Working Paper Series, Working Paper 293*. http://www.bepress.com/uwbiostat/paper293.

MARSCHNER, I. C. (2010). Stable computation of maximum likelihood estimates in identity link Poisson regression. *Journal of Computational and Graphical Statistics* **19**, 666–683.

MCLACHLAN, G. J. AND KRISHNAN, T. (2008). *The EM Algorithm and Extensions*, 2nd edition. Hoboken, NJ: Wiley.

MCNUTT, L. A., WU, C., XUE, X. AND HAFNER, J. P. (2003). Estimating the relative risk in cohort studies and clinical trials of common outcomes. *American Journal of Epidemiology* **157**, 940–943.

MORETON-JONES, T., DIGGLE, P., PARKER, L., DICKINSON, H. O. AND BINKS, K. (2000). Additive isotonic regression models in epidemiology. *Statistics in Medicine* **19**, 849–859.

R DEVELOPMENT CORE TEAM (2010). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

SPEIGELMAN, D. AND HERTZMARK, E. (2005). Easy SAS calculations for risk or prevalence ratios and differences. *American Journal of Epidemiology* **162**, 199–200.

WACHOLDER, S. (1986). Binomial regression in GLIM: estimating risk ratios and risk differences. *American Journal of Epidemiology* **123**, 174–184.

YU, B. AND WANG, Z. (2008). Estimating relative risks for common outcome using PROC NLP. *Computer Methods and Programs in Biomedicine* **90**, 179–186.

ZOU, G. (2004). A modified Poisson regression approach to prospective studies with binary data. *American Journal of Epidemiology* **159**, 702–706.

[*Received February 7, 2011; revised August 4, 2011; accepted for publication August 5, 2011*]