Bernoulli **6**(3), 2000, 447–455

On the semi-parametric efficiency of logistic regression under case–control sampling

NORMAN E. BRESLOW[1]*, JAMES M. ROBINS[2] and JON A. WELLNER[1,3]**

[1]*Department of Biostatistics, University of Washington, Seattle WA 98195-7232, USA.*
**E-mail: norm@biostat.washington.edu*
[2]*Department of Epidemiology, Harvard School of Public Health, Boston, MA 02115-6096, USA.*
E-mail: robins@hsph.harvard.edu
[3]*Department of Statistics, University of Washington, Seattle WA 98195-7232, USA.*
***E-mail: jaw@stat.washington.edu*

Using modern theory for semi-parametric models, we provide details for an argument of Robins *et al.* showing efficiency of the standard logistic regression estimator applied to data from case–control studies. Our elaboration of this argument, and of a related one by Bickel *et al.*, includes a constructive new proof of the result.

Keywords: biased sampling; epidemiology; influence function

1. Introduction

Logistic regression is widely used for the analysis of data from case–control studies in epidemiology (Breslow 1996). Two principal reasons for this popularity are that the regression coefficients in the logistic model have a desired interpretation in terms of log-odds ratios; and that odds ratios in models for disease probabilities are estimable from case–control samples. Anderson (1972) made the remarkable discovery that maximum likelihood estimates of logistic regression parameters from case–control samples could be obtained by fitting the standard logistic regression model to the case–control data, ignoring the outcome-dependent nature of the sampling and treating case–control status as a 'random' outcome variable. His results were limited, however, by the requirement that the explanatory variables were discrete so that the joint distribution of outcome and explanatory variables could be specified using a finite number of parameters. Prentice and Pyke (1979) removed this restriction and demonstrated that, whatever the marginal distribution of the explanatory variables, the regression coefficients obtained by fitting the standard logistic model were nonparametric maximum likelihood estimates (NPMLEs). Further argument is required, however, to conclude that the NPMLE achieves the efficiency bounds of modern semi-parametric theory.

Cosslett (1981) showed that the variance of the NPMLEs of parameters in binary

1350–7265

response models achieved a semi-parametric lower bound under case–control sampling. He specifically excluded multiplicative intercept models of which the logistic is the paradigm, however, because of the non-identifiability of the intercept. Robins *et al.* (1994) provided a proof of semi-parametric efficiency for the standard logistic regression coefficients under case–control sampling in the course of a treatise on missing data problems. Their arguments are elaborated further below. Rabinowitz (1997) showed that efficient estimates in non-multiplicative intercept models result if one first enlarges the parameter space to include a multiplicative intercept term.

Bickel *et al.* (1993, Section 4.4) considered case–control sampling as the first of four concrete examples that illustrated their general theory of biased sampling models. They derived a formula for the efficient scores that involved projecting the parametric logistic regression scores onto the sum of two linear function spaces, one of which is one-dimensional. Although they noted that this projection can be evaluated explicitly, they stated that 'the formula is uninstructive'. One purpose of this paper is to show that, on the contrary, calculation of the efficient scores for the biased sampling model using the Bickel *et al.* approach confirms that the standard estimator achieves semi-parametric efficiency. We also argue that this calculation could have been avoided entirely had they adopted the more abstract approach of Robins *et al.*

2. The standard logit model

Let Y denote a binary outcome variable taking values $y = 1$ (for diseased) and $y = 0$ (for non-diseased) and let Z denote a p-vector of explanatory variables. As a binary response model we assume the logistic relationship

$$\Pr(Y = y|Z = z) = f(y|z;\, \theta) = \frac{\exp\{y(\alpha + z^{\mathrm{T}}\beta)\}}{1 + \exp(\alpha + z^{\mathrm{T}}\beta)},$$

where $\theta = (\alpha, \beta^{\mathrm{T}})^{\mathrm{T}}$. Nothing is assumed regarding the marginal distribution H of Z except that it belongs to the collection $\mathscr{H}$ of distributions that have densities h with respect to some measure m. This defines a semi-parametric random sampling model $\mathscr{Q} = (\mathscr{Q}_{(\theta,H)} : \theta \in \mathbb{R}^{p+1}, H \in \mathscr{H})$. We suppose that the data (y_i, z_i), $i = 1, \ldots, n$, constitute a random sample from the joint distribution $Q(Y, Z)$ with density

$$q(y, z;\, \theta,\, h) = f(y|z;\, \theta)h(z). \tag{1}$$

The scores for the parametric part of the model, $\dot{\mathbf{l}}_\theta = (\dot{\mathbf{l}}_\alpha, \dot{\mathbf{l}}_\beta^{\mathrm{T}})^{\mathrm{T}}$, are given by

$$\dot{\mathbf{l}}_\theta = Z^{\mathrm{e}}\{Y - \mathrm{E}(Y|Z;\, \theta)\} \tag{2}$$

where $Z^{\mathrm{e}} = (1, Z^{\mathrm{T}})^{\mathrm{T}}$. The 'tangent space' of scores for the nonparametric part of the model equals $L_2^0(H) = \{a = a(Z)|\mathrm{E}(a) = 0,\ \mathrm{var}(a) < \infty\}$. Since the parametric scores (2) have conditional mean zero given Z, they are orthogonal to $L_2^0(H)$ and thus are the efficient scores for θ in the semi-parametric model where h is unknown. This also follows trivially from Proposition 2 of Bickel *et al.* (1993, Section 4.3). Using standard parametric theory (Bickel *et al.* 1993, Section 2.4), the efficient scores for the odds-ratio parameters β of primary interest are

$$\mathbf{l}_\beta^* = \dot{\mathbf{l}}_\beta - I_{\beta\alpha} I_{\alpha\alpha}^{-1} \dot{\mathbf{l}}_\alpha$$

$$= \left[Z - \frac{\mathrm{E}\{Z \operatorname{var}(Y|Z)\}}{\mathrm{E}\{\operatorname{var}(Y|Z)\}} \right] \{Y - \mathrm{E}(Y|Z)\}, \tag{3}$$

where $I_{\theta\theta} = \mathrm{E}(\dot{\mathbf{l}}_\theta \dot{\mathbf{l}}_\theta^{\mathrm{T}}) = \mathrm{E}\{Z^{\mathrm{e}} \operatorname{var}(Y|Z)(Z^{\mathrm{e}})^{\mathrm{T}}\}$ denotes the Fisher information. The information for β at $Q = Q_{\alpha,\beta,H}$ in the model $\mathscr{Q}$ is thus given by

$$I(Q|\beta, \mathscr{Q}) = \mathrm{E}_Q \left[\left(Z - \frac{\mathrm{E}_Q\{Z \operatorname{var}_Q(Y|Z)\}}{\mathrm{E}_Q\{\operatorname{var}_Q(Y|Z)\}} \right)^{\otimes 2} \operatorname{var}_Q(Y|Z) \right]. \tag{4}$$

The usual logistic regression coefficients $\hat{\theta} = (\hat{\alpha}, \hat{\beta}^{\mathrm{T}})^{\mathrm{T}}$, obtained by applying standard computer programs to the data $\{(y_i, z_i),\ i = 1, \ldots, n\}$, solve the score equations corresponding to (2), namely

$$\sum_{i=1}^{n} z_i^{\mathrm{e}} \{y_i - \mathrm{E}(Y|Z = z_i; \theta)\} = 0. \tag{5}$$

The influence function and asymptotic variance for $\hat{\beta}$ alone are determined by (3) and (4). Compare with equation (20) of Bickel *et al.* (1993, p. 111).

3. The biased sampling model

Whereas the standard model assumes random sampling from (Y, Z), the model for the case–control or retrospective study involves sampling from Z given Y. Specifically, suppose that n_1 cases are drawn from the conditional distribution $(Z|Y = 1)$ and n_0 controls are drawn from $(Z|Y = 0)$. Because separate samples of fixed size are drawn from two subpopulations, this set-up does not strictly correspond to the theory developed by Bickel *et al.* (1993), for which the observations are independent and identically distributed (i.i.d.). Thus we modify the usual definition of the case–control study slightly so that it involves a simple random sample of size n from a biased sampling model, as follows. First, select a case or a control with probabilities λ_1 and $\lambda_0 = 1 - \lambda_1$, respectively. Then sample Z from the appropriate conditional distribution given $Y = 1$ or $Y = 0$. A similar modified sampling design was proposed for choice-based sampling in econometrics by Manski and Lerman (1977) and for case–control studies in epidemiology by Weinberg and Sandler (1991), who call it the randomized recruitment design. It is also known as Bernoulli sampling. The essential difference between it and the usual two-sample retrospective design is that the numbers of cases and controls, n_1 and $n_0 = n - n_1$, are random variables that result from binomial sampling with probability λ_1. The asymptotic distributions of the resulting estimators are the same, whether the subsample sizes are regarded as fixed or random. McNeney (1998) demonstrates that efficiency properties under the i.i.d. set-up also extend to subsamples of fixed size.

The semi-parametric model $\mathscr{P}$ just described is a special case of Example 1 of Bickel *et al.* (1993, Section 4.4): $\mathscr{P} = \{P_{\lambda_1,\theta,H} : \theta \in \mathbb{R}^{p+1},\ H \in \mathscr{H}\}$ where $P_{\lambda_1,\theta,H}$ has density

$$p(y, z; \theta, h) = \lambda_y \frac{f(y|z; \theta)h(z)}{\int f(y|u; \theta)h(u)\,\mathrm{d}m(u)} \tag{6}$$

and the marginal probabilities $P(Y = y)$ are fixed at λ_y by the experimenter. The constant term α is not identifiable in $\mathscr{P}$, which confirms that α is not estimable from case–control data, and the efficient score $\mathbf{l}_\alpha^*$ is identically zero (Bickel *et al.* 1993, p. 118). As will be shown in the next section, this model is strictly contained in $\mathscr{Q}$ and can be extended to $\mathscr{P}^* = \mathscr{Q}$ by allowing λ_1 to vary freely in (0,1).

As shown in equation (15) of Bickel *et al.* (1993, Section 4.4), the efficient score for β is

$$\mathbf{l}_\beta^*(Y, Z) = ZY - \mathrm{E}(ZY) - ACE(ZY|Z, Y). \tag{7}$$

Here $ACE(\cdot|Z, Y)$ denotes the orthogonal projection onto the direct sum of two Hilbert spaces, $L_2^0(H)$ as defined earlier and, since Y takes only two values, the one-dimensional linear space spanned by $Y - \mathrm{E}(Y)$. (All expectations in this section are taken with respect to the biased sampling distribution $\mathscr{P} \in \mathscr{P}$.) According to Appendix A.4 of Bickel *et al.* (1993), especially equations (32)–(37), the projection is given by

$$ACE(b|Z, Y) = \mathrm{E}(b|Z) - \mathrm{E}(b) + \lambda(b)[Y - \mathrm{E}(Y|Z)],$$

where

$$\lambda(b) = \frac{\mathrm{E}[b\{Y - \mathrm{E}(Y|Z)\}]}{\mathrm{E}\{\mathrm{var}(Y|Z)\}}.$$

Applying this formula with $b(Y, Z) = ZY$, we have

$$\lambda(b) = \frac{\mathrm{E}[ZY\{Y - \mathrm{E}(Y|Z)\}]}{\mathrm{E}\{\mathrm{var}(Y|Z)\}} = \frac{\mathrm{E}\{Z\,\mathrm{var}(Y|Z)\}}{\mathrm{E}\{\mathrm{var}(Y|Z)\}}.$$

Inserting these expressions into (7) yields

$$\begin{aligned}\mathbf{l}_\beta^* &= ZY - \mathrm{E}(ZY) - Z\mathrm{E}(Y|Z) + \mathrm{E}(ZY) - \frac{\mathrm{E}\{Z\,\mathrm{var}(Y|Z)\}}{\mathrm{E}\{\mathrm{var}(Y|Z)\}}\{Y - \mathrm{E}(Y|Z)\} \\ &= \left[Z - \frac{\mathrm{E}\{Z\,\mathrm{var}(Y|Z)\}}{\mathrm{E}\{\mathrm{var}(Y|Z)\}}\right]\{Y - \mathrm{E}(Y|Z)\}. \end{aligned} \tag{8}$$

The information for β at $P = P_{\alpha,\beta,H}$ in the model $\mathscr{P}$ is thus given by

$$I(P|\beta, \mathscr{P}) = \mathrm{E}_P\left[\left(Z - \frac{\mathrm{E}_P\{Z\,\mathrm{var}_P(Y|Z)\}}{\mathrm{E}_P\{\mathrm{var}_P(Y|Z)\}}\right)^{\otimes 2} \mathrm{var}_P(Y|Z)\right]. \tag{9}$$

Comparison of equations (3) and (8) shows that the efficient score for β in the random sampling model $\mathscr{Q}$ has exactly the same form as the efficient score for β in the biased sampling model $\mathscr{P}$. Consequently, the information for β also has the same form (equations (9) and (4)). These are precisely the identities anticipated by Robins *et al.* (1994) from the fact that $\mathscr{P}^*$ and $\mathscr{Q}$ correspond to two different parametrizations of the same model. The only difference is that expectations in the random sampling model are taken with respect to Q as defined in equation (1), whereas expectations in the biased sampling model are taken

with respect to P as defined by equation (6). The scores for one model need not, and generally will not, have expectation zero under the other model. We next consider in some detail the arguments of Robins *et al.*

4. Alternate parametrizations and the Robins *et al.* approach

The semi-parametric model (1) can be characterized as the set $\mathcal{Q}$ of all distributions of (Y, Z) that have finite second moments and for which the logarithm of the odds ratio is a linear function of Z (Prentice and Pyke 1979). More explicitly, using the well-known invariance property of the odds ratio (Cornfield 1951), (1) is equivalent to

$$\begin{aligned} \mathrm{OR}(z) &= \frac{f(Y=1|Z=z)f(Y=0|Z=0)}{f(Y=0|Z=z)f(Y=1|Z=0)} \\ &= \frac{f(Z=z|Y=1)f(Z=0|Y=0)}{f(Z=z|Y=0)f(Z=0|Y=1)} = \exp(z^{\mathrm{T}}\beta). \end{aligned} \tag{10}$$

Writing the joint distribution as the marginal of Y times the conditional of Z given Y, it follows that the densities of distributions in $\mathcal{Q}$ may be re-expressed

$$q(y, z; \theta, h) = \pi_y c_y \, \mathrm{e}^{yz^{\mathrm{T}}\beta} g(z) \tag{11}$$

where $c_y^{-1} = \int \exp(yu^{\mathrm{T}}\beta) g(u) \, \mathrm{d}m(u)$. The parameters (α, β, h) and (π, β, g) are related via

$$\pi_y = \Pr(Y = y) = \int f[y, u; \theta(\alpha, \beta)] h(u) \, \mathrm{d}m(u) \quad \text{and} \quad g(z) = \frac{h(z)}{1 + \mathrm{e}^{\alpha + \beta^{\mathrm{T}} z}}.$$

As noted by Prentice and Pyke (1979), (1) and (11) are precisely equivalent, and are also equivalent to (10), provided that the two sets of parameters are unrestricted. Re-expression of the densities in the form (11) provides the reparametrization $\mathcal{P}^*$ of $\mathcal{Q}$ obtained by extending the biased sampling model, as mentioned earlier.

Somewhat more generally (Bickel *et al.* 1993, Section 3.3), the odds-ratio parameter β may be viewed as the value of a mapping $\nu: \mathcal{Q} \to \mathbb{R}^p$. Robins, *et al.* correctly point out that this alone is sufficient to conclude that the semi-parametric efficient scores, influence function and variance bound for the common interest parameters β are identical regardless of which equation is used to define the model. Indeed, as shown explicitly in Proposition 2 of Bickel *et al.* (1993, Section 3.4), the efficient influence function $\tilde{\mathbf{l}}_\beta$ equals $\dot{\nu}(Q)$, where $\dot{\nu}$ is the *pathwise derivative* of ν. Regardless of the parametrization, a regular estimator with this influence function is the semi-parametric efficient estimator of β. Consequently, as argued by Robins *et al.*, if one has demonstrated already that $\hat{\beta}$ is semi-parametric efficient for model (1), it follows that $\hat{\beta}$ is semi-parametric efficient in model (11), and vice versa.

Efficient influence functions and estimators are usually not determined by taking functional derivatives but rather by working with the scores that arise from the semi-parametric model. The advantage of (1) for random sampling is that the covariates z are ancillary for θ and hence, as noted earlier, efficient estimation of β need only consider the parametric part of the model specified by $f(y|z; \theta)$. The advantage of the alternate

parametrization only becomes evident for biased sampling. The biased sampling model (6) is equivalent to

$$p(y, z; \beta, g) = \lambda_y c_y e^{yz^T\beta} g(z), \tag{12}$$

which is identical to (11) except that $\pi_1 = \Pr(Y = 1)$ is now fixed at λ_1. In fact an equivalence class of random sampling models (1) with parameters $(\gamma, \beta, \tilde{h})$ generates the same biased sampling model (12). Roeder *et al.* (1996) calculated explicitly the relationship, depending on the associated marginal probabilities π_1 and $\tilde{\pi}_1$, that must hold between two different members of this class with parameters (α, β, h) and $(\gamma, \beta, \tilde{h})$, respectively. One member has $\tilde{\pi}_1 = \lambda_1$, the sampling fraction specified by the biased sampling design.

The advantage of the alternate parametrization for biased sampling is that the case–control indicator y is ancillary for (β, g). Robins *et al.* (1994) conclude in their Lemma 6.1 that the efficient scores, influence functions and variance bounds for β are therefore identical whether one treats λ_1 as a free parameter to be estimated from the data or instead fixes it at the known true value. (This result is implied more formally by Corollary 1 to Theorem 4.4.1 of Bickel *et al.* (1993).) Inferences made about β from the biased sampling model $\mathscr{P}$ (12) therefore are identical to those from the alternate parametrization $\mathscr{P}^*$ of the random sampling model (11) applied to the case–control data. As already noted, these are in turn identical to the inferences made by applying the original model (1) to these same data. This is the remarkable result first established by Anderson (1972) for discrete Z and later by Prentice and Pyke (1979) for arbitrary Z through their derivation of $\hat{\beta}$ as the NPMLE in the biased sampling model.

Although we now know already that this must equal (3), the alternate parametrization also leads to a simple proof of Bickel *et al.*s formula (our equation (7)) for the efficient score. One readily calculates from (12) that the β score is $\dot{\mathbf{l}}_\beta = Y[Z - \mathrm{E}(Z|Y)]$ and that the nuisance tangent space is $\dot{\mathscr{P}}_2 = \{a(Z) - \mathrm{E}[a(Z)|Y] : a(Z)\varepsilon L_2^0(P)\}$. Following Bickel *et al.* (1993, Section 4.4), and using the fact that $\dot{\mathbf{l}}_\beta$ is orthogonal to the space spanned by $Y - \mathrm{E}(Y)$, the projection of $\dot{\mathbf{l}}_\beta$ onto $\dot{\mathscr{P}}_2$ is simply $ACE(\dot{\mathbf{l}}_\beta|Z, Y)$ and the formula follows.

The Appendix contains explicit calculations which confirm that application of the estimating equations (2) to the case–control sample leads to a consistent, asymptotically linear estimate for β whose variance achieves the semi-parametric lower bound. According to what has been argued above, a 'corollary' is the well-known fact (see, for example, Chamberlain 1987) that the same estimator is consistent, asymptotically linear and efficient under simple random sampling.

Appendix: Consistency, asymptotic linearity and efficiency

Throughout this Appendix we make use of the reparametrization $P_{\alpha,\beta,H} = \mathscr{Q}_{\gamma,\beta,\tilde{H}}$ that equates the biased sampling model with a specific member of the equivalence class of random sampling models. Suppose the data $\{(y_i, z_i), i = 1, \ldots, n\}$ are a random sample from $P_0 = P_{\alpha_0,\beta_0,H_0} \in \mathscr{P}$, corresponding to $Q_0 = Q_{\gamma_0,\beta_0,\tilde{H}_0} \in \mathscr{Q}$. Let P_n be the empirical

distribution of $\{(y_i, z_i): i = 1, \ldots, n\}$, and let $D(\tilde{\theta}, P_n)$ be the term of the log-likelihood corresponding to $\tilde{\theta}$:

$$D(\tilde{\theta}, P_n) \equiv D(\gamma, \beta, P_n) \equiv P_n \log f(y|z; \gamma, \beta) = \frac{1}{n}\sum_{i=1}^{n} \log f(y_i|z_i; \gamma, \beta).$$

Define $\hat{\theta}$ to be the maximizer of $D(\tilde{\theta}, P_n)$ and note that $D(\tilde{\theta}, P_n)$ is strictly concave in $\tilde{\theta}$ if $z_1, \ldots, z_n$ are not all in a linear subspace of $\mathbb{R}^p$. It follows from Theorem 7.4.1 of Bickel *et al.* (1993, p. 325), that the estimator $\hat{\theta}$ satisfies the score equations (5), and is consistent for $\tilde{\theta}_0 = (\gamma_0, \beta_0)$, provided that Z is not concentrated on any hyperplane in $\mathbb{R}^p$ under the marginal distribution $\tilde{H}$. Hypothesis (2) of the theorem follows from concavity of the functions in

$$\mathscr{F} = \{\log f(y|z; \tilde{\theta}) = y(\gamma + z^{\mathrm{T}}\beta) - \log(1 + e^{\gamma + z^{\mathrm{T}}\beta}): \tilde{\theta} \in K \subset \mathbb{R}^{p+1}\}$$

for K compact via Theorem II.1 of Andersen and Gill (1982). Alternatively, it follows from a standard Glivenko–Cantelli theorem for the class $\mathscr{F}$.

Next we examine the asymptotic linearity of $\hat{\beta}$ obtained by solving the system of equations (5) with θ replaced by $\tilde{\theta} = (\gamma, \beta)$. These equations can be rewritten as

$$\sum_{i=1}^{n} z_i \left[y_i - \frac{\exp\{\hat{\gamma}(\beta) + z_i^{\mathrm{T}}\beta\}}{1 + \exp\{\hat{\gamma}(\beta) + z_i^{\mathrm{T}}\beta\}} \right] = 0 \tag{13}$$

and

$$\sum_{i=1}^{n} \left[y_i - \frac{\exp(\gamma + z_i^{\mathrm{T}}\beta)}{1 + \exp(\gamma + z_i^{\mathrm{T}}\beta)} \right] = 0, \tag{14}$$

where $\hat{\gamma}(\beta)$ solves (14). Dividing equation (14) by n and taking the limit shows that, with probability one, $\lambda_1 = \int f(1|z; \lim_n \hat{\gamma}, \beta) p_0(z)\, \mathrm{d}m(z)$ for any limit point of $\hat{\gamma}$ and thus that $\hat{\gamma}(\beta) \to \gamma(\beta)$ where $\gamma(\beta)$ satisfies $\lambda_1 = \int f(1|z; \gamma(\beta), \beta)\tilde{h}_0(z)\, \mathrm{d}m(z)$. Set $\tilde{p}(y|z; \beta) \equiv f(y|z; \gamma(\beta), \beta)$. Linearizing the same equation as a function of $\hat{\gamma}$ at $\gamma(\hat{\beta})$,

$$\hat{\gamma}(\hat{\beta}) - \gamma(\hat{\beta}) = \frac{\sum_i \{y_i - \tilde{p}(1|z_i; \hat{\beta})\}}{\sum_i \tilde{p}(1|z_i; \hat{\beta})\tilde{p}(0|z_i; \hat{\beta})} + o_p(n^{-1/2}).$$

Similarly linearizing equation (13), we find

$$0 = \sum_{i=1}^{n} z_i\{y_i - \tilde{p}(1|z_i;\hat{\beta})\} - \sum_{i=1}^{n} z_i\tilde{p}(1|z_i;\hat{\beta})\tilde{p}(0|z_i;\hat{\beta})\{\hat{\gamma}(\hat{\beta}) - \gamma(\hat{\beta})\} + o_p(n^{1/2})$$

$$= \sum_{i=1}^{n}\left\{z_i - \frac{\sum_{j=1}^{n} z_j\tilde{p}(1|z_j;\hat{\beta})\tilde{p}(0|z_j;\hat{\beta})}{\sum_{j=1}^{n}\tilde{p}(1|z_j;\hat{\beta})\tilde{p}(0|z_j;\hat{\beta})}\right\}\{y_i - \tilde{p}(1|z_i;\hat{\beta})\} + o_p(n^{1/2})$$

$$= \sum_{i=1}^{n}\{z_i - C\}\{y_i - \tilde{p}(1|z_i;\hat{\beta})\} + o_p(n^{1/2})$$

$$= \sum_{i=1}^{n}\Psi(y_i, z_i;\hat{\beta}) + o_p(n^{1/2}),$$

where

$$\Psi(Y, Z;\beta) = (Z - C)\{Y - \tilde{p}(1|Z;\beta)\}$$

and

$$C = \frac{\mathrm{E}\{Z\,\mathrm{var}(Y|Z)\}}{\mathrm{E}\{\mathrm{var}(Y|Z)\}}$$

with $\mathrm{var}(Y|Z;\beta) = \tilde{p}(1|Z;\beta)\tilde{p}(0|Z;\beta)$. This shows that $\hat{\beta}$ is an asymptotic M-estimator (Bickel *et al.* 1993, Sections 7.2, 7.3) with influence function $-\{\mathrm{E}\dot{\Psi}(Y, Z;\beta_0)\}^{-1}\Psi(Y, Z;\beta_0)$, where

$$\dot{\Psi}(Y, Z;\beta_0) = \frac{\partial}{\partial\beta^{\mathrm{T}}}\Psi(Y, Z;\beta)|_{\beta=\beta_0} = -(Z - C)\tilde{p}(1|Z;\beta_0)\tilde{p}(0|Z;\beta_0)(Z^{\mathrm{T}} + \partial\gamma/\partial\beta^{\mathrm{T}}).$$

Note first that $\Psi(Y, Z;\beta_0)$ equals the efficient score $\mathbf{l}^*_\beta(Y, Z)$ for the biased sampling model as derived in equation (8). Next, since $\mathrm{E}\{(Z - C)\,\mathrm{var}(Y|Z)\} = 0$, we have the expected identity

$$-\mathrm{E}\{\dot{\Psi}(Y, Z)\} = \mathrm{var}\,\Psi(Y, Z) = \mathrm{E}\{(Z - C)\,\mathrm{var}(Y|Z)(Z - C)^{\mathrm{T}}\}.$$

Thus the influence function for $\hat{\beta}$ equals the efficient influence function $\{\mathrm{E}(\mathbf{l}^*_\beta\mathbf{l}^{*\mathrm{T}}_\beta)\}^{-1}\mathbf{l}^*_\beta$ and $\sqrt{n}(\hat{\beta} - \beta)$ has an asymptotic normal distribution whose variance attains the semi-parametric lower bound:

$$\{\mathrm{E}(\mathbf{l}^*_\beta\mathbf{l}^{*\mathrm{T}}_\beta)\}^{-1} = [\mathrm{E}\{(Z - C)\,\mathrm{var}(Y|Z)(Z - C)^{\mathrm{T}}\}]^{-1}. \tag{15}$$

Acknowledgements

This work was supported in part by grants from the US National Science Foundation and Public Health Service.

References

Andersen, P.K. and Gill, R.D. (1982) Cox's regression model for counting processes: a large sample study. *Ann. Statist.*, **10**, 1100–1120.

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

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

Breslow, N.E. (1996) Statistics in epidemiology: the case–control study. *J. Amer. Statist. Assoc.*, **91**, 14–28.

Chamberlain, G. (1987) Asymptotic efficiency in estimation with conditional moment restrictions. *J. Econometrics*, **34**, 305–334.

Cornfield, J. (1951) A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix. *J. Nat. Cancer Inst.*, **11**, 1269–1275.

Cosslett, S.R. (1981) Maximum likelihood estimator for choice-based samples. *Econometrica*, **49**, 1289–1316.

Manski, C.F. and Lerman, S.R. (1977) The estimation of choice probabilities from choice-based samples. *Econometrica*, **45**, 1977–1988.

McNeney, W.B. (1998) Asymptotic efficiency in semiparametric models with non-i.i.d. data. Ph.D. thesis, University of Washington.

Prentice, R.L. and Pyke, R. (1979) Logistic disease incidence models and case–control studies. *Biometrika*, **66**, 403–411.

Rabinowitz, D. (1997) A note on efficient estimation from case–control data. *Biometrika*, **84**, 486–488.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1994) Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.*, **89**, 846–866.

Roeder, K., Carroll, R.J. and Lindsay, B.G. (1996) A semiparametric mixture approach to case–control studies with errors in covariables. *J. Amer. Statist. Assoc.*, **91**, 722–732.

Weinberg, C.R. and Sandler, D.P. (1991) Randomized recruitment in case–control studies. *Amer. J. Epidemiology*, **134**, 421–432.

Received May 1997 and revised December 1998