UW Biostatistics Working Paper Series

7-19-2006

Relative Risk Regression in Medical Research: Models, Contrasts, Estimators, and Algorithms

Thomas Lumley
University of Washington, tlumley@u.washington.edu

Richard Kronmal
University of Washington, kronmal@u.washington.edu

Shuangge Ma
Yale University, shuangge.ma@yale.edu

By far the most popular regression model for binary data is logistic regression. Logistic regression plays a central role in observational epidemiology because the odds ratios that it estimates are identifiable from case–control samples [Cornfield, 1956], and because for rare events the odds ratios approximate the relative risk. Logistic regression is also used to model common events in cross-sectional or longitudinal studies, where the relative risk could be directly estimated and is not close to the odds ratio.

The argument that relative risks will often provide a more useful summary of associations has been made repeatedly over at least two decades [eg Wacholder, 1986, Sinclair & Bracken 1994, Davies et al 1998, Skov et al 1998, McNutt et al 2003, Greenland 2004, Liberman 2005, Katz 2006] and we will not belabor it here. Perhaps even more important as a reason for preferring relative risks in summarising associations in binary data is the difficulty of explaining the correct interpretation of odds ratios. Unfortunately, even researchers who clearly understand the distinction between relative risks and odds ratios will find it difficult to ensure that the correct interpretation of an odds ratio is communicated to a general audience. The difficulties are illustrated well by [Schulman et al, 1999]. The authors estimated an odds ratio of 0.6 for effect of race on referral for angiography in a well-designed and well-executed study and reported this odds ratio in a major medical journal. The study was widely discussed in the news media as if it reported a relative risk of 0.6, when in fact the relative risk was 0.93 [Schwartz et al, 1999]. It appears unavoidable that the reported associations will be interpreted as relative risks, arguing that they should in fact be relative risks.

Several estimators of the relative risk have been proposed over the years, often on more than one occasion. Barros & Hirakata (2003) give a comprehensive listing of proposals up to their date of writing and compare performance in simulations. Our review differs from previous reports in focusing on the estimating equations solved by various proposed estimators of the relative risk and in evaluating the efficiency and robustness tradeoffs they make. Previous reviews have often considered software rather than estimators. This appears more straightforward but frequently leads to confusion. For example, Deddens & Petersen (2004) commented that using Cox model software for binary outcomes seemed inappropriate because the Cox model is for survival data, and Ma & Wong (1999) argued that three relative risk estimators were based on models whose assumptions were not satisfied by the data. This misses the point. If the estimator is a good one then the fact that it can be easily computed by abusing software designed for the Cox model is an advantage, not a disadvantage. Aesthetically and pedagogically there may be cause for complaint, but the objection is best removed by persuading software vendors to add a new 'relative risk regression' interface to the same estimation code, and again the fact that no new estimation code is required should make this easier.

We describe the relative risk regression model and review a number of estimation algorithms that have been proposed. We show that the estimators that give consistent results and valid standard errors can be seen as a series of robust generalizations of the maximum likelihood estimator. We compare the efficiency of these estimators and give some guidelines for implementation in popular software. In a companion paper (Lumley & Kronmal 2006) we provide more algorithmic and mathematical detail, including implementations of all these algorithms in the R statistical environment (R Development Core Team, 2006)

In the interests of terminological clarity we note that relative risk regression is also called 'prevalence ratio' or 'prevalence rate ratio' regression by some authors. We assume that binary outcome data are actually of interest, i.e., that time-to-event information is either unavailable or inappropriate to the substantive question at hand. We also note that our discussion in this paper is entirely concerned with inference about relative risks as a description of associations between variables, and not with prediction. The literature on relative risk regression has generally this same focus. Criteria for building and evaluating models for the purpose of prediction are completely different. Good discussions of modelling for prediction or classification can be found in Hastie et al (2001), which discusses more automated methods, and Harrell (2001), which focuses on approaches requiring more detailed control by the analyst.

1 The relative risk regression model

Relative risks arise naturally from the regression model

$$\log P[Y=1|X] \equiv \log\mu = \eta \equiv \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{1}$$

in which e^{β_i} is a relative risk contrasting levels of X_i that differ by 1. If $P[Y=1|X]$ is small then

$$\log P[Y=1|X] \approx \log\frac{P[Y=1|X]}{1-P[Y=1|X]} \equiv \operatorname{logit} P[Y=1|X],$$

and if this is true for for all observed values of X the relative risk regression model is very close to the logistic regression model

$$\operatorname{logit} P[Y=1|X] = \operatorname{logit}\mu = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_p X_p$$

If P[Y=1] is larger than 10-15% for any observed values of X (Greenland) then α and β will differ noticeably, with $|\alpha| > |\beta|$.

Like the logistic regression model, the relative risk regression model is a generalized linear model (McCullagh & Nelder, 1989), with log link and variance function $V(\mu) = \mu(1-\mu)$. Unlike the logistic regression model, the relative risk model requires constraints on β to ensure that fitted probabilities remain in the interval [0,1].

The maximum likelihood estimator, and all the other consistent estimators that have been proposed, solve equations of the form

$$\sum_{i=1}^{n} \frac{\partial\mu_i}{\partial\beta} w(\mu_i)(Y_i-\mu_i) = \sum_{i=1}^{n} x_i\mu_i w(\mu_i)(Y_i-\mu_i) = 0. \tag{2}$$

with different choices of weight function $w(\cdot)$, at least in situations where a solution to this equation exists in the parameter space being considered.

The estimating equations are unbiased for any choice of $w(\cdot)$, so all these estimators are consistent, and are asymptotically Normal as long as β is in the interior of the parameter space being considered (McCullagh & Nelder, 1989).

2 Estimators of the relative risk

2.1 Maximum likelihood estimation

The statistically natural estimator for the relative risk is the maximum likelihood estimator in model 1. A number of authors have noted that this is a generalized linear model and argued that standard generalized linear model software should be used to fit it (Wacholder 1986, Skov et al 1998, Robbins et al 2002). In practice, difficulties arise.

There is at most one solution (Wedderburn, 1976) to the likelihood equations

$$\sum_{i=1}^{n} x_i \mu_i \frac{1}{\mu_i(1-\mu_i)}(Y_i - \mu_i) = 0, \tag{3}$$

with all $\mu_i \leq 1$, and if it exists it is the maximum likelihood estimator. There may be multiple solutions with some $\mu_i > 1$ and, in addition, if any $\mu_i \approx 1$ the estimating function will be dominated by observation i, and estimation software may either fail or falsely report convergence.

Some software (eg, Stata) attempts to provide the solution to equation 3 even if it has some $\mu_i > 1$, though a more or less clear warning may be given that something unexpected has occurred. Other software (eg R) attempts to solve the constrained optimization problem needed to give the true maximum likelihood estimator.

Both the MLE and the solution to the likelihood equations are consistent for β and asymptotically efficient under model 1. They are also design-consistent under sampling from a population even when model 1 is not true. That is, as sample size increases they converge to the values of β that would be obtained by applying the same estimator to the whole population.

Any consistent sequence of solutions to equation 3 will always be asymptotically Normal. The true maximum likelihood estimator will be asymptotically Normal only when the true β lies in the interior of the parameter space. If the maximum likelihood estimator is on the boundary of the parameter space or the solution to equation 3 has some $\mu_i > 1$, the usual model-based standard errors from model 1 will be incorrect. The solution to equation 3 is computationally simpler, but has the disadvantage that the weight $1/\mu(1-\mu)$ given to some observations may be very large or may be negative. From a methodological viewpoint the meaning of negative weights is unclear. A practical disadvantage is that ordinarily reliable statistical software may not have been as thoroughly tested with negative weights.

2.1.1 Algorithms for the MLE

Most standard statistical software will not automatically produce the true maximum likelihood estimator, because it is not readily obtained from the same Fisher scoring algorithm used for more popular generalized linear models.

We will argue below that the maximum likelihood estimator is insufficiently robust to model misspecification and that other estimators are often preferable. Algorithms for computing the maximum likelihood estimator, or a good approximation to it, are still of interest.

COPY algorithm Deddens & Petersen (2003) describe one algorithm for obtaining something close to the maximum likelihood estimator. They proposed taking C copies of the data and 1 copy with Y set to $1 - Y$ and fitting the relative risk model to these modified data. They call this the COPY algorithm. The same effect could be obtained with orders of magnitude less computational effort by taking just one copy of the data with outcome Y and one with with outcome $1 - Y$ and using weights C and 1 respectively. This is particularly important in our research applications, which are to cohort studies of thousands of individuals. Essentially the same weighted estimator was proposed by Carter & Lipsitz (2006), motivated by the common practice of adding a small constant to cell counts to remove the problem of zero counts when analysing categorical data.

Writing β_C for the maximum likelihood estimator in the modified data, as $C \to \infty$, β_C converges to the solution to equation 3 and as $C \to 1$, $\beta_C \to (-\log 2, 0, 0, \ldots, 0)$. The weight C must be chosen large enough that little bias results and small enough that the solution to equation 3 lies within the permissible parameter space. Deddens & Petersen recommend $C = 1000$ and show that it works well in some simulated examples. We show in examples below that (presumably when the model is not exactly true) it may be necessary to use much smaller values of C and the resulting estimator need not be close to the MLE.

Truncating fitted values Another approach to obtaining convergence to an approximate mle for the log-binomial model is to simply truncate the range of μ (Wacholder, 1986). A threshold near 1, such as 0.999, is chosen and μ is set to $\min(\mu, 0.999)$ after each iteration for the purpose of computing working residuals and working weights for the next iteration. The resulting estimator is the MLE when the MLE is sufficiently far in the interior of the parameter space. When the MLE is on the boundary of the parameter space this estimator is not the MLE, and need not be consistent or asymptotically Normal, although it is almost as non-robust as the MLE. Baumgarten et al (1989) also warn that this algorithm may be sensitive to starting values and to the tolerance for convergence.

Searching the boundary Deddens & Petersen (2003) also mention (without details) a method for finding the true maximum likelihood estimator by searching the boundary of the parameter space. With two predictors, as in the example they mention, this is straightforward as either a graphical or an automated approach. If the MLE is not in the interior of the parameter space then it must have fitted probability equal to 1 at one corner or one edge of the convex hull of the observed predictors x_i, and a search for this edge or corner is easy. In higher dimensions, however, it appears much more difficult to characterize and search the extreme points.

Step-halving The `glm` function in R augments the usual Fisher scoring algorithm with step-halving. That is, $\hat\beta_{OLD}$ is first updated to $\hat\beta_{NEW}$ by Fisher scoring. If $\hat\beta_{NEW}$ is outside the valid parameter space we set

$$\hat\beta_{NEW} := \frac{1}{2}\left(\hat\beta_{NEW} + \hat\beta_{OLD}\right)$$

and repeat this until $\hat{\beta}_{NEW}$ is inside the parameter space.

In theory this algorithm will always converge to the MLE. In practice it is fairly reliable, but if an edge of the parameter space is almost perpendicular to the gradient of the the loglikelihood, the estimate will move only very slowly along the edge and convergence may be declared before the MLE is reached.

Other constrained optimization methods Many algorithms for constrained optimization have been developed in the numeric analysis literature [eg Boyd & Vandeberghe, Chapter 11]. These are not readily available to users of statistical software, but would be a natural choice for software vendors wishing to add maximum-likelihood relative risk regression to their offerings. One example of such an algorithm is the adaptive log-barrier algorithm of Lange (1994), which is built in to R and which we have used to compute the maximum likelihood estimator when testing other algorithms.

2.2 Poisson working model

Poisson regression software is another natural choice for fitting a log-linear model, since it estimates incidence rate ratio and since most medical applications of the Poisson distribution arise via the Poisson approximation to the binomial distribution. This approach has been proposed by Traissac et al (1999), McNutt et al (2003), Zou (2004), and Carter et al (2005). The estimating equations are those for a generalized linear model with log link and variance proportional to mean

$$\sum_{i=1}^{n} x_i \mu_i \frac{1}{\mu_i}(Y_i - \mu_i) = 0.$$

The estimating equations for Poisson regression are unbiased when the response variable is binary rather than Poisson, and thus lead to consistent estimation of the relative risk. Software for Poisson regression is widely available, although one important package, SPSS, does not currently provide it.

This estimator is often described as 'assuming a Poisson model', which is unfortunate choice of terminology since it is neither reasonable nor necessary to assume that the binary variable Y_i has a Poisson distribution. When referring to the relationship between this estimator and the maximum likelihood estimator for a Poisson model a better term might be 'using a Poisson working model', by analogy with the working correlation models used in GEE (Zeger & Liang 1986).

When use to estimate relative risks from binary data, Poisson regression gives standard errors that are too large, because the variance of a Poisson random variable is always larger than that of a binary variable with the same mean. This bias can be removed by using model-robust standard error estimates. Zou (2004) and Carter et al (2005) suggested using the model-robust sandwich estimator; Barros & Hirakata (2003) show in some examples that two standard scale adjustments for overdispersion can give reasonably accurate standard errors if the model-robust estimator is not available and a bootstrap is too difficult to implement.

An aesthetically pleasing property of using the Poisson working model is that as the outcome becomes rarer the estimator approaches both the log-binomial and the logistic regression estimator. All three of these are asymptotically equivalent and thus fully efficient for rare events.

2.3 Nonlinear least squares

Nonlinear least squares estimation involves finding the relative risk estimates than minimize $\sum_i (Y_i - \mu_i)^2$. The estimating equations that result from differentiating this objective function are the same as those for a generalized linear model with log link and constant variance

$$\sum_{i=1}^{n} x_i \mu_i (Y_i - \mu_i) = 0.$$

These would be the likelihood equations for data with a Gaussian distribution, and a 'Gaussian working model' may be the clearest description of these estimator when computations are done using standard generalized linear model software.

As with the working Poisson model, using a variance function other than the binomial results in less efficient estimates and biased standard errors, but still gives consistent estimators of the relative risk. In constrast to the working Poisson model, the standard errors estimates may be either too large or too small, depending on the range of fitted probabilities and the skewness of the predictor variables. The bias in the standard errors is often very small for nonlinear least squares; the usual standard error formula already incorporates an estimated dispersion term analogous to those studied for Poisson regression by Barros & Hirakata (2003). As with the Poisson working model, correct standard errors can be obtained from a robust 'sandwich' variance estimator or from a jackknife or bootstrap.

2.4 Using Cox regression software

Cox regression estimates hazard ratios, which are relative risks at an instant. This motivated the use of Cox regression software for relative risk regression. In reality, different events occur at different instants of time and the hazard ratio will lie between the odds ratio and the relative risk. To make the hazard ratio equal to the relative risk a fictitious data set is constructed where every individual has the same observation time and so all events occur at the same instant (Lee & Chia 1993, Lee, 1994)). This data construction poses problems for the Cox model, which is mathematically and computationally simple only in continuous time, when no tied events occur. These complications have led some authors (Ma & Wong 1999) to conclude, incorrectly, that the method is invalid.

The first popular approximation for handling tied event times in genuine survival data was given by Breslow (1974). When every individual has the same artificial observation time this approximation results in the same estimating equations as Poisson regression, and so gives the same consistent estimates, the same upwardly-biased model-based standard errors, and the same consistent model-robust standard errors. The Breslow approximation is the default in many statistical packages, including SPSS, SAS, and Stata.

Many packages also provide, either as an option (Stata, SAS) or as the default (S-PLUS, R), a more accurate approximation due to Efron (1977). Some packages also offer the exact partial or marginal likelihoods. These more accurate methods of handling ties are often recommended for genuine time-to-event data when the number of ties is large (Hertz-Picciotto & Rockhill, 1997), but their use in relative risk regression would cause serious bias. Using the exact partial likelihood gives a hazard ratio estimate that is identical to the conditional logistic regression estimator of the odds ratio. Using the exact marginal likelihood or the Efron approximation results in estimators that are not consistent for either the odds ratio or the relative risk.

Because of the dependence of the results on the method used to handle ties, and because there is, even at best, no advantage over the use of Poisson regression software we do not recommend the use of Cox regression software to estimate relative risks. This technique might be in principle be useful when Poisson regression software is not available, but we know of no software package that provides Cox regression with model-robust standard errors and does not provide Poisson regression.

2.5 Scaling by the average prevalence

In a model with a single, binary predictor the relative risk can be converted to the odds ratio by

$$RR = \frac{OR}{(1-p_0) + p_0 \times OR} \tag{4}$$

where p_0 is the probability of the event in the unexposed (or referent) group.

Zhang & Yu(JAMA 1998) propose using equation 4 for more general logistic regression models, and Zocchetti et al (1997) give a similar proposal. The advantage of this proposal is its simplicity, and equation 4 is certainly useful when a study has published odds ratios and a reader needs an approximate translation to relative risks. However, as noted by authors including McNutt et al (2003), equation 4 does not give a consistent estimator of the relative risk. Even at an operational level it may not be clear what referent group to use to compute p_0, for example in a model with continuous predictors or with interactions.

2.6 Duplication of cases

Schouten et al suggested duplicating each observation that has Y=1, setting Y=0 for the duplicate. They argued that in this new data set $P[Y=1] \equiv \nu = \mu/(1+\mu)$ and so

$$\operatorname{logit} \nu = \log \frac{\nu}{1-\nu} = \log \frac{\mu/(1+\mu)}{1-\mu/(1+\mu)} = \log \mu$$

so that logistic regression could be used to estimate β, though this estimation procedure will not guarantee $\mu \leq 1$ without additional constraints.

If we write (x_{ij}, Y_{ij}) with $j=0$ for the original data and $j=1$ for the duplicate, the logistic regression estimating equations are

$$\sum_{i,j} x_{ij}(Y_{ij} - \nu_{ij}) = 0$$

Writing these equations in terms of the original data we obtain yet another set of generalized linear model estimating equations

$$\sum_{i=1}^{n} x_i \left(Y_i - (1+Y_i)\frac{\mu_i}{1+\mu_i} \right) = \sum_{i=1}^{n} x_i \frac{1}{1+\mu_i}(Y_i - \mu_i) = 0. \tag{5}$$

The estimating equations are unbiased, and so the solution is consistent for β and asymptotically Normal. Model-robust standard errors are needed in this case to handle correlation from the duplicated observations Y_{i0}, Y_{i1} rather than misspecification of the marginal variance function. This use of sandwich estimators is analogous to that in GEE, leading Skov et al (1998) to call equation 5 the GEE–logistic procedure.

3 Robust estimators in the generalized linear model.

Equation 2 yields consistent, asymptotically Normal estimators of β for all the $w(\cdot)$ we consider, so the choice of $w(\cdot)$ presents a tradeoff between efficiency and robustness. The efficient estimator of β is obtained with

$$w(\mu) = \frac{1}{\mu(1-\mu)}.$$

While it is not unusual for an efficient estimator to lack robustness, the extent of the non-robustness is greater than usual. A single point with μ close to 1 can have arbitrarily large influence despite having bounded covariate values. If gross outliers in x are plausible, either because of measurement error or because of the presence of small subpopulations where the risk relationship is truly different, the efficient estimator is likely to be too sensitive.

Figure 1 shows the net observation weight $\mu w(\mu)$ for the four consistent estimators. It is clear that the weight becomes very much larger for the efficient estimator, where $\mu w(\mu) = 1/(1-\mu)$. The other three estimators have relatively stable weight functions and it is not obvious how their performance will vary. Wacholder's estimate, which truncates μ at a value $1-\epsilon$, such as 0.999, has a weight function that follows that of the MLE but is truncated at $1/\epsilon$. For $1-\epsilon = 0.999$ the truncation point is well off the top of the figure.

3.1 Efficiency

The price for robustness is usually a moderate reduction in efficiency, so we explored the loss of efficiency from using the robust weight functions rather than maximum likelihood estimation. To gain more insights, we consider a simplified case with $p = 1$, i.e., one dimensional covariate. In this case, the relative efficiency for estimating β is invariant to location-scale transformation of X. So the asymptotic relative efficiencies (of different estimates) will be completely determined by the distribution of μ, given the one-to-one relationship between μ and X with fixed coefficients.

We use the following simple simulation study to investigate the relative efficiencies for differently distributed P with one dimensional X. Consider a simple regression model $\log(P[Y=1|X]) = \beta_0 + \beta_1 X$, with $\beta_0 = -1$ and $\beta = 1$. We assume the probability μ has a Beta distribution $B(\theta_1, \theta_2)$

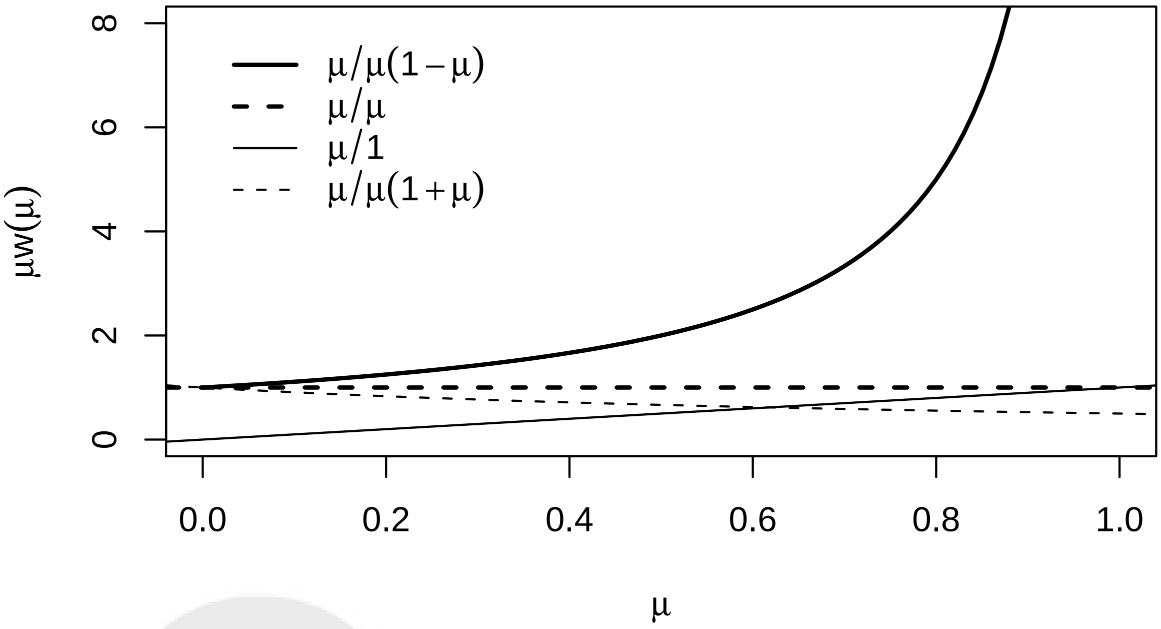

Figure 1: Net observation weight $\mu w(\mu)$ vs μ for the four consistent estimators of the relative risk

distribution. Different combinations of θ_1 and θ_2 yield a rich family of probability distributions. For example, $\mu \sim B(0.2, 0.2)$ has a bowl-shape distribution, with a majority of μ close to 0 and 1; $\mu \sim B(1, 1)$ is uniformly distributed between 0 and 1; $\mu \sim B(2, 2)$ is bell-shaped, with mode at 1/2; $\mu \sim B(0.2, 2)$ is clustered around 0 and $\mu \sim B(2, 0.2)$ is clustered around 1.

In our simulation study, we first generate 500,000 realizations of μ from the appropriate Beta distribution. Once μ is generated, we can compute corresponding X from the equation $\log(P[Y = 1|X]) = \mu = \beta_0 + \beta_1 X$ and $(\beta_0, \beta_1) = (-1, 1)$. We can then obtain the asymptotic variances of the estimates of β_1 for different approaches by approximating the sandwich variance estimates with their empirical counterparts based on the 500,000 realizations. Asymptotic relative efficiencies can be then computed.

In Figures 2–5, we show the relative efficiencies for the three approaches with respect to the MLE as a function of θ_2 for different, fixed θ_1. It can be seen that under all simulated settings, the duplication of cases approach is less efficient than the Poisson regression; the relative efficiency of the nonlinear least squares approach with respect to the Poisson regression depends on the distribution of μ; Both the Poisson and the nonlinear least squares approaches are very efficient (with relative efficiency greater than 0.8), unless μ is extremely clustered around 0 and/or 1.

3.2 Gross error sensitivity

To illustrate the sensitivity of the MLE we performed a simulation based on $\mu \sim B(2, 2)$ as above, but contaminating the covariates with 0.5% and 1% gross errors in x. Two contaminating distributions were used. In the first, the contaminating values of x were taken so that $\mu \sim B(0.5, 0.5)$. This means that at the true value of β the fitted values are all in $[0, 1]$. The second contaminating distribution was $logNormal(0.5, 1)$, which has the same median as the true x but a larger range, and where some erroneous x values will lead to fitted $\mu > 1$.

In Table 1 we show the bias and mean squared error for estimates from the contaminated and uncontaminated data with three of the four weight functions in Figure 1. The MLE is more efficient in the absence of contamination, but loses its efficiency advantage in the presence of very small amounts of contamination even when the fitted values stay in $[0, 1]$. When errors produce fitted values outside $[0, 1]$ the MLE is much more biased and has much large MSE than the alternatives.

3.3 Example:

In this section, we illustrate the relative risk regression using data from the Multi-Ethnic Study of Atherosclerosis (MESA), a multi-center study of sub-clinical cardiovascular disease (Bild et al., 2002). The sample consists of 6,814 men and women aged 4584, who are Caucasian, African-American, Hispanic, or Chinese-American. The response to be considered here is the presence of coronary artery calcium (CAC), a measure of the presence of coronary artery disease, determined by the use of computed tomography (CT). For illustrative purposes, we restrict the predictors of interest to age, gender, and high density lipoprotein cholesterol (HDL). Approximately 50% of the MESA participants have CAC present on the CT scan.

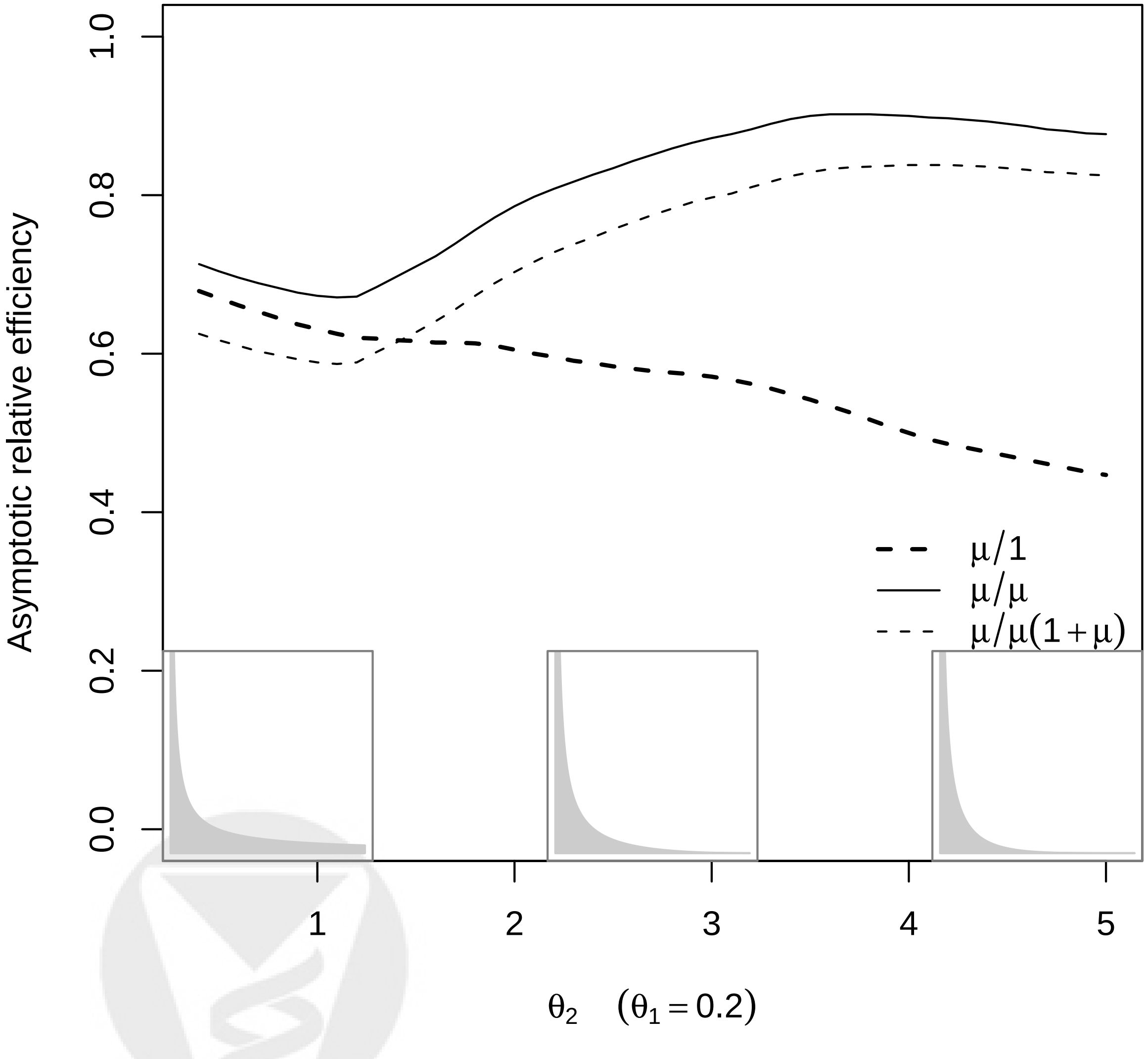

Figure 2: Efficiency of three alternative estimators relative to the MLE, when $\mu \sim B(0.2, \theta_2)$. Inset graphs show probability density for $B(0.2, 1)$, $B(0.2, 3)$, $B(0.2, 5)$

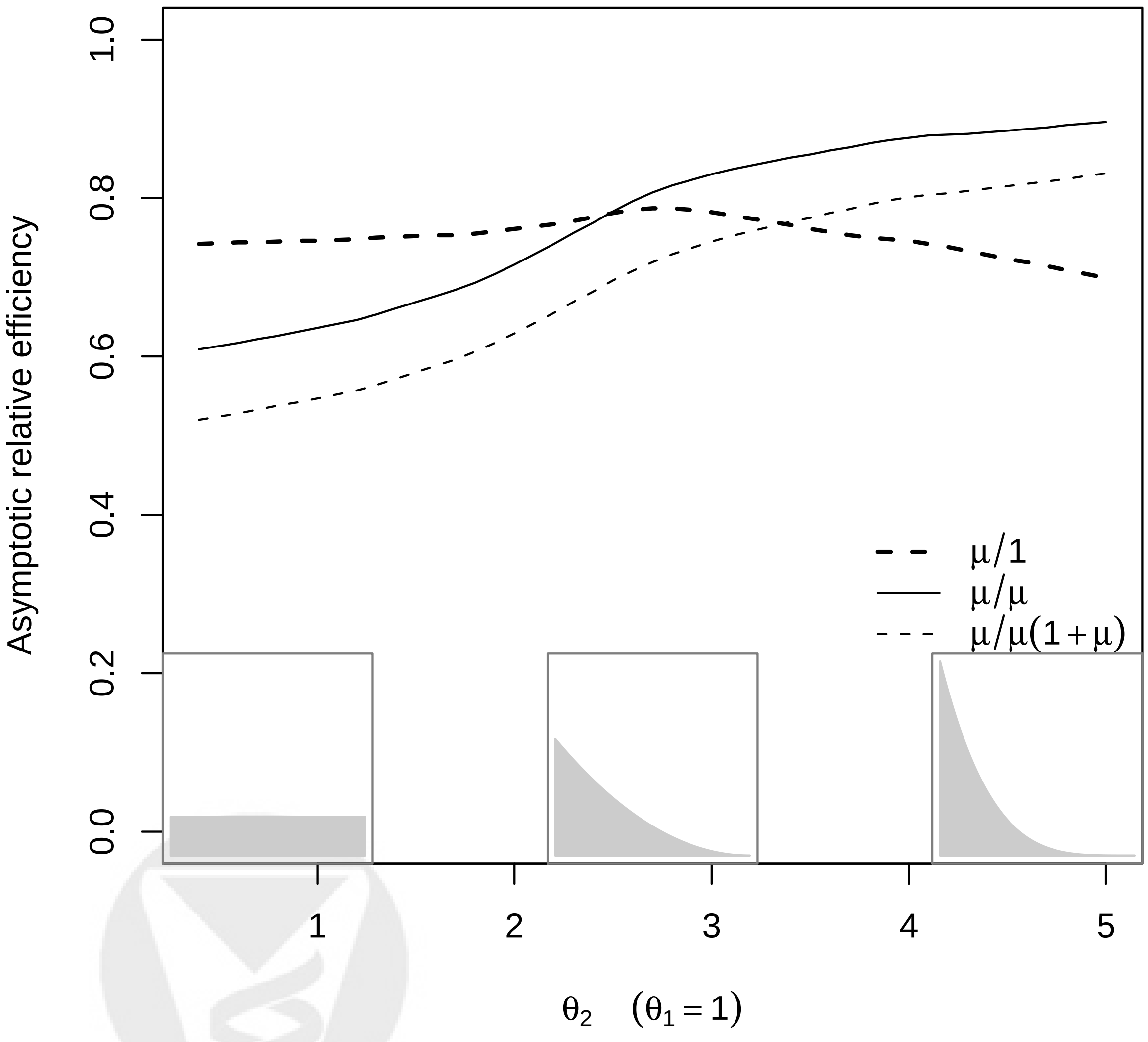

Figure 3: Efficiency of three alternative estimators relative to the MLE, when $\mu \sim B(1, \theta_2)$. Inset graphs show probability density for $B(1, 1)$, $B(1, 3)$, $B(1, 5)$

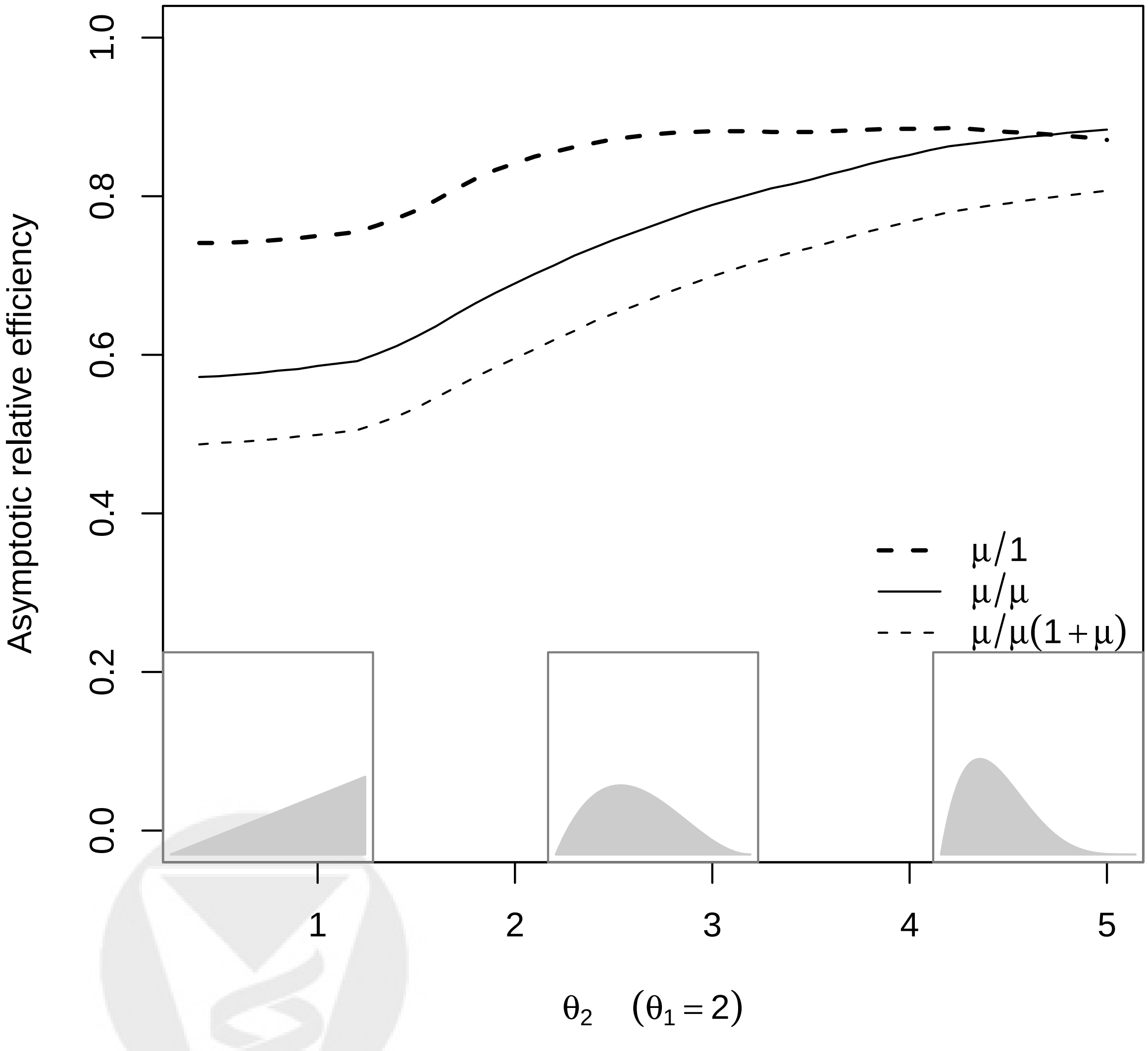

Figure 4: Efficiency of three alternative estimators relative to the MLE, when $\mu \sim B(2, \theta_2)$. Inset graphs show probability density for $B(2, 1)$, $B(2, 3)$, $B(2, 5)$

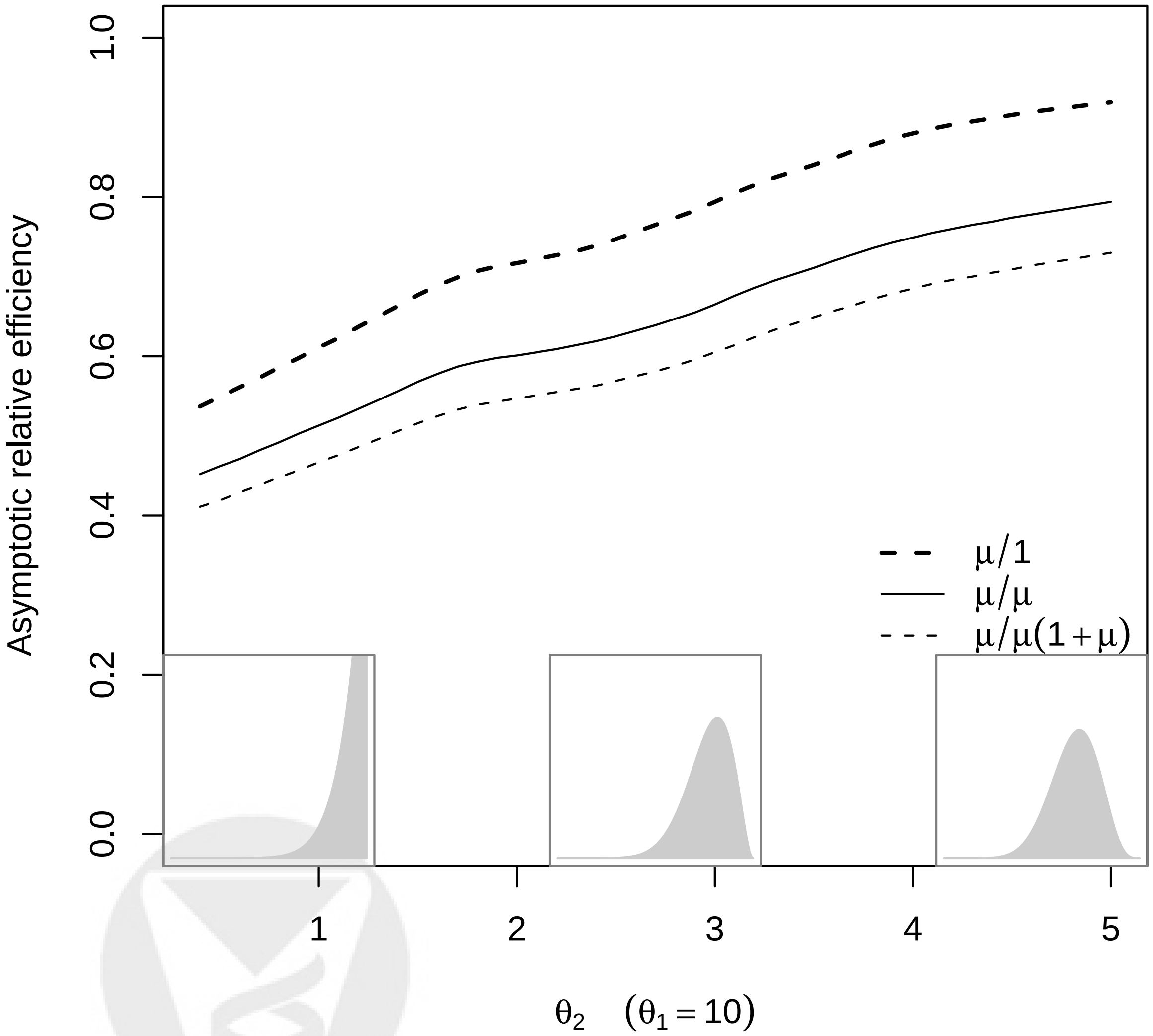

Figure 5: Efficiency of three alternative estimators relative to the MLE, when $\mu \sim B(10, \theta_2)$. Inset graphs show probability density for $B(10, 1)$, $B(10, 3)$, $B(10, 5)$

Table 1: Median bias and MSE of regression coefficient from three estimators under gross error contamination, based on 1000 replications

	Weight function		
	Binomial	Poisson	Constant
	$\mu/(\mu(1-\mu))$	$\mu/1$	μ/μ
Uncontaminated			
(bias)	-0.007	0.001	0.003
(MSE)	0.0035	0.0065	0.0051
Beta(0.5,0.5)			
0.5% (bias)	-0.031	-0.026	-0.009
(MSE)	0.0076	0.0079	0.0057
1.0% (bias)	-0.05	-0.05	-0.02
(MSE)	0.0090	0.0110	0.0058
Lognormal			
0.5% (bias)	-0.42	-0.02	-0.07
(MSE)	0.20	0.0072	0.0135
1.0% (bias)	-0.52	-0.05	-0.14
(MSE)	0.27	0.0089	0.0271

Table 2: Relative risk and odds ratio estimates for risk of coronary calcification

	Relative Risk				
	MLE	Poisson	NLS	Scaled OR	Odds ratio
Age (vs < 55)					
55–64	1.84	1.79	1.84	1.46	2.69
65–74	2.58	2.50	2.62	1.73	6.19
75–84	3.07	3.13	3.34	1.89	15.91
Male	1.37	1.37	1.45	1.43	2.47
HDL (vs lowest quintile)					
2nd	0.98	0.95	0.91	0.87	0.78
3rd	0.96	0.92	0.87	0.82	0.69
4th	0.95	0.92	0.88	0.83	0.71
5th	0.91	0.87	0.83	0.77	0.62

Table 3: Relative risk and odds ratio estimates for risk of coronary calcification

	Relative Risk				
	MLE	Poisson	NLS	Scaled OR	Odds ratio
Age (/10 yrs)	2.02	1.98	1.88	1.46	2.70
$(Age - 50)^2$	0.90	0.91	0.92	1.00	0.99
Male	1.36	1.44	1.36	1.43	2.48
$\log HDL$	0.89	0.77	0.82	0.68	0.51

Table 2 shows results from the maximum likelihood, non-linear least squares, working Poisson model, and logistic regression. In these analyses, age and HDL are categorized into 5 intervals. In this model there was no difficulty in estimating the MLE by the usual Fisher scoring algorithm. All fitted values were below 1 for the MLE and non-linear least squares estimates; the working Poisson model gave fitted values up to 1.08.

The first three RR model estimates are similar and would lead to roughly the same quantitative and qualitative inferences. The scaled odds ratio estimate of the relative risk, using the mean fitted value as p_0 in equation 4, is noticeably different, in particular for age. Using the fitted value when all covariates are zero as p_0 gives a much worse approximation to the relative risk. As expected, in this example the odds ratios from logistic regression grossly overestimate the RRs and would likely lead to the impression by the unsophisticated reader of very strong risks associated with all of the variables.

Table 3 shows the various estimates for the different models with age, age squared and ln(HDL) as continuous variables (age centered at 50). In this example the MLE is on the boundary of the parameter space, but is not far from the two quasilikelihood estimates.

Estimation of the MLE failed in Stata so the MLE was computed in R. The R `glm` function and the log-barrier constrained optimizer gave relative risks differing by 0.2–2%. One observation has $\hat{\mu} = 1$ at the MLE, an 81-year old man with the lowest observed HDL. The COPY algorithm required $C = 25$ for the final estimate to be in the interior of the parameter space and $C = 2$ for Fisher scoring to stay inside the parameter space with starting value -1 for the intercept and 0 for all other parameters. At $C = 25$ the COPY algorithm gave very similar results to nonlinear least squares, but at $C = 2$ there was substantial bias, eg a coefficient of 1.31 for age and 1.18 for male gender.

In the figures presented below, all of the curves shown are for the unadjusted models. Figure 6(a) shows plots with age as the predictor variable of a lowess estimator of the RR function, along with the RR functions from the maximum likelihood, non-linear least squares and logistic regression. Each estimator has a comparison value of age 50 for the RR computation. The lowess estimator is computed by first estimating the probability of CAC presence by using CAC coded as 0 for presence and 1 for absence and then computing from the estimated probabilities, the RRs for each study participant. As observed in table 3, the odds ratio curve provides an extremely poor estimate of the RR. Although it is not readily apparent from the figure, the RR curves deviate somewhat from the lowess curve. Figure 6(b) shows the maximum likelihood curve for the linear model and additionally the maximum likelihood and non-linear least squares models including both a linear and quadratic term. Clearly the RR model curves with the quadratic term now closely approximate

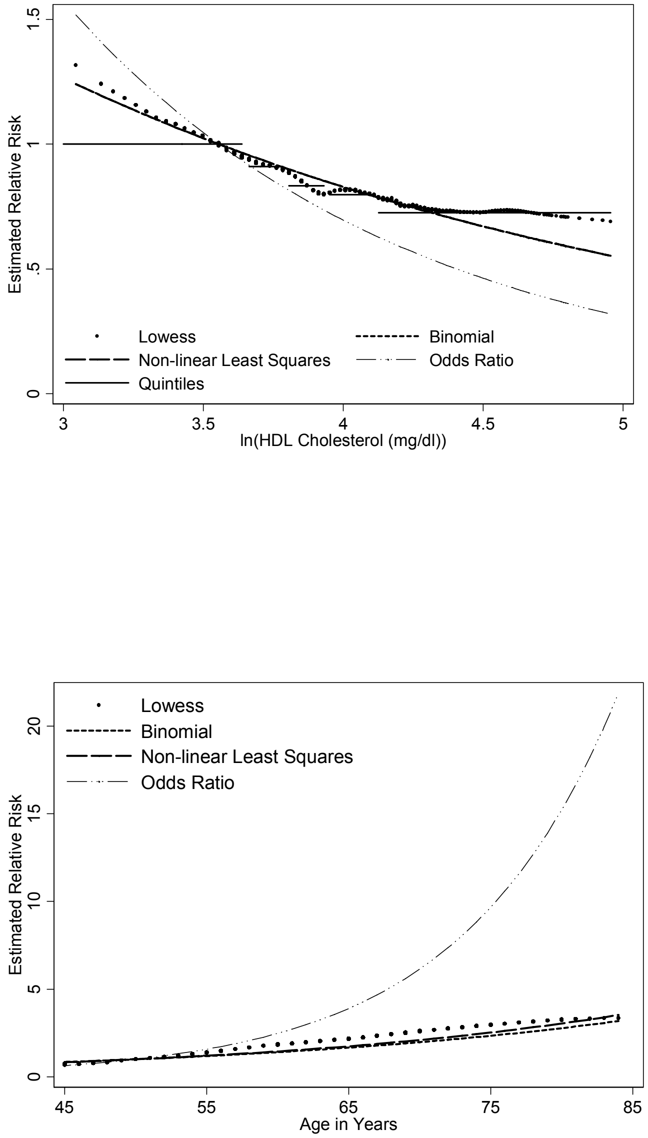

Figure 6: Fitted risks relative to age 50 from models and smoothing: (a) linear term in age (b) linear and quadratic terms in age.

the lowess curve. Also overlaid on the plot for comparison to the models in table 3, is the estimator of the RRs for age in categories.

In Figure 7 the same curves as described above except with ln(HDL) as the predictor and the 10th percentile value of ln(HDL) as the comparator. In this example the RR models provide excellent estimates of the lowess estimated RR curve. Again the odds ratios are clearly not reasonable estimates of the RRs.

4 Implementations

We consider the statistical packages SAS, SPSS (version 11), Stata (version 9), S-PLUS (version 7.0.3), and R (version 2.2.1).

SAS Computations using SAS (SAS Institute Inc, 2004) `PROC GENMOD` are described in detail by Spiegelman & Hertzmark (2005). In brief, the log-binomial estimator is obtained simply by specifying binomial distribution and log link. To compute valid standard errors for other weight functions they add the lines

```
class id;
repeated subject=id/type=ind;
```

which asks for estimates suitable for longitudinal data. With one observation per individual this gives model-robust standard errors without changing the point estimates of relative risk.

SPSS (SPSS Inc, 2001) does not provide routines for Poisson regression, but does provide nonlinear least squares. The nonlinear least squares routines do not compute the model-robust standard error, but the model-based standard error is likely to be adequate unless covariate effects are very strong.

Stata (StataCorp 2005) provides model-robust standard errors for nearly all its regression models. The efficient weights and the Poisson and Gaussian working models are available by

```
glm y x, link(log) eform robust binomial
glm y x, link(log) eform robust poisson
glm y x, link(log) eform robust gaussian
```

The `binomial` option performs unconstrained estimation. The additional optimization options `difficult` and `search` may be helpful in finding the MLE when it is inside the parameter space but relatively close to the boundary. Version 9 of Stata also provides a relative risk regression command, using Wacholder's method of truncation at 0.999 to handle $\mu > 1$.

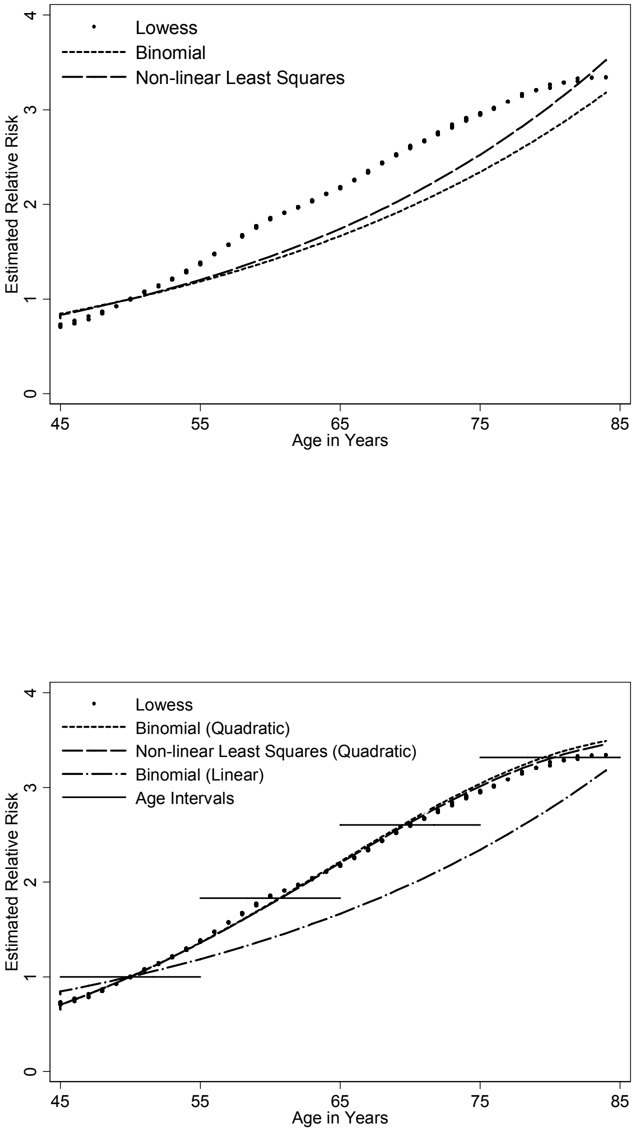

Figure 7: Fitted risks relative to age 50 from models and smoothing: (a) linear term in age (b) linear and quadratic terms in age.

S-PLUS and R: In S-PLUS (Insightful Corporation 2005) and R (R Development Core Team, 2006) the `glm` function takes as an argument a family object. The Poisson regression and non-linear least squares weights are available by `family=poisson()` and `family=gaussian(log)`. In R, `family=binomial(log)` attempts constrained estimation of the mle. It is usually successful if the number of iterations is set high enough. In S-PLUS `family=binomial(log)` is interpreted as an abbreviation for `family=binomial(logit)` and performs *logistic* regression, not relative risk regression.

In addition, R code for all the estimators we discuss, including the adaptive barrier algorithm for the maximum likelihood estimator, is available from the authors.

5 Conclusions

Estimators obtained from solving equation 2 always have the possibility of producing some fitted $\mu > 1$. Constrained optimization, which will often result in some points having very high influence, is the only way to avoid this. There has been ongoing controversy about the practical and theoretical importance of the constraint $\mu \leq 1$.

Deddens and coworkers stress the fact that the MLE cannot lie outside the parameter space, but if the MLE lies on the boundary it is likely that applying the model to new data will lead to fitted probabilities outside [0,1]. They also correctly point out that if model 1 is true, the limiting value of β must lie in the parameter space and that the finite-sample MLE cannot be far outside. This is the motivation for the COPY algorithm and demonstrates why it is successful, especially in simulated data where the model holds exactly.

On the other hand, the only point on which there appears to exist consensus is that the main reason for choosing the relative risk regression model is the greater interpretability of relative risks. If we choose the model primarily based on the contrasts we are interested in, we are surely precluded from assuming that it fits perfectly; we can only ensure that it is a good approximation to the bulk of the data. Even more than usual, Peter McCullagh's aphorism 'models play the same role in statistics as in fashion: as idealizations of reality' applies.

In our experience in large epidemiologic studies it is relatively common for the log relative risk to be linear in an exposure variable over nearly all the range of the data, but to have a few outlying measurements that do not fit the linear model. In these cases the MLE produces a more misleading summary of the relationship than an estimator that allows $\mu > 1$ for a handful of observations.

When the purpose of regression modeling is to estimate a contrast that summarizes the effect of some exposure or intervention there should be a strong preference for expressing the contrast in a form that is easy to communicate. This principle suggests that, other things being equal, we should prefer to estimate relative risks rather than odds ratios when modeling common events. Discussions of relative risk regression have often evaluated proposals solely by their performance in small simulations or in a few examples. We have shown that consistency, and even relative efficiency, of many of these proposals can be evaluated analytically when the estimator is characterized by the estimating equation it solves, rather than the software or algorithm used to obtain it.

Relative risk regression is now a feasible technique for most public health or clinical researchers. While some previously proposed estimation algorithms give inconsistent relative risk estimates or invalid standard errors, there are convenient and widely available techniques that do give valid estimation.

As an interim approach, we recommend using the maximum likelihood estimator when it lies in the interior of the parameter space. If the maximum likelihood estimator is on the boundary of the parameter space this likely indicates the model does not fit perfectly. We would then recommend using an unconstrained estimator, such as nonlinear least squares, and investigating the fit of the model.

If the model fits reasonably well over the bulk of the data, but produces a few outliers with estimated probabilities exceeding one, we would report the nonlinear least squares results. Spiegelman & Hertzmark (2005) give a similar recommendation, defaulting to the Poisson working model rather than nonlinear linear squares, and their software will be helpful for SAS users.

If the fit is generally poor we would investigate transformations of the data, interactions, and other standard approaches to model criticism. Blizzard & Hosmer (2006) found that formal goodness-of-fit tests for the log-binomial model had low power, but other model diagnostics for generalized linear models should still be useful.

We hope that relative risk regression commands will be implemented in standard software, making these circuitous approaches unnecessary. The fact that nearly 20 years after Wacholder (1986) we are still seeing new proposals for tricking software into fitting the relative risk regression model implies either that researchers do not care about the difference between odds ratios and relative risks or that the statistical software industry is not listening to them.

Acknowledgements

This research was supported in part by contract N01-HC-95159 from the National Heart, Lung, and Blood Institute. The example data are from the MESA study, which is supported by contracts N01-HC-95159 through N01-HC-95166 and N01-HC-95169 from the National Heart, Lung, and Blood Institute. The authors thank the other investigators, the staff, and the participants of the MESA study for their valuable contributions. A full list of participating MESA investigators and institutions can be found at http://www.mesa-nhlbi.org.

References:

Barros AJ, Hirakata VN. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. BMC Med Res Methodol. 2003 Oct 20;3(1):21.

Baumgarten M, Seliske P, Goldberg MS. Warning regarding the use of GLIM macros for the estimation of risk ratios. Am J Epidemiol. 1989 Nov;130(5):1065.

Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Rouz AV, Folsom AR, Greenland P, Jacobs Jr DR, Kronmal R, Liu K, Nelson JC, O'leary D, Saad MF, Shea S, Szklo M, Tracy RP. The Multi-Ethnic Study of Atherosclerosis (MESA): objectives and design. Am J Epidemiol. 2002;156(9):87181.

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

Boyd S, Vandenberghe L (2004) *Convex Optimization* Cambridge: Cambridge University Press.

Carter RE, Lipsitz SR, Tilley BC. (2005) Quasi-likelihood estimation for relative risk regression models.

Carter RE, Lipsitz SR. (2006) Sampling weighted relative risk regression. International Biometric Society (ENAR) Spring Meeting.

Davies HT, Crombie IK, Tavakoli M.(1998) When can odds ratios mislead? BMJ. 316(7136):989-91.

Deddens & Petersen. (2003) Estimation of prevalence ratios when PROC GENMOD does not converge. Proceedings of the 28th Annual SAS Users Group International Conference, paper 270—28. Cary NC, SAS Institute Inc (`http://www2.sas.com/proceedings/sugi28270-28.pdf`).

Deddens & Petersen. Re: "Estimating the relative risk in cohort studies and clinical trials of common outcomes" Am J Epidemiol 2004; 159:213-214.

Efron B, (1977) The efficiency of Cox's likelihood function for censored data. Journal of the American Statistical Association 72, 557-565.

Greenland S. Model-based Estimation of Relative Risks and Other Epidemiologic Measures in Studies of Common Outcomes and in Case-Control Studies. American Journal of Epidemiology Volume: 160 Number: 4 Page: 301 – 305

Harrell FE (2001). *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis.* New York:Springer

Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning.* New York: Springer.

Hertz-Picciotto I, Rockhill B.(1997) Validity and efficiency of approximation methods for tied survival times in Cox regression. Biometrics. Sep;53(3):1151-6.

Insightful Corporation (1988-2005). S-PLUS version 7.0 for Linux (computer software). Seattle, Washington: Insightful Corporation.

Katz, KA. (2006) The (Relative) Risks of Using Odds Ratios. Arch Dermatol. 2006;142:761-764.

Lange K. (1994) An adaptive barrier method for convex programming. *Methods Applications Analysis* 1:392–402.

Lee J (2004) Odds ratio or relative risk for cross-sectional data? [letter]. International Journal of Epidemiology, 23:201-203.

Lee J, Chia KS (1993): Estimation of prevalence rate ratios for cross sectional data: an example in occupational epidemiology [letter]. British Journal of Industrial Medicine 50:861-862.

Liberman AM. (2005).How Much More Likely? The Implications of Odds Ratios for Probabilities. American Journal of Evaluation, Vol. 26, No. 2, 253-266

Lumley T, Kronmal RA (2006) Algorithms and estimators for relative risk regression. Technical report No FIXME. Department of Biostatistics, University of Washington

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, London.

Ma S, Wong CM. Estimation of prevalence proportion rates. (Letter). Int J Epidemiol 1999;28:175.

McNutt L-A, Wu C, Xue X, Hafner JP (2003) Estimating relative risk in cohort studies and clinical trials of common events. American Journal of Epidemiology. 157: 940-943.

R Development Core Team (2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Robbins AS, Chao SY, Fonseca VP.What's the relative risk? A method to directly estimate risk ratios in cohort studies of common outcomes. Ann Epidemiol. 2002 Oct;12(7):452-4.

SAS Institute Inc., SAS 9.1.3 Help and Documentation, Cary, NC: SAS Institute Inc., 2000-2004.

Schulman KA, Berlin JA, Harless W, et al. (1999) The effect of race and sex on physicians' recommendations for cardiac catheterization. N Engl J Med 1999;340:618-626.

Schwartz LM, Woloshin S, Welch HG. (1999) Misunderstandings about the effects of race and sex on physicians' referrals for cardiac catheterization.1: N Engl J Med. 1999 Jul 22;341(4):279-83.

Sinclair JC, Bracken MB (1994). Clinically useful measures of effect in binary analyses of randomized trials. J Clin Epidemiol. Aug;47(8):881-9.

Skov T, Deddens J, Petersen MR, Endahl L. Prevalence proportion ratios: estimation and hypothesis testing. Int J Epidemiol. 1998 Feb;27(1):91-5.

Spiegelman D, Hertzmark E. (2005) Easy SAS Calculations for Risk or Prevalence Ratios and Differences. American Journal of Epidemiology 162 (3):199-200.

SPSS Inc (2001) SPSS for Windows, Rel. 11.0.1. Chicago: SPSS Inc.

StataCorp. (2005). Stata Statistical Software: Release 9. College Station, TX: StataCorp LP.

Traissac P, Martin-Prevel Y, Delpeuch F, Maire B: Régression logistique vs autres modèles linéaires généralisés pour l'estimation de rapports de prévalences. Rev Epidemiol Sante Publique 1999, 47:593-604.

Wacholder S. (1986) Binomial regression in GLIM: estimating risk ratios and risk differences. Am J Epidemiol. 1986 Jan;123(1):174-84.

Wedderburn, RWM (1976). On the Existence and Uniqueness of Maximum Likelihood Estimates for Certain Generalized Linear Models. Biometrika 63, 27–32.

Zocchetti C, Consonni D, Bertazzi PA. Relationship between prevalence rate ratios and odds ratios in cross-sectional studies. Int J Epidemiol. 1997 Feb;26(1):220-3.

Zocchetti C, Consonni D, Bertazzi PA. Estimation of prevalence rate ratios from cross-sectional data. Int J Epidemiol. 1995 Oct;24(5):1064-7.

Zou G (2004) A modified Poisson regression approach to prospective studies with binary data. American Journal of Epidemiology. 154: 702-706