Biometrika (1999), **86**, 1, *pp.* 27–43
© 1999 Biometrika Trust
Printed in Great Britain

Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials

By PETER B. GILBERT
Department of Biostatistics, Harvard University, Boston, Massachusetts 02115, U.S.A.
pgilbert@hsph.harvard.edu

SUBHASH R. LELE
Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205, U.S.A.
slele@welchlink.welch.jhu.edu

AND YEHUDA VARDI
Department of Statistics, Rutgers University, New Brunswick, New Jersey 08903, U.S.A.
vardi@stat.rutgers.edu

Summary

The following problem is treated: given s possibly selection biased samples from an unknown distribution function, and assuming that the sampling rule weight functions for each of the samples are mathematically specified up to a common unknown finite-dimensional parameter, how can we use the data to estimate the unknown parameters? We propose a simple maximum partial likelihood method for deriving the semiparametric maximum likelihood estimator. A discussion of assumptions under which the selection bias model is identifiable and uniquely estimable is presented. We motivate the need for the methodology by discussing the generalised logistic regression model (Gilbert, Self & Ashby, 1998), a semiparametric selection bias model which is useful for assessing from vaccine trial data how the efficacy of an HIV vaccine varies with characteristics of the exposing virus. We show through simulations and an example that the maximum likelihood estimator in the generalised logistic regression model has satisfactory finite-sample properties.

Some key words: Biased sampling model; Confidence interval; Generalised logistic regression model; Human immunodeficiency virus vaccine efficacy trial; Hypothesis testing; Partial likelihood; Profile likelihood; Semiparametric model; Weighted distribution.

1. Introduction

Biological vaccine efficacy is commonly defined as VE = 1 − RR, where RR is the relative risk among vaccinated and unvaccinated persons of infection resulting from a single exposure to the infectious pathogen. Since the human immunodeficiency virus (HIV) exhibits broad genotypic and phenotypic diversity, it is important to investigate how the vaccine efficacy of a candidate HIV vaccine may vary with characteristics of the exposing HIV. Measurements of characteristics of viruses isolated from study participants infected during a preventive placebo-controlled vaccine efficacy trial can be used to make this assessment.

The difficulty is that only those HIV types which cause an infection can be observed, so that the available dataset is from an improper subset of trial participants. As a consequence, HIV type-specific exposure-adjusted infection rates in the placebo and vaccine groups cannot be directly calculated. Thus, there is a selection bias in assessing type-specific vaccine protection from viral data extracted from participants who become infected while enrolled in preventive vaccine trials. Berman et al. (1997) called this problem 'sieve' analysis, where the sieve is the vaccine's strain-specific barrier to infection.

When there are K distinct HIV types of interest and failure times are measured, competing risks methods may be considered (Prentice et al., 1978). However, the rarity of HIV infection implies that over 90% of participants will be censored by not becoming infected during the trial. Secondly, defining and measuring the infection time is problematic because several months may pass before maximum vaccine immunity is achieved, and the infection time endpoint is infrequently ascertained and may be interval censored.

Gilbert et al. (1998) used the simpler endpoint infection status at study termination in the aforementioned case in which viral variation is described by K categories. Here assume viruses are ordered by a continuous distance from the prototype virus or viruses used in the vaccine preparation. This distance may be, for instance, a protein dissimilarity score such as the Hamming distance based on the amino acid sequence of some viral coding region.

Let Y be a random variable denoting the distance of an observed infecting strain. Let F denote the 'baseline' distribution of Y among infected trial participants who received placebo, and let F_v denote the distribution of Y among infected vaccinated trial participants. We observe samples from F and from F_v. The generalised logistic regression model relates these two distributions in the following way:

$$F_v(y) = \frac{\int_0^y \exp\{g(u,\theta)\}\,dF(u)}{\int_0^\infty \exp\{g(u,\theta)\}\,dF(u)} \quad (y \in [0,\infty)), \tag{1·1}$$

where $g(y,\theta)$ is a given function and θ is an unknown d-dimensional parameter. We require $g(0,\theta)=0$ for identifiability. Simple practical choices of $g(y,\theta)$ are a linear form, $g(y,\theta)=y\theta$, and a quadratic form, $g(y,\theta)=y\theta_1+y^2\theta_2$. Note that F_v is a weighted version of the baseline distribution F (Patil, Rao & Zelen, 1988).

The model generalises a categorical linear logit model: if Y is categorical, with possible values $y=0,\ldots,K$, then (1·1) equivalently expresses

$$\log\left\{\frac{\mathrm{pr}(Y=y\mid F_v)}{\mathrm{pr}(Y=0\mid F_v)}\right\} = \log\left\{\frac{\mathrm{pr}(Y=y\mid F)}{\mathrm{pr}(Y=0\mid F)}\right\} + g(y,\theta).$$

The primary motivation for the generalised logistic regression model is the useful interpretation of the 'differential vaccine efficacy parameter' θ. Gilbert et al. (1998) identified assumptions of exposure and infection dynamics of the circulating HIV strains in the trial population under which the log-odds-ratio $g(y,\theta)$ equals $\log\{\mathrm{RR}(y)/\mathrm{RR}(0)\}$, where $\mathrm{RR}(y)$ is the relative probability among vaccinees and non-vaccinees that a single exposure to a strain of distance y leads to transmission. Therefore, the relationship between vaccine protection and strain distance can be directly assessed: for any two strain distances y_1 and y_2, $\log\{\mathrm{RR}(y_1)/\mathrm{RR}(y_2)\}$ is estimated by $g(y_1,\hat{\theta})-g(y_2,\hat{\theta})$. Thus the model approximately accounts for the sieve selection bias by enabling estimation of an interpretable parameter.

The two-sample generalised logistic regression model (1·1) also generalises easily to an s-sample situation where s corresponds to the number of groups in the trial. Let F denote the distribution of infecting distances observed in the sth group, which we take to be the baseline distribution. Let θ be a d-dimensional parameter and $g_i(y, \theta)$, for $i = 1, \ldots, s-1$, be given functions satisfying $g_i(0, \theta) = 0$. Then an s-sample model is formed by combining the sample $y_s = (y_{s1}, \ldots, y_{sn_s})$ from F with the $s-1$ samples $y_i \equiv (y_{i1}, \ldots, y_{in_i})$, for $n_i \geqslant 1$, from the distribution function

$$F_i(y) = \frac{\int_0^y \exp\{g_i(u, \theta)\}\, dF(u)}{\int_0^\infty \exp\{g_i(u, \theta)\}\, dF(u)} \quad (y \in [0, \infty)). \tag{1·2}$$

A practical choice for the functions $\{g_i\}$ is $g_i(y, \theta) = \sum_{k=1}^d g_{ik}(y)\theta_k$, where g_{ik} are given functions of y independent of θ, and satisfy $g_{ik}(0) = 0$.

Although the unknown baseline distribution F might be given a parametric form, for the HIV application we prefer to leave the model flexible with respect to the baseline distribution, and to view the generalised logistic regression model as semiparametric, where the regression relationship is parametric and the baseline distribution is left nonparametric. Model (1·2) is an example of an s-sample semiparametric selection bias model as shown in § 2.

Selection bias models or weighted distributions arise in many other practical situations. Patil et al. (1988) and Patil & Rao (1977, 1978) give excellent reviews of the statistical properties of these models and describe various situations in wildlife management, biology, fisheries, geology, etc. where these distributions arise. Of the many areas of application we highlight only case-control studies in biostatistics (Prentice & Pyke, 1979) and stratified or truncated regression (Bhattacharya, Chernoff & Yang, 1983).

Nonparametric inference about F was first considered by Cox (1969) for the case of length biased sampling, where the weight function is $w(y) = y$ and there is only one sample. Vardi (1982, 1985) generalised this procedure to the case of s independent samples from the same population but each with a different, but known, weight function $w_i(.)$.

Thus far, estimation of F has been considered mainly for the fully parametric case where both the weights and F belong to parametric families, and for the fully nonparametric case where the weights are completely specified but F is nonparametric. Sun & Woodroofe (1997) introduced a semiparametric approach by estimating the weight nonparametrically and F parametrically for the one-sample model. Here we leave F nonparametric, but assume that the weight function of the ith sample is $w_i(., \theta)$, for some parameter θ, and the sampling density of the ith sample is proportional to $w_i(., \theta)f(.)$. We provide a simple procedure for simultaneously estimating θ and F.

In § 2, we introduce the notation and mathematical formulation of the problem. In § 3 we discuss identifiability issues. The model is identifiable for most choices of weight functions when there are two or more samples, but rarely so in the one-sample case. In § 4, we present the maximum likelihood estimation procedure, and discuss conditions on the weights and data under which a unique maximum exists. In § 5, a simulation study of the maximum likelihood estimate $(\hat{\theta}, \hat{F})$ in the two- and three-sample generalised logistic regression model is presented, illustrating consistency and other results. An example based on real Thai HIV-1 sequence data is given in § 6. The Appendix contains proofs of theorems displayed in the paper.

2. Notation and mathematical formulation

Suppose we observe $I_1, \ldots, I_n$, independently and identically distributed with $\text{pr}(I_j = i) = \lambda_i$, with $\lambda_1 + \ldots + \lambda_s = 1$, where the selection probabilities λ_i are not necessarily known. For example s is the number of vaccine groups and I indicates the group. Conditional on $I = i$, we observe $y_i \equiv (y_{i1}, \ldots, y_{in_i})$, for $n_i \geqslant 1$, a random sample of size n_i from the cumulative distribution function

$$F_i(y, \theta, F) = W_i(\theta, F)^{-1} \int_{-\infty}^{y} w_i(u, \theta)\, dF(u) \quad (i = 1, \ldots, s), \tag{2·1}$$

where θ is an unknown real or vector-valued parameter, F is an unknown cumulative distribution function, and

$$W_i(\theta, F) \equiv \int_{-\infty}^{\infty} w_i(u, \theta)\, dF(u)$$

is the ith normalising function. The weight functions $w_i(., \theta)$ are assumed to be nonnegative and of a known parametric form. The normalising functions $W_i(\theta, F)$ are assumed to be finite and positive for all θ in the parameter space $\Theta \subset R^d$.

The model arising from (2·1) is an s-sample selection bias model. Comparing (1·2) and (2·1) shows that the s-sample generalised logistic regression model is an s-sample selection bias model. For instance, the two-sample model has weight functions w_1 and w_2, corresponding to the vaccine and placebo groups, respectively, given by $w_1(y, \theta) = \exp\{g(y, \theta)\}$ and $w_2(y, \theta) = 1$.

Following the notation of Vardi (1985), denote the size of the ith sample by n_i, the total sample size by $n = n_1 + \ldots + n_s$, and the ith sampling fraction by $\lambda_{ni} = n_i/n$. Then a semiparametric likelihood estimate of θ and F maximises the likelihood

$$L(\theta, F \mid y) = \prod_{i=1}^{s} \prod_{j=1}^{n_i} \frac{w_i(y_{ij}, \theta) f(y_{ij})}{\int w_i(u, \theta)\, dF(u)} \tag{2·2}$$

with respect to $\theta \in \Theta$ and F over all distribution functions.

Since F is constrained only to be a distribution function the maximisation problem is of infinite dimension. The key result of this paper is that, subject to identifiability and estimability conditions, this maximisation problem is equivalent to maximising a partial likelihood which turns out to be a maximisation problem of fixed dimension.

3. Identifiability

First suppose $\theta \in \Theta$ is known. Suppose the sample space equals $[y : w_i(y, \theta) > 0$ for some $i \in \{1, \ldots, s\}]$, and that

$$\int I\{w_i(y, \theta) > 0\} I\{w_k(y, \theta) > 0\}\, dF(y) > 0$$

for all $i, k \in \{1, \ldots, s\}$. As in Gill, Vardi & Wellner (1988, p. 1072), these conditions are necessary and sufficient for F to be identifiable when θ is known. In what follows, assume these conditions hold for all $\theta \in \Theta$. Note that they automatically hold if all of the weight functions are strictly positive.

Now consider identifiability of unknown θ and F. When there is only one sample, the identifiability issue is complicated. For example, suppose $w(y, \theta) = |y|^{\theta}$. Then the pairs (θ, f) and $(\theta + c, |y|^{-c} f(y))$ give rise to the same likelihood function for any value of c, assuming the integral exists, and so the pair (θ, F) is not identifiable. This example is typical of the single-sample case. If w is such that either $w(y, \theta)/w(y, \theta + c)$ or

$w(y, \theta)/w(y, \theta c)$ is constant in θ, then the pair (θ, F) is nonidentifiable. In fact, for any w, the pair $(w(y, \theta), F)$ and the pair $(h(y)w(y, \theta), h^{-1}(y)F(dy)/\int h^{-1}(u)F(du))$, for an arbitrary $h > 0$ for which the integral is finite, give the same likelihood. Thus it is often impossible to estimate both θ and F uniquely.

The situation is different when the domain of $w(y, \theta)$ depends on θ, in which case it is usually possible to estimate consistently θ and the conditional distribution of F. For instance, in the simple case in which F is supported on the whole real line and $w(y, \theta) = I\{y < \theta\}$, θ is consistently estimated by the largest observation, and $I\{y < \theta\}F(y)/F(\theta)$ is consistently estimated by the sample distribution function; clearly there are many F's for which $I\{y < \theta\}F(y)/F(\theta)$ is the same. Theorem 1 gives a sufficient condition for identifiability in the one-sample case. Let $D(\theta)$ be the collection of y's for which $w(y, \theta)$ is strictly positive.

THEOREM 1. *Let $s = 1$. If the domain $D(\theta)$ depends on θ, in that $F\{D(\theta) - D(\tilde{\theta})\} > 0$ or $F\{D(\tilde{\theta}) - D(\theta)\} > 0$ for all $\tilde{\theta}, \theta \in \Theta$ with $\theta \neq \tilde{\theta}$, then the one-sample selection bias model is identifiable.*

For the s-sample case, with $s \geqslant 2$, identifiability is less of a problem. When one of the weight functions is independent of θ, the class of identifiable models can be characterised by a simple condition on the weight functions.

THEOREM 2. *Let $s \geqslant 2$, with w_s independent of θ. Then the s-sample selection bias model is identifiable if and only if the following condition holds: for all $\tilde{\theta}, \theta \in \Theta$ with $\tilde{\theta} \neq \theta$, there is at least one weight function w_i, for $i \in \{1, \ldots, s-1\}$, such that $w_i(y, \tilde{\theta})$ and $w_i(y, \theta)$ are linearly independent as functions of y.*

A large class of weight functions satisfy the condition of Theorem 2. For instance, consider an s-sample model with w_s known and one weight function, w_1, say, of the form $w_1(y, \theta) = \exp\{g(y)h(\theta)\}$. If the real-valued function g is nonconstant with $g(0) = 0$, and the real-valued function h is one-to-one, then the condition holds. Next consider model (1·2), with

$$g_i(y, \theta) = \sum_{k=1}^{d} g_{ik}(y)\theta_k, \quad g_{ik}(0) = 0 \quad (i = 1, \ldots, s-1, k = 1, \ldots, d).$$

It is straightforward to show that the condition of Theorem 2 holds if and only if, for some $i \in \{1, \ldots, s-1\}$, the functions $\{g_{i1}, \ldots, g_{id}\}$ are linearly independent. This holds for all interesting generalised logistic regression models. Another selection bias model satisfying the condition is one with a dth degree polynomial weight function with known intercept and unknown coefficients $\theta = (\theta_1, \ldots, \theta_d)'$.

In the general situation, if one does not want to assume that one of the weight functions is completely known, the following condition is sufficient for identifiability.

THEOREM 3. *Suppose there exist weight functions w_i and w_k, for $i, k \in \{1, \ldots, s\}$, $i \neq k$, such that the functions $w_i(y, \theta)w_k(y, \tilde{\theta})$ and $w_i(y, \tilde{\theta})w_k(y, \theta)$ are linearly independent in y for all $\tilde{\theta}, \theta \in \Theta$ with $\tilde{\theta} \neq \theta$. Then the s-sample selection bias model is identifiable.*

Since the identifiability condition of Theorem 2 or 3 only involves two of the s weight functions, the class of identifiable models grows with s. In fact, any s-sample model is identifiable if two of its weights compose an identifiable two-sample model. It also holds that a two-sample model is identifiable if one of its parametric weights forms a one-sample identifiable model.

The condition in Theorem 3 is not necessary. For a counter-example, consider the two-sample model with $w_1(y, \theta) = I\{y > \theta\}$, $w_2(y, \theta) = cI\{y > \theta\}$. For $c > 1$ this model is identifiable by Theorem 1 and the remark in the preceding paragraph. In general, it appears that, if all pairs of weight functions are linearly dependent in y for each fixed θ, the model will not be identifiable unless the domain of each weight function depends on θ.

In what follows, we assume that the selection bias model satisfies the conditions of Theorem 2, so that $s \geqslant 2$, w_s is known, and the model is identifiable. Under these conditions we describe a simple procedure for deriving the maximum likelihood estimates.

4. Maximum likelihood estimation

4·1. *The likelihood*

Rewrite the likelihood in the following form. Let $t_1, \ldots, t_h$ be the distinct observed Y values, with multiplicities $r_1, \ldots, r_h$. Let n_{ij} $(i = 1, \ldots, s, j = 1, \ldots, h)$ be the number of observations from the ith group with value t_j. Then the likelihood of the data (2·2) can be rewritten as

$$L(\theta, F \mid y) = \prod_{i=1}^{s} \prod_{j=1}^{h} \left\{ \frac{w_i(t_j, \theta) f(t_j)}{W_i(\theta, F)} \right\}^{n_{ij}}. \tag{4·1}$$

As in Vardi (1985), clearly $L = 0$ if any t_j is a continuity point of F, while $L > 0$ if $f(t_j) > 0$ $(j = 1, \ldots, h)$. Furthermore, for any given θ, if F assigns positive mass to a Borel set outside $\{t_1, \ldots, t_h\}$, then F can be replaced with a distribution G satisfying $L(\theta, F) \leqslant L(\theta, G)$. Thus, in order to find an F that maximises (4·1), we can restrict our search to the class of discrete F's which have positive jumps at each of the points $t_1, \ldots, t_h$, and only there. Put $p_1 = f(t_1), \ldots, p_h = f(t_h)$, and denote the likelihood function by $L(\theta, p \mid y) \equiv \mathrm{pr}_{(\theta, p)}(y_1, \ldots, y_s)$. Then our problem is to maximise

$$L(\theta, p \mid y) = \prod_{i=1}^{s} \prod_{j=1}^{h} \left\{ \frac{w_{ij}(\theta) p_j}{W_i(\theta, p)} \right\}^{n_{ij}} \tag{4·2}$$

subject to $\theta \in \Theta$ and $\sum_{j=1}^{h} p_j = 1$, for $p_j > 0$, where we put

$$p = (p_1, \ldots, p_h), \quad w_{ij}(\theta) = w_i(t_j, \theta), \quad W_i(\theta, p) = \sum_{j=1}^{h} w_{ij}(\theta) p_j \quad (i = 1, \ldots, s, j = 1, \ldots, h).$$

4·2. *Maximum partial likelihood estimation procedure*

Let F_M be the mixture distribution over the s samples, with density defined by $f_M(t_j, \theta) = \sum_i \lambda_{ni} \{w_{ij}(\theta) p_j / W_i(\theta, p)\}$. The full likelihood can then be factorised as

$$L(\theta, p \mid y) = \prod_{i=1}^{s} \prod_{j=1}^{h} \left\{ \frac{w_{ij}(\theta) p_j / W_i(\theta, p)}{f_M(t_j, \theta)} \right\}^{n_{ij}} \times \prod_{i=1}^{s} \prod_{j=1}^{h} \{f_M(t_j, \theta)\}^{n_{ij}}.$$

Consider the first term, which we denote by $L_1(\theta, p \mid y)$. Set $V_i = W_i(\theta, F)$, suppressing the dependence of V_i on θ and F. Set $V = (V_1, \ldots, V_s)'$. Then this partial likelihood simplifies to

$$L_1(\theta, V \mid y) = \prod_{i=1}^{s} \prod_{j=1}^{h} \left\{ \frac{w_{ij}(\theta) V_i^{-1}}{\sum_{k=1}^{s} \lambda_{nk} w_{kj}(\theta) V_k^{-1}} \right\}^{n_{ij}}. \tag{4·3}$$

As we show in the proof of Theorem 4, we can behave as if the parameters $\{V_i\}$ are free, subject to the constraint that they are positive. For estimability one of the weight functions must be constant, so suppose $w_s(y, \theta) = 1$ for all y and θ, which forces $V_s = 1$. This corresponds to the assumption of strong connectivity of a certain graph $\mathscr{G}(\theta)$ for all $\theta \in \Theta$, defined in Vardi (1985) and restated here. For $i, k \in \{1, \ldots, s\}$, we say there is a directed path from a vertex i to a vertex k, as $i \rightarrow k$, if and only if $w_{i1}(\theta) n_{k1} + \ldots + w_{ih}(\theta) n_{kh} > 0$. The graph $\mathscr{G}(\theta)$ defined on the s vertices $\{1, \ldots, s\}$ is said to be strongly connected if, for every pair (i, k), there exists a directed path from i to k and a directed path from k to i.

The following procedure then yields the joint maximum likelihood estimate $(\hat{\theta}, \hat{F})$.

Step 1. Maximise L_1 over θ and V, subject to $\theta \in \Theta$, $V_1 > 0$, $V_2 > 0, \ldots, V_{s-1} > 0$, $V_s = 1$ to obtain $(\hat{\theta}, \hat{V})$.

Step 2. Compute Vardi's nonparametric maximum likelihood estimator $\hat{F} \equiv \hat{F}(\hat{\theta})$ from data with 'known' weight functions $w_i(., \hat{\theta})$, with $w_s(., \hat{\theta}) \equiv 1$.

Step 3. Then $\hat{W}_i = \hat{V}_i = \int w_i(y, \hat{\theta})\, d\hat{F}(y)$ $(i = 1, \ldots, s)$ automatically.

Step 1 can be accomplished via profile likelihood. For fixed $\theta \in \Theta$, let

$$\hat{V}(\theta) = (\hat{V}_1(\theta), \ldots, \hat{V}_{s-1}(\theta), 1)'$$

be the unique solution of

$$H_i\{V_1(\theta), \ldots, V_s(\theta)\} \equiv V_i^{-1}(\theta) \sum_{j=1}^{h} \frac{r_j w_{ij}(\theta)}{\sum_{k=1}^{s} n_k w_{kj}(\theta) V_k^{-1}(\theta)} = 1 \quad (i = 1, \ldots, s-1) \qquad (4{\cdot}4)$$

in the region $V_1(\theta) > 0, \ldots, V_{s-1}(\theta) > 0$, $V_s \equiv 1$. Vardi (1985) proved that there exists a unique solution to (4·4) if and only if the graph $\mathscr{G}(\theta)$ is strongly connected. The estimator $\hat{\theta}$ is the argument which maximises the profile partial likelihood L_{pr} defined by

$$L_{\text{pr}}(\theta) = \prod_{i=1}^{s} \prod_{j=1}^{h} \left\{ \frac{w_{ij}(\theta) \hat{V}_i^{-1}(\theta)}{\sum_{k=1}^{s} \lambda_{nk} w_{kj}(\theta) \hat{V}_k^{-1}(\theta)} \right\}^{n_{ij}}. \qquad (4{\cdot}5)$$

With $\hat{V} = \hat{V}(\hat{\theta})$, Step 2 proceeds by setting $\hat{p} = p(\hat{\theta})$, where

$$p_j(\theta) \propto \frac{r_j}{\sum_k n_k w_{kj}(\theta) \hat{V}_k^{-1}} \quad (j = 1, \ldots, h).$$

Thus the semiparametric estimator of F is

$$\hat{F}(y) = \frac{n^{-1} \sum_{j=1}^{h} I\{t_j \leqslant y\} \{r_j / \sum_{k=1}^{s} n_k w_{kj}(\hat{\theta}) \hat{V}_k^{-1}\}}{n^{-1} \sum_{j=1}^{h} \{r_j / \sum_{k=1}^{s} n_k w_{kj}(\hat{\theta}) \hat{V}_k^{-1}\}}.$$

The above procedure, based on Vardi (1985), is computationally attractive because it requires computation of $\hat{F}$ on only one occasion; once $(\hat{\theta}, \hat{V})$ is obtained, $\hat{p}$ is obtained through substitution only. Thus, in essence, the procedure only requires maximising a function depending on finite-dimensional parameters and solving a system of equations of fixed dimension $(s - 1)$. Theorem 4 asserts that this procedure indeed yields the maximum likelihood estimator.

THEOREM 4. *Suppose* $s \geqslant 2$ *with* $w_s \equiv 1$ *and that the identifiability condition of Theorem* 2 *holds. Further suppose that the graph* $\mathscr{G}(\theta)$ *is strongly connected for all* $\theta \in \Theta$. *Then, if*

$(\hat{\theta}, \hat{V}, \hat{F})$ obtained from Steps 1–3 *maximises the partial likelihood* (4·3), *it maximises the full likelihood* (4·1).

4·3. *Estimability and uniqueness*

In this section we study conditions on the weight functions and the data which guarantee that problem (4·2) has a unique maximum. Assume the weight functions are differentiable in θ. Let the d-vector $\dot{w}_i(y, \theta)$ denote the derivative of w_i with respect to θ evaluated at (y, θ).

Under strong connectivity of the graph $\mathscr{F}(\theta)$ for all $\theta \in \Theta$, the problem will have a unique maximum if the logarithm of the profile partial likelihood (4·5) is strictly concave on Θ. Theorem 5 gives a sufficient condition for this to hold for two-sample models, which is satisfied by all generalised logistic regression models imagined to be interesting.

Theorem 5. *Suppose $s = 2$, $w_2 \equiv 1$, $\dot{w}_1(y, \theta)/w_1(y, \theta)$ is not degenerate at the d-unit vector, there is at least one observation in each group, and $(\partial^2/\partial\theta^2)\log\{w_1(y, \theta)\} = 0$, the $d \times d$ zero matrix. Then, if the graph $\mathscr{F}(\theta)$ is strongly connected for all $\theta \in \Theta$, the logarithm of the profile partial likelihood* (4·5) *is strictly concave on Θ.*

The conclusion of Theorem 5 also holds for s-sample selection bias models with weight functions satisfying $\dot{w}_1(y, \theta)/w_1(y, \theta) = \ldots = \dot{w}_{s-1}(y, \theta)/w_{s-1}(y, \theta)$ for all y and θ.

Now consider arbitrary $s \geqslant 2$ with $w_s \equiv 1$. We have been unable to determine conditions under which the log profile partial likelihood is strictly concave in θ or under which the log partial likelihood is strictly concave jointly in θ and V. Theorem 6 provides a partial solution, giving sufficient conditions on the weight and data for L_1 to have a unique maximum marginally in V for fixed θ and in θ for fixed V.

Theorem 6. (i) *For fixed $\theta \in \Theta$, suppose the graph $\mathscr{F}(\theta)$ is strongly connected. Then $L_1(V \mid \theta, y)$ has a unique maximum.*

(ii) *Suppose $(\partial^2/\partial\theta^2)\log\{w_i(y, \theta)\} = 0$ $(i = 1, \ldots, s-1)$ and, for all $i, k \in \{1, \ldots, s-1\}$ with $i < k$, and all y and θ,*

$$\dot{w}_i(y, \theta)/w_i(y, \theta) \leqslant \dot{w}_k(y, \theta)/w_k(y, \theta)$$

if and only if $\dot{w}_i(y, \theta) \leqslant \dot{w}_k(y, \theta)$. Then $\log L_1(\theta \mid V, y)$ is strictly convave on $\Theta \subset R$, so that it has a unique maximum.

The condition in (ii) guarantees $\log\{L_1(\theta \mid V, y)\}$ strictly convave on Θ, but, as seen in the full proof, is not at all necessary. All that must happen is that, for $i < k$,

$$\{\dot{w}_k(t_j, \theta)/w_k(t_j, \theta)\} - \{\dot{w}_i(t_j, \theta)/w_i(t_j, \theta)\}, \quad \dot{w}_k(t_j, \theta) - \dot{w}_i(t_j, \theta)$$

tend to have the same sign for $j = 1, \ldots, h$.

5. Simulations

We study the performance of $(\hat{\theta}, \hat{F})$ through computer simulations of the two- and three-sample generalised logistic regression models. We investigate bias and estimation of variance via observed inverse generalised Fisher information and via the bootstrap. The bootstrap is attractive because it is first-order asymptotically correct; see P. B. Gilbert's 1996 University of Washington Ph.D. thesis. We also study the power of likelihood ratio, Wald and score tests of the hypothesis $\{H_0 : \theta = 0\}$ of uniform vaccine protection, and the coverage accuracy of confidence intervals derived from these test statistics.

We take Y to be the percent amino acid difference between an observed infecting virus and the global subtype B consensus virus in the V3 loop region of the HIV-1 envelope gene. This definition is motivated by the functional importance of the V3 loop. For the two-sample generalised logistic regression model (1·1), take g to be linear, scaled by the maximum observed subtype B distance of 35%. Thus the log relative risk ratio is given by $\log\{\mathrm{RR}(y)/\mathrm{RR}(0)\} = \frac{1}{35}y\theta$. We consider three values of θ: 0, 2 and 4. The value $\theta = 2$ implies the relative risk at $y = 35$ is $\mathrm{RR}(35) = \exp(2)\,\mathrm{RR}(0) = 7{\cdot}39\,\mathrm{RR}(0)$, while the value $\theta = 4$ implies $\mathrm{RR}(35) = 54{\cdot}60\,\mathrm{RR}(0)$. For the three-sample model (1·2) with two vaccine groups and a placebo group $\theta = (\theta_1, \theta_2)'$ is two-dimensional, where $g_1(y, \theta_1)$ and $g_2(y, \theta_2)$ describe dependency of vaccine efficacy on distance for the first and second vaccines, respectively. We take $g_1 = g_2 = g$.

Four baseline distribution functions F are considered, the first three members of the beta family. These are $\mathrm{Un}(0, 35)$, $N(\mu, \sigma^2)$ and $\mathrm{Ex}(\mu/2)$. Real HIV-1 sequence data are used to guide the selection of μ and σ^2. The global subtype B V3 loop consensus and 159 subtype B V3 loop sequences from United States infections are available from the public sequence databank of the Los Alamos National Laboratory; see Los Alamos National Laboratory technical report MS K710 by B. Foley and B. Korber. The parameters μ and σ^2 are taken to be the sample mean and sample variance of the 159 computed distances, giving $\hat{\mu} = 11{\cdot}57\%$ and $\hat{\sigma}^2 = 7{\cdot}10\%^2$. The frequency distribution of these distances is depicted in Fig. 1(a). The fourth baseline distribution is constructed from observed distances of HIV-1 isolates in Thailand, where HIV-1 subtypes B and E circulate. From 30 Thai subtype B sequences and 64 Thai subtype E sequences from the Los Alamos library, and assuming a relative prevalence of 40% B and 60% E (Vaniyapongs et al., 1996), we constructed the empirical baseline distribution, illustrated in Fig. 1(b). For this Thai distribution we take $g(y, \theta) = \frac{1}{70}y\theta$, as the maximum observed distance is nearly 70%.

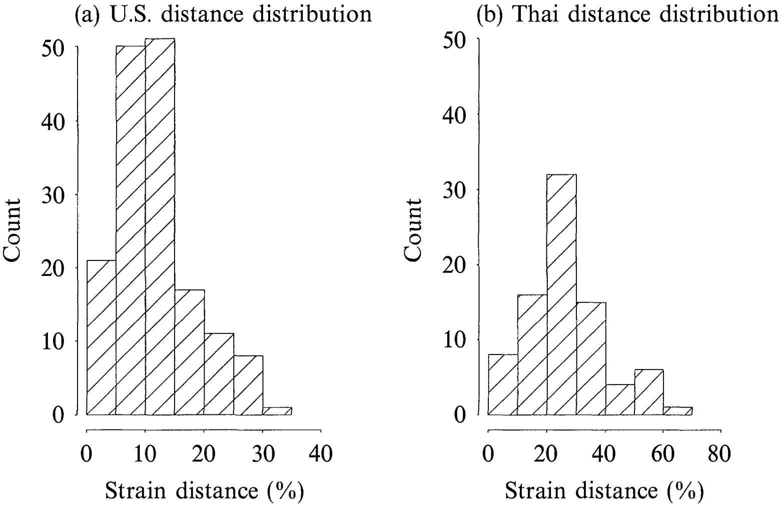

Fig. 1. (a) The distribution of the V3 loop amino acid distance between 159 U.S. subtype B sequences and the global subtype B consensus sequence. (b) The distribution of 94 V3 loop amino acid distances of infecting strains in Thailand.

Four sample sizes in terms of the number of infections occurring during the trial are considered, representing vaccines with overall efficacy 0% and 50%.

For each combination of the above parameter values, the statistics were computed 1000 times. Details of the computations can be found in P. B. Gilbert's dissertation. We now present a fraction of the simulation results; for a complete report see the dissertation.

The maximisation algorithm converged for 98·2% of generated datasets. Figure 2 displays the log profile partial likelihood of (4·5) as a function of θ for a representative sample of 16 of these datasets, 8 for the two-sample problem and 8 for the three-sample problem. The log profile partial likelihood is visibly strictly concave in all cases, corroborating Theorem 5.

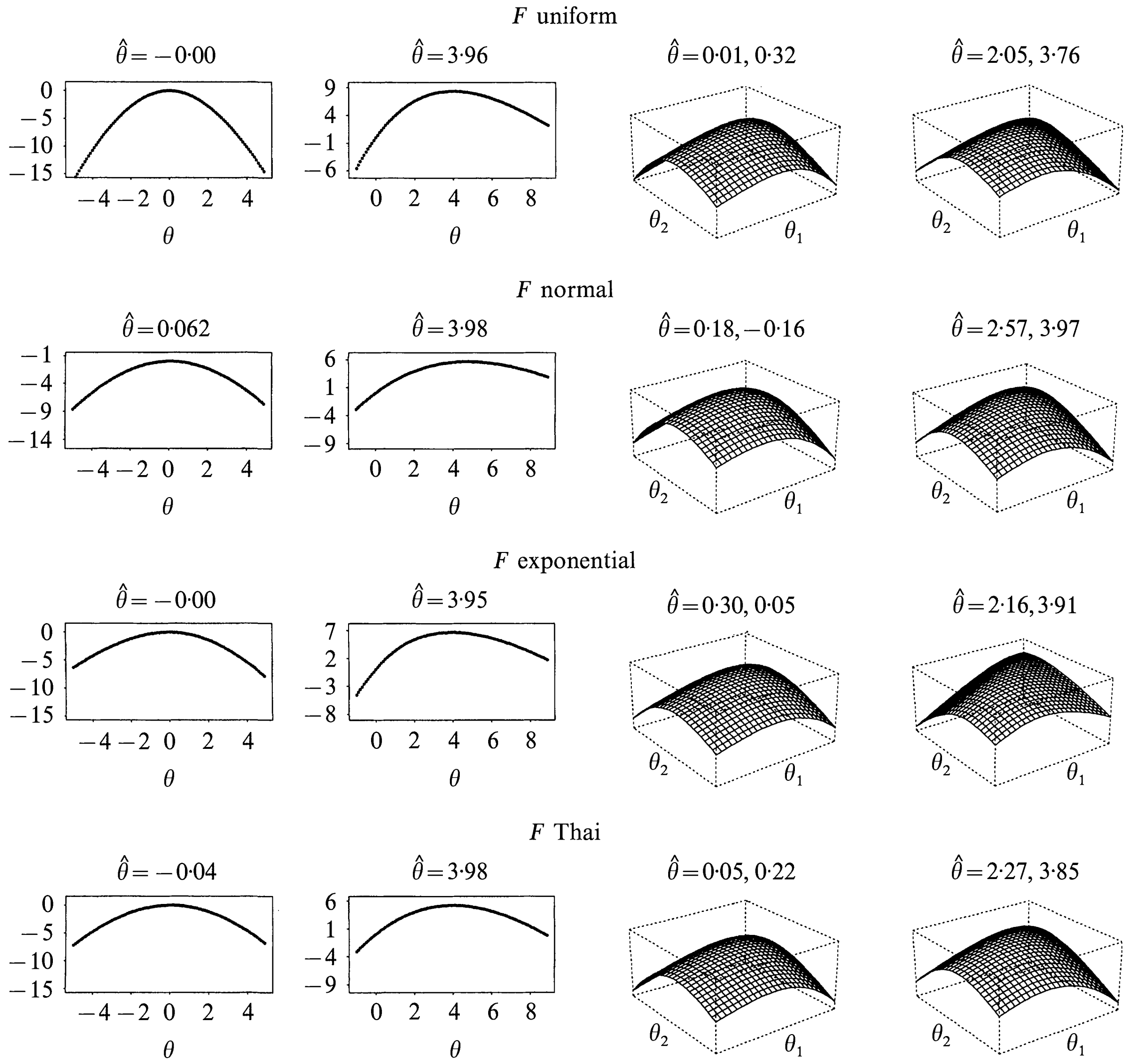

Fig. 2. Plots of the log profile partial likelihood versus θ for a spectrum of generated datasets. The obtained $\hat\theta$ is written above each plot. The first two columns are plots for the two-sample problem representing sample size $n_p = 100$, $n_v = 50$; the second two columns are plots for the three-sample problem representing sample size $n_p = 100$, $n_{v1} = 50$, $n_{v2} = 25$.

Tables 1–3 display simulation results for the two-sample problem. As reported in Table 1, $\hat\theta$ is unbiased in large samples, demonstrating asymptotic consistency. In small samples $\hat\theta$ is positively biased, most noticeably when $\theta = 0$ or $\theta = 4$, and when F is not distributed uniformly. In all cases the finite-sample and observed generalised Fisher information variance estimates are close in magnitude, demonstrating consistency of the

Table 1. *Bias and variance of the maximum likelihood estimator $\hat{\theta}$; finite-sample variance s^2, observed generalised Fisher information variance estimate* var_F, *bootstrap variance estimate* var_B

n_p	n_v	θ	Bias	s^2	var_F	var_B	Bias	s^2	var_F	var_B
				F uniform				F normal		
100	100	0	−0·01	0·24	0·24	0·20	−0·03	0·47	0·50	0·33
50	25	0	−0·00	0·71	0·76	0·52	0·24	1·21	1·56	0·96
100	100	2	0·02	0·30	0·29	0·30	−0·01	0·52	0·54	0·57
50	25	2	0·08	0·97	0·95	1·01	0·20	1·63	1·71	1·78
100	100	4	0·08	0·49	0·46	0·48	0·01	0·66	0·67	0·72
50	25	4	0·27	1·87	1·67	1·77	0·29	2·28	2·14	2·67
				F exponential				F Thai		
100	100	0	0·07	0·67	0·79	0·47	0·01	0·49	0·52	0·38
50	25	0	0·38	1·48	2·49	1·66	0·21	1·26	1·63	0·97
100	100	2	0·05	0·64	0·64	0·67	0·04	0·54	0·53	0·54
50	25	2	0·10	1·68	1·91	1·92	0·12	1·80	1·65	1·55
100	100	4	0·06	0·64	0·65	0·67	0·09	0·67	0·62	0·66
50	25	4	0·23	1·94	1·87	2·27	0·25	2·19	1·96	2·23

information variance estimator. The bootstrap variance estimate is also close to the finite-sample variance, but tends to be slightly smaller when $\theta = 0$ and slightly larger when $\theta = 2$ or 4. Note that the variance estimates increase with θ. They are of comparable magnitude for F normally, exponentially and Thai distributed, and about 50% lower for F uniformly distributed.

For distances generated from F uniformly, normally or Thai distributed, size and power are very similar for the likelihood ratio, Wald and score tests. As shown in Table 2, they have nominal size. When F is exponentially distributed, the score test is more powerful than the likelihood ratio and Wald tests, which are slightly conservative. Generally power is equal for F normal, exponential and Thai, and higher for F uniform. For example, consider a trial with 150 infections, 100 in the placebo group. With size 0·05, the power to detect $\theta = 2$ is 0·65, 0·66, 0·66 and 0·89 for F normal, exponential, Thai and uniform, respectively.

Table 2. *Power of likelihood ratio* (LR), *Wald and score tests of* $H_0 : \theta = 0$ *with* $\alpha = 0{\cdot}05$

			F uniform	F normal	F exponential			F Thai
n_p	n_v	θ	LR	LR	LR	Wald	Score	LR
100	100	0	0·05	0·05	0·02	0·02	0·04	0·05
50	25	0	0·04	0·03	0·02	0·02	0·04	0·03
100	100	2	0·97	0·81	0·79	0·77	0·86	0·84
50	25	2	0·61	0·40	0·37	0·33	0·51	0·41
100	100	4	1·00	1·00	1·00	1·00	1·00	1·00
50	25	4	0·99	0·93	0·99	0·97	0·99	0·92

Table 3 shows that the confidence intervals derived from the score statistics are symmetric about the true value $\theta = 0$ and $\theta = 2$, and mildly skewed to the right about $\theta = 4$.

The confidence intervals derived from the likelihood ratio and Wald statistics have generally similar performance, not shown.

Table 3. *Score statistic confidence intervals about θ, $\alpha = 0{\cdot}05$*

n_p	n_v	θ	F uniform	F normal	F exponential	F Thai
100	100	0	−1·16, 1·12	−1·55, 1·58	−1·38, 1·84	−1·63, 1·88
50	25	0	−1·68, 2·14	−1·81, 3·35	−1·71, 3·17	−1·60, 3·07
100	100	2	0·93, 3·03	0·66, 3·49	0·52, 3·59	0·62, 3·58
50	25	2	0·28, 4·01	−0·15, 4·42	0·18, 4·92	−0·12, 4·68
100	100	4	2·77, 5·50	2·58, 5·53	3·01, 4·78	2·53, 6·11
50	25	4	1·81, 6·60	1·67, 7·26	2·33, 6·41	1·59, 6·68

Wald-based and profile likelihood-based confidence intervals have nearly identical coverage probabilities, and are within one or two percent of the correct probabilities, not shown. They are most accurate when $\theta = 2$ or 4, and slightly conservative for each distribution when $\theta = 0$.

Gaussian quantile–quantile plots of $\hat{\theta}$, not shown, illustrate approximate asymptotic normality of $\hat{\theta}$. The distribution of $\hat{\theta}$ is skewed to the right relative to a normal distribution when the sample size is small and $\theta = 4$, most so when F is exponential.

Figure 3 contains plots of $\hat{F}$ for each baseline distribution, with 95% confidence bands

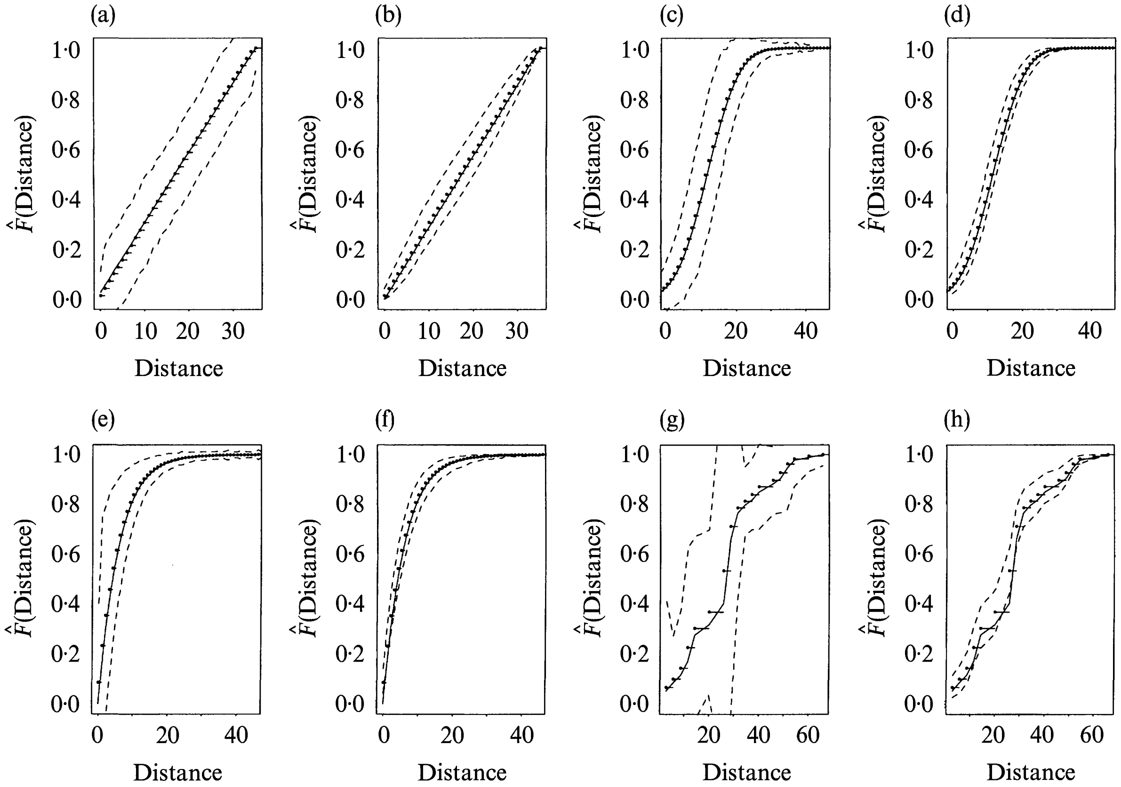

Fig. 3. Datasets generated with $n_p = 100$, $n_v = 50$ and $\theta = 2$. (a), (b) F uniform; (c), (d) F normal; (e), (f), F exponential; (g), (h), F Thai. (a), (c), (e) and (g) show the mean of F across the 1000 replications, with 95% symmetric asymptotic normal approximation confidence bands. The true distribution is depicted by a solid line. (b), (d), (f) and (h) include 95% bootstrap confidence bands.

derived from the bootstrap and from an asymptotic normality approximation using the observed Fisher variance estimate. For all parameter settings the true distribution and its average are superimposed, so evidently $\hat{F}$ is unbiased. The sample variability of $\hat{F}$, not shown, is slightly larger than as indicated by the bootstrap bands, and slightly smaller than as indicated by the normal approximation bands. The shape of the bootstrap confidence bands matches the shape of the sample variability. The estimator $\hat{F}$ behaved very similarly for all other sample sizes and true θ values.

For the three-sample problem, $\hat{\theta}$ performs similarly as in the two-sample problem when the sample sizes are balanced across the groups, but poorer for unbalanced sample sizes. In the unbalanced case $\hat{\theta}$ exhibits positive bias, and the asymptotic variance estimate tends to be greater than the finite-sample variance. Moreover, the likelihood ratio and Wald tests are conservative, although the score test performs as well as in the two-sample problem. The estimator $\hat{F}$ performs as well as in the two-sample problem.

In conclusion, $(\hat{\theta}, \hat{F})$ has satisfactory finite-sample properties. Performance of estimates, tests and confidence intervals for θ is best when the true baseline distribution F is uniform, acceptable when F is normal or Thai, and appreciably worse when F is heavily skewed, i.e. exponential. The semiparametric estimator $\hat{F}$ performed well in all cases.

6. Example

We now illustrate how the generalised logistic regression model can be applied to data arising from a large-scale preventive HIV vaccine trial. Since such a trial has not yet been conducted, we use a pseudo-example in the setting where the first international Phase III trial is underway, in Bangkok. The placebo dataset is formed by randomly sampling 100 distances from the empirical distribution $\hat{F}$ depicted in Fig. 1(b). To construct data for the vaccine group, we sample under the assumption that the vaccine is 50% effective against strains within 10% of prototype, but efficacy declines with V3 loop amino acid sequence divergence, with efficacies 40%, 30%, 20%, 10% and 0% against strains with distances in the ranges 11–20%, 21–30%, 31–40%, 41–50% and 51%+, respectively.

Define the two-sample generalised logistic regression model as in the simulations, with $g(y, \theta) = \frac{1}{70} y\theta$. We fit this model to the generated dataset, which had 69 infections in the vaccine group. We obtain $\hat{\theta} = 1{\cdot}28$, with 95% Wald and profile likelihood-based confidence intervals $(-0{\cdot}26, 2{\cdot}82)$ and $(-0{\cdot}24, 2{\cdot}85)$. The normal approximation and bootstrap variance estimates of $\hat{\theta}$ are 0·62 and 0·59, which are in the range expected from the simulation study. The likelihood ratio, Wald and score statistics are respectively 2·70, 2·65 and 2·89, which all narrowly miss rejecting the null hypothesis of uniform protection at the 0·05 significance level.

According to the model, the estimated vaccine relative risk varies with distance according to the function

$$\widehat{\mathrm{RR}}(y) = \exp(\tfrac{1}{70} y\theta)\, \widehat{\mathrm{RR}}(0) = \exp(0{\cdot}018y)\, \widehat{\mathrm{RR}}(0).$$

To illustrate the interpretation, set $\bar{y}_B = 11{\cdot}24\%$ and $\bar{y}_E = 34{\cdot}87\%$, the average distances of Thai subtype B and E sequences computed from the Los Alamos library. The model estimates that the vaccine protects

$$\mathrm{RR}(34{\cdot}87)/\mathrm{RR}(11{\cdot}24) = \exp\{0{\cdot}018(34{\cdot}87 - 11{\cdot}24)\} = 1{\cdot}53$$

times better against exposing strains with distance $\bar{y}_B$ than against exposing strains with distance $\bar{y}_E$. As seen in Fig. 4(a), the true relative risk ratio is closely estimated for distances

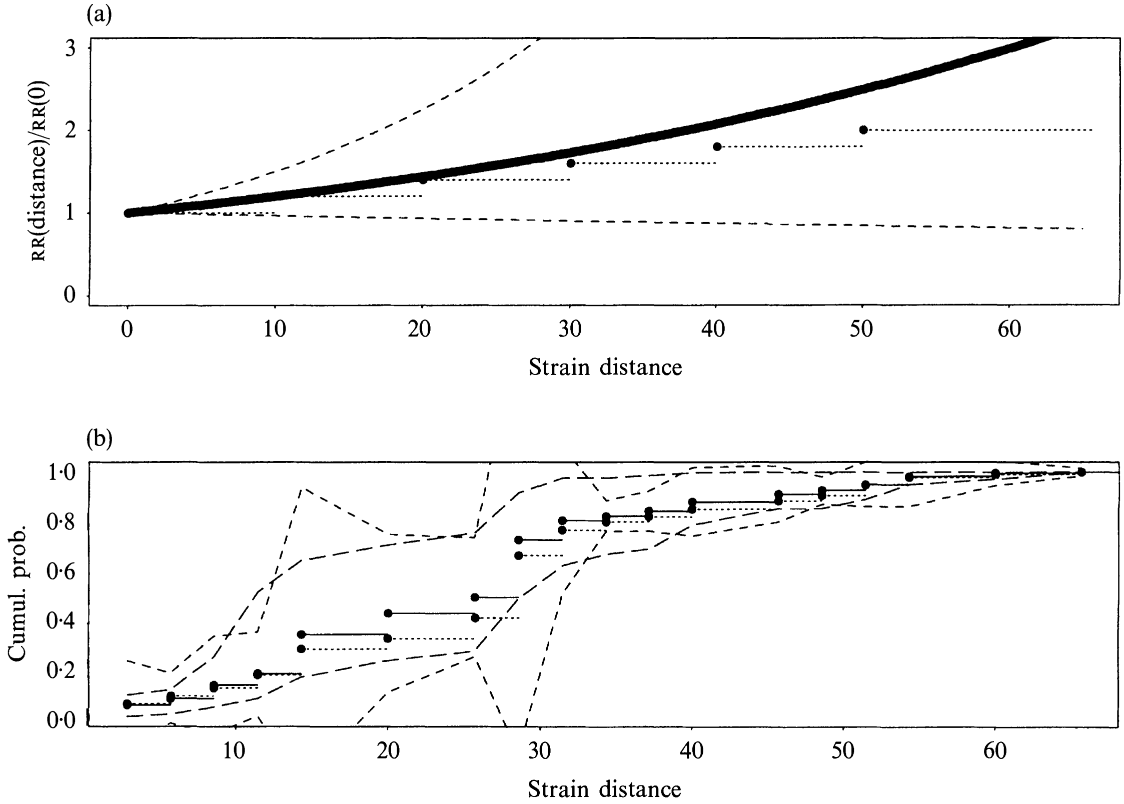

Fig. 4. (a) Vaccine protection: estimated ratio of relative risks $\widehat{\text{RR}}(y)/\widehat{\text{RR}}(0)$ versus strain distance y displayed as a solid line. The broken lines are profile likelihood-based confidence intervals, and the dotted line step function is the true relative risk ratio. (b) $\hat{F}$ shown as solid lines, with 95% asymptotic normal approximation confidence bands as short dashed lines and 95% bootstrap confidence bands as long dashed lines. The true F is portrayed as dotted lines.

of less than 30% and overestimated for large distances. Figure 4(b) shows that $\hat{F}$ slightly underestimates F.

In conclusion, the model detects decreasing vaccine protection with increasing strain distance, but fails to describe accurately the precise dependency of the relative risk ratio on distance. This is because of the strong parametric form of the weight function w_1 imposed by the linear function g. Instead, w_1 could be given a richer parametric form, or estimated by kernel methods or smoothing splines. Alternatively, if $f = dF/d\mu$ is given a known or parametric form, the distribution function

$$W_1(y) = \int_{-\infty}^{y} w_1(z)\,dz \Bigg/ \int_{-\infty}^{\infty} w_1(z)\,dz$$

could be estimated nonparametrically by a maximum partial likelihood procedure like Steps 1–3 with f playing the role of the weight function.

7. Discussion

Although we have taken F to be the cumulative distribution function of a random variable defined on $\mathcal{Y} = R$, all the results presented here hold generally for $\mathcal{Y}$ a sample space with a σ-field of subsets $\mathscr{B}$. If F is an unknown probability measure on $(\mathscr{Y}, \mathscr{B})$, and $w_1, \ldots, w_s$ are nonnegative, measurable weight functions defined on $\mathscr{Y}$, then the corresponding biased probability measures $F_1, \ldots, F_s$ are modelled by

$$F_i(A) \equiv W_i(\theta, F)^{-1} \int_A w_i(u, \theta)\, dF(u) \quad (A \in \mathscr{B};\ i = 1, \ldots, s).$$

An application is the generalised logistic regression model of a multivariate distance $Y \equiv (Y_1, \ldots, Y_k)' \in \mathscr{Y} \subset [0, \infty)^k$. This model is practically important since HIV vaccine protection is likely to vary by variations in several attributes of the virus. The multivariate generalised logistic regression model allows assessment of how vaccine protection depends jointly on distances and marginally on each distance adjusted for the other distances. An example of a useful multivariate model is the two-sample bivariate model specified by

$$w_1(y_1, y_2, \theta) = \exp(y_1\theta_1 + y_2\theta_2 + y_1y_2\theta_3), \quad w_2 \equiv 1.$$

The generalised logistic regression model has many other applications, for example to assess differential protection of a vaccine against any pathogen that exhibits variation, and, for any treatment comparison trial with a failure time endpoint, to assess how treatment efficacy varies by the time Y since treatment initiation.

Elsewhere we will describe the desirable large sample properties of $(\hat{\theta}, \hat{F})$.

In conclusion, the semiparametric maximum likelihood estimator shares many properties with the semiparametric maximum likelihood estimator in Cox's proportional hazards model. These include a simple computational procedure through maximisation of a smooth log profile partial likelihood, comparable finite-sample properties and optimal asymptotic properties. This is not surprising, as the s-group proportional hazards model, defined by $\lambda(y, \theta \mid i) = \exp(\theta_i)\lambda(y \mid s)$ for $i = 1, \ldots, s$, $\theta = (\theta_1, \ldots, \theta_s)'$, $\theta_s \equiv 0$, has the analytic form of an s-sample selection bias model, albeit with weight functions depending on the infinite-dimensional parameter F, with $w_i(y, \theta, F) = \{1 - F(y)\}^{\exp(\theta_i) - 1}$, where F is the cumulative distribution function of Y for the sth group.

Acknowledgement

The authors are grateful to the editor, associate editor and a referee for valuable comments on the paper.

Appendix

Proofs

We sketch the proofs of Theorems 1, 2, 3 and 5, and present more fully the proofs of Theorems 4 and 6. See P. B. Gilbert's dissertation for details of all the proofs.

The proofs of Theorems 1–3 are straightforward, using the Radon–Nikodym Theorem (Ash, 1972, p. 63). The proof of Theorem 5 proceeds by directly verifying, using the Cauchy–Schwarz inequality and strong connectivity of the graph $\mathscr{F}(\theta)$, that the Hessian of the logarithm of the profile partial likelihood is negative definite unless the vector $\dot{w}_1(y, \theta)/w_1(y, \theta)$ is degenerate at the unit vector.

Proof of Theorem 4. Evidently $\sup_{\theta, p} \log\{L(\theta, p \mid y)\}$ equals

$$\sup_\theta \left(\sum_{i=1}^{s} \sum_{j=1}^{h} n_{ij} \log w_{ij}(\theta) + \sup_p \left[\sum_{j=1}^{h} r_j \log p_j(\theta) - \sum_{k=1}^{s} n_k \log \left\{ \sum_{j=1}^{h} w_{kj}(\theta) p_j(\theta) \right\} \right] \right),$$

where the suprema are for $\theta \in \Theta$ and p being a discrete probability measure. Since the graph $\mathscr{F}(\theta)$ is strongly connected, by Vardi (1985, p. 186) the inner maximisation has a unique solution, given by

$$p_j(\theta) \propto r_j \Big/ \sum_k n_k w_{kj}(\theta) V_k^{-1}(\theta) \quad (j=1,\ldots,h),$$

where $(V_1(\theta), V_2(\theta), \ldots, V_{s-1}(\theta), 1)'$ is the unique solution of (4·4). Therefore, it suffices to show that, when $p(\theta)$ is substituted into $L(\theta, p \mid y)$, we obtain the partial likelihood function L_1 of (4·3) up to a constant which does not depend on the parameters.

To this end write

$$q_j(\theta) = r_j \Big/ \left\{\sum_k n_k w_{kj}(\theta) W_k^{-1}(\theta, p)\right\}, \quad p_j(\theta) = \gamma(\theta) q_j(\theta) = \left\{\sum_j q_j(\theta)\right\}^{-1}. \qquad (\text{A}\cdot 1)$$

The form of $p_j(\theta)$ and equations (A·1) imply that $V_i(\theta) = \gamma(\theta) W_i(\theta, p)$ $(i = 1, \ldots, s)$ and $\gamma(\theta) = W_s^{-1}(\theta, p)$. Since $W_s(\theta, p) = 1$ is independent of θ and F by hypothesis, it follows that

$$\frac{w_{ij}(\theta) p_j(\theta)}{W_i(\theta, p)} = \frac{r_j w_{ij}(\theta) W_i^{-1}(\theta, p)}{\sum_k n_k w_{kj}(\theta) W_k^{-1}(\theta, p)}.$$

Substituting this into (4·2) shows that the full likelihood equals the partial likelihood (4·3) times the constant $n^{-1} \prod_j r_j^{r_j}$. □

Proof of Theorem 6. (i) As in Pollard (1990, § 14), it is easily shown that, for fixed θ, $l_1(.\mid \theta, y) \equiv \log L_1(.\mid \theta, y)$ is concave in the transformed variables $Z_i \equiv \log(V_i)$ $(i = 1, \ldots, s)$. Moreover, it is constant along lines through the origin, since it is homogeneous of degree zero. By the concavity, the maximising line is obtained by setting

$$(\partial/\partial V_i) l_1(V \mid \theta, y) = 0 \quad (i = 1, \ldots, s),$$

and putting $V_s = 1$ fixes a unique point along the maximising line.

(ii) This is proved by algebraically manipulating the second derivatives of $l_1(.\mid V, y)$ into an expression which is seen by inspection to be negative under the condition. □

References

Ash, R. B. (1972). *Measure, Integration, and Functional Analysis.* New York: Academic Press.

Berman, P. W., Gray, A., Ashby, M., Eastman, D., Wrin, T., Vennari, J. A., Francis, D., Gregory, T., Fast, P., Schwartz, D., Gorse, G. & McElrath, M. J. (1997). Genetic and immunologic characterization of viruses infecting MN-rgp 120 vaccinated volunteers. *J. Inf. Dis.* **176**, 384–97.

Bhattacharya, P. K., Chernoff, H. & Yang, S. S. (1983). Nonparametric estimation of the slope of a truncated regression. *Ann. Statist.* **11**, 505–14.

Cox, D. R. (1969). Some sampling problems in technology. In *New Developments in Survey Sampling*, Ed. N. L. Johnson and M. Smith, Jr., pp. 506–17. New York: Wiley Interscience.

Gilbert, P. B., Self, S. G. & Ashby, M. (1998). Statistical methods for assessing differential vaccine protection against HIV types. *Biometrics* **54**, 799–814.

Gill, R. D., Vardi, Y. & Wellner, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *Ann. Statist.* **16**, 1069–112.

Patil, G. P. & Rao, C. R. (1977). The weighted distributions: A survey of their applications. In *Applications of Statistics*, Ed. P. R. Krishnaiah, pp. 383–405. Amsterdam: North-Holland.

Patil, G. P. & Rao, C. R. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics* **34**, 179–89.

Patil, G. P., Rao, C. R. & Zelen, M. (1988). Weighted distributions. In *Encyclopedia of Statistical Sciences* **9**, Ed. S. Kotz and N. L. Johnson, pp. 565–71. New York: Wiley.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*, **2**, NSF-CBMS Regional Conference Series in Probability and Statistics 2. Hayward, CA: Inst. Math. Statist. and Am. Statist. Assoc.

Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., Fluornoy, N., Farewell, V. T. & Breslow, N. E. (1978). The analysis of failure time in the presence of competing risks. *Biometrics* **34**, 541–54.

PRENTICE, R. L. & PYKE, R. (1979). Logistic disease incidence models and case control studies. *Biometrika* **66**, 403–11.

SUN, J. & WOODROOFE, M. (1997). Semi-parametric estimates under biased sampling. *Statist. Sinica* **7**, 545–75.

VANIYAPONGS, T., KAMPANARTSANYAKORN, C., VANICHSENI, S., APAIWONGSE, O., RAKTAM, S., KASEMSOOK, R., WASI, C., KITAYAPORN, D., MASTRO, T. D., DES JARLAIS, D. C., HEYWARD, W. L. & ESPARZA, J. (1996). Feasibility of an HIV-1 vaccine efficacy trial among injecting drug users (IDUs) in Bangkok, Thailand. In *XI International Conference on AIDS*, Vancouver, B.C., Ed. R. Sussel, V. Guillet and N. Ruedy, abstract We.C.214.

VARDI, Y. (1982). Nonparametric estimation in the presence of length bias. *Ann. Statist.* **10**, 616–20.

VARDI, Y. (1985). Empirical distributions in selection bias models. *Ann. Statist.* **13**, 178–203.

[*Received September* 1996. *Revised March* 1998]