Statistica Sinica **10**(2000), 1179-1198

ESTIMATING THE DISTRIBUTION OF AGE AT ONSET FROM CASE-CONTROL FAMILY DATA

Emilia Bagiella and Daniel Rabinowitz

Columbia University

Abstract: Motivated by a design for discovering associations between inherited childhood risk factors and adult-onset disease, the problem of estimating the distribution of age at onset of the disease from a case-control sample of subjects who have and have not yet experienced onset is examined. An embedding of the distribution function of age at onset in a multiplicative intercept model simplifies estimation by allowing the distribution function of age at enrollment to be conditioned out of the likelihood. A class of estimators of the distribution function of age at onset is developed and it is argued that a member of the class is efficient. Standard error calculations and an approach to approximating the efficient member of the class are described.

Key words and phrases: Efficiency, Lynden-Bell estimator, multiplicative intercept, one sample problem, truncated data.

1. Introduction

Exploring associations between an adult-onset disease and risk factors that appear during childhood can contribute to the understanding of the disease's etiology and to the development of early screening and intervention programs. A prospective cohort study in which children are enrolled and followed until onset of the disease would allow straightforward analysis of the association between childhood measurements and the disease. However, if onset of the disease occurred late in life, the followup time might be prohibitively long and, if the disease were rare, a prohibitively large sample might be required. A case-control study in which adults with and without the disease are enrolled would avoid these problems. See, for example, Breslow and Day (1980). However, such a design would not be useful unless childhood measurements were available retrospectively.

Shea (1994), in the context of exploring associations between childhood serum lipid levels and early onset cardiovascular disease, proposed an alternative study design that avoids the problems associated with long followup and rare disease. That design calls for a case sample of families with a parent who has experienced onset of the disease and a control sample of families with disease-free parents. Only families with children eligible for measurement of the risk factors

are enrolled. The risk factors measured in the children are used as surrogates for the parent's childhood values.

Use of the alternative design is predicated on three conditions. The first condition is that having a child eligible for enrollment is independent of the disease. The second is that enrollment of the case families and of the control families are independent of children's risk factors. The third is that a heritable factor induces the association between the disease and the risk factors through its influence on both. The first and second conditions are necessary to avoid selection bias. The third condition ensures that the association between a parent's disease and the parent's childhood risk factor values is reflected, through inheritance, in the association between the children's risk factors and the parent's disease. It is this condition that provides the basis for using the children's risk factors as surrogates for the parents' childhood values.

The statistical analysis of data from the alternative design involves several issues. Among them are bivariate failure times (brought about through pairs of parents) and measurement error in the predictors (brought about through imperfect correlation between the parents' childhood risk factors and the children's risk factors). The issue of bivariate failure times is not generally a concern as analyses would often be stratified on the parents' sex. An accounting for the measurement error issue is beyond the scope of this paper. As a first step towards a more general setting, efforts here are restricted to the one sample problem of estimating the marginal distribution of age at onset. Attention is focussed on characterizing the information available in data from the alternative design, and on developing a strategy for efficiently extracting the information. The methods may be used directly as a building block in sub-group comparisons.

A statistical formulation of the one sample problem in the case of one parent per family is that the data are drawn from a population of independent age at enrollment and age-at-onset pairs. Cases are drawn from the sub-population where age at onset precedes age at enrollment. Controls are drawn from the sub-population where age at enrollment precedes age at onset. In the cases, both age at onset and age at enrollment are observed. In the controls, only age at enrollment is observed. The assumption of independence between age at enrollment and age at onset is equivalent to an assumption of negligible change in the distribution of age at onset over the range of age at enrollment represented in the population.

The data are naturally thought of as divided into two parts: information from the case families and information from the controls. However, another division proves quite useful. In this, one part of the data is the age at enrollment information in both the cases and controls, together with case-control indicators. The other part is conditional age-at-onset information given age at enrollment

data in the cases. The case-control part may be thought of as a case-control sample where the dichotomous outcome is presence or absence of disease and the covariate is age at enrollment. The truncated part may be thought of as a sample of right truncated failure time data, where the truncation time is age at enrollment and the truncated failure time is the age at onset.

The case-control part of the data identifies the logit of the distribution of age at onset up to an additive constant. See, for example, Breslow and Storer (1985). The truncated data part identifies the marginal distribution of age at onset up to a multiplicative constant. See, for example, Woodroofe (1985). Together, therefore, the two parts identify the marginal distribution of age at onset. Extracting the available information is a matter of efficiently combining the two parts of the data.

Because the first part of the data is a case-control sample, it is helpful to examine the usual strategy for the statistical analysis of case-control data. The usual strategy is to assume a multiplicative intercept model for the conditional expectation, given the covariates, of the dichotomous outcomes, and to proceed as if the data had been obtained through cross-sectional, rather than case-control, sampling. Efficient estimates of the logit of the conditional expectation, up to an additive constant, result from this strategy, and the usual standard error calculations based on the assumption of cross-sectional sampling are correct as well. The advantage of the strategy is that, with cross-sectional sampling, the covariates are ancillary for the conditional expectation and the marginal distribution of the covariates may be conditioned out of the likelihood without loss of information. See, for example, Anderson (1972), Scott and Wild (1986) and Weinberg and Wacholder (1993). If a non-multiplicative intercept model for the conditional expectation is specified, then a possible strategy is to imbed the specified model in a multiplicative intercept model and to proceed as if the data had been obtained through cross-sectional sampling. The marginal distribution of the covariates may be conditioned out of the likelihood and the logit of the true conditional expectation is efficienctly estimated. See, for example, Breslow and Storer (1985), Manski and McFadden (1981) or Cosslett (1981).

The methodology developed here rests on a generalization of the usual strategy for the analysis of case-control data. As with the usual strategy, the generalization involves proceeding as if the case-control indicators had been obtained through cross-sectional rather than case-control sampling with the conditional expectation of the indicators following a multiplicative intercept model. In the generalization, the conditional distribution of age at onset given age at enrollment in the truncated data part of the likelihood is left unchanged. As with the usual strategy for case-control data, the advantage to developing estimators in the context of the new model with cross-sectional sampling from the multiplicative intercept model, is that the marginal distribution of age at enrollment may

be conditioned out of the likelihood. In the following, the new model with cross-sectional sampling from the multiplicative intercept model will be referred to as the *modified version* of the model, while the true model with the case-control sampling will be referred to as the *original version* of the model.

The remainder of the paper is organized as follows. In the second section, notation is defined. In the third, the validity of developing estimators in the context of the modified version of the model is justified. In the fourth, a class of estimators and standard error calculations for the estimators are developed in the context of the modified version of the model. Direct calculations in an appendix show that the estimators and standard error calculations developed in the context of the modified version of the model are valid in the original model. In the fifth section the issue of efficiency is examined, and in the sixth section the results of some small simulation experiments are presented.

2. Notation and Likelihoods

This section describes notation and likelihoods for the original and modified versions of the model. First, the original version of the model is described. Let T and A denote a generic pair of independent age at onset and age at enrollment times, and let Y denote the indicator that onset precedes enrollment, $Y = 1_{\{T \leq A\}}$. Let F denote the distribution function of age at onset, and let G denote the distribution function of age at enrollment. Let n denote the number of cases, and let m denote the number of controls in the original version of the model.

Settings in which onset does not always eventually occur are of interest. Let π denote the probability that onset eventually occurs, $\pi = \lim_{t \longrightarrow \infty} F(t)$, and let F_0 denote the conditional distribution of age at onset, given that onset occurs, $F_0(t) = F(t)/\pi$.

In the original version of the model, cases are a sample of pairs, $(T_i, A_i),\ i = 1, \ldots, n$, from the conditional distribution of age at onset and age at enrollment, given that onset occurs before enrollment. Controls are a sample, $A_i,\ i = n+1, n+2, \ldots, n+m$, from the conditional distribution of age at enrollment, given that onset has not occurred before enrollment. The likelihood for the cases is $\prod_{i=1}^{n} dG(A_i)dF(T_i)/\int dG(a)F(a)$, and the likelihood for the controls is $\prod_{i=1}^{m} dG(A_i)(1-F(A_i))/\int dG(a)(1-F(a))$. Let $Y_i,\ i = 1, \ldots, n+m$, be the indicator of case-control status. The joint likelihood may be partitioned into the likelihood for the case-control part of the data and the likelihood for the truncated data part,

$$\prod_{i=1}^{n+m} \frac{dG(A_i)F(A_i)^{Y_i}(1-F(A_i))^{1-Y_i}}{\int dG(a)F(a)^{Y_i}(1-F(a))^{1-Y_i}} \times \prod_{j=1}^{n} \frac{dF_0(T_j)}{F_0(A_j)}. \tag{1}$$

The modified version of the model is described next. Let A, T, F, F_0 and π be as in the original version of the model. It is convenient to have different notation for the marginal distribution function of the age at enrollment in the modified and original versions of the model: let H denote the marginal distribution function of age at enrollment in the modified version of the model. Let $\Phi(\alpha, \pi, F_0)$ be defined implicitly by $\text{logit}\,(\Phi(\alpha, \pi, F_0)) = \alpha + \text{logit}(\pi F_0)$. In the modified version of the model, the case-control part of the data is $n + m$ age at enrollment and case-control indicator pairs, (A_i, Y_i), obtained through cross-sectional sampling, with the conditional expectation of Y_i given A_i equal to $\Phi(\alpha, \pi, F_0(A_i))$. The conditional distribution of the truncated age-at-onset part of the data in the modified version of the model is the same as in the original version of the model. In the modified version of the model, the Y_i, and therefore the numbers of cases and controls, are random. Let N and M denote the random numbers of cases and of controls, respectively, in the modified version of the model.

The likelihood in the modified version is $\prod_{i=1}^{n+m} dH(A_i)\Phi(\alpha, \pi, F_0(A_i))^{Y_i}(1 - \Phi(\alpha, \pi, F_0(A_i)))^{1-Y_i} \prod_{j:Y_j=1} dF_0(T_j)/F_0(A_j)$. It follows from the form of the likelihood that, in the modified version of the model, the A_i are ancillary for F. The distribution function H may therefore be conditioned out of the likelihood without loss of efficiency. Note that the parameter space in the modified version of the model is augmented by the multiplicative intercept α.

Finally, it will be convenient to define some backward-in-time counting process notation. For $i = 1, \ldots, n$, let $N_i(u)$ denote minus the indicator that $T_i < u$ and $Y_i = 1$. Let $\delta_i(u)$ be the indicator that T_i is less than u and that A_i is greater than or equal to u. Let Λ be the cumulative hazard in backwards time, $\Lambda(t) = \int_\infty^t -dF_0(s)/F_0(s)$. Let $\mathcal{F}_t$ denote the σ-algebra generated by $Y_i, A_i, N_i(s),\ s > t,\ i = 1, \ldots, n+m$. The compensator of $dN_i(s)$ with respect to the filtration $\{\mathcal{F}_s\}_{s=\infty}^0$ is $d\Lambda(s)\delta_i(s)$. See, for example, Andersen, Borgan, Gill and Keiding (1992). Let $R(t)$ denote the cardinality of the risk set at time t, $R(t) = \sum_{j=1}^{n+m} \delta_i(t)$.

3. Justification for the Modified Version

In this section, a justification for developing methods in the context of the modified version of the model is presented. First some motivation is provided. Then, two results for the behavior of methods developed in the context of the modified model are described. The main point is that methods developed in the context of the modified version of the model may be applied to data that follow the original version of the model.

Motivation for developing methods in the context of the modified version of the model rests on two facts. The first is that in the modified version of the model, the pair (N, M) is ancillary for F. The second is that the conditional

likelihood given $(N, M) = (n, m)$ in the modified version of the model is, up to a reparameterization, equivalent to the likelihood in the original model. These two facts provide motivation because ancillarity suggests that natural approaches to inference in the modified version of the model would condition on (N, M), and because the equivalence of the likelihoods therefore suggests that conditional methods for the modified version of the model would be valid in the original model.

Before moving to the two results about the behavior, in the original model, of estimators developed in the context of the modified model, the two facts that underlie the motivation, the ancillarity result and the equivalence result, are verified. To verify ancillarity, first note that the likelihood in the modified version of the model may be written as the product of the marginal likelihood of the numbers of cases and controls, $\rho^N(1-\rho)^M$, and the conditional likelihood of the ages at enrollment and onset,

$$\prod_{i=1}^{n+m} \frac{dK(A_i)F(A_i)^{Y_i}(1-F(A_i))^{1-Y_i}}{\int dK(a)F(a)^{Y_i}(1-F(a))^{1-Y_i}} \prod_{j:Y_j=1} \frac{dF_0(T_j)}{F_0(A_i)}, \tag{2}$$

where ρ is defined by $\rho = \int dH(a)\Phi(\alpha, \pi, F_0(a))$, and K is defined by

$$dK(a) = \frac{dH(a)}{1-F(a)+e^{\alpha}F(a)} \Big/ \int \frac{dH(u)}{1-F(u)+e^{\alpha}F(u)} \, .$$

To verify ancillarity, therefore, it suffices to show that the range of ρ is unrestricted by the values of K and F. And, by allowing α to range from $-\infty$ to ∞ with F fixed and with

$$dH(a) = \frac{dK(a)(1-F(a)+\exp(\alpha)F(a))}{\int dK(u)(1-F(u)+\exp(\alpha)F(u))},$$

it may be observed that K remains fixed and that ρ ranges from 0 to 1. To verify the equivalence of the likelihoods, note that, from the form of the conditional likelihood (2) in the modified version of the model and the likelihood (1) in the original version, it suffices to show that the range of K is the same as the range of G. And, it may be verified directly that as H ranges over all distribution functions, so does K.

This section concludes with statements and justifications of two results for the behavior, in the original version of the model, of estimators developed in the context of the modified version of the model. The first result is that if an estimator of F is unbiased in the modified version of the model, then it is unbiased in the original version as well. The second result is that with unbiased estimators of F, standard error calculations which are accurate in the modified version of

the model are, under certain regularity conditions, also accurate in the original version.

The approach taken here to justifying these two results relies on a correspondence between (n, m) and G in the original version of the model, and α and H in the modified version. The correspondence is that, for fixed F, for every (n, m) and G in the original version of the model, there are corresponding values of α and H so that $\mathrm{E}N = n$ and $\mathrm{E}M = m$. Then the conditional distribution given $(N, M) = (n, m)$ in the modified version of the model is the same as the distribution in the original version of the model.

The relevance of the correspondence is that when methods developed in the context of the modified version of the model are applied to data from the original version of the model, the resulting estimators behave probabilistically as they would have conditionally given $(N, M) = (n, m)$ in the modified version of the model at the corresponding values of the parameters. Results that hold conditionally in the modified version of the model for all H and α must therefore also hold in the original version for all G and (n, m).

It may be verified by direct calculation that the corresponding values of H and α are defined by

$$dH(a) = \frac{n}{n+m}\frac{dG(a)F(a)}{\int dG(u)F(u)} + \frac{m}{n+m}\frac{dG(a)(1-F(a))}{\int dG(u)(1-F(u))} \tag{3}$$

and

$$\exp(\alpha) = \frac{n}{m}\frac{\int (1-F(u))dG(u)}{\int F(u)dG(u)}. \tag{4}$$

See, for example, Prentice and Pyke (1979), Cosslett (1981) or Hsieh, Manski and McFadden (1985).

Now, justification for the two results is presented. For the first, that estimators which are unbiased in the modified version of the model are unbiased in the original model, note that, in the modified version of the model, for fixed F and K, (N, M) is complete sufficient for ρ. It follows that, in the modified version of the model, for fixed F and K, any unbiased estimator of F is conditionally, given (N, M), unbiased. In particular, the result holds with $K = G$. But, with $K = G$, the conditional distribution of the estimator in the modified version of the model is the same as the distribution of the estimator in the original model.

Justification for the second result is similar. In the modified version of the model, let $\widehat{\theta}$ be an unbiased estimator of $F(t)$, and let $\widehat{\sigma}^2$ be a consistent estimator of the variance of $\widehat{\theta}$. Suppose also that the conditional variance of $\widehat{\theta}$ does not vary substantially over typical values of (N, M). (Such a stability condition would be achieved, for example, by asymptotically linear variance estimators.) From consistency of $\widehat{\sigma}^2$, it follows that $\widehat{\sigma}^2 \approx \mathrm{E}\{\mathrm{Var}\{\widehat{\theta} \mid (N, M)\}\} +$

$\mathrm{Var}\{\mathrm{E}\{\widehat{\theta}\,|(N,M)\}\}$. From the conditional unbiasedness of $\widehat{\theta}$ derived above and the stability of the conditional variance, the right hand side above may be approximated by $\mathrm{E}\{\mathrm{Var}\{\widehat{\theta}\,|(N,M) \approx \mathrm{E}(N,M)\}\}$. And, with H and α given by (3) and (4), the conditional variance is the variance of $\widehat{\theta}$ in the original version of the model.

4. Estimation

In this section a class of estimators for F, an asymptotic expansion for the estimators, and standard error calculations are described. The class of estimators is developed in the context of the modified version of the model. As discussed in the introduction and in the previous section, estimators developed in the context of the modified version of the model are valid in the original version of the model. The estimators depend on arbitrary weight functions. In the next section, optimal choice of the weights is discussed.

At the core of the estimators proposed here is the Lynden-Bell estimator for truncated data, $\widehat{F}_{LB}(t) = \prod_{i:Y_i=1,\, T_i \geq t}(1 - 1/R(T_i))$. The Lynden-Bell estimator is consistent for F_0. See, for example, Chen, Chao and Lo (1995).

The class of estimators of $F(t)$ proposed here takes the form $\widehat{F}(t) = \widehat{\pi}\widehat{F}_{LB}(t) + \widehat{\pi}\sum_{i=1}^{n+m}(Y_i - \Phi(\widehat{\alpha}, \widehat{\pi}, \widehat{F}_{LB}(A_i)))\gamma_t(A_i)$, where $\widehat{\pi}$ and $\widehat{\alpha}$ are the solutions to a pair of estimating equations $0 = \sum_{i=1}^{n+m}(Y_i - \Phi(\alpha, \pi, \widehat{F}_{LB}(A_i)))\beta(A_i)$. Here, β is a two dimensional column vector of weight functions, and γ_t is a weight function that may depend on t.

Before describing the asymptotic expansion of the estimators, some justification for the class of estimators in the context of the modified version of the model is presented. In the modified version of the model, the right hand side of the estimating equation for π and α, evaluated at the true values of α, π and F_0, has expectation zero. Since $\widehat{F}_{LB}$ is consistent for F_0, this suggests that the solutions to the estimating equations should be consistent for π and α. This in turn suggests that $\widehat{\pi}\widehat{F}_{LB}(t)$ should be consistent for $F(t)$. The second addend in $\widehat{F}(t)$ is included to increase the efficiency of the estimator. Although the second addend contributes additional variance, it can also contribute negative covariance with the Lynden-Bell estimator. With a judicious choice of γ_t, the net effect of the addend is to reduce the overall variance of the estimator.

Now, turn to an asymptotic expansion for the estimator. The expansion is derived in the appendix using first order Taylor expansions of the estimator and of the estimating equation around the true values of F_0, α and π, an expansion of the Lynden-Bell estimator, and substitutions based on law of large numbers results. The expansion takes the form

$$\widehat{F}(t) - F(t) \approx \sum_{i=1}^{n+m} (Y_i - \Phi(\alpha, \pi, F_0(A_i))\, \mu_t(A_i)$$

$$+ \sum_{i:Y_i=1} \int_{\infty}^{0} (dN_i(s) - d\Lambda(s)\delta_i(s))\, \psi_t(s),$$

where $\mu_t(a)$ and $\psi_t(s)$ are defined by (8) and (9) in the appendix. As derived in the appendix, the expansion is valid in the original version of the model, if in the modified version of the model, H and α are interpreted as the values given by (3) and (4).

Now, the asymptotic expansion is used to develop standard error calculations. In the modified version of the model, the first addend in the expansion is a sum of independent identically distributed variables with variance

$$(n+m) \int dH(a)\Phi(\alpha, \pi, F_0(a))(1 - \Phi(\alpha, \pi, F_0(a)))\mu_t^2(a). \tag{5}$$

The variance of the second term is the expected predictable quadratic variation of the stochastic integral, $\int_{\infty}^{0} d\Lambda(s) ds \mathrm{E} R(s) \psi_t^2(s)$. The two terms are uncorrelated, as the second term has conditional expectation zero given the Y_i and the A_i. In the original version of the model, the first term is not a sum of identically distributed variables. Nevertheless, as verified in the appendix, if H and α are interpreted as the values given by (3) and (4), then the variance formula is correct in the original version of the model.

The formulae for the variances of the two addends in the asymptotic expansion suggest that the variance of $\widehat{F}(t)$ may be consistently estimated by

$$(n+m) \int d\widehat{H}(a)\Phi(\widehat{\alpha}, \widehat{\pi}, \widehat{F}(a)/\widehat{\pi})(1 - \Phi(\widehat{\alpha}, \widehat{\pi}, \widehat{F}(a)/\widehat{\pi}))\widehat{\mu}_t^2(a)$$
$$+ \int_{\infty}^{0} d\widehat{\Lambda}(s) R(u) \widehat{\psi}_t^2(s),$$

where $\widehat{\mu}_t(a)$ and $\widehat{\psi}_t(s)$ are defined by (10) and (11) in the appendix, $\widehat{\Lambda}$ is the truncated data version of the Nelson-Aalen estimator, $d\widehat{\Lambda}(s) = \sum_{i:Y_i=1} dN_i(s)/R(s)$, and $\widehat{H}$ is the empirical distribution function of the A_i. Bagiella (1997) discusses regularity conditions for the normalized estimator $(n+m)^{1/2}(\widehat{F}(t) - F(t))$ to converge in distribution, as a process in t, to a continuous Gaussian process.

5. Efficiency

This section is concerned with the asymptotic efficiency of the class of estimators. The main result is that it is possible to choose weight functions γ_t and β so that $\widehat{F}(t)$ achieves the semiparametric efficiency bound for estimation of $F(t)$. An approach to estimating the optimal weight functions is also described.

To show that a regular estimator achieves the semiparametric efficiency bound, it suffices to show that the estimator is asymptotic to an element of the

tangent space. See, for example, Stein (1956), Newey (1990) or Bickel, Klaassen, Ritov and Wellner (1993). So, to determine the optimal weight functions it suffices to find weight functions for which the corresponding estimator is in the tangent space. In the appendix, such weight functions are characterized as the solution to an integral equation by comparing the form of scores in parametric submodels to the asymptotic expansion of the estimator.

Although the concern here is with efficiency in the original version of the model, the following argument shows that it suffices that the estimator be asymptotic to an element of the tangent space for the modified version of the model at $K = G$. That is, if an estimator is efficient in the context of the modified version of the model, then it is efficient when applied in the context of the original version of the model. Thus, when deriving the integral equations for the optimal weight functions, the derivation is made in the context of the modifed version of the model.

From the form of the conditional likelihood in (2), it is apparent that, in the original version of the model, the derivative of the log-likelihood for the modified version of the model, along a parametric submodel at $K = G$, is equal to an element of the tangent space for the original version of the model plus a term proportional to

$$N/\rho - M/(1-\rho). \tag{6}$$

The additional term is the derivative of the marginal log-likelihood for the Y_i, and is therefore uncorrelated, at $K = G$ in the modified version of the model, with the score from the conditional likelihood. It follows that if an estimator is in the tangent space for the modified version of the model, and if the estimator is uncorrelated with (6), then the estimator is in the tangent space for the original version of the model. To complete the argument, it is sufficient to show that asymptotically unbiased estimators in the modified version of the model are asymptotically uncorrelated with (6). (See, for example, Bickel, Klaassen, Ritov and Wellner (1993)). But this follows directly from the fact that the correlation is the derivative with respect to ρ, with F and K fixed, of the expectation of the estimator.

It is shown in the appendix that equating the form of scores in parametric submodels with the asymptotic expansion of the estimator results in integral equations for γ_t and β: $\gamma_t(a) = b_t(a) + \int \gamma_t(u)W(a,u)du$ and $\beta(a) = (1, F_0(a)/[F(a)(1-F(a))])^T + \int \beta(u)W(a,u)du$, where W and b_t are defined by (13) and (12) in the appendix. The form of the integral equations suggests estimating the optimal weight functions as the solutions to the matrix equation approximations to the integral equations: $\gamma_t(a) = \widehat{b}_t(a) + \sum_{i=1}^{n+m} \gamma_t(A_i)\widehat{W}(a, A_i)$,

$a \in \{A_1, \ldots, A_{n+m}\}$, and

$$\beta(a) = \begin{pmatrix} 1 \\ \frac{\widehat{F}(a)/\widehat{\pi}}{\widehat{F}(a)(1-\widehat{F}(a))} \end{pmatrix} + \sum_{i=1}^{n+m} \beta(A_i)\widehat{W}(a, A_i), \quad a \in \{A_1, \ldots, A_{n+m}\},$$

where $\widehat{W}$ and $\widehat{b}_t$ are defined in the appendix in (15) and (14) respectively. See, for example, Kress (1989).

6. Simulation Results

The results of some simulation experiments that explore the moderate-sample behavior of the estimator and the efficiency gain over the Lynden-Bell estimator are reported. With degenerate distributions, the Lynden-Bell estimator is not consistent. So, to compare the proposed estimator to the Lynden-Bell estimator, attention was restricted to the case where π is known to be 1 and therefore need not be estimated. In this case, the estimator reduces to $\widehat{F}(t) = \widehat{F}_{LB}(t) + \sum_{i=1}^{n+m}(Y_i - \Phi(\widehat{\alpha}, \widehat{F}_{LB}(A_i)))\gamma_t(A_i)$, where $\widehat{\alpha}$ is the solution to the estimating equation $0 = \sum_{i=1}^{n+m}(Y_i - \Phi(\alpha, \widehat{F}_{LB}(A_i)))\beta(A_i)$, and the estimates of the optimal $\gamma_t(A_i)$ and $\beta(A_i)$ are obtained as solutions of the matrix equations $\gamma_t(a) = \widehat{b}_t(a) + \sum_{i=1}^{n+m}\gamma_t(A_i)\widehat{W}(a, A_i)$, $a \in \{A_1, \ldots, A_{n+m}\}$ and $\beta(a) = 1 + \sum_{i=1}^{n+m}\beta(A_i)\widehat{W}(a, A_i)$, $a \in \{A_1, \ldots, A_{n+m}\}$, where $\widehat{b}_t$ and $\widehat{W}$ are defined as $\widehat{b}_t(a) = \widehat{F}_0(t)(1 - \widehat{F}_0(a))[\int_\infty^{a\vee t} d\widehat{\Lambda}(s)/R(s)]$ and $\widehat{W}(a, u)du =$

$$\frac{\widehat{F}_0(u)}{1 - \widehat{F}_0(a)}\Phi_F(\widehat{\alpha}, \widehat{F}_0(u))d\widehat{H}(u)\int_\infty^{a\vee u}\frac{d\widehat{\Lambda}(s)}{R(s)}.$$

Fortran and the IMSL subroutine library, (IMSL, Inc. (1991)), were used to perform the simulations. The survival times T_i and the enrollment times A_i were randomly generated as Uniform on $(0, 1)$. Cases and controls were sampled separately with $T_i \le A_i$ for the cases and $T_i > A_i$ for the controls. Three experiments were carried out, and in each, one thousand replications were computed. In the first set of replications, there were 200 cases and 200 controls. In the second set there were 200 cases, and 400 controls. In the third experiment there were 200 cases and 100 controls. Estimates of the survival function were calculated at five different values of t: $0.10, 0.25, 0.50, 0.75, 0.90$.

Results of the three sets of simulations are shown in Tables 1, 2 and 3. The sample mean and standard deviations of the Lynden-Bell estimator, the new estimator, the difference, and the sample correlation between the Lynden-Bell estimator and the difference are tabulated. It may be observed that the estimator appears reasonably consistent, though apparently with some systematic bias. The bias is most likely related to the non-linearity of the estimating equations. The relative efficiency of the estimator relative to the Lynden-Bell estimator

ranges from negligible to fairly substantial. See, for example, the case $t = 0.90$ in Table 2. The amount of variance decrease obtained by including the sum over the case-control portion of the data is linked to the amount of (negative) correlation between the Lynden-Bell estimator and the sum. This is born out by the empirical correlations found in the simulations. It should be kept in mind, perhaps, that the Lynden-Bell estimator is not even consistent for π strictly less than 1.

Table 1. Mean and standard deviation of F_{LB}, $\widehat{F}$, and the difference between the two, and the correlation between the difference and F_{LB}, for n=200, m=200.

	$F_{LB}(t)$		$\widehat{F}(t)$		difference		correlation
	mean	std dev	mean	std dev	mean	std dev	
$t = 0.10$	0.10133	0.02129	0.10389	0.02129	0.00256	0.00555	-0.1316
$t = 0.25$	0.25253	0.03722	0.25799	0.03652	0.00547	0.01090	-0.2100
$t = 0.50$	0.50029	0.06339	0.50876	0.05693	0.00848	0.02674	-0.4403
$t = 0.75$	0.75545	0.08095	0.76353	0.06160	0.00808	0.04102	-0.6687
$t = 0.90$	0.91067	0.07673	0.91866	0.05493	0.00799	0.04023	-0.7271

Table 2. Mean and standard deviation of F_{LB}, $\widehat{F}$, the difference between the two, and the correlation between the difference and F_{LB}, for n=200, m=400.

	$F_{LB}(t)$		$\widehat{F}(t)$		difference		correlation
	mean	std dev	mean	std dev	mean	std dev	
$t = 0.10$	0.10068	0.02037	0.10371	0.02029	0.00303	0.00636	-0.1679
$t = 0.25$	0.25041	0.03915	0.25798	0.03757	0.00757	0.01637	-0.3039
$t = 0.50$	0.50506	0.06318	0.51474	0.05729	0.00968	0.02302	-0.4262
$t = 0.75$	0.75393	0.08085	0.76380	0.06164	0.00987	0.03861	-0.6772
$t = 0.90$	0.91076	0.07354	0.91877	0.05079	0.00892	0.04277	-0.7405

Table 3. Mean and standard deviation of F_{LB}, $\widehat{F}$, the difference between the two, and the correlation between the difference and F_{LB}, for n=200, m=100.

	$F_{LB}(t)$		$\widehat{F}(t)$		difference		correlation
	mean	std dev	mean	std dev	mean	std dev	
$t = 0.10$	0.10067	0.02024	0.10297	0.01994	0.00778	0.07777	-0.2306
$t = 0.25$	0.25374	0.03883	0.25844	0.03872	0.00470	0.01660	-0.2204
$t = 0.50$	0.50128	0.06376	0.50829	0.05755	0.00701	0.02225	-0.4402
$t = 0.75$	0.75282	0.08357	0.75773	0.06645	0.00491	0.03528	-0.6565
$t = 0.90$	0.90733	0.07820	0.91418	0.06082	0.00685	0.03705	-0.6538

7. Discussion

Researchers who wish to compare childhood risk factors with the age at onset of adult-onset disease must contend with either long follow-up times and large cohorts, or with difficulties in retrospectively obtaining childhood measurements from adults. In order to side-step these difficulties, a family based case-control sampling design may be used in which children's risk factors are used as surrogates for their parents' childhood values. While the children's values will generally not be perfect surrogates, the design represents a practical solution to what otherwise might be insurmountable difficulties.

Estimation of the distribution of age at onset in the whole sample or when making sub-group comparisons must contend with the case-control aspect of the sampling design. In standard case-control data, the usual approach is to assume a multiplicative intercept model (or to imbed a given model in a multiplicative intercept framework) and then to proceed as if the data had been obtained through random sampling. This approach allows that the infinite dimensional nuisance parameter, the distribution of the covariates, need not be estimated. The generalization of this approach, advocated for the sampling plan considered here, involves imbedding a portion of the model in a multiplicative intercept framework, leaving the remaining portion of the model unchanged, and then proceeding as if the data had been obtained through random sampling. The generalization allows that the infinite dimensional nuisance parameter, the distribution of the age at enrollment, need not be estimated.

The modified version of the model obtained through the imbedding is not proposed as the model for the data. It is simply an analytic technique used to develop estimators. A justification for the technique follows along the same lines as the justification for using prospective logistic regression likelihoods for standard case-control data. Similarly, the efficiency of estimators developed in the context of randomly sampled data from the modified version of the model for case-control data from the original version of the model may be demonstrated by a generalization of the efficiency arguments for the result of using prospective logistic regression likelihoods for standard case-control data.

Here, a family of estimators and standard errors are developed in the context of prospective data from the modified version of the model. Direct calculations are used to verify that the estimators and standard errors are unbiased in the context of case-control data in the original version of the model. It is shown that a particular member of the class is semi-parametric efficient in the context of the modified version of the model, and it is shown how the weight functions that result in the optimal member may be estimated.

Underlying the theory is a correspondence between the modified version of the model and the original version of the model. For any given point in the

parameter space of the original model and for any given ratio of the number of cases to the number of controls, there is a point in the modified version of the model for which the conditional distribution of the data, given the number of cases and controls, is the same as the true distribution of the data, and for which the ratio of the expected number of cases to the expected number of controls is the same as the true ratio. Furthermore, with randomly sampled data in the modified version of the model, the numbers of cases and controls are ancillary. Heuristically, estimators developed in the context of the modified version of the model when applied to data from the original version of the model estimate the corresponding parameters. Furthermore, reasonable estimators condition on the ancillary statistics, the numbers of cases and controls. Thus, estimators for the modified version of the model are correct for the original version of the model.

The approach to estimation is based on splitting the data into two portions, a case-control portion and a truncated data portion. The truncated data portion is used to estimate the conditional distribution of age-at-onset given that onset occurs. The estimator is then combined with the case-control portion of the data to estimate the marginal probability that onset occurs. The family of estimators is obtained by adding a weighted sum of asymptotically unbiased estimators of zero.

Acknowledgement

This work was partially supported by grant GM55978 from the National Institutes of General Medical Sciences.

Appendix

This appendix presents some details about several results. The first of the results is a derivation of the asymptotic expansion of the estimator in the modified version of the model. The second verifies that the estimating equation, in the original version of the model, has expectation zero at α defined by (4). The third shows that the empirical distribution of the A_i in the original version of the model has H, defined by (3), for its expectation. It may be observed from the derivation of the expansion, that the latter two results imply that the expansion is correct in the orginal version of the model. The fourth result is to verify that the variance formula is correct in the original version of the model. The fifth is a derivation of the form of scores in parametric submodels, and the sixth is a derivation of the integral equations for the optimal weight functions. It will be useful to adopt the notation $\Phi_{\alpha,\pi}$ for the gradient of Φ with respect to its first two arguments, and Φ_F for its derivative with respect to its last.

To derive the expansion of the estimator, begin by neglecting the remainder term in a first order Taylor expansion of $\widehat{F}(t) - F(t)$,

$$\begin{aligned}
&\pi \sum_{i=1}^{n+m} (Y_i - \Phi(\alpha, \pi, F_0(A_i)))\gamma_t(A_i) \\
&+\pi(\widehat{F}_{LB}(t) - F_0(t)) - \pi \sum_{i=1}^{n+m} (\widehat{F}_{LB}(A_i) - F_0(A_i))\Phi_F(\alpha, \pi, F_0(A_i))\gamma_t(A_i) \\
&- \left(\pi \sum_{i=1}^{n+m} \Phi_{\alpha,\pi}(\alpha, \pi, F_0(A_i))\gamma_t(A_i) - (0, F_0(t)) \right) \begin{pmatrix} \widehat{\alpha} - \alpha \\ \widehat{\pi} - \pi \end{pmatrix}.
\end{aligned}$$

A similar expansion of the estimating equation leads to

$$\begin{aligned}
\begin{pmatrix} \widehat{\alpha} - \alpha \\ \widehat{\pi} - \pi \end{pmatrix} \approx &\left(\sum_{i=1}^{n+m} \beta(A_i)\Phi_{\alpha,\pi}(\alpha, \pi, F_0(A_i)) \right)^{-1} \\
&\left(\sum_{i=1}^{n+m} (Y_i - \Phi(\alpha, \pi, F_0(A_i)))\beta(A_i) \right. \\
&\left. - \sum_{i=1}^{n+m} \Phi_F(\alpha, \pi, F_0(A_i)))\beta(A_i)(\widehat{F}_{LB}(A_i) - F_0(A_i)) \right).
\end{aligned}$$

Then, substitute and make the law of large numbers approximation

$$\begin{aligned}
&\left(\pi \sum_{i=1}^{n+m} \Phi_{\alpha,\pi}(\alpha, \pi, F_0(A_i))\gamma_t(A_i) - (0, F_0(t)) \right) \\
&\qquad \left(\sum_{i=1}^{n+m} \beta(A_i)\Phi_{\alpha,\pi}(\alpha, \pi, F_0(A_i)) \right)^{-1} \\
\approx &\left(\pi \int dH(a)\Phi_{\alpha,\pi}(\alpha, \pi, F_0(a))\gamma_t(a) - (0, F_0(t)) \right) \\
&\qquad \left(\int dH(a)\beta(a)\Phi_{\alpha,\pi}(\alpha, \pi, F_0(a)) \right)^{-1}. \qquad (7)
\end{aligned}$$

This results in the approximation for the estimator

$$\begin{aligned}
&\sum_{i=1}^{n+m} (Y_i - \Phi(\alpha, \pi, F_0(A_i))) (\pi\gamma_t(A_i) - B\beta(A_i)) \\
&+ \pi \left(\widehat{F}_{LB}(t) - F_0(t) \right) \\
&- \sum_{i=1}^{n+m} \left(\widehat{F}_{LB}(A_i) - F_0(A_i) \right) (\pi\gamma_t(A_i) - B\beta(A_i)) \Phi_F(\alpha, \pi, F_0(A_i)),
\end{aligned}$$

where B is given by (7). Finally, making the approximation $\widehat{F}_{LB}(x) - F_0(x) \approx$

$$F_0(x) \sum_{i:Y_i=1} \int (dN_i(s) - d\Lambda(s)\delta_i(s)ds) \frac{1_{\{s>x\}}}{\mathrm{E}R(s)}$$

(see, for example, Andersen *et al.* (1992)), changing the order of integration and summation and applying another law of large numbers approximation results in

$$\widehat{F}(t) - F(t) \sim \sum_{i=1}^{n+m} (Y_i - \Phi(\alpha, \pi, F_0(A_i)) \mu_t(A_i) + \sum_{i:Y_i=1} \int_{\infty}^{0} (dN_i(s) - d\Lambda(s)\delta_i(s)) \psi_t(s),$$

where

$$\mu_t(a) = \pi\gamma_t(a) - B\beta(a), \tag{8}$$

and $\psi(s) =$

$$\frac{F_0(t)1_{\{s>t\}} - (n+m)\int dH(a)F_0(a)1_{\{s>a\}}\mu_t(a)\Phi_F(\alpha, \pi, F_0(a))}{\mathrm{E}R(s)} \tag{9}$$

Estimates of μ_t and ψ_t that may be used in computing standard errors are given by

$$\widehat{\mu}_t(a) = \widehat{\pi}\gamma_t(a) - \widehat{B}\beta(a) \tag{10}$$

and

$$\widehat{\psi}_t(s) = \frac{\widehat{F}(t)1_{\{s>t\}}/\widehat{\pi}}{R(s)} - \frac{(n+m)\int d\widehat{H}(a)\widehat{F}(a)1_{\{s>a\}}\widehat{\mu}_t(a)\Phi_F(\widehat{\alpha}, \widehat{\pi}, \widehat{F}(a)/\widehat{\pi})}{R(s)} \tag{11}$$

where $\widehat{B} = (\widehat{\pi}\sum_{i=1}^{n+m} \Phi_{\alpha,\pi}(\widehat{\alpha}, \widehat{\pi}, \widehat{F}(A_i)/\widehat{\pi})\gamma_t(A_i) - (0, \widehat{F}(t)/\widehat{\pi}))(\sum_{i=1}^{n+m} \beta(A_i)\Phi_{\alpha,\pi}(\widehat{\alpha}, \widehat{\pi}, \widehat{F}(A_i)/\widehat{\pi}))^{-1}$.

Now it is verified that the estimating equation has expectation zero in the original model.

$$\begin{aligned}
&\mathrm{E}\left\{\sum_{i=1}^{n+m} (Y_i - \Phi(\alpha, \pi, F_0(A_i)))\,\beta(A_i)\right\} \\
&= n\mathrm{E}\left\{\left(1 - \frac{e^{\alpha}F(A)}{1 - F(A) + e^{\alpha}F(A)}\right)\beta(A)\middle| Y = 1\right\} \\
&\qquad + m\mathrm{E}\left\{\left(-\frac{e^{\alpha}F(A)}{1 - F(A) + e^{\alpha}F(A)}\right)\beta(A)\middle| Y = 0\right\}
\end{aligned}$$

$$
\begin{aligned}
&= n \int \frac{1 - F(a)}{1 - F(a) + e^{\alpha} F(a)} \beta(a) \frac{F(a) dG(a)}{\int F(u) dG(u)} \\
&\qquad + m \int \frac{-e^{\alpha} F(a)}{1 - F(a) + e^{\alpha} F(a)} \beta(a) \frac{(1 - F(a)) dG(a)}{\int (1 - F(u)) dG(u)} \\
&= \left(\frac{n}{\int F(a) dG(a)} - \frac{m e^{\alpha}}{\int (1 - F(a)) dG(a)} \right) \int \frac{(1 - F(a)) F(a) dG(a)}{1 - F(a) + e^{\alpha} F(a)} \beta(a).
\end{aligned}
$$

But, the first multiplicand in the last line is 0 by (4).

Now, the validity in the original version of the model of the variance formula (5) is verified. The conditional distribution of the truncated part of the data is the same in the original version of the model as in the modified version, so it suffices to show that, in the original version of the model, with α and H defined by (4) and (3),

$$
\begin{aligned}
&n \mathrm{Var} \left\{ (Y - \Phi(\alpha, \pi, F_0(A)) \mu_t(A) \mid Y = 1 \right\} \\
&+ m \mathrm{Var} \left\{ (Y - \Phi(\alpha, \pi, F_0(A)) \mu_t(A) \mid Y = 0 \right\} \\
&= (n + m) \int dH(a) \Phi(\alpha, \pi, F_0(a)) (1 - \Phi(\alpha, \pi, F_0(a))) \mu_t^2(a).
\end{aligned}
$$

The argument has two steps. The first step is to show that $0 = \mathrm{E}\{(Y - \Phi(\alpha, \pi, F_0(A))\mu_t(A) \mid Y = 1\} = \mathrm{E}\{(Y - \Phi(\alpha, \pi, F_0(A))\mu_t(A) \mid Y = 0\}$. The second step is to compute $n\mathrm{E}\{(Y - \Phi(\alpha, \pi, F_0(A))^2 \mu_t^2(A) | Y = 1\} + m\mathrm{E}\{(Y - \Phi(\alpha, \pi, F_0(A))^2 \mu_t^2(A) | Y = 0\}$. For the first step, it may be calculated that

$$
\begin{aligned}
&\mathrm{E} \left\{ (Y - \Phi(\alpha, \pi, F_0(A)) \mu_t(A) \mid Y = y \right\} = \\
&\int (y - \Phi(\alpha, \pi, F_0(a)) \mu_t(a) \frac{dG(a) (\Phi(\alpha, \pi, F_0(a))^y (1 - \Phi(\alpha, \pi, F_0(a))^{1-y}}{\int dG(u) (\Phi(\alpha, \pi, F_0(u))^y (1 - \Phi(\alpha, \pi, F_0(u))^{1-y}}.
\end{aligned}
$$

Up to multiplication by a function of y, this quantity is $\int dG(a) \Phi(\alpha, \pi, F_0(a))(1 - \Phi(\alpha, \pi, F_0(a)))\mu_t(a)$. Straightforward but tedious computations starting from the definitions of μ_t and B show that this last quantity is zero. For the second step, it may be calculated that

$$
\begin{aligned}
&n \mathrm{E} \left\{ (Y - \Phi(\alpha, \pi, F_0(A))^2 \mu_t^2(A) \mid Y = 1 \right\} \\
&+ m \mathrm{E} \left\{ \left(Y - \Phi(\alpha, \pi, F_0(A))^2 \mu_t^2(A) \mid Y = 0 \right\} \right. \\
&= n \int \left(\frac{1 - F(a)}{1 - F(a) + e^{\alpha} F(a)} \mu_t(a) \right)^2 \frac{F(a) dG(a)}{\int F(u) dG(u)} \\
&+ m \int \left(\frac{-e^{\alpha} F(a)}{1 - F(a) + e^{\alpha} F(a)} \mu_t(a) \right)^2 \frac{(1 - F(a)) dG(a)}{\int (1 - F(u)) dG(u)} \\
&= \int \left(\frac{m e^{\alpha} F(a)}{\int (1 - F(u)) dG(u)} + \frac{n e^{-\alpha} (1 - F(a))}{\int F(u) dG(u)} \right) \frac{e^{\alpha} F(a) (1 - F(a))}{(1 - F(a) + e^{\alpha} F(a))^2} \mu_t^2(a) dG(a) \\
&= (n + m) \int dH(a) \Phi(\alpha, \pi, F_0(a)) (1 - \Phi(\alpha, \pi, F_0(a))) \mu_t^2(a).
\end{aligned}
$$

Now, the form of scores for submodels in the modified version of the model is computed. Because the A_i are ancillary for F in the modified version of the model, the log conditional likelihood given the A_i is taken as the starting point:

$$\begin{aligned}
&\sum_{i=1}^{n+m} Y_i \log(\Phi(\alpha,\pi,F_0(A_i))) + (1-Y_i)\log(1-\Phi(\alpha,\pi,F_0(A_i))) \\
&+ \sum_{j:Y_j=1} \log\left(\frac{dF_0(T_i)}{F_0(A_i)}\right) \\
&= \sum_{i=1}^{n+m} Y_i \log(\Phi(\alpha,\pi,\exp(-\int_{\infty}^{A_i} d\Lambda(s)))) \\
&+(1-Y_i)\log(1-\Phi(\alpha,\pi,\exp(-\int_{\infty}^{A_i} d\Lambda(s)))) \\
&+ \sum_{j:Y_j=1} \int_{A_i}^{T_i} d\Lambda(s) + \log(d\Lambda(T_i)).
\end{aligned}$$

Differentiating along a parametric submodel, $\theta \longrightarrow (\alpha_\theta, F_\theta)$, results in

$$\begin{aligned}
&\sum_{i=1}^{n+m} (Y_i - \Phi(\alpha,\pi,F_0(A_i)))\left(C + \frac{DF_0(A_i) + \pi F_0(A_i)\int_{\infty}^{A_i} d\Lambda(s)\kappa(s)}{F(A_i)(1-F(A_i))}\right) \\
&+ \sum_{j:Y_j=1} \int_{\infty}^{0} (dN_i(s) - d\Lambda(s)\delta_i(s))\kappa(s),
\end{aligned}$$

where $\kappa(s)$ is the derivative with respect to θ of $\log(d\Lambda(s))$, C is the derivative of α and D is the derivative of π.

Finally, a derivation of the integral equations is presented. Matching terms in the form of scores for parametric submodles and the asymptotic expansion result in a pair of equations:

$$\mu_t(a) \;=\; C + \frac{DF_0(a) + \pi F_0(A_i)\int_{\infty}^{a} d\Lambda(s)\kappa(s)}{F(a)(1-F(a))}$$

$$\frac{F_0(t)1_{\{s>t\}} - (n+m)\int dH(u)F_0(u)1_{\{s>u\}}\mu_t(u)\Phi_F(\alpha,\pi,F_0(u))}{\mathrm{E}R(s)} \;=\; \kappa(s).$$

Substituting the left hand side of the second equation for $\kappa(s)$ in the first equation results in $\mu_t(a) \;=\; C + \frac{DF_0(a)}{F(a)(1-F(a))} + \pi b_t(a) + \int W(a,u)du\mu_t(a)$, where

$$b_t(a) \;=\; \frac{F_0(t)F_0(a)}{F(a)(1-F(a))}\int_{\infty}^{a\vee t} \frac{d\Lambda(s)}{\mathrm{E}R(s)}, \tag{12}$$

and $W(a,u)du$ is

$$\frac{-(n+m)F_0(u)F_0(a)}{F(a)(1-F(a))}dH(u)\Phi_F(\alpha,\pi,F_0(s))\int_\infty^{a\vee u}\frac{d\Lambda(s)}{\mathrm{E}R(s)}. \tag{13}$$

It follows that it is sufficient for γ_t and β to satisfy $\gamma_t(a)=b_t(a)+\int\gamma_t(u)W(a,u)du$ and

$$B\beta(a) \;=\left(\begin{array}{c} C \\ \frac{DF_0(A)}{F(a)(1-F(a))}\end{array}\right)+\int B\beta(u)W(a,u)du.$$

Since linear transformations of β leave the estimates of π and α invariant, and since C and D are arbitrary, the equation for β may be replaced by

$$\beta(a) \;=\left(\begin{array}{c} 1 \\ \frac{F_0(a)}{F(a)(1-F(a))}\end{array}\right)+\int\beta(u)W(a,u)du.$$

Given preliminary estimates, $\tilde\alpha$ and $\tilde\pi$ of α and π based on an arbitrary β, estimates of W and b_t that may be used in the matrix equation approximation to the integral equations are given by

$$\widehat{b}_t(a)=\frac{\widehat{F}(t)/\widehat{\pi}\widehat{F}_0(a)}{\widehat{F}(a)(1-\widehat{F}(a))}\int_\infty^{a\vee t}\frac{d\widehat{\Lambda}(s)}{R(s)} \tag{14}$$

and

$$\widehat{W}(a,u)du=\frac{-(n+m)\widehat{F}(u)/\widehat{\pi}\widehat{F}_0(a)}{\widehat{F}(a)(1-\widehat{F}(a))}d\widehat{H}(u)\Phi_F(\widehat{\alpha},\widehat{\pi},\widehat{F}(s)/\widehat{\pi})\int_\infty^{a\vee u}\frac{d\widehat{\Lambda}(s)}{R(s)}. \tag{15}$$

References

Andersen, P. K., Borgan, O., Gill, R. and Keiding, N. (1992). *Statistical Models Based on Counting Processes.* Springer-Verlag, New York.

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika* **59**, 19-35.

Bagiella, E. (1997). Estimating a survival distribution from case-control family data. Columbia University Department of Biostatistics Doctoral Dissertation.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore, Maryland.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research: The Analysis of Case-control Studies.* Vol. 1. World Health Organization.

Breslow, N. E. and Storer, B. E. (1985). General relative risk functions for case-control studies. *Amer. J. Epidemiology* **122**, 149-162.

Chen, K., Chao, M. T. and Lo, S. H. (1995). On strong uniform consistency of the Lynden-Bell estimator for truncated data. *Ann. Statist.* **23**, 440-449.

Cosslet, S. (1981). Efficient estimation of discrete choice models. In *Structural Anal. Discrete Data* (Edited by C. F. Manski and D. McFadden). M.I.T. Press, Cambridge, Massachusetts.

Hsieh, D. A., Manski, C. F. and McFadden, D. (1985). Estimation of response probabilities from augmented retrospective observations. *J. Amer. Statist. Assoc.* **80**, 651-662.

IMSL, Inc. (1991). IMSL MATH/LIBRARY *User's Manual*, Version 2.0. IMSL, Huston.

Kress, R. (1989). *Linear Integral Equations.* Springer-Verlag, New York.

Manski, C. F. and McFadden, D. (1981). Alternative estimators and sample designs for discrete choice analysis, In *Structural Anal. Discrete Data* (Edited by C. F. Manski and D. McFadden). M.I.T. Press, Cambridge, Massachusetts.

Newey, W. K. (1990). Semiparametric efficiency bounds. *J. Appl. Econometrics* **5**, 99-135.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403-412.

Shea, S. (1994). *Biochemical/Genetic Markers for Premature Atherogenesis.* Department of Health and Human Service, Public Health Service Grant R01-HD32195.

Scott, A. J. and Wild, C. J. (1986). Fitting logistic models under case-control or choice based sampling. *J. Roy. Statist. Soc. Ser. B* **48**, 170-182.

Stein, C. (1956). Efficient non-parametric testing and estimation. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 187-195.

Weinberg, C. R. and Wacholder, S. (1993). Prospective analysis of case-control data under general multiplicative-intercept risk models. *Biometrika* **80**, 461-465.

Woodroofe, M. (1985). Estimating a distribution function with truncated data. *Ann. Statist.* **13**, 163-177.

Division of Biostatistics, School of Public Health, Columbia University, New York, NY 10032, U.S.A.

Department of Statistics, Columbia University, New York, NY 10027, U.S.A.

(Received October 1998; accepted December 1999)