ON THE ESTIMATION OF RELATIVE RISKS VIA LOG BINOMIAL REGRESSION

Bernardo Borba de ANDRADE[1]
Hélène CARABIN[2]

▪ *ABSTRACT: Given the well known convergence difficulties in fitting log binomial regression with standard GLM software, we implement a direct solution via constrained optimization which avoids the circumventions found in the literature. The use of a log binomial model is motivated by our interest in directly estimating relative risks adjusted for confounders. A Bayesian log binomial regression model is also discussed for a dataset of epidemiological interest. We developed R functionality to illustrate our proposal.*

▪ *KEYWORDS: Log binomial regression; constrained maximum likelihood; quasilikelihood; relative risks.*

1 Introduction

Generalized linear models (GLM) can be used to estimate probabilities or odds adjusted for continuous and discrete covariates. The main interest is typically in estimating the association between a set of possible risk factors and a dichotomous outcome using *odds ratios* or *risks ratios*. Denoting the outcome of interest by $y \in \{0, 1\}$ and letting $p = \Pr(y = 1|\text{risk factors})$, we can represent a GLM by

$$h(p) = f(\text{risk factors}),$$

where h is known as the *link function* and f is a linear predictor.

[1]Universidade Federal do Rio Grande do Norte – UFRN, Centro de Ciências Exaras e da Terra – CCET, Departamento de Estatística, CEP: 59078-970, Natal, RN, Brazil. E-mail: *bba@ccet.ufrn.br*
[2]University of Oklahoma HSC, COPH, Department of Biostatistics & Epidemiology, OK 73104, Oklahoma City, USA. E-mail: *hcarabin@ouhsc.edu*

Given two probabilities of interest, p_1 and p_2, adjusted odds ratios (OR),

$$\frac{p_1/(1-p_1)}{p_2/(1-p_2)},$$

are directly obtained from logistic regression, whereas adjusted relative risks (RR),

$$\frac{p_1}{p_2},$$

are naturally estimated by another GLM, known as *log binomial regression* (LBR). Logistic regression has link function given by $\text{logit}(p) = \log[p/(1-p)]$ whereas LBR has $\log(p)$ as its link.

In logistic regression, when the maximum likelihood (ML) estimate exists (Geyer, 2009b) for an example where the ML estimate does not exist but it can be defined in an unconventional way), it can often be found by the usual optimization procedures implemented in GLM software and the estimated probabilities are valid, that is $0 \leq \hat{p} \leq 1$, with no constraints on the parameter space. The use of a log link, however, entails some technical difficulties. Without constraints on the parameter space, LBR may estimate probabilities greater than unit, preventing common statistical packages from successfully fitting the given model.

In the special case of very low cumulative incidence, $p \approx 0$, the OR approaches the RR. This is also the case in time-matched case-control studies where the controls are matched on the time of occurrence of the cases, rendering the OR equal to the incidence rate ratio and to the RR. However, when the disease is common, the OR will overestimate the RR and hence the incidence rate ratio. The fact that the OR and the RR can be close under some special instances combined with the convergence problems found in the direct implementation of LBR has led many to use logistic regression and incorrectly interpret the OR as RR even when the incidence is moderate or high. Coutinho *et al.* (2008) apply LBR and alternative methods to estimate adjusted RR for outcomes with low, moderate and high prevalence in cross-sectional data. They report that logistic regression provided highly discrepant results (in line with the simulation study of McNutt *et al.*, 2003) and they also reported convergence problems while using LBR (as also reported by Carter *et al.*, 2005).

The two measures of association, RR and OR, have been extensively compared in the epidemiological literature (*e.g.* Axelson *et al.*, 1994; Lee, 1994; Stromberg, 1994; Hughes, 1995; Osborn and Cattaruzza, 1995; and, more recently, Reichenheim and Coutinho, 2010). See also Fisher (1954) for an early comparison of different models for binary responses with discussion by other famous statisticians. Martuzzi and Elliott (1998) emphasize correct interpretation rather than uniform superiority of one measure over the other: "In practice, the two measures will give almost identical results for low-prevalence diseases; and, *if interpreted correctly*, they will lead to similar conclusions also for non-rare conditions."

RR is, in general, much more easily interpreted than OR. One may argue that direct interpretation is not enough to justify the additional computational

requirements needed to estimate adjusted RR. However, the fact that there is abundant literature (*e.g.* Martuzzi and Elliott, 1998; Barros and Hirakata, 2003; Deddens *et al.*, 2003; McNutt *et al.*, 2003; Zou, 2004; Carter *et al.*, 2005) on how to circumvent the often difficult estimation of RR by ML, shows that the OR is not universally preferred among applied researchers when the outcome of interest is not rare.

Bearing the above paragraphs in mind, we will assume in this paper that the RR, not the OR, is the chosen measure of association.

Convergence problems when fitting LBR have been reported in the literature and they have motivated alternative methods designed to circumvent the convergence issues. The alternatives proposed include: simple truncation of fitted values, the COPY algorithm, tweaking Cox regression software, adapting logistic regression by scaling by the average prevalence, stratified analysis, duplication of cases and notably, robust quasilikelihood (QL). For studies comparing different subsets of the above alternatives, see Barros and Hirakata (2003), McNutt *et al.* (2003), Zou (2004), Carter *et al.* (2005), Blizzard and Hosmer (2006), Lumley *et al.* (2006) and Coutinho *et al.* (2008). Our main conclusion from these studies has been towards critical use of current software and recommendation of the QL approach for its simplicity, desired asymptotic properties and easy extension to multilevel analyses. In addition, we recommend that, if the data analyst has access to a flexible computing system, constrained optimization routines be explored so those circumventions can be avoided.

We note that the epidemiological literature has focused on the above-mentioned circumventions, rather than on allowing LBR to be fitted by constrained optimization, as opposed to the unconstrained algorithms on which GLM software is typically based. As far as we know, this approach first appeared quite recently in the computational biology literature (Yu and Wang, 2008) where nonlinear programming (in SAS) was implemented to find ML estimates of the coefficients in LBR. The literature on regression models with parameter constraints includes the work of Judge and Takayama (1966) when linear inequalities where first treated in linear regression models and Geyer (1991) where logistic regression restricted by ordered parameters is treated.

The three most popular link functions for binary responses are the probit, the logit and the log. The log link, on which we focus, has a relatively short history and has attracted recent interest especially in the epidemiological literature. Approximate inference for the RR dates back to at least as far as Cornfield (1951, 1956). Estimation of adjusted RR via LBR has only recently been investigated. Wacholder (1986) is perhaps one of the earliest references. The complementary log-log, $\text{cloglog}(p) = \log(-\log(1-p))$, is a less popular link function which can be used in some instances to estimate the RR in the form of incidence ratios, as discussed by Martuzzi and Elliott (1998).

Setting aside some earlier use of the logistic curve in modeling population growth, the probit model, where the link function is given by the inverse of the cumulative standard normal curve, $\Phi^{-1}(p)$, was developed before the logit

approach (Stigler, 1986; Cramer, 2011). The probit model can be traced back to the psychometric work of the German scholar G.T. Fechner around the 1860's. Probit analyses developed quickly with the work of Gaddum (1933) and Bliss (1934), who actually introduced the term "probit" (short for *probability unit*). Biomedical and pharmacological research provided fertile ground for the development of probit regression since it allowed for direct estimation of tolerance levels (*e.g.* lethal doses) and of similar measures of pharmacological interest. From a numerical perspective, probit analyses where, at that time, cumbersome and ML estimation difficult. The probit has also been present in the econometric literature at least since the 1950's (Cramer, 2011). We do note that a very early study on the estimation of a GLM with random effects belongs to the econometric literature and it is motivated by a random-intercept probit regression (Guilkey and Murphy, 1985).

Despite earlier uses (at least as early as Wilson, 1925), the logit gained traction around the 1950's (Berkson, 1944, 1951; Cornfield, 1951, 1956; Cox, 1958). After a natural period of dissemination, the analytical properties and mathematical tractability of the logit model became widely recognized and by the early 1970's it was already as popular as the probit (Cramer, 2011). In the early 1970's the economist Daniel McFadden combined logit regression with discrete choice models, initiating an active area of research and leading to McFadden's Nobel Prize in 2006. Though a natural model in certain econometric and epidemiological contexts (*e.g.* discrete choice models, cohort studies with low-incidence outcomes, time-matched case-control studies), logistic regression may not be appropriate in several instances where the probit and the log will yield direct estimation of quantities of interest. However, logistic regression is currently the most popular GLM due, to a great extent, to its numerical and analytical tractability. In this paper we explore available constrained optimization routines in R to make LBR a viable choice over logistic regression.

We start by reviewing the basics of ML and QL estimation in the context of LBR and then illustrate how the R system can be used to fit log binomial models with constrained optimization. We will also present a simple Bayesian analysis since the parameter constraints imposed by LBR can be easily incorporated in a Bayesian model.

2 Log binomial regression

2.1 Maximum likelihood via IWLS

In a LBR model we consider binary responses, y_i, $i = 1, \ldots, n$, with means, $\Pr\{y_i = 1|\mathbf{x}_i\} = p_i$, explained by a set of k covariates, whose values for the i-th individual are represented by the row vector $\mathbf{x}_i' = (x_{i1}, \ldots, x_{ik})$. The mean is linked to a linear predictor by

$$\log(p_i) = \mathbf{x}_i'\boldsymbol{\beta}, \tag{2.1}$$

so the vector of coefficients $\boldsymbol{\beta}$ is directly related to RR as explained next. The ratio

$$\frac{\Pr\{y_i = 1|\mathbf{x}_i = \mathbf{a}\}}{\Pr\{y_i = 1|\mathbf{x}_i = \mathbf{b}\}} = \exp\left((\mathbf{a} - \mathbf{b})'\boldsymbol{\beta}\right),$$

measures the relative size of the effect of observing $\mathbf{x} = \mathbf{a}$ over $\mathbf{x} = \mathbf{b}$. Therefore, the RR between two levels of x_j that differ by one unit is $\exp(\beta_j)$, keeping all other regressors constant.

The resulting log-likelihood with independent observations is given by

$$\ell(\boldsymbol{\beta}) = \sum_i y_i \log\left(p_i(\boldsymbol{\beta})\right) + \sum_i (1 - y_i) \log\left(1 - p_i(\boldsymbol{\beta})\right), \tag{2.2}$$

with parameter space B so that $0 \leq p_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}) \leq 1$, *i.e.*,

$$B = \{\boldsymbol{\beta} \in \mathbb{R}^k : \mathbf{x}_i'\boldsymbol{\beta} \leq 0, \text{for all } i\}. \tag{2.3}$$

We shall denote the $(k \times 1)$-vector of first derivatives of ℓ and the $(k \times k)$-matrix of second derivatives of ℓ by $\nabla\ell(\boldsymbol{\beta})$ and $\nabla^2\ell(\boldsymbol{\beta})$, respectively. These are given by

$$\nabla\ell(\boldsymbol{\beta}) = \sum_i \mathbf{x}_i \frac{y_i - p_i(\boldsymbol{\beta})}{1 - p_i(\boldsymbol{\beta})}, \tag{2.4}$$

and

$$\nabla^2\ell(\boldsymbol{\beta}) = \sum_i \mathbf{x}_i\mathbf{x}_i' \frac{p_i(\boldsymbol{\beta})(y_i - 1)}{(1 - p_i(\boldsymbol{\beta}))^2}. \tag{2.5}$$

Blizzard and Hosmer (2006) discuss expressions (2.4) and (2.5) in more detail, analyzing their use in obtaining standard errors and providing diagnostic tools for LBR. We observe, for later reference in Section 2, that when a valid estimate ($\hat{\boldsymbol{\beta}} \in B$) is obtained, observed Fisher information can be used to estimate standard errors (Efron and Hinkley, 1978),

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = -\left[\nabla^2\ell(\hat{\boldsymbol{\beta}})\right]^{-1}. \tag{2.6}$$

In the case of LBR, the right hand side of (2.6) is minus the inverse of (2.5) evaluated at the MLE. Furthermore, asymptotically valid standard errors for the relative risks can be readily calculated. The asymptotically valid standard error for the estimated RR associated with the j-th covariate, x_j, is simply

$$\mathrm{stderr}(\mathrm{RR}_j) = \sqrt{\mathrm{RR}_j \times \widehat{\mathrm{Var}}(\hat{\beta}_j)}, \tag{2.7}$$

where $\widehat{\mathrm{Var}}(\hat{\beta}_j)$ is the j-th entry in the diagonal of (2.6).

ML estimation of GLM is routinely done by *iteratively weighted least squares* (IWLS). IWLS is an adaptation of the Newton-Raphson algorithm for solving $\nabla\ell(\boldsymbol{\beta}) = 0$ in which the data-dependent matrix $\nabla^2\ell(\boldsymbol{\beta})$ needed at each iteration is

replaced by its expectation, a procedure often referred to as *Fisher scoring*. IWLS is an unconstrained maximization procedure and therefore inadequate to fit LBR. Statistical packages use different flavors of IWLS and results are typically better when safeguards are implemented so that iterates remain in the feasible region given by (2.3).

Deddens *et al.* (2003) considered a small dataset ($n = 10$) where $x_i = i$ and $y_i = 0$ for $i = 1, 2, 3, 4, 6$ and $y_i = 1$ otherwise, to illustrate convergence failure of SAS 9.2 GENMOD (using the default starting values). Extending their small experiment to R and Stata we observed that several trials using sensible starting values failed to converge, even though R did get very close to the MLE with many of the starting values used, despite failing to self-start. R `glm` and SAS GENMOD use different variations of IWLS. Stata `glm` allows for IWLS and also four other maximization routines (NR (default), BHHH, DFP, BFGS). We attribute the better performance of R to the fact that it uses safeguards (step halving) to keep the iterations inside the parameter space. Even though safeguards do not make up for true constrained optimization – and convergence to solutions on the boundary will occur, if at all, very slowly – safeguards allow the algorithm to get closer to the solution than do bare versions of IWLS. These results are based on one toy dataset but are illustrative of the limitations of unconstrained optimization in the context of LBR. We also note that QL estimates based on the Poisson and normal estimating equations (see below) were outside the parameter space, as depicted in Figure 1, along with a valid (inside B) Bayesian estimate (see Section 3.3). With only one covariate it is easy to visualize the parameter space. In order to satisfy (2.3) we must have $\beta_0 + \beta_1 x_i \leq 0$ for all i, that is, $\beta_1 \leq -\beta_0/x_i$, for all i. If we consider the two possible signs for β_0, given that $x_i > 0$, for all i, and that $\max(x_i) = 10$ and $\min(x_i) = 1$ we see that the restricted parameter space is given by

$$\{\beta_1 \leq -\beta_0/10, \beta_0 \geq 0\} \cap \{\beta_1 \leq -\beta_0, \beta_0 < 0\}.$$

2.2 Quasi-likelihood

Robust QL is perhaps the most common alternative estimator of RR when GLM software fails to converge with a binomial response and log link. Recent discussions of this alternative to fitting LBR can be found in Carter *et al.* (2005), Blizzard and Hosmer (2006) and Lumley *et al.* (2006). QL enjoys desirable asymptotic properties and it is available through regular GLM software. The basic idea is twofold: deliberate misspecification of the response as either Poisson or normal (rather than binomial) and use of robust standard errors (White, 1982).

To make distinctions clearer we may write the gradient (2.4) as $\nabla\ell(\boldsymbol{\beta}) = \sum_i s_i(\boldsymbol{\beta})$, where

$$s_i(\boldsymbol{\beta}) = \mathbf{x}_i \left[y_i - p_i(\boldsymbol{\beta})\right], \tag{2.8}$$

under a Poisson model, and

$$s_i(\boldsymbol{\beta}) = \mathbf{x}_i p_i(\boldsymbol{\beta}) \left[y_i - p_i(\boldsymbol{\beta})\right], \tag{2.9}$$

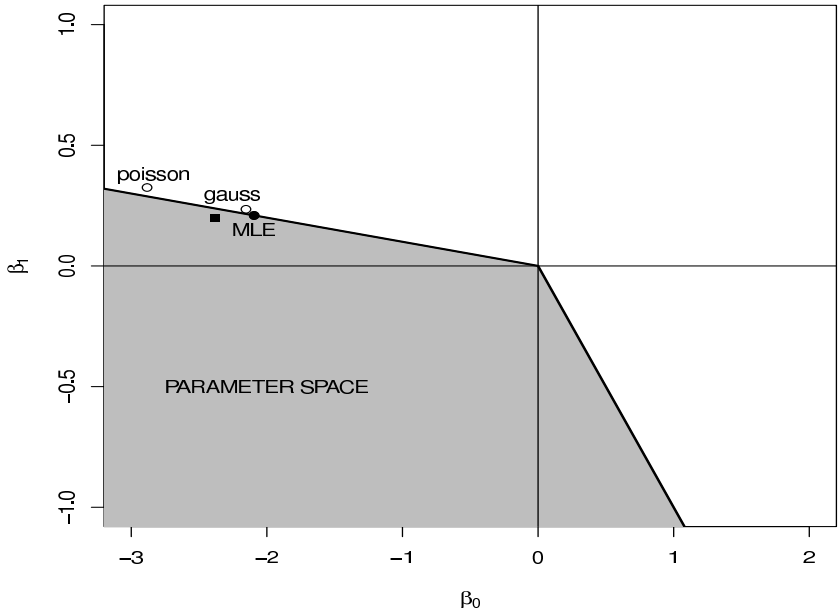

Figure 1 - Parameter space for log-binomial model with toy data from Deddens *et al.* (2003). Estimates (circles) correspond to the MLE, Poisson QL, normal QL and posterior mean from Bayesian analysis (square) with diffuse priors.

under a normal model, whereas under the correctly specified LBR (see (2.4)), we have

$$s_i(\boldsymbol{\beta}) = \mathbf{x}_i \frac{y_i - p_i(\boldsymbol{\beta})}{1 - p_i(\boldsymbol{\beta})}. \tag{2.10}$$

In any case, the estimating equations to be solved are $\nabla \ell(\boldsymbol{\beta}) = 0$. In comparing (2.10) to either (2.8) or (2.9) we see that the case $p = 1$ does not prevent the computation of the gradient when Poisson or normal models are used (in those two cases $p = \mathrm{E}(y)$ need not be bounded by 0 and 1). Since the Poisson and normal specifications do not require the mean response to be bounded, convergence issues arise a lot less frequently but at a price: there is no guarantee that QL estimates will yield valid probabilities. Robust standard errors are obtained from the diagonal of

$$\left[\nabla^2 \ell(\hat{\boldsymbol{\beta}})\right]^{-1} \left[\sum_i s_i(\hat{\boldsymbol{\beta}}) s_i(\hat{\boldsymbol{\beta}})'\right]^{-1} \left[\nabla^2 \ell(\hat{\boldsymbol{\beta}})\right]^{-1}, \tag{2.11}$$

which is readily available through SAS GENMOD or Stata `glm`. Robust standard errors are not directly available through R `glm` but rather through contributed packages as illustrated in Section 3.1.

2.3 Maximum likelihood via constrained optimization

Despite the popularity of QL in fitting LBR when convergence difficulties are met, ML may still be viable under constrained optimization routines. To our knowledge, such routines are not available in standard GLM software. We have used the *adaptive barrier algorithm* (Nocedal and Wright, 2006) to fit LBR. This algorithm allows for linear constraints such as the ones in (2.3) and it is already part of the optimization tools in R.

Suppose an interior solution, $\hat{\boldsymbol{\beta}}$, is found. Then, asymptotic variances of estimates can be easily computed since, under the usual conditions assumed for the existence of a ML estimate in a GLM, the ML estimate is asymptotically unbiased and normal with variance given by the inverse of Fisher information (see eqs. (2.6) and (2.7)).

This procedure has worked in problems where the `glm` function failed to converge. We thus recommend that proper constrained optimization routines be explored to fit LBR and if an interior solution is found, standard ML theory can be used for inference without need for QL or other circumventions.

By using constrained optimization, we were able to obtain ML estimates for the LBR coefficients. The Hessian (2.5) was needed to calculate the variance (2.6) and subsequently the standard errors (2.7). An illustration with epidemiological data is given in Section 3 and simulation results are given next.

Simulations

Blizzard and Hosmer (2006) conducted a simulation study to compare the performance of LBR, an expanded-data method based on logistic regression and a QL approach (log Poisson regression). Their simulations "showed that when the log binomial model could be fit, it provided a better estimator than either the expanded data logit model or the Poisson regression model. The Poisson regression estimator had slightly better properties than the expanded data logit model estimator, but the differences between these two estimators were not great."

We note that in their simulations, LBR was fitted by routine GLM software and therefore, fitting problems were identified as expected. For instance, in one of their simulation settings, a total of 2450 samples were required to obtain 1000 successful samples (each of size $n = 500$) without convergence problems. Three other settings presented the same need to generate a larger number of samples in order to obtain the desired 1000 succesful ones. There were eight settings with four problematic ones. The simulated model was

$$\Pr(y = 1|x) = \exp(\beta_0 + \beta_1 x),$$

with $x \sim \text{Uniform}(0, a)$, where a was chosen so that $\Pr(y = 1|x = a)$ and, consequently, $\Pr(y = 1)$, was as high as desired. In the four problematic settings, $\Pr(y = 1|x = a)$ was high, with maximum values over the simulated samples above 0.90, whereas in the other four settings, where there was no need to generate more than 1000 samples, the same probability did not exceed 0.53 accross simulations.

$\Pr(y = 1)$ was around 0.20 in the problematic settings and around 0.12 in the straightforward ones. For details see Blizzard and Hosmer (2006).

We considered the four problematic settings to illustrate the constrained optimization approach. We generated 1000 samples ($n = 500$) under each setting and applied both constrained optimization (as implemented by the R functionality provided in the Appendix, function `lbreg`) and IWLS (with safeguards as implemented in the R function `glm`). Constrained optimization succeeded in all samples under all four settings whereas IWLS success rates where between 45% and 76%. For a SAS implementation of restricted optimization in the context of LBR see Yu and Wang (2008). Table 1 brings the specifications (β_0, β_1 and a) of the problematic settings, the success rate (% of samples where convergence was reached) of R function `glm` for the 1000 samples we generated and the maximum probability fitted under constrained optimization (which succeed in all 4000 samples). The starting value for both `glm` and `lbreg` was ($\beta_0 = -5, \beta_1 = 0.01$) which is inside the parameter space for all four settings, $B = \{\beta_0 \leq 0, \beta_0/6 \leq \beta_1 \leq -\beta_0/a\}$, $a = 0.5, 1, 2, 6$.

Table 1 - Simulation results using four settings designed by Blizzard and Hosmer (2006) where LBR fitting problems were present under common GLM software but absent under constrained optimization (R function `lbreg`; see Appendix) with convergence rate of 100% in all four settings

β_0	β_1	a	IWLS success rate	Average max($\hat{p}$) estimated under constrained optimization
-0.35667	0.70808	0.5	54%	0.98
-0.69315	0.65200	1	76%	0.96
-1.20397	0.56687	2	45%	0.92
-2.30259	0.38376	6	55%	0.98

3 Analysis of STEP data

The example mentioned in Section 2.1 referring to Deddens *et al.* (2003) could be considered special in that the MLE lied on the boundary of the parameter space. Even when the MLE lies in the interior of (2.3), which is to be expected in large samples under the usual regularity conditions (one such condition being that the true parameter value is in the interior of the parameter space), GLM software may still fail to converge.

As an example, we consider a subset of the clustered dataset analyzed in Tallo *et al.* (2008). We have taken the 1764 observations of the largest cluster in the dataset. The data come from an epidemiological study aimed at identifying socio-demographic and community-based factors associated with participation in a mass drug delivery program for schistosomiasis in the Philippines. Table 2 summarizes the dataset.

The model of interest has a binary outcome, y_i, which is coded one if the i-th individual comes to the treatment site and zero otherwise. The probability of participation is modeled as a function of three categorical covariates: two dichotomous regressors, sex and participation in a previous study (called STEP), age (four levels) and exam result for *Schistosoma japonicum* infection (three levels). We are thus interested in the following LBR,

$$\Pr\{y_i = 1|\mathbf{x}_i\} = p_i(\boldsymbol{\beta}), \quad i = 1, \ldots, 1764, \tag{3.1}$$

where,

$$p_i = \exp\left(\beta_0 + \beta^{\text{sex}} x_{i1} + \beta^{\text{step}} x_{i2} + \mathbf{x}'_{i3}\boldsymbol{\beta}^{\text{age}} + \mathbf{x}'_{i4}\boldsymbol{\beta}^{\text{sj}}\right). \tag{3.2}$$

Routine GLM software (R `glm`, SAS GENMOD, Stata `glm`) failed to converge (several starting values were used and the maximal number of iterations was set at values five times greater than the defaults) with only R giving a final iterate close to the solution but stuck at a boundary value (see Table 3). In R `glm`, a safeguard (step halving) is implemented to keep iterates inside the valid domain and the algorithm can easily reach the maximum number of iterations if the solution is near or at the boundary. In this case, standard errors reported by R `glm` are senseless and therefore they are not included in Table 3. In R `glm`, the maximal number of IWLS iterations is 25 by default. Setting this value to 100 caused no relevant change in the final iterate. We emphasize that the first column of Table 3 is merely illustrative of how a simple safeguard allowed IWLS to get close to the solution as obtained by constrained optimization. This final iterate of a non-convergent run can also be useful as an initial value for trials with optimization routines other than IWLS.

3.1 QL approach

Table 3 displays QL estimates based on (2.8) and robust standard errors (2.11) which we calculated in R using the tools provided by Zeileis (2006). We note that 10 out of the 1764 estimated probabilities were outside the valid domain and the maximum estimated probability was 1.58. These 10 invalid fitted probabilities were all associated with the group `sj2` (positive for parasitological exam) which has only 28 observations (see Table 2).

3.2 Constrained optimization

Table 3 shows RR estimated by constrained ML along with 95% confidence intervals based on (2.7) and on asymptotic normality. The results are similar to QL but a noticeable difference occurs for the estimate of the RR associated with `sj2`. The confidence interval obtained by ML includes the null $\text{RR}_{sj2} = 1$ whereas the QL results point to RR_{sj2} significantly above unit. The ML estimates seem closer to the null which is also the case when comparing a RR with an OR estimate in a 2×2 table for a frequent outcome.

Table 2 - Unadjusted relative risks associated with sociodemographic characteristics, among members of the largest village of Samar province, the Philippines, 2004 – STEP data

Variables	Count	RR
Age in years		
age1: (5, 10] (reference)	372	–
age2: (10, 16]	354	0.92
age3: (16, 40]	589	0.47
age4: > 40	449	0.58
Sex		
sex1: Female (reference)	921	–
sex2: Male	843	0.96
STEP participation		
step1: No (reference)	158	–
step2: Yes	1606	1.64
Parasitological exam		
sj1: No exam (reference)	1629	–
sj2: Positive	28	2.34
sj3: Negative	107	1.79

3.3 Bayesian estimation

Bayesian estimation of unadjusted RR dates back to at least Aitchison and Bacon-Shone (1981) (see also Gupta *et al.*, 1997) but only recently has Bayesian modeling been proposed for the estimation of adjusted RR via LBR (Chu and Cole, 2010).

Besides constrained ML, another way to obtain valid estimates of the coefficients in a log binomial model is through Bayesian inference. Bayesian regression analysis when the parameter space is unconstrained (*e.g.* normal, Poisson or binomial data with canonical links) is straightforward (Gelman and Hill, 2006) with typical prior distributions for $\boldsymbol{\beta}$ being unrestricted.

Consider a LBR with an unrestricted prior, say $\pi(\boldsymbol{\beta}) \sim$ multivariate normal, with hyperparameters fixed. The associated proper prior would be $\pi^*(\boldsymbol{\beta}) = \pi(\boldsymbol{\beta})\mathbb{I}_B(\boldsymbol{\beta})/c$ where $\mathbb{I}_B$ is the indicator function of the set B given by (2.3) and $c = \int_B \pi(\boldsymbol{\beta})d\boldsymbol{\beta}$. Denote the likelihood by $L(\text{data}|\boldsymbol{\beta})$. The Gibbs sampler or the Metropolis-Hastings algorithm can be used to draw samples from the unnormalized posterior $L(\text{data}|\boldsymbol{\beta})\pi(\boldsymbol{\beta})\mathbb{I}_B(\boldsymbol{\beta})$ without knowledge of c or other normalizing constants. The resulting estimates (posterior means) will be inside B and the optimization convergence problems associated with classical estimation are no longer present. One does, however, face the technicalities of MCMC sampling. Here we have used a very simple sampler, the random walk Metropolis algorithm as implemented in R by function `metrop` (Geyer, 2009a), which requires little user

Table 3 - Estimated relative risks and corresponding 95% confidence intervals under different methods using the STEP data: IWLS (LBR using R function `glm` with binomial family and log link – final iterate at boundary value without convergence); Poi-QL (quasilikelihood in the form of log Poisson regression – QL estimates are outside valid domain); Constr.Opt. (LBR using adaptive barrier algorithm as implemented in the R function `constrOptim`)

	IWLS	Poi-QL	low95	up95	Constr.Opt.	low95	up95
sex	0.92	0.94	0.84	1.04	0.94	0.85	1.04
step	0.92	0.86	0.59	1.12	0.94	0.58	1.30
age2	0.92	0.91	0.79	1.02	0.98	0.86	1.10
age3	0.49	0.47	0.37	0.57	0.51	0.40	0.63
age4	0.61	0.59	0.48	0.70	0.64	0.51	0.76
sj2	1.99	2.85	2.35	3.35	1.89	0.64	3.15
sj3	1.70	1.93	1.54	2.33	1.65	1.16	2.13

input compared to other ergodic samplers (Robert and Casella, 1999).

Since a large dataset is available, we are confident that we can estimate the quantities of interest with independent diffuse normal priors for $\boldsymbol{\beta}$. We started our sampler at the MLE. After short pilot trials we set the scale parameter in the normal proposal at 0.03 in order to achieve an acceptance rate of about 25%; the 20–30% mark is a guideline in the MCMC literature (Robert and Casella, 1999). We have used the method of batch means among several options such as subsampling, block bootstrap or regeneration, to calculate Monte Carlo standard errors. Correlograms and trace plots were used to determine the length of batches and our final run (after pilot trials) used 10^4 iterations with batches of length 100. Visual inspection of the output did not point to any obvious convergence problems. Precision of the estimates was measured by Monte Carlo standard errors which were low compared to the magnitudes involved (see Table 4). Visualization of the posterior distribution is illustrated for RR_{step} in Figure 2.

Comparison with classical estimation

The data showed a very strong confounding effect between participation in STEP and the results of the parasitological exam for *S. japonicum* (Tallo *et al.*, 2008). This is why all multivariable estimates of the effect of STEP are very different to those from the unadjusted estimation. The QL estimates were generally away from the null, which would result in overestimating the impact of risk factors, especially for participation in STEP and the result of the parasitological exam. Interestingly, where the LBR and Bayesian models suggest very similar strengths of association between the parasitological exam results and being present during mass treatment, the QL model estimates are closer to the biased unadjusted estimates. The LBR and Bayesian methods lead to very similar estimates of the strength

Table 4 - Posterior means of adjusted RR for the STEP data and corresponding 95% Bayesian credible intervals; MCSE = Monte Carlo standard error

	RR	bci.low	bci.up	MCSEx100
sex	0.94	0.86	1.03	0.09
step	0.96	0.79	1.13	0.24
age2	0.96	0.85	1.07	0.13
age3	0.51	0.43	0.59	0.10
age4	0.63	0.53	0.72	0.12
sj2	1.73	1.46	2.01	0.39
sj3	1.59	1.27	1.92	0.47

of association between all variables and being present for the mass treatment. The difference in those two methods lies in the precision of the estimates. Where the LBR method would have led to the conclusion that being positive to the parasitological exam for *S. japonicum* had a marginal effect on being present for the mass treatment (the 95%CI includes 1), the Bayesian would have concluded that this effect was important (the 95%CI does not include 1). This is most likely due to the small number of individuals ($n = 28$) in this village who had a positive result to the parasitological test.

Concluding remarks

Log-binomial regression yield direct estimates of RR in models with both categorical and continuous regressors. However, care must be taken with the fact that the parameter space is constrained and routine GLM software may not work. GLM software with reliable constrained optimization routines is necessary. To the authors' knowledge, these routines are not directly implemented in the GLM functionality available in either R 2.12.1, SAS 9.2 or Stata 11. To this end we developed R functionality to estimate LBR which allows for the linear constraints imposed by the log link. We illustrated this new functionality in situations where the standard GLM function failed. Our functions make ML viable so the data analyst need not resort to the alternative methods found in the literature. However, if the user's software does not allow flexible programming and does not have built-in constrained optimization tools, the QL alternative has been recommended but we warn that there is no guarantee that the resulting estimates will be inside the parameter space, as was the case in our examples, even with a large sample size. QL is an asymptotically valid method that can be implemented via regular GLM software and it can be easily extended to mixed models via generalized estimating equations.

Bayesian inference is an alternative to obtain estimates inside the parameter space even when unconstrained priors are employed and this approach has also been illustrated. Results were similar to that of constrained ML but with shorter confidence (credible) intervals.

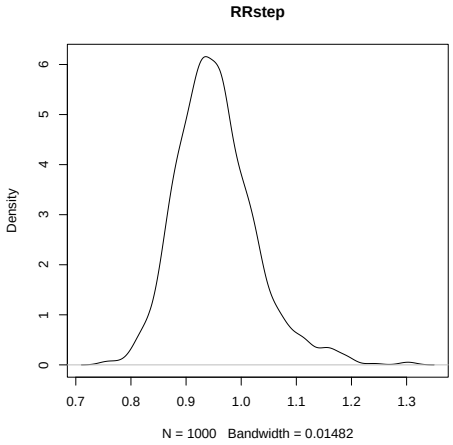

Figure 2 - Estimated posterior distribution for RR_{step}.

Acknowledgments

We are grateful to the helpful comments of Joanlise Andrade (Dep. de Estatística, UFRN) and to an anonymous referee for valuable feedback. We thank two authors of Tallo *et al.* (2008) for providing the STEP dataset (data collection funded by the NIH/NSF Ecology of Infectious Diseases programme, NIH Grant R01 TW01582). The first author was partly supported by FINATEC/UnB (Fomento-4/09) during his stay at the Department of Economics, University of Brasília, while this work was being completed.

ANDRADE, B. B.; CARABIN, H. Estimação do risco relativo via regressão log-binomial. *Rev. Bras. Biom.*, São Paulo, v.29, n.1, p.25-46, 2011.

- RESUMO: Sendo bem conhecidas as dificuldades em se ajustar um modelo de regressão log-binomial com software para modelos lineares generalizados, nós implementamos uma solução direta via otimização restrita que evita os métodos indiretos propostos na literatura. O uso de um modelo log binomial é motivado pelo interesse em se estimar diretamente riscos relativos ajustados por covariáveis. Também se discute a implementação de um modelo bayesiano para um conjunto de dados de interesse epidemiológico. Desenvolvemos programas em R para ilustrar essas abordagens de estimação.

- PALAVRAS-CHAVE: Regressão log-binomial; máxima verossimilhança restrita; quasi-verossimilhança; risco relativo.

References

AITCHISON, J.; BACON-SHONE, J. Bayesian relative risk analysis. *Am. Stat.*, Alexandria, v.35, p.254-257, 1981.

AXELSON, O.; FREDRIKSSON, M.; EKBERG, K. Use of the prevalence ratio v the prevalence odds ratio as a measure of risk in cross sectional studies. *Occup. Environ. Med.*, London, v.51, n.8, p.574, 1994.

BARROS, A. J. D.; HIRAKATA, V. N. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. *BMC Med. Res. Methodol.*, London, v.3, n.21, p.1-13, 2003.

BERKSON, J. Application of the logistic function to bio-assay. *J. Am. Stat. Assoc.*, Alexandria, v.39, p.357-365, 1944.

BERKSON, J. Why I prefer logits to probits. *Biometrics*, Washington, v.7, n.4, p.327-339, 1951.

BLISS, C. I. The method of probits. *Science*, Washington, v.79, p.38-39, 409-410, 1934.

BLIZZARD, L.; HOSMER, D. W. Parameter estimation and goodness-of-fit in log binomial regression. *Biom. J.*, Weinheim, v.48, n.1, p.5-22, 2006.

CARTER, R. E.; LIPSITZ, S. R.; TILLEY, B. C. Quasi-likelihood estimation for relative risk regression models. *Biostatistics*, Oxford, v.6, n.1, p.39-44, 2005.

CHU, H.; COLE, S. R. Estimation of risk ratios in cohort studies with common outcomes: a bayesian approach. *Epidemiology*, Durham, v.21, n.6, p.855-862, 2010.

CORNFIELD, J. A method of estimating comparative rates from clinical data. *J. Nat. Cancer Inst.*, Oxford, v.11, p.1269-1275, 1951.

CORNFIELD, J. A statistical problem arising from retrospective studies. In: NEYMAN, J. (Ed.) *Proceedings of the Trird Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press, 1956. p.135-148.

COUTINHO, L. M. S.; SCAZUFCA, M.; MENEZES, P. R. Métodos para estimar razão de prevalência em estudos de corte transversal. *Rev. Saúde Pública*, São Paulo, v.42, n.6, p.992-998, 2008.

COX, D. R. The regression analysis of binary sequences. *J. R. Stat. Soc.*, Series B, London, v.20, p.215-242, 1958.

CRAMER, J. S. *Logit models from economics and other fields*. Cambridge: Cambridge University Press, 2011. 184p.

DEDDENS, J. A; PETERSEN, M. R.; LEI, X. Estimation of prevalence ratios when PROC GENMOD does not converge. *SUGI 28 Proceedings*, Seattle, Paper 270, 2003.

EFRON, B.; HINKLEY, D. V. Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika*, Oxford, v.65, n.6, p.457-487, 1978.

FISHER, R. A. The analysis of variance with various binomial transformations. *Biometrics*, Washington, v.10, n.1, p.130-151, 1954.

GADDUM, J. H. *Reports on biological standard III. Methods of biological assay depending on a quantal response.* London: Medical Research Council, 1933. (Special Report Series n.183).

GELMAN, A.; HILL, J. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press, 2006. 625p.

GEYER, C. J. Constrained maximum likelihood exemplified by isotonic convex logistic regression. *J. Am. Stat. Assoc.*, Alexandria, v.86, n.415, p.717-724, 1991.

GEYER, C. J. *MCMC: Markov Chain Monte Carlo*. R package version 0.7-2, available from CRAN. 2009a. Disponível em: $< http://cran.r-project.org >$

GEYER, C. J. Likelihood inference in exponential families and directions of recession. *Electron. J. Stat.*, Beachwood, v.3, n.2, p.259-289, 2009b.

GUILKEY, D. K.; MURPHY, J. L. Estimation and testing in the random effects probit model. *J. Econometrics*, Amsterdam, v.61, p.166-181, 1985.

GUPTA, R. C.; ALBANESE, R. A.; PENN, J. W.; WHITE, T. J. Bayesian estimation of relative risk in biomedical research. *Environmetrics*, New York, v.8, p.133-143, 1997.

HUGHES, K. Odds ratios in cross-sectional studies. *Int. J. Epidemiol.*, Oxford, v.24, n.2, p.463-464, 1995.

JUDGE, G. G.; TAKAYAMA, T. Inequality restrictions in regression analysis. *J. Am. Stat. Assoc.*, Washington, v.61, p.166-181, 1966.

LEE, J. Odds ratio or relative risk for cross-sectional data? *Int. J. Epidemiol.*, Oxford, v.23, n.1, p.201-203, 1994.

LUMLEY, T; KRONMAL, R.; MA, S. Relative risk regression in medical research: models, contrasts, estimators, and algorithms. *UW Biostatistics Working Paper Series*, Seattle, Paper 293, 2006.

MARTUZZI, M.; ELLIOTT, P. Estimating the incidence rate ratio in cross-sectional studies using a simple alternative to logistic regression. *Ann. Epidemiol.*, Raleigh, v.8, p.52-55, 1998.

McNUTT, L. A.; WU, C.; XUE, X.; HAFNER, J. P. Estimating relative risk in cohort studies and clinical trials of common events. *Am. J. Epidemiol.*, Oxford, v.157, p.940-943, 2003.

NOCEDAL, J.; WRIGHT, S. *Numerical optimization*. 2.ed. New York: Springer, 2006. 664p. (Springer Series in Operations Research and Financial Engineering).

OSBORN. J.; CATTARUZZA, M. S. Odds ratio and relative risk for cross-sectional data. *Int. J. Epidemiol.*, Oxford, v.24, n.2, p.464-465, 1995.

REICHENHEIM, M.; COUTINHO, E. Measures and models for causal inference in cross-sectional studies: arguments for the appropriateness of the prevalence odds

ratio and related logistic regression. *BMC Med. Res. Methodol.*, London, v.10, n.66, 2010.

ROBERT, C. P.; CASELLA, G. *Monte Carlo statistical methods*. New York: Springer, 1999. 507p.

SCHWARTZ, L. M.; WOLOSHIN, S.; WELCH, H. G. Misunderstandings about the effects of race and sex on physicians' referrals for cardiac catheterization. *New Engl. J. Med.*, Waltham, v.341, n.4, p.279-283, 1999.

STIGLER, S. M. *The history of statistics*. Cambridge: Harvard University Press, 1986. 432p.

STROMBERG, U. Prevalence odds ratios v.s. prevalence ratio. *Occup. Environ. Med.*, London, v.51, p.143-144, 1994.

TALLO, V. L; CARABIN, H.; ALDAY, P. P.; BALOLONG, E. JR.; OLVEDA, R. M.; MCGARVEY, S. T. Is mass treatment the appropriate schistosomiasis elimination strategy? *Bull. World Health Organ.*, Geneva, v.86, n.10, p.765-771, 2008.

WACHOLDER S. Binomial regression in GLIM: estimating risk ratios and risk differences. *Am. J. Epidemiol.*, Oxford, v.123, n.1, p.174-184, 1986.

WHITE, H. Maximum likelihood estimation of misspecified model. *Econometrica*, New York, v.50, p.1-25, 1982.

WILSON, E. B. The logistic or autocatalytic grid. *Proceeding of the National Academy of Sciences*, Washington, v.11, p.451-456, 1925.

YU, B.; WANG, Z. Estimating relative risks for common outcome using PROC NLP. *Comp. Methods Progr. Biomed.*, Amsterdam, v.90, n.2, p.179-186, 2008.

ZEILEIS, A. Object-oriented computation of sandwich estimators. *J. Stat. Softw.*, Los Angeles, v.16, p.1-16, 2006.

ZOU, G. A modified Poisson regression approach to prospective studies with binary data. *Am. J. Epidemiol.*, Oxford, v.154, n.7, p.702-706, 2004.

Received in 05.08.2010.

Approved after revised in 08.04.2011.

Appendix

A R Source Code

We have written the function `lbreg` as a replacement to

```
glm( formula, family = binomial("log"), data )
```

It is based on the adaptive barrier algorithm provided through `constrOptim` rather than IWLS as is the case of `glm`. The supporting functions calculate the negative of (2.2) and the Hessian (2.5). These are the main functions. We have also added `print`, `summary`, and `relrisk` output methods to `lbreg` (only `relrisk` shown). Usage of this functionality is illustrated in the next section. Function `lupost` is the log unnormalized posterior described in Section 3.3.

```
####  Log Binomial GLM Using Constrained Optimization  ####

####  Negative Loglikelihood

negll <- function(beta, x, y)
{
  b <- matrix(beta, nrow = length(beta), ncol = 1)
  eta <- x %*% as.matrix(beta)
  llik <- dbinom(y, size = 1, prob = exp( eta ), log = TRUE )
  .value <- -sum(llik)
  return(.value)
}

####   Hessian

hess <- function(beta, x, y)
{
  b <- matrix(beta, nrow = length(beta), ncol = 1)
  p <- exp( x %*% as.matrix(beta) )
  z <- p*(y-1) / (1-p)^2
  hh <- 0
  for (i in 1:nrow(x)) {
    hh <- hh + x[i,] %*% t(x[i,]) * z[i]
  }
  return( hh )
}

####   LBR

lbreg0 <- function(x, y, start)
{
  out <- constrOptim(theta = start, f = negll, grad=NULL,
            ui = -x, ci = rep(0,nrow(x)),
            x = x, y = y)
  coef <- out$par
  vcov <- -solve( hess(beta=out$par, x=x, y=y) )
  colnames(vcov) <- rownames(vcov) <- names(coef) <- colnames(x)
  list(coefficients = coef,
               vcov = vcov,
               sigma = sqrt(sum(y - exp(x%*%coef))),
               df = nrow(x) - ncol(x),
               loglik= -out$value,
```

```
              convergence = out$convergence,
              barrier.value = out$barrier.value,
              outer.iterations = out$outer.iterations
    )
}

lbreg <- function(x, y, start, ...) UseMethod("lbreg")

lbreg.default <- function(x, y, start, ...)
{
   x <- as.matrix(x)
   y <- as.matrix(y)
   out <- lbreg0(x, y, start)
   out$fitted.values <- as.vector(exp(x %*% out$coefficients))
   out$residuals <- y - out$fitted.values
   out$call <- match.call()
   class(out) <- "lbreg"
   return( out )
}


relrisk <- function(object, alpha, ...) UseMethod("relrisk")

relrisk.default <- function(object, alpha=0.05, ...)
{
  rr <- exp(coef(object))
  se <- sqrt( diag(object$vcov) )
  zcrit <- qnorm(1 - alpha/2)
  TAB <- cbind( RR = rr,
         l95=rr-zcrit*sqrt(rr)*se,
         u95=rr+zcrit*sqrt(rr)*se  )
  Output <- list(call=object$call,
                 risk=TAB)
  class(Output) <- "relrisk"
  return(Output)
}

relrisk.glm <- function(object, alpha=0.05, robust=TRUE...)
{
  rr <- exp(coef(object))
  ## Robust SEs
  require(lmtest)
  require(sandwich)
  if(robust){
    se <- coeftest(object, vcov = sandwich)[,2] }else{
    se <- sqrt(diag(vcov(object)))
  }
  zcrit <- qnorm(1 - alpha/2)
  TAB <- cbind( RR = rr,
         l95=rr-zcrit*sqrt(rr)*se,
         u95=rr+zcrit*sqrt(rr)*se  )
  Output <- list(call=object$call,
                 risk=TAB)
  class(Output) <- "relrisk"
  return(Output)
}
```

```
###  Log unnormalized posterior

lupost <- function(b, x, y, m0, v0)
{
    eta <- x %*% as.matrix(b)
    lik <- dbinom(y, size = 1, prob = exp( eta ) )
    loglik <- sum( ifelse(is.nan(lik), -Inf, log(lik)) )
    lprior <- sum( dnorm( b, m0, sqrt(v0), log=TRUE ) )
    return(loglik + lprior)
}

# Bayes function based on 'metrop' (Geyer, 2009a)

RRbayes <- function(x, y, m0, v0, initial, nbatch, blen, scale, ...)
{
   ini.bad <- is.infinite( lupost(b=initial,x=x,y=y,m0=m0,v0=v0) )
   if(ini.bad) stop("initial not in feasible region")
   require(mcmc)
   Out <- metrop( lupost, initial = initial,
              nbatch = nbatch, blen = blen,
              outfun = function(z, ...)  exp(z),
              x = x, y = y,
              v0 = v0, m0 = m0,
              scale = scale )
   postrr <- apply(Out$batch, 2, mean)
   quant <- t( apply(Out$batch, 2, quantile, prob=c(.025,.975)) )
   names(quant) <- c("q0.025","q0.975")
   bci <- data.frame(RR=postrr, quant,
                     row.names=colnames(x))
   mcse <- apply(Out$batch, 2, sd) / sqrt(Out$nbatch)
   names(mcse) <- colnames(x)
   return( list(Out=Out, Summary=bci, Precision=mcse) )
}
```

B Implementation with STEP Data

```
> source("lbr.R")  # source code containing the functions above
> X <- read.csv(file = "schistosoma.csv")  # data from Tallo et al. (2008), available upon request

> dim(X)

[1] 1764    5

> head(X)

      y sex step age sj
12677 1   0    0   2  1
12678 0   1    0   4  1
12679 1   1    1   1  1
12680 0   1    0   1  1
12681 0   1    0   3  1
12682 0   1    0   4  1

> sapply(X, class)
```

```
        y       sex      step       age        sj
"numeric" "numeric" "numeric"  "factor"  "factor"
```

```
> modGLM <- glm(y ~ sex + step + age + sj, data = X,
+     family = binomial(link = "log"), start = rep(-1,
+         8))
> modGLM$boundary
```

```
[1] TRUE
```

```
> modGLM$converged
```

```
[1] FALSE
```

```
> relrisk.glm(modGLM, robust = FALSE)
```

```
Call:
glm(formula = y ~ sex + step + age + sj, family = binomial(link = "log"),
    data = X, start = rep(-1, 8))

               RR   l95   u95
(Intercept) 0.545 0.471 0.619
sex         0.922 0.830 1.013
step        0.922 0.830 1.013
age2        0.925 0.805 1.045
age3        0.491 0.381 0.601
age4        0.608 0.487 0.729
sj2         1.991 1.842 2.140
sj3         1.695 1.477 1.914
```

```
> modQL <- glm(y ~ sex + step + age + sj, data = X,
+     family = poisson(link = "log"), start = rep(-0.5,
+         8))
> modQL$converged
```

```
[1] TRUE
```

```
> relrisk.glm(modQL, robust = TRUE)
```

```
Call:
glm(formula = y ~ sex + step + age + sj, family = poisson(link = "log"),
    data = X, start = rep(-0.5, 8))

               RR   l95   u95
(Intercept) 0.588 0.511 0.665
sex         0.943 0.842 1.045
step        0.855 0.592 1.119
age2        0.905 0.788 1.023
age3        0.469 0.367 0.572
age4        0.589 0.475 0.702
sj2         2.851 2.353 3.350
sj3         1.932 1.539 2.326
```

```
> 100 * sum(fitted(modQL) >= 1)/1764
```

```
[1] 0.5669
```

```
> modLBR <- lbreg(y ~ sex + step + age + sj, start = rep(-1,
+     8), data = X)
> summary(modLBR)
```

```
Call:
lbreg.formula(formula = y ~ sex + step + age + sj, data = X,
```

```
    start = rep(-1, 8))

            Estimate Std.Err z.value p.value
(Intercept)  -0.5804  0.0562  -10.34 < 2e-16
sex          -0.0575  0.0515   -1.12  0.2644
step         -0.0575  0.1892   -0.30  0.7612
age2         -0.0250  0.0621   -0.40  0.6875
age3         -0.6674  0.0800   -8.34 < 2e-16
age4         -0.4523  0.0780   -5.80 6.7e-09
sj2           0.6379  0.4655    1.37  0.1706
sj3           0.4986  0.1926    2.59  0.0096

Log Likelihood: -1108
Convergence code: 0
Barrier Value: 0.1648
Outer Iterations: 6
```

```
> relrisk(modLBR)
```

```
Call:
lbreg.formula(formula = y ~ sex + step + age + sj, data = X,
    start = rep(-1, 8))

               RR   l95   u95
(Intercept) 0.560 0.477 0.642
sex         0.944 0.846 1.042
step        0.944 0.584 1.304
age2        0.975 0.855 1.096
age3        0.513 0.401 0.625
age4        0.636 0.514 0.758
sj2         1.893 0.637 3.148
sj3         1.646 1.162 2.131
```

```
> # Bayesian estimation
> modBayes <- RRbayes(x=xx, y=X$y, m0=0, v0=100,
initial=coef(modLBR), nbatch=1000, blen=100, scale=0.03)
> modBayes$Summary
(...)          #   see TABLE 3
> plot(density(modBayes$Out$batch[,3]))     #  see FIGURE 2
```