SORT 27 (1) January-June 2003, 79-94

Indirect inference for survival data

Bruce W. Turnbull* and Wenxin Jiang

Cornell University

Abstract

In this paper we describe the so-called "indirect" method of inference, originally developed from the econometric literature, and apply it to survival analyses of two data sets with repeated events. This method is often more convenient computationally than maximum likelihood estimation when handling such model complexities as random effects and measurement error, for example; and it can also serve as a basis for robust inference with less stringent assumptions on the data generating mechanism. The first data set concerns recurrence times of mammary tumors in rats and is modeled using a Poisson process model with covariates and frailties. The second data set involves times of recurrences of skin tumors in individual patients in a clinical trial. The methodology is applied in both parametric and semi-parametric regression analyses to accommodate random effects and covariate measurement error.

MSC: 62N01, 62N02, 62P10

Keywords: Estimating equations, frailty, hazard rate regression, indirect inference, measurement error, naive estimators, overdispersion, quasi-likelihood, random effects, robustness

1 Introduction

Methods of *indirect inference* (Gourieroux, Monfort and Renault, 1993) have been developed and used in the field of econometrics where they have proved valuable for parameter estimation in highly complex models. This paper recasts the basic technique in a likelihood-flavoured approach and illustrates some applications in biostatistics, in particular for survival and repeated events data.

We begin by illustrating the steps involved in the indirect method in the following simple pedagogic example.

* *Address for correspondence:* B. W. Turnbull. School of Operations Research and Department of Statistical Science, 227 Rhodes Hall, Cornell University, Ithaca, NY 14853-3801, USA. turnbull@orie.cornell.edu

Received: October 2002
Accepted: January 2003

Example 1: *exponential survival with censoring.* Consider lifetimes $\{T_1,\ldots,T_n\}$, which are independent and identically distributed (i.i.d.) according to an exponential distribution with mean θ. The data are subject to Type I single censoring after fixed time c. Thus the observed data are $\{Y_1,\ldots,Y_n\}$, where $Y_i = \min(T_i,c)$, $(i = 1,\ldots,n)$. We consider indirect inference based on the intermediate statistic $\hat{s} = \overline{Y}$. This choice can be considered either as the basis for a method of moments estimator or as the MLE (maximum likelihood estimator) for a misspecified model M′ in which the presence of censoring has been ignored. The naive estimator $\overline{Y}$ in fact consistently estimates not θ but the "naive" or "auxiliary" parameter

$$s(\theta) = \theta\,[1-\exp(-c/\theta)], \tag{1}$$

the expectation of $\overline{Y}$. The equation (1) is an example of what may be termed as a "bridge relation" (Jiang and Turnbull, 2001) or a "binding relation" (Gourieroux, *et al.*, 1993). We can see the obvious effect of the misspecification, namely that $\hat{s}$ underestimates θ. However a consistent estimator $\hat{\theta}$ of θ as $n \longrightarrow \infty$ can be obtained by solving (1) for θ with $s(\theta)$ replaced by $\hat{s} = \overline{Y}$. That is, $\hat{\theta} = s^{-1}(\hat{s})$. (Note that $s(\cdot)$ is strictly increasing on $\Re^+$ and thus invertible). We note also that $\hat{\theta}$ is not the MLE of θ which is $n\overline{Y}/[\sum_{i=1}^n I(Y_i < c)]$.)

More generally, a consistent estimator can be constructed based on an intermediate statistic $\hat{s}$ that does not need to have the interpretation of a 'naive' estimator. For example, above we could have chosen perhaps $\hat{s} = \overline{Y^2} = n^{-1}\sum_{i=1}^n Y_i^2$ so that $\hat{\theta} = s^{-1}(\hat{s})$ where s^{-1} is the inverse function of $s(\theta) \equiv E(\overline{Y^2}|\theta)$. In fact,the dimension of $\hat{s}$ can be greater than that of θ — e.g., we could take $\hat{s} = (\overline{Y},\overline{Y^2})^T$ in the above example. Now a consistent estimator $\hat{\theta}$ of θ can be found by using weighted least squares,

$$\hat{\theta} = \arg\min_{\theta}\{\hat{s} - s(\theta)\}^T A\{\hat{s} - s(\theta)\},$$

where an optimal choice of A is the inverse of the estimated variance matrix of $\hat{s}$, as we will discuss later. This is the principal idea of indirect inference– statistical inference of θ based on an indirect data "summary" $\hat{s}$. The choice of $\hat{s}$ is not unique, but in most applications there will natural one to use as we shall see.

We will term $\hat{\theta}$ as the "indirect MLE", since it can be viewed as the MLE using an approximate likelihood based on the indirect data summary $\hat{s}$. We will also see how to obtain the standard error for $\hat{\theta}$.

2 Indirect inference

In general, the indirect MLE has properties similar to those of the usual MLE: consistency, asymptotic normality, and certain efficiency properties. In addition, chi-squared goodness-of-fit tests can be based on the indirect likelihood.

Advantages of the indirect method include ease of computation; robustness; and informativeness on the effect of model misspecification. We will summarize the framework of indirect inference below.

2.1 The basic approach

Suppose we have a data set consisting of n independent units. The essential ingredients of the indirect approach, when reformulated in a likelihood-flavoured treatment, are as follows.

- There is a hypothesized true model M for data generation, with distribution $P^{(\theta)}$ which depends on an unknown parameter θ of interest which is of dimension p.
- One first computes an *intermediate* or *auxiliary* statistic $\hat{s}$ of dimension $q \geq p$, which is asymptotically normal with mean $s(\theta)$, say, under model M.
- An *indirect likelihood* $L(\theta|\hat{s})$ is then constructed based on the normal approximation, so that, apart from an additive constant,

$$-2\log L(\theta|\hat{s}) = \{\hat{s} - s(\theta)\}^T v^{-1} \{\hat{s} - s(\theta)\} \; = \; H(\theta), \text{ say}, \tag{2}$$

  where v is a consistent estimate of the asymptotic variance $\widehat{var}(\hat{s})$. A typical choice might be the 'robust' or 'sandwich formula', when $\hat{s}$ solves an estimating equation (see e.g. Carroll, Ruppert and Stefanski 1995, Section A.3).
- This indirect likelihood is then maximized to generate an indirect maximum likelihood estimate (*indirect MLE*) or *adjusted* estimate $\hat{\theta}(\hat{s})$ for θ. In the case when the dimension (q) of the intermediate statistic equals that (p) of the parameter θ and $s(\theta)$ is invertible, it can be seen from (2) that maximization of the indirect likelihood is equivalent to solving the "bridge" or "binding" equation $s(\theta) = \hat{s}$ for θ, because then (2) can be made zero.

In the "indirect" analysis of the pedagogic example of Section 1, M is the i.i.d. exponential model with censoring: $Y_i = \min(T_i, c)$ and $P^{(\theta)}(T_i \leq t) = 1 - e^{-t/\theta}$, for $t \in [0, \infty)$, $i = 1, \ldots, n$. In the initial approach, the intermediate statistic was $\hat{s} = n^{-1} \sum_{i=1}^n Y_i$, which is asymptotically normal as $n - \longrightarrow \infty$ by the central limit theorem, with asymptotic mean $s(\theta) = \theta\,[1 - \exp(-c/\theta)]$. The indirect likelihood $L(\theta|\hat{s})$ is given by

$$-2\log L(\theta|\hat{s}) = \{n^{-1} \sum_{i=1}^n Y_i - s(\theta)\}^T v^{-1} \{n^{-1} \sum_{i=1}^n Y_i - s(\theta)\}$$

where one can substitute the robust estimate $\hat{v} = \{n(n-1)\}^{-1} \sum_{i=1}^n (Y_i - \overline{Y})^2$ for the asymptotic variance $v = var(\hat{s})$. Finally the adjusted estimate (or indirect MLE) is $\hat{\theta} = s^{-1}(\hat{s})$.

In summary, in this indirect approach, the data are first summarized by the intermediate statistic $\hat{s}$. Its asymptotic mean s is referred to as the *auxiliary parameter*. The auxiliary parameter is related to the original parameter by a relation $s = s(\theta)$, termed the *bridge relation* or *binding function*.

The starting point is the choice of an intermediate statistic $\hat{s}$. This can be chosen as some set of sample moments, or the solution of some estimating equations, or the MLE based on some convenient model M′, say, termed the *auxiliary* (or *naive*) model. If the last, then the model M′ is a simpler but misspecified or partially misspecified model. As stated previously, the choice of an intermediate statistic $\hat{s}$ is not necessarily unique; however in any given situation there is often a natural one to use.

2.2 Intermediate statistics arising from estimating equations

Most intermediate statistics can be defined implicitly as a solution, $s = \hat{s}$, of a (q-dimensional) estimating equation of the form $G(\mathbf{W}, s) = 0$, say. [Clearly this includes any statistic $\hat{s} = \hat{s}(\mathbf{W})$ that has an explicit expression as a special case, by taking $G = s - \hat{s}(\mathbf{W})$.] The estimating equation could be the normal equation from a least-squares analysis, or the score equation based on some likelihood function.

In such situations there is a parallel formulation of indirect inference in 'implicit form'. For instance, one can state the 'bridge relation' $s(\theta)$ implicitly as $F(\theta, s) = 0$ where $F(\theta, s) \equiv E_{\mathbf{W}|\theta} G(\mathbf{W}, s)$, which is the limiting version of the estimating equation $G(\mathbf{W}, \hat{s}) = 0$. Correspondingly, in the definition of the indirect likelihood L, H can be (asymptotically) equivalently defined by $H(\theta, \hat{s}) = F(\theta, \hat{s})^T v^{-1} F(\theta, \hat{s})$. Here v is (a sample estimate of) the avar of $F(\theta, \hat{s})$, which can be evaluated by the delta method (e.g. Bickel and Doksum (2001), Sec. 5.3.2), and found to be the same as $var(G)$ evaluated at $s = s(\theta)$ (the auxiliary parameter). Then we define the *adjusted estimator* (or the *indirect MLE*) $\hat{\theta}$ to be the maximizer of L, or the minimizer of H.

2.3 Properties of indirect MLE

In general, the indirect MLE has a set of properties analogous to those of the usual MLE. These include, under appropriate regularity conditions:

(i) (Indirect Score Function). The asymptotic mean and variance of the indirect likelihood score function satisfy the usual relations $E(\nabla_\theta \log L) = 0$ and $var(\nabla_\theta \log L) + E(\nabla_\theta^2 \log L) = 0$.

(ii) (Asymptotic Normality). The adjusted estimator $\hat{\theta}$ is asymptotically normal (AN) with mean θ, and with asymptotic variance (avar) estimated by $-(\nabla_\theta^2 \log L)^{-1}$ or $2(\nabla_\theta^2 H)^{-1}$ where consistent estimates are substituted for parameter values.

(iii) (Tests). Likelihood-ratio statistics based on the indirect likelihood for testing simple and composite null hypotheses have the usual asymptotic χ^2 distributions.
(iv) (Efficient use of indirect data). The adjusted estimator has smallest avar among all consistent AN estimators $f(\hat{s})$ of θ, which are constructed from the naive estimator $\hat{s}$ by continuously differentiable mappings f.

These results can be found in the references cited in Section 2.5, and are summarized in Jiang and Turnbull (2001, Proposition 1).

When different intermediate statistics are used, the asymptotic efficiency can be different. In general the indirect MLE is not as efficient as the MLE based on the true model M; although there are situations that can be identified where the efficiency will be high, as in the example of Section 3.1.

In a special case when the dimension of the intermediate statistic (q) equals that (p) of the parameter θ, and $s(\cdot)$ is a diffeomorphism on the parameter space Θ of θ, maximization of L is equivalent to the bias correction $\hat{\theta} = s^{-1}(\hat{s})$ (from solving $F(\theta, \hat{s}) = 0$), which is AN and consistent for θ. See, e.g., Kuk (1995), Turnbull *et al.* (1997) and Jiang *et al.* (1999) for biostatistical applications.

When $q < p$, there are more unknown true parameters than 'naive parameters'. In this case the bridge relation is many-to-one and does not in general permit the construction of adjusted estimates. It is mainly of interest for investigating the effects of misspecification when the naive estimators are constructed under misspecified models. However, in such situations it may be possible to construct consistent estimates for a subset of true parameters, which may be of interest. In other situations, some components of the higher-dimensional true parameter are known or can be estimated from other outside data sources. This enables the other components to be consistently estimated by inverting the bridge relation. Examples of this kind arising from errors-in-variables regression models are given in Sections 3.2 and 3.3.

2.4 Why consider the indirect method?

This indirect approach offers the following advantages:

1. *Ease of computation.* The indirect method is typically computationally simpler and more convenient. For example, when $\hat{s}$ is based on some simplified model M′, it can often be computed with available standard computer software.
2. *Informativeness on the effect of model misspecification.* When $\hat{s}$ is a 'naive estimate' obtained from a naive model M′ neglecting certain model complexities, the approach is very informative on the effect of model misspecification — the bridge relation $s = s(\theta)$ provides a dynamic correspondence between M′ and M. For example, in errors-in-variable regression, such a relation is sometimes termed an

'attenuation relation' (see e.g., Carroll, Ruppert and Stefanski 1995, Chapter 2), and tells how regression coefficients can be underestimated when neglecting the measurement error in a predictor.

3. *Robustness.* The validity of the inference based on an intermediate statistic essentially relies on the correct specification of its asymptotic mean. This is typically a less demanding assumption than the correct specification of a full probability model, which would be generally needed for a direct likelihood inference to be valid. Therefore inferences based on the adjusted estimate $\hat{\theta}$ can remain valid despite some departure of the data generation mechanism from the hypothesized true model M.

2.5 Bibliography and notes

The above very brief exposition of the indirect method of inference represents a summary of results that have appeared in the econometric and statistical literature in varying forms and generality and tailored for various applications. Examples include: the generalized method of moments (GMM: Hansen 1982); the method of linear forms and minimum χ^2 (Ferguson 1958); the regular best asymptotic normal estimates that are functions of sample averages (Chiang 1956, Theorem 3); simulated method of moments and indirect inference [McFadden (1989), Pakes and Pollard (1989), Gourieroux *et al.* (1993), Gallant and Tauchen (1996, 1999) Gallant and Long (1997)]. Newey and McFadden (1994, Chapters 6 and 8) discuss two-stage parametric and nonparametric estimation in the GMM context, where some 'nuisance' parameter, possibly infinite dimensional, is estimated from a preliminary consistent method.

Applications of GMM in the settings of generalized estimating equations from biostatistics are discussed in Qu, Lindsay and Li (2000). McCullagh and Nelder (1989, p. 341), as referred to by Qin and Lawless (1994, p. 315), consider optimal linear combination of estimating equations, as is traditionally done in GMM literature. Qin and Lawless (1994) also provide an alternative but asymptotically equivalent way of combining estimating equations using empirical likelihood.

The theory of estimators obtained from misspecified likelihoods goes back at least as far as Cox (1962), Berk (1966) and Huber (1967) and is summarized in the comprehensive monograph by White (1994). The use of $\hat{s}$ (based on an auxiliary model M′) in indirect inference about θ (under model M) appears recently in the field of econometrics to treat complex time series and dynamic models, see, e.g., Gourieroux *et al.* (1993) and Gallant and Tauchen (1996, 1999); as well as in the field of biostatistics to treat regression models with random effects and measurement error, see e.g., Kuk (1995), Turnbull, Jiang and Clark (1997), and Jiang *et al.* (1999). This bibliography is far from exhaustive. A thorough review and synthesis of the methods of indirect inference are given in Jiang and Turnbull (2001).

3 Three applications with survival data

In this section we discuss three applications with recurrent event data which use models of increasing order of complexity:

3.1 A Poisson process regression model with random effects ("frailties" or "unexplained heterogeneity");
3.2 A Poisson process regression model with random effects and covariate measurement error;
3.3 A semi-parametric intensity rate regression model with random effects and measurement error.

The first example uses mammary tumor recurrence times from a rodent carcinogenicity experiment. The remaining two examples use data on skin cancer recurrences in the Nutritional Prevention of Cancer (NPC) trial — a long-term randomized clinical trial for cancer prevention (Clark *et al.* 1996).

3.1 Animal carcinogenicity data: multiple times to tumor

Gail, Santner and Brown (1980, Table 1) present data on multiple mammary tumor incidence times from an experiment conducted by Thompson *et al.* (1978). Forty-eight female rats which remained tumor-free after sixty days of pre-treatment of a prevention drug (retinyl acetate) were randomized with equal probability into two groups. In Group 1 they continued to receive treatment ($Z = 1$), in Group 2 they received placebo ($Z = 0$). All rats were followed for an additional 122 days and the time of any newly diagnosed mammary tumor was recorded. The numbers of tumors diagnosed in individual rats ranged from 0 to 13. The objective of the study was to estimate the effect of the preventive treatment (Z) on tumor recurrence.

Suppose we consider a model in which the tumors occur over time in a given subject (rat) according to a Poisson process with a constant intensity rate which depends on treatment Z, a fixed effect, and on subject, a random effect. If we define Y to be the number of tumors diagnosed in a particular rat during the 122 day followup time, the model M specifies that, given Z and ϵ, Y is Poisson distributed with mean $\exp(\alpha + Z\beta + \epsilon)$. Here the assigned treatment Z is observed, but ϵ represents an unobserved random effect modeled as normally distributed with zero mean and constant variance σ^2, independent of Z. This random effect or "unexplained heterogeneity" could be considered to be caused by omitted covariates. We observe $n = 48$ i.i.d. pairs $W_i = (Y_i, Z_i), i = 1, \ldots, n$. The likelihood for the observed data involves integration over ϵ and is difficult to compute. (However it is possible – see below.) Instead, we start by taking the indirect approach with an auxiliary statistic $\hat{s} = (\hat{a}, \hat{b}, \hat{t}^2)^T$, where

$(\hat{a},\hat{b})$ are the regression coefficient estimates maximizing a naive log-likelihood $R = \sum_1^n\{Y_i(a+Z_ib) - e^{a+Z_ib}\}$, and $\hat{t}^2 = n^{-1}\sum_{i=1}^n Y_i^2$ is the second sample moment. Here the auxiliary parameter is $s = \text{plim}(\hat{a},\hat{b},\hat{t}^2)^T$, whereas the true parameter to be estimated is $\theta = (\alpha,\beta,\sigma^2)^T$. The use of the naive log-likelihood R corresponds to a simplified model M′ in which the presence of the random effect ϵ is neglected. The second sample moment is included in the intermediate statistic to provide information for estimation of the variance parameter. Therefore $\hat{s}$ is solved from the estimating equation $G(\mathbf{W},s) = 0$, where (formally) $G = (n^{-1}\partial_a R, n^{-1}\partial_b R, \hat{t}^2 - t^2)^T$, i.e.

$$G = n^{-1}\sum_{i=1}^n (Y_i - e^{a+Z_ib},\ Z_i(Y_i - e^{a+Z_ib}),\ Y_i^2 - t^2)^T \quad = n^{-1}\sum_{i=1}^n g_i,\ \text{say.}$$

The solution $\hat{s} = (\hat{a},\hat{b},\hat{t}^2)^T$ can be computed easily. For the rat carcinogenicity data we obtain the auxiliary estimates $\hat{a} = 1.7984$; $\hat{b} = -0.8230$; $\hat{t}^2 = 31.875$. The asymptotic variance $var(\hat{s})$ can be estimated by the sandwich formula (see e.g. Carroll, Ruppert and Stefanski 1995, Section A.3)

$$v = (\nabla_s G)^{-1}\widehat{var}(G)(\nabla_s G)^{-T}|_{s=\hat{s}}$$

where $\widehat{var}(G) = n^{-2}\sum_{i=1}^n g_i g_i^T|_{s=\hat{s}}$, $\nabla_s G$ is a 3×3 matrix with elements $(\nabla_s G)_{jk} = \partial_{s_k} G_j$, $j,k = 1,2,3$, and $A^{-T} = (A^{-1})^T$ for a generic matrix A.

The indirect likelihood $L(\theta|\hat{s})$, up to an additive constant, satisfies

$$-2\log L(\theta|\hat{s}) = \{\hat{s} - s(\theta)\}^T v^{-1}\{\hat{s} - s(\theta)\},$$

where $s(\theta)$ is the asymptotic mean or large sample almost sure limit of $\hat{s}$. Since $\hat{s}$ solves the estimating equation $G = 0$, its limit is the solution of the limiting estimating equation $F(\theta,s) = E_{\mathbf{W}|\theta}G(\mathbf{W},s) = 0$, which can be explicitly solved to obtain $s = s(\theta)$. This yields the bridge equation:

$$\begin{aligned} s \ &= \ \text{plim}(\hat{a},\hat{b},\hat{t}^2)^T \\ &= \ s(\theta) = \left(\alpha + \sigma^2/2,\ \beta,\ \frac{1}{2}(1+e^{\beta})e^{\alpha+\frac{1}{2}\sigma^2} + \frac{1}{2}(1+e^{2\beta})e^{2(\alpha+\sigma^2)}\right)^T. \end{aligned}$$

Because dim(s) = dim(θ) = 3 and $s(\theta)$ is a smooth invertible mapping, the indirect MLE $\hat{\theta} = \arg\max_\theta L(\theta|\hat{s})$ can be obtained by solving $\hat{s} = s(\theta)$, which gives the adjusted estimates $\hat{\theta} = (\hat{\alpha},\hat{\beta},\hat{\sigma}^2) = s^{-1}(\hat{s})$. Thus $\hat{\beta} = \hat{b}$, and $\hat{\alpha} = \hat{a} - \hat{\sigma}^2/2$ where $\hat{\sigma}^2 = \log\left\{\frac{2\hat{t}^2 - e^{\hat{a}}(1+e^{\hat{b}})}{e^{2\hat{a}}(1+e^{2\hat{b}})}\right\}$. For the rat data, this leads to adjusted estimates $\hat{\alpha} = 1.6808(0.1589)$; $\hat{\beta} = -0.8230(0.1968)$; $\hat{\sigma} = 0.4850(0.1274)$.

The estimated standard errors shown in parentheses are obtained using the delta method formula: $\widehat{var}(\hat{\theta}) = (\nabla_\theta s)^{-1}v(\nabla_\theta s)^{-T}|_{\theta=\hat{\theta}}$, and then taking the square roots of the 3 diagonal elements of this matrix. It is noted that this delta method expression is equivalent to deriving the variance by

$$\{-\nabla_\theta^2 \log L(\theta|\hat{s})\}^{-1}|_{\theta=\hat{\theta}}$$

based on the 'indirect likelihood Fisher information', where ∇_θ^2 represents the Hessian. This follows because the jkth element of the Hessian is, for $j,k = 1,2,3$,

$$\begin{aligned}
\{-\nabla_\theta^2 \log L(\theta|\hat{s})\}_{jk}|_{\theta=\hat{\theta}} &= -\partial_{\theta_j}\partial_{\theta_k} \log L(\theta|\hat{s})|_{\theta=\hat{\theta}} \\
&= (\partial_{\theta_j} s^T) v^{-1} (\partial_{\theta_k} s)|_{\theta=\hat{\theta}} - (\partial_{\theta_j}\partial_{\theta_k} s^T) v^{-1} (\hat{s}-s)|_{\theta=\hat{\theta}} \\
&= (\partial_{\theta_j} s^T) v^{-1} (\partial_{\theta_k} s)|_{\theta=\hat{\theta}} + 0 \\
&= \{\widehat{var}(\hat{\theta})^{-1}\}_{jk}.
\end{aligned}$$

If we wish to obtain the MLE of $\theta = (\alpha,\beta,\sigma^2)$ based on model M, then it can be found by a somewhat tedious iterative numerical maximization of the true likelihood which involves numerical integration over the distribution of ϵ. These estimates are: $\hat{\alpha}_{ML} = 1.6717$ (0.1560); $\hat{\beta}_{ML} = -0.8125$ (0.2078); $\hat{\sigma}_{ML} = 0.5034$ (0.0859). For the MLEs, the estimated standard errors are based on the inverse of the Fisher information matrix, evaluated at the corresponding estimate values.

The estimated standard errors suggest that the efficiency of indirect estimation of the treatment effect parameter β is high here in this example. Related results (Cox, 1983; Jiang *et al.*, 1999) show that such high efficiency is achievable if the follow-up times are about the same across different subjects (which is true here), or if the overdispersion is small. Also it should be noted that the adjusted estimator $\hat{\beta}$ is robust, in the sense that it remains consistent, essentially as long as the mean function $E(Y|Z,\epsilon)$ is correctly specified and ϵ and Z are independent. (Its standard error estimate from the sandwich formula is also model-independent and robust.) In particular, the consistency property does not depend on the specification of a complete probability model, namely that Y is Poisson and ϵ is normal. Thus the indirect estimator enjoys a robustness advantage over the MLE.

The indirect approach, although formulated from the different perspective of using naive model plus method of moments, is intimately related to the work of Breslow (1990) based on quasi-likelihood and method of moments. Breslow used a different linear combination of Y_i's based on quasi-likelihood (Wedderburn, 1974; McCullagh and Nelder, 1989), which enjoy general efficiency properties among linear estimating equations. However, (i) our approach can be interpreted as basing inference on the simple moments $n^{-1}\sum Y_i$, $n^{-1}\sum Z_i Y_i$ and $n^{-1}\sum Y_i^2$ (which can be easily seen from the estimating equation $G = 0$), and (ii) our approach shows clearly, by the use of bridge relations, the sensitivity and robustness of parameter estimates to the omission of over-dispersion in modeling. Also note that here we used a log-normal distribution to model the random effects and the variance parameter also enters the mean model (unconditional on ϵ), whereas Breslow (1990) focused on the examples such as ones with gamma multiplicative random effects in which the mean model does not change.

For the only comparable parameter β (the treatment effect), the Breslow method (from his equations (1), (2) and (7)) gives exactly the same answer as our adjusted analysis: $\hat{\beta}_{\text{Breslow}} = -0.8230(0.1968)$. This is because, for this special two-group design, both methods essentially use the log(frequency ratio) to estimate the treatment effect.

3.2 Skin cancer recurrence data from the NPC trial: parametric modeling

Clark *et al.* (1996) have described the results of the "Nutritional Prevention of Cancer" (NPC) trial. This trial, begun in 1983, studied the long-term safety and efficacy of a daily 200μg nutritional supplement of selenium (Se) for the prevention of cancer. It was a double-blind, placebo-controlled randomized clinical trial with $n = 1312$ patients accrued and followed for up to about ten years. Here we shall consider a particular primary endpoint — namely squamous cell carcinoma (SCC) of the skin. The results for this endpoint are of particular interest because Clark *et al.* (1996) found a negative (but not statistically significant, P = 0.15) effect of selenium (Se) supplementation. This was opposite to previous expectations, and contrasted sharply with findings of highly significant positive benefits of the selenium supplementation in preventing a number of other types of cancers. However in their analysis, Clark *et al.* used only data on the time to *first* occurrence of SCC in each subject and employed a Cox model that ignored patient heterogeneity (i.e. that assumed a common baseline hazard) and ignored that some explanatory covariates were measured with error.

We consider the recurrences of SCC over time, measured from date of randomization, for patients $i = 1,...,n$ as n i.i.d. discrete point processes $\{Y_i(t)\}$. Here $Y_i(t)$ is the observed number of recurrences for patient i on day t (usually zero or one). Time t is measured in days on a discrete time scale $t = 1,\ldots,K$, where $K = 4618$ days, the longest followup time. The indicator variable $H_i(t)$ is one if patient i is still on study ("at risk") on day t and zero otherwise. For illustration purposes, we will consider only two explanatory variables, namely treatment assignment indicator a and baseline Se level x. The latter is an important predictor, measured prior to randomization in each patient, but is contaminated with measurement error so that the observed value is recorded as z not x. In the parametric approach, we postulate an independent Poisson process model with constant baseline mean event rate as the underlying data generating mechanism: for $i = 1,2,\ldots,n; t = 1,...,K$, $Y_i(t)$ are independent Poisson random variables with mean $E[Y_i(t)] = H_i(t)\psi_i\lambda\exp(a_i\gamma + x_i\beta)$.

Here the $\{\psi_i\}$ represent subject-specific random effects or "frailties", which modulate the constant baseline mean rate λ. In this framework, the sufficient statistics is $Y_i \equiv \sum_{t=1}^{K} H_i(t)Y_i(t)$ which follows Poisson distribution with mean $\tau_i\psi_i\lambda\exp(a_i\gamma + x_i\beta)$, with $\tau_i = \sum_{t=1}^{K} H_i(t)$ being the length of follow-up for patient i.

When a conjugate distribution Gamma(mean 1, variance ν) for ψ_i is used, the integration over the unobservable ψ_i can be carried out analytically, so that unconditional

on ψ_i:

> Y_i follows a negative binomial distribution with mean μ_i and variance $\mu_i + \nu\mu_i^2$,

where $\mu_i = \tau_i \lambda \exp(a_i\gamma + x_i\beta)$. We refer to this as our base model "M(para)".

In Turnbull *et al.* (1997), the intended x_i is the long-term average of the baseline Se level (in log-scale), which is subject to measurement error and temporal fluctuation. An error-contaminated version $z_i = x_i + u_i$ is observed, where x_i and u_i are assumed to be independent normal with zero means (after centering) and respective variances σ_x^2 and σ_u^2. A naive analysis ignoring measurement error would involve a negative binomial regression of Y_i on (a_i, z_i), instead of on (a_i, x_i). The auxiliary model is then:

> M′(para): Y_i is negative binomial with mean q_i and variance $q_i + vq_i^2$,

where $q_i = \tau_i m \exp(a_i g + x_i b)$ and $s = (g, b, m, v)$ is the naive / auxiliary parameter corresponding to the parameter $\theta = (\gamma, \beta, \lambda, \nu)$ used in M(para).

Table 1: *Statistical analyses for several models of NPC trial SCC data.*

Model	Treatment estimate	(s.e.)	Baseline Se estimate	(s.e.)
1) Parametric: Constant Intensity				
a) Naive (Model M′ (para))	$\hat{g}$=0.122	(0.059)	$\hat{b}$=−0.725	(0.145)
b) Adjusted (Model M (para))	$\hat{\gamma}$=0.122	(0.125)	$\hat{\beta}$=−2.181	(0.963)
2) Semi-parametric				
a) Naive (Model M′ (semi-par))	$\hat{g}$=0.117	(0.059)	$\hat{b}$=−0.690	(0.146)
b) Adjusted (Model M (semi-par))	$\hat{\gamma}$=0.117	(0.125)	$\hat{\beta}$=−2.076	(0.963)

Such a naive analysis based on M′(para) was run and the resulting estimator $\hat{s}$ for s forms our intermediate statistic. Computer packages for negative binomial regression can be used for this task, e.g. the procedure *nbreg* in STATA 5.0 (StataCorp 1997). The bridge relation $s(\theta)$ as a consistent limit of $\hat{s}$ when the true parameter is θ was shown (Turnbull *et al.*, 1997) to include an implicit equation for solving for v, as well as the following explicit formulae:

$$g = \gamma, \quad b = \pi\beta \quad \text{and} \quad m = \lambda \exp(0.5\beta^2 \sigma_{x|z}^2),$$

where $\pi = \sigma_x^2/(\sigma_x^2 + \sigma_u^2)$ is the attenuation coefficient, and $\sigma_{x|z}^2 = \pi\sigma_u^2$, which were obtained from an internal validation study (Turnbull *et al.*, 1997).

This bridge relation is then inverted to obtain a consistent adjusted estimator $\hat{\theta}$ for the true parameter $\theta = (\gamma, \beta, \lambda, \nu)$. Robust sandwich variance estimates were used to obtain standard errors. Details of the calculations are given by Turnbull *et al.*, (1997). Inference

on the regression parameters of interest, (γ,β), are summarized in lines 1a and 1b of Table 1 and compared with the results from the semi-parametric approach described next.

3.3 Skin cancer recurrence data from the NPC trial: semi-parametric modelling

Jiang *et al.* (1999) consider a semi-parametric approach to analyze the NPC study, for the purpose of removing the following assumptions used in the parametric approach: (i) Constant mean rate λ; (ii) Poisson distribution assumption on $Y_i(t)$ conditional on the random effects; (iii) Gamma distribution assumption on the frailties ψ_i.

Specifically, now we assume a model M(semi-par) for the observed mean response:

$$E[Y_i(t)] = H_i(t)\psi_i\lambda(t)\exp(a_i\gamma + x_i\beta), \text{ for all } i = 1,2,\ldots,n,\ t = 1,\ldots,K. \qquad (3)$$

Without loss of generality we may take $E[\psi_i] = 1$. Note that only the mean responses are modeled (not the higher moments) and the Poisson assumption is removed. Instead of the constant baseline mean rate λ, we use a nonparametric baseline mean rate $\lambda(t)$. There is no distributional assumption on frailties $\{\psi_i\}$ either. The semi-parametric approach is therefore considerably more flexible.

Here the parameter of interest is $\theta = (\gamma,\beta,\lambda(\cdot))$, where $\lambda(\cdot) = (\lambda(1),\ldots,\lambda(K))$. This is clearly a complex model, particularly because the frailties $\{\psi_i\}$ are unobserved, and only the surrogate z_i is observed in place of x_i. Jiang *et al.* (1999) proposed an indirect inference approach based on the auxiliary model M′(semi-par) given by nonhomogeneous Poisson process model with multiplicative intensity $m(t)\exp(a_ig + z_ib)$. Note M′(semi-par) is simpler; it ignores the presence of frailties and measurement error. This leads to consideration of the intermediate statistic $\hat{s} = (\hat{g},\hat{b},\hat{m}(\cdot))$. Here $(\hat{g},\hat{b})^T$ is the Cox (1972) partial likelihood estimate and $\hat{m}(t)$ is a discrete intensity estimate for $\lambda(t)$ that corresponds to the Nelson-Aalen estimate of the cumulative intensity (see Andersen *et al.* 1993, Sec.VII.2.1). Standard computer software can be employed to compute these estimates — e.g. in Splus Release 6 (Insightful Corp. 2001). The auxiliary or 'naive' estimator $\hat{s}$ is computed ignoring both the random effect (by taking ψ_i to be its mean 1) and the measurement error (by taking x_i to be z_i). The dimensionality of $\hat{s}$ and θ are equal and so the $\hat{\theta}$ can be obtained from the bridge relation $s = s(\theta)$. Under the same Gaussian additive model for the measurement error as described in the last section, they go on to find the auxiliary or 'naive' parameter $s = (g,b,m(\cdot))$, the asymptotic mean of $\hat{s}$, leading to the bridge relations:

$$g = \gamma, \quad b = \pi\beta, \quad m(t) = \lambda(t)\exp(0.5\beta^2\sigma^2_{x|z}).$$

This bridge relation is then inverted to obtain a consistent adjusted estimator $\hat{\theta}$ for the true parameter $\theta = (\gamma,\beta,\lambda(\cdot))$. Robust sandwich variance estimates were used to obtain standard errors. Details of the calculations are given by Jiang *et al.* (1999). The results are summarized in lines 2a and 2b of Table 1.

Note that there is a qualitative difference between the estimates of treatment effect: in the general model M(semi-par), the treatment is no longer statistically significant. The results based on model M(semi-par) (line 1b) are robust against misspecifications of models on the response $\{Y_i(t)\}$– only a very general model for the mean need be postulated (cf. Lawless and Nadeau 1995). Assumptions on higher moments, such as those that might be imposed by the Poisson distribution, are not needed for valid inference.

When we compare the results of the previous parametric analysis described in Section 3.2 as displayed in lines 1a and 1b of Table 1, we find that the results are similar. This suggests that the much simpler constant intensity function model may well be adequate here.

4 Acknowledgements

The authors were partially supported by a grant from the U.S. National Institutes of Health and a grant from the National Science Foundation.

5 References

Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.

Berk, R.H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Annals of Mathematical Statististics*, 37, 51-58.

Bickel, P.J. and Doksum, K.A. (2001). *Mathematical Statistics*, 2nd Ed. Upper Saddle River, New Jersey: Prentice Hall.

Breslow, N. (1990). Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. *Journal of American Statistical Association*, 85, 565-571.

Carroll, R.J., Ruppert, D. and Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models*. London: Chapman and Hall.

Chiang, C.L. (1956). On regular best asymptotically normal estimates. *Annals of Mathematical Statististics*, 27, 336-351.

Clark, L.C. , Combs, G.F., Turnbull, B.W., Slate, E.H., Chalker, D.K., Chow, J., Davis, L.S., Glover, R.A., Graham, G.F., Gross, E.G., Krongrad, A., Lesher, J.L., Park, H.K., Sanders, B.B., Smith, C.L., Taylor, J.R. and the Nutritional Prevention of Cancer Study Group. (1996). Effects of Selenium Supplementation for Cancer Prevention in Patients with Carcinoma of the Skin: A Randomized Clinical Trial. *Journal of American Medical Association*, 276 (24), 1957-1963. (Editorial: p1984-5.)

Cox, D.R. (1962). Further results on tests of separate families of hypotheses. *J. R. Statist. Soc.* B, 24, 406-424.

Cox, D.R. (1972). Regression models and life-tables (with discussion). *J.R. Statist. Soc.* B, 34, 187-207.

Cox D.R. (1983). Some remarks on overdispersion. *Biometrika*, 70, 269-274.

Ferguson, T.S. (1958). A method of generating best asymptotic normal estimates with application to the estimation of bacterial densities. *Annals of Mathematical Statististics*, 29, 1046-1062.

Gail, M.H., Santner, T.J. and Brown, C.C. (1980). An analysis of comparative carcinogenesis experiments based on multiple times to tumor. *Biometrics*, 36, 255-266.

Gallant, A.R. and Long, J.R. (1997). Estimating stochastic differential equations efficiently by minimum chi-squared. *Biometrika*, 84, 125-141.

Gallant, A.R. and Tauchen, G. (1996). Which moments to match? *Econometric Theory*, 12, 657-681.

Gallant, A.R. and Tauchen, G. (1999). The relative efficiency of method of moments estimators. *Journal of Econometrics*, 92, 149-172.

Gourieroux, C., Monfort, A. and Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics* 8S, 85-118.

Hansen, L.P. (1982). Large sample properties of generalised method of moments estimators. *Econometrica*, 50, 1029-1054.

Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Procceedings of the Fifth Berkeley Symposium on Probability and Statistics I*, pp. 221-233. Berkeley: Univ. of California Press.

Insightful Corporation (2001). *S-PLUS 6*. Seattle, Washington.

Jiang, W. and Turnbull, B.W. (2001). The indirect method — robust inference based on intermediate statistics. *Technical Report, Department of Statistics, Northwestern University*, Evanston, IL, USA.

Jiang, W., Turnbull, B.W. and Clark, L.C. (1999). Semiparametric Regression Models for Repeated Events with Random Effects and Measurement Error. *Journal of American Statistical Association*, 94 111-124.

Kuk, A.Y.C. (1995). Asymptotically unbiased estimation in generalised linear models with random effects. *Journal of Royal Statistical Association* B, 57, 395-407.

Lawless J.F. and Nadeau, C. (1995) Some simple robust methods for the analysis of recurrent events. *Technometrics*, 37, 158-168.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. New York: Chapman and Hall.

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 57, 995-1026.

Newey, W.K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. in *Handbook of Econometrics*, edited by Engle, R.F. and McFadden, D.L., Elsevier, New York.

Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica*, 57, 1027-57.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Annals of Statistics*, 22, 300-325.

Qu, A., Lindsay, B.G. and Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika*, 87, 823-836.

StataCorp (1997). Stata Statistical Software, Release 5.0. Stata Corporation, College Station, Texas.

Thall, P.F. and Vail, S.C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics*, 46, 657-71.

Thompson, H.F., Grubbs, C.J., Moon, R.C. and Sporn, M.B. (1978). Continual requirement of retinoid for maintenance of mammary cancer inhibition. *Proceedings of the Annual Meeting of the American Association for Cancer Research*, 19, 74.

Turnbull, B.W., Jiang, W. and Clark, L.C. (1997). Regression models for recurrent event data: parametric random effects models with measurement error. *Statistics in Medecine*, 16, 853-64.

Wedderburn, R.W.M. (1974). Quasi-likelihood, generalized linear models and the Gauss-Newton method. *Biometrika*, 61, 439-447.

White, H. (1994). *Estimation, Inference and Specification Analysis*. Cambridge: Cambridge University Press.

Resum

Es descriu en aquest treball l'anomenat mètode indirecte d'inferència. Aquest mètode es va desenvolupar inicialment en la literatura econòmica i nosaltres l'apliquem a l'anàlisi de la supervivència de dos conjunts de dades amb esdeveniments repetits. Aquest mètode acostuma a ser més convenient computacionalment que el mètode de màxima versemblança quan el model inclou, per exemple, complexitats tals com efectes aleatoris i errors de mesura, i també pot servir com a base per a inferències robustes sota hipòtesis menys estrictes sobre el mecanisme que ha generat les dades. El primer conjunt de dades conté temps de recurrència de tumors mamaris en rates i es modela fent servir un procés de Poisson amb covariàncies i fragilitats (frailties). El segon conjunt de dades involucra temps de recurrència de tumors de pell en individus d'un assaig clínic. S'aplica la metodologia a anàlisis de regressió, tant paramètrics com semiparamètrics, que acomoden efectes aleatoris i errors de mesura en les covariàncies.

MSC: 62N01, 62N02, 62P10

Paraules clau: Efectes aleatoris; equacions d' estimació; error de mesura; estimadors naive; frailty; inferència indirecta; quasi-versemblança; regressió de la taxa de risc; sobredispersió; robustesa