Statistica Sinica **20** (2010), 1529-1550

PARAMETER ESTIMATION FROM AN OUTCOME-DEPENDENT ENRICHED SAMPLE USING WEIGHTED LIKELIHOOD METHOD

Qing Kang, Paul I. Nelson[1] and Christopher I. Vahl[2]

[1]*Kansas State University and* [2]*Elanco Animal Health*

Abstract: An outcome-dependent enriched (ODE) sample results from adding a random sample to a stratified sample, where the stratification is based on levels of a categorical outcome. In biometrics, such a sample can be generated by combining data from a cohort study with data from an independent case-control study. Suppose that the probability of an outcome is determined by covariates according to a given model. For the case where the marginal distributions of the outcome and predictors are both unknown, Morgenthaler and Vardi (1986) proposed a weighted likelihood (WL) method to estimate the model parameters from an ODE sample. Here, we derive and study the asymptotic properties of the WL estimator. Simulation and an asymptotic comparison demonstrate that when the presumed model is correct, the performance of the WL method is often comparable to the asymptotically efficient profile likelihood (PL) method. If the model is misspecified, the WL method has a nice interpretation and is more robust than the PL method. This leads us to recommend use of the WL method, especially for the situation where the fitness of the presumed model is in doubt and the sample size is large.

Key words and phrases: Case-control sample, central limit theorem, choice-based sample, empirical distribution, generalized linear model, semiparametric likelihood.

1. Introduction

Consider a categorical outcome variable Y whose distribution depends on a vector $\mathbf{X}$ of covariates according to a model specified up to an unknown parameter $\boldsymbol{\theta}$. Assume that (i) the marginal probability of Y is unknown; (ii) the marginal mass/density function of $\mathbf{X}$ is unknown and free of $\boldsymbol{\theta}$. Although $\boldsymbol{\theta}$ can be estimated from a random sample using the likelihood method, the variance of its maximum likelihood estimator (MLE) is large if some outcome levels are rarely observed. This motivates a stratified sampling design based on values of Y. For example, Doll and Hill (1950) investigated the risk factors for lung cancer by querying hospital patients admitted with lung cancer (case) and those without cancer (control). The other advantage of Y-stratified sampling comes from its cost. In practice, units in the population are often naturally grouped

according to their outcome levels. Manski and Lerman (1977) therefore recommended studying the choice of transportation modes by interviewing travelers at a train station and auto drivers at a parking lot, rather than phoning people at home. However, statistical inference from a Y-stratified sample is very limited. For a binomial outcome, the intercept term in the logit linear model is not identifiable from a Y-stratified sample (Prentice and Pyke (1979); Chen (2003)). Depending on whether $\mathbf{X}$ is discrete or continuous, the intercept term in probit and complementary-log-log linear models is either unidentifiable or poorly determined (Cosslett (1993, Sec. 3.5); Chen (2003)). As a result, it is impossible to predict Y given $\mathbf{X}$ from a Y-stratified sample. The situation is even worse for a multinomial outcome. Only the baseline-category logit model (Agresti (2002, Sec. 7.1.1)) is feasible. Applying cumulative logit models for an ordinal outcome frequently causes the likelihood function to fail to converge or to carry a large estimation bias.

When the outcome levels are not very rare, problems with model selection and parameter estimation can be ameliorated by enriching the Y-stratified sample with an independent random sample of $\mathbf{X}$ and Y. In econometrics, Y often corresponds to a consumer's choice. Cosslett (1981a, 1993) defines the combination of random and Y-stratified samples as a choice-based/endogenously-stratified enriched sample. In biometrics, Y usually represents a patient's health status. The combined sample is known as an outcome-dependent two-component sample (Zhou et al. (2002); Wang and Zhou (2006)). To promote cross-communication between these two disciplines, we refer to such a combined sample as Outcome-Dependent Enriched (ODE). It is worth mentioning that even if the outcome has some rare levels, ODE sampling is still feasible as long as the sample size for the random sample component is sufficiently large. For instance, Doll started a perspective study for the effect of smoking on the mortality of British doctors in 1951. Among those male participants, over twenty-five thousand passed away by 2001, and 1,052 of these deaths were due to lung cancer (Doll et al. (2004, Table 1)). Combining this cohort data with Doll and Hill's case-control data yields an ODE sample.

Many semiparametric likelihood methods have been proposed to estimate $\boldsymbol{\theta}$ from a two-stage outcome-dependent (TSOD) sample. The first stage of this sampling scheme collects a random sample of Y and, perhaps, a partial measurement of $\mathbf{X}$, to define strata. The second stage draws a stratified sample from the first-stage sample and measures $\mathbf{X}$ in detail. For a summary of these methods, see Lawless, Kalbfleisch and Wild (1999). Note that the ODE sample considered here is different from the TSOD sample because the random and stratified components in the ODE sample are independent. There are two published likelihood-based methods for estimating $\boldsymbol{\theta}$ from an ODE sample when no

prior knowledge about the marginal distributions of $\mathbf{X}$ and Y is available. One is Cosslett's (1981a) profile likelihood (PL) method. The other is Morgenthaler and Vardi's (1986) weighted likelihood (WL) method where the weights are based on Vardi's (1985) nonparametric MLE (NPMLE) of the joint distribution of $\mathbf{X}$ and Y. Consistency, efficiency, and asymptotic normality were obtained for the PL estimator (Cosslett (1981b)). However, Morgenthaler and Vardi (1986) only sketched a derivation of the limiting normality of the WL estimator and their heuristic approach leads to the omission of a positive-semi-definite term that can result in severely underestimated standard errors (SE), as noted in Remarks 1 and 2 after the proofs of Lemma 1 and Theorem 1, respectively. One of our main contributions is to rigorously derive the limiting distribution of the WL estimator via a new technique based on conditioning. Results of this new approach allow us to compare the WL method with the asymptotically efficient PL method and provide practitioners with some guidance about their performance. Because they often behave very similarly when the presumed model is correct, and because the WL estimator is more robust to model misspecification, we recommend its usage, especially when fitness of the presumed model is uncertain and the sample size is large.

Section 2 specifies the WL objective function. In Section 3, we present the asymptotic properties of its MLE. Section 4 evaluates the WL and PL methods under properly specified generalized linear models. Section 5 studies the robustness of these two methods when the presumed model is incorrect. In Section 6, we illustrate these two methods using an ODE sample drawn from a national survey. Additional comments on the analysis of an ODE sample are made in Section 7.

2. Semiparametric Likelihood Functions

Suppose that Y has $(K+1)$ levels and its distribution is determined by $\mathbf{X}$ according to a working model of the form $\Pr(Y = k \mid \mathbf{X} = \mathbf{x}) = \rho(k, \mathbf{x}; \boldsymbol{\theta})$, $k = 0, \ldots, K$. For convenience, we take $\mathbf{X}$ to be a continuous random vector with values in Euclidean space, denoted by $\mathscr{X}$. Accordingly, all densities are with respect to Lebesgue measure on $\mathscr{X}$. The following notation is used to describe the actual population, whether or not the model is correct:

$$g(\mathbf{x}) : \text{the marginal density function of } \mathbf{X};$$
$$f(\mathbf{x}, y) : \text{the joint distribution of } \mathbf{X} \text{ and } Y.$$

Statistical inference about $\boldsymbol{\theta}$ is made from an ODE sample of size N, denoted by $\{(\mathbf{X}_i, Y_i)\}_{i=1}^N$. The following notation is used to describe this sample:

$$R = \{\text{Indices of units that belong to the random sample component}\};$$

$$
\begin{aligned}
n_R &= \text{Size of } R;\\
m_{+k} &= \textstyle\sum_{i\in R} I(Y_i = k), \quad \sum_{k=0}^{K} m_{+k} = n_R;\\
S_k &= \{\text{Indices of units in the stratified sample that belong to stratum with } Y = k\};\\
n_k &= \text{Size of } S_k,\ n_R + \textstyle\sum_{k=0}^{K} n_k = N.
\end{aligned}
$$

Note that for $i \in S_k$, $Y_i \equiv k$; while for $i \in R$, Y_i is a categorical random variable with $(K+1)$ possible levels. The PL method estimates $\boldsymbol{\theta}$ by first expressing the full log-likelihood as

$$
l_N(\boldsymbol{\theta}, g(\cdot)) = \sum_{i\in R} \log(\rho(Y_i, \mathbf{X}_i; \boldsymbol{\theta}) g(\mathbf{X}_i)) + \sum_{k=0}^{K} \sum_{i\in S_k} \log\Big(\rho(Y_i, \mathbf{X}_i; \boldsymbol{\theta}) \frac{g(\mathbf{X}_i)}{\pi_{+k}}\Big). \quad (2.1)
$$

Here, $\pi_{+0}, \ldots, \pi_{+K} \in (0,1)$ denote the marginal probabilities of Y, and $\sum_{k=0}^{K} \pi_{+k} = 1$. They are unknown and considered to be nuisance parameters. The presumed model connects $\boldsymbol{\theta}$ and $g(\mathbf{x})$ with the restriction $\pi_{+k} = \int_{\mathscr{X}} f(\mathbf{x}, k) d\mathbf{x} = \int_{\mathscr{X}} \rho(k, \mathbf{x}; \boldsymbol{\theta}) g(\mathbf{x}) d\mathbf{x}$, $k = 0, \ldots, K$. Under a nonparametric framework, the density function of continuous $\mathbf{X}$ is estimated by a discrete distribution, which puts probability mass over the observed data points. Fixing $\boldsymbol{\nu} = (\boldsymbol{\theta}, \pi_{+1}, \ldots, \pi_{+K})$, Cosslett (1981a) showed that (2.1) is maximized at $\tilde{g}(\mathbf{x}, \boldsymbol{\nu}) = (NT(\mathbf{x}; \boldsymbol{\nu}))^{-1} \sum_{i=1}^{N} I(\mathbf{X}_i = \mathbf{x})$, where

$$
T(\mathbf{x}; \boldsymbol{\nu}) = \frac{n_R}{N} + \sum_{k=0}^{K} \Big[\frac{n_k}{N} \frac{\rho(k, \mathbf{x}; \boldsymbol{\theta})}{\pi_{+k}}\Big].
$$

Recall that the ODE sample arises from combining a random sample of $\mathbf{X}$ and Y with a Y-stratified sample of $\mathbf{X}$. Regarding $\mathbf{X}$, this is equivalent to mixing $g(\mathbf{x})$ and $f(\mathbf{x}, y; \boldsymbol{\theta})/\Pr(Y = k)$ with weights n_R/N and n_k/N, $k = 0, \ldots, K$. Provided the assumed working model is correct, the resulting mixture distribution of $\mathbf{X}$ can be represented by $g(\mathbf{x})T(\mathbf{x}; \boldsymbol{\nu})$. So, a heuristic rational for $\tilde{g}(\mathbf{x}; \boldsymbol{\nu})$ is that it is the empirical estimate of $g(\mathbf{x})T(\mathbf{x}; \boldsymbol{\nu})$, $N^{-1} \sum_{i=1}^{N} I(\mathbf{X} = \mathbf{x})$, divided by $T(\mathbf{x}; \boldsymbol{\nu})$. Replacing $g(\mathbf{x})$in (2.1) with $\tilde{g}(\mathbf{x}; \boldsymbol{\nu})$, the final objective function of the PL method is

$$
l_N^P(\boldsymbol{\nu}) = \sum_{i\in R} \log\Big(\frac{\rho(Y_i, \mathbf{X}_i; \boldsymbol{\theta})}{T(\mathbf{X}_i; \boldsymbol{\nu})}\Big) + \sum_{k=0}^{K} \sum_{i\in S_k} \log\Big(\frac{\rho(Y_i, \mathbf{X}_i; \boldsymbol{\theta})}{T(\mathbf{X}_i; \boldsymbol{\nu})\pi_{+k}}\Big). \quad (2.2)
$$

Cosslett (1981b) proved that the PL estimator of $\boldsymbol{\theta}$ is efficient in the sense that it is consistent and its asymptotic variance achieves the Cramér-Rao lower bound for all asymptotically unbiased estimators under the same restrictions.

Morgenthaler and Vardi (1986) recognized that the stratified sample component in an ODE sample can be viewed as a biased sample where the biasing

is caused by truncation on Y. They applied Vardi's (1985) general methodology for analyzing biased samples and derived the NPMLE of $f(\mathbf{x}, Y = k) = \Pr(Y = k) f(\mathbf{x} \mid Y = k)$ to be

$$\tilde{f}(\mathbf{x}, y) = \frac{m_{+k}}{n_R} \frac{\sum_{i=1}^{N} I(\mathbf{X}_i = \mathbf{x}, Y_i = k)}{n_k + m_{+k}}. \tag{2.3}$$

An intuitive explanation for (2.3) is that the first term estimates $\Pr(Y = k)$ solely based on the random sample component, and the second term pools information from the random and the stratified components together to estimate $f(\mathbf{x} \mid Y = k)$. Let Θ be the parameter space of $\boldsymbol{\theta}$. It is well known that the true value of $\boldsymbol{\theta}$, denoted by $\boldsymbol{\theta}^*$, minimizes the Kullback-Leibler divergence between $f(\mathbf{x}, y)$ and $\rho(y, \mathbf{x}; \boldsymbol{\theta}) g(\mathbf{X})$ (Kullback (1997)), i.e.,

$$\boldsymbol{\theta}^* \in \operatorname*{argmin}_{\theta \in \Theta} \Big\{ \sum_{k=0}^{K} \int_{\mathscr{X}} \log \Big(\frac{f(\mathbf{x}, k)}{\rho(k, \mathbf{x}; \boldsymbol{\theta}) g(\mathbf{x})} \Big) f(\mathbf{x}, k) d\mathbf{x} \Big\}.$$

By replacing $f(\mathbf{x}, y)$ with Vardi's NPMLE of $f(\mathbf{x}, y)$ given at (2.3), it naturally follows that one should maximize the weighted log-likelihood type function

$$l_N(\boldsymbol{\theta}) = \sum_{i=1}^{N} \{ \tilde{f}(\mathbf{X}_i, Y_i) \log(\rho(Y_i, \mathbf{X}_i; \boldsymbol{\theta}) g(\mathbf{X}_i)) \}.$$

Because $g(\mathbf{x})$ is free of $\boldsymbol{\theta}$, we can estimate $\boldsymbol{\theta}$ without specifying a parametric form for $g(\mathbf{x})$. The WL objective function is

$$l_N^W(\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{k=0}^{K} \Big\{ \frac{m_{+k}}{n_R} \frac{I(Y_i = k)}{n_k + m_{+k}} \log(\rho(Y_i, \mathbf{X}_i; \boldsymbol{\theta})) \Big\}. \tag{2.4}$$

Weighting is customary in analyzing complex survey data, where a pseudo log-likelihood function is adjusted by the sampling weight of each unit. Unlike the pseudo likelihood approach, the weights in the WL method are based on Vardi's NPMLE of $f(\mathbf{x}, y)$. According to Vardi (1985), necessary and sufficient conditions for the NPMLE of $f(\mathbf{x}, y)$ to exist under ODE sampling are that

(i) a random sample component is present in the ODE sample, i.e., $n_R \neq 0$;

(ii) the random sample component contains all levels of Y, i.e., $m_{+k} \neq 0$ for all $k = 0, \ldots, K$.

These conditions also serve as the prerequisites for the application of the WL method.

The WL method is easy to interpret because, as noted before, it minimizes the Kullback-Leibler divergence between the model and the empirical joint distribution of $\mathbf{X}$ and Y. The WL method also benefits from its computational

ease. Note that (2.4) has only $\boldsymbol{\theta}$ as its argument, and a model-free estimate of π_{+k} can be obtained from (2.3), namely $\tilde{\pi}_{+k} = m_{+k}/n_R$. In contrast, the PL method estimates $\pi_{+1}, \ldots, \pi_{+K}$ and $\boldsymbol{\theta}$ simultaneously using the presumed model. It is interesting to see that when the ODE sample contains only a random sample component, both the WL and PL objective functions reduce to the standard likelihood function from a random sample.

3. Asymptotic Properties

Because application of the WL method requires the random sample component be present in the ODE sample, a crucial restriction in our asymptotic derivation is that as $N \to \infty$, $n_R/N \to \lambda_R$ with $\lambda_R \neq 0$. Let the generic $1 \times q$ parameter vector $\boldsymbol{\theta}$ range over the parameter space Θ, an open subset of q-dimensional Euclidean space. Let $\hat{\boldsymbol{\theta}}^W$ stand for the estimator that maximizes (2.4). Sufficient conditions for deriving the asymptotic properties of $\hat{\boldsymbol{\theta}}^W$ can be sorted into three groups. The first conditions, for all $k = 0, \ldots, K$, $\boldsymbol{\theta} \in \Theta$, are common in general likelihood theory.

(C1) The first- and second-partial derivatives of $\rho(k, \mathbf{x}; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, denoted by the $1 \times q$ vector $\nabla \rho(k, \mathbf{x}; \boldsymbol{\theta})$ and the $q \times q$ matrix $\nabla^2 \rho(k, \mathbf{x}; \boldsymbol{\theta})$, exist and are continuous a.e. with respect to $g(\mathbf{x})$. For any constant $q \times 1$ vector $\boldsymbol{a}$, $\int_{\mathscr{X}} |\nabla \log(\rho(k, \mathbf{x}; \boldsymbol{\theta}))\boldsymbol{a}|^3 f(\mathbf{x}, k) d\mathbf{x} < \infty$.

(C2) For any triplet of elements in $\boldsymbol{\theta}$, $(\theta_i, \theta_j, \theta_l)$, $\int_{\mathscr{X}} |\partial^3 \log(\rho(k, \mathbf{x}; \boldsymbol{\theta}))/\partial\theta_i \partial\theta_j \partial\theta_l| f(\mathbf{x}, k) d\mathbf{x} < \infty$.

The second conditions apply specifically to categorical outcomes and require that $\rho(k, \mathbf{x}; \boldsymbol{\theta})$, $k = 0, \ldots, K$, be a valid probability model. In particular,

(C3) For all $k = 0, \ldots, K$, $\boldsymbol{\theta} \in \Theta$, $0 < \rho(k, \mathbf{x}; \boldsymbol{\theta}) < 1$ and $\sum_{k=0}^{K} \rho(k, \mathbf{x}; \boldsymbol{\theta}) = 1$ a.e. with respect to $g(\mathbf{x})$.

(C4) The model is identifiable, i.e., if $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$, then

$$\sum_{k=0}^{K} \int_{\mathscr{X}} \log\Big(\frac{\rho(k, \mathbf{x}; \boldsymbol{\theta}_2)}{\rho(k, \mathbf{x}; \boldsymbol{\theta}_1)}\Big) \rho(k, \mathbf{x}; \boldsymbol{\theta}_1) g(\mathbf{x}) d\mathbf{x} \neq 0.$$

The identifiability condition used by Cosslett (1981b) requires that if $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$, there must be a region $\mathscr{L} \subseteq \mathscr{X}$ such that for some $k \in \{0, \ldots, K\}$,

$$\int_{\mathscr{L}} \rho(k, \mathbf{x}; \boldsymbol{\theta}_1) g(\mathbf{x}) d\mathbf{x} \neq \int_{\mathscr{L}} \rho(k, \mathbf{x}; \boldsymbol{\theta}_2) g(\mathbf{x}) d\mathbf{x}.$$

By Jensen's Inequality, Cosslett's condition implies (C4). The following condition establishes the asymptotic variance of $\hat{\boldsymbol{\theta}}^W$.

(C5) For all $k = 0, \ldots, K$ the components of the $1 \times q$ vector $\nabla\rho(k, \mathbf{x}; \boldsymbol{\theta})$ are linearly independent a.e. with respect to $g(\mathbf{x})$.

Conditions (C1)−(C5) are comparable to what Cosslett (1981b) used to obtain consistency of the PL estimator. As Cosslett pointed out, these conditions are not as restrictive as they may first appear. For example, if we choose $\rho(k, \mathbf{x}; \boldsymbol{\theta})$, $k = 0, \ldots, K$, to be a generalized linear model with appropriate parameterization, these conditions simplify into moment regularities.

Taking the first- and second-partial derivatives at (2.4) with respect to $\boldsymbol{\theta}$ gives rise to the score type function $\nabla l_N^W(\boldsymbol{\theta})$ and the Hessian function $\nabla^2 l_N^W(\boldsymbol{\theta})$. They are

$$\nabla l_N^W(\boldsymbol{\theta}) = \sum_{i=1}^{N}\sum_{k=0}^{K}\Big\{\frac{m_{+k}}{n_R}\frac{I(Y_i = k)}{n_k + m_{+k}}\nabla \log(\rho(Y_i, \mathbf{X}_i; \boldsymbol{\theta}))\Big\},$$
$$\nabla^2 l_N^W(\boldsymbol{\theta}) = \sum_{i=1}^{N}\sum_{k=0}^{K}\Big\{\frac{m_{+k}}{n_R}\frac{I(Y_i = k)}{n_k + m_{+k}}\nabla^2 \log(\rho(Y_i, \mathbf{X}_i; \boldsymbol{\theta}))\Big\}.$$

Each term inside the summands here depends on m_{+k}, Y_i and $\mathbf{X}_i$. Obtaining the asymptotic distribution of $\nabla l_N^W(\boldsymbol{\theta})$ is challenging because it is not a sum of independent terms. However, we recognize that conditional on $\mathbf{Y}_N = (Y_1, \ldots, Y_n)'$, the vector $\mathbf{M}_N = (m_{+0}, \ldots, m_{+K})'$ is a constant and $\nabla l_N^W(\boldsymbol{\theta})$ becomes a sum of independent, but not identical, functions of $\mathbf{X}_i$. Proving the asymptotic normality of $\nabla l_N^W(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^*$, the true value of $\boldsymbol{\theta}$, involves first applying Liapounov's Central Limit Theorem (CLT) for triangular arrays (Lehmann (1999, Thm. 2.7.2)) to the conditional distribution of $\mathbf{Y}_N$, and then using dominated convergence to remove the conditioning. The following lemma summarizes our findings. Note that the operator '$\times 2$' stands for the outer product of a vector with itself and $\mathbf{0}^q$ stands for $1 \times q$ a vector of zeros.

Lemma 1. *Suppose that the outcome variable Y has $(K+1)$ possible levels, its joint distribution with predictor $\mathbf{X}$ is $f(\mathbf{x}, y)$, and $\Pr(Y = k \mid \mathbf{X} = \mathbf{x}) = \rho(k, \mathbf{x}; \boldsymbol{\theta}^*)$, $\boldsymbol{\theta}^* \in \Theta$. As $N \to \infty$, $n_R/N \to \lambda_R$, $\lambda_R \neq 0$, and $n_k/N \to \lambda_k$, $k = 0, \ldots, K$. Set*

$$\begin{aligned}
\mathbf{A}_k(\boldsymbol{\theta}) &= \int_{\mathscr{X}} \nabla \log(\rho(k, \mathbf{X}; \boldsymbol{\theta}))f(\mathbf{x}, k)d\mathbf{x},\\
\mathbf{B}_k(\boldsymbol{\theta}) &= \int_{\mathscr{X}} \{\nabla \log(\rho(k, \mathbf{X}; \boldsymbol{\theta}))\}^2 f(\mathbf{x}, k)d\mathbf{x},\\
\pi_{+k}^* &= \Pr(Y = k) = \int_{\mathscr{X}} f(\mathbf{x}, k)d\mathbf{x},
\end{aligned} \tag{3.1}$$

$$\mathbf{V}_a(\boldsymbol{\theta}) = \sum_{k=0}^{K} \Big\{ \frac{\lambda_k}{\lambda_R \pi^*_{+k}(\lambda_k + \lambda_R \pi^*_{+k})} \mathbf{A}_k(\boldsymbol{\theta})^{\times 2} \Big\},$$
$$\mathbf{V}_b(\boldsymbol{\theta}) = \sum_{k=0}^{K} \Big\{ \frac{\pi^*_{+k}}{\lambda_k + \lambda_R \pi^*_{+k}} \mathbf{B}_k(\boldsymbol{\theta}) \Big\},$$
$$\mathbf{H}(\boldsymbol{\theta}) = -\sum_{k=0}^{K} \mathbf{B}_k(\boldsymbol{\theta}).$$

Consider the likelihood function given at (2.4). *If* (C1) *and* (C3) *are satisfied, then as* $N \to \infty$,

$$\sqrt{N} \nabla l_N^W(\boldsymbol{\theta}^*) \xrightarrow{D} \text{Normal}(\mathbf{0}^q, \mathbf{V}_a(\boldsymbol{\theta}^*) + \mathbf{V}_b(\boldsymbol{\theta}^*)),$$
$$\nabla^2 l_N^W(\boldsymbol{\theta}^*) \xrightarrow{a.e.} \mathbf{H}(\boldsymbol{\theta}^*).$$

Proof. When we condition on $\mathbf{Y}_N = (Y_1, \ldots, Y_n)'$, $\mathbf{M}_N = (m_{+0}, \ldots, m_{+K})'$ is a constant and there are $(n_k + m_{+k})$ independent $\mathbf{X}$ terms in the ODE sample at the $Y = k$ level. First, we employ Liapounov's CLT and the Cramer-Wold device to establish the asymptotic normality of $\nabla l_N^W(\boldsymbol{\theta})$, for any $\boldsymbol{\theta} \in \Theta$, conditioning on $\mathbf{Y}_N$. Note that, $\nabla l_N^W(\boldsymbol{\theta})$ can be re-expressed as

$$\nabla l_N^W(\boldsymbol{\theta}) = \sum_{k=0}^{K} \Big[\frac{m_{+k}}{n_R(n_k + m_{+k})} \Big\{ \sum_{i \in R, Y_i = k} \nabla \log(\rho(k, \mathbf{X}_i; \boldsymbol{\theta})) + \sum_{i \in S_k} \nabla \log(\rho(k, \mathbf{X}_i; \boldsymbol{\theta})) \Big\} \Big].$$

Under (C1), the conditional mean and variance of $\nabla l_N^W(\boldsymbol{\theta})$ are

$$E(\nabla l_N^W(\boldsymbol{\theta}) \mid \mathbf{Y}_N) = \sum_{k=0}^{K} \Big\{ \frac{m_{+k}}{n_R} E(\nabla \log(\rho(k, \mathbf{X}; \boldsymbol{\theta})) \mid Y = k) \Big\} \Big\}$$
$$= \sum_{k=0}^{K} \Big\{ \frac{m_{+k}}{n_R \pi^*_{+k}} \mathbf{A}_k(\boldsymbol{\theta}) \Big\},$$
$$\text{Var}\,(\nabla l_N^W(\boldsymbol{\theta}) \mid \mathbf{Y}_N) = \sum_{k=0}^{K} \Big\{ \big(\frac{m_{+k}}{n_R}\big)^2 \frac{1}{n_k + m_{+k}} \text{Var}\,(\nabla \log(\rho(k, \mathbf{X}; \boldsymbol{\theta})) \mid Y = k) \Big\}$$
$$= \sum_{k=0}^{K} \Big[\big(\frac{m_{+k}}{n_R}\big)^2 \frac{1}{n_k + m_{+k}} \Big\{ \frac{1}{\pi^*_{+k}} \mathbf{B}_k(\boldsymbol{\theta}) - \frac{1}{(\pi^*_{+k})^2} \mathbf{A}_k(\boldsymbol{\theta})^{\times 2} \Big\} \Big].$$

Let $\boldsymbol{a}$ be any $q \times 1$ non-zero vector of constants. To apply Liapounov's CLT, we examine the third moment of those terms summed by $\nabla l_N^W(\boldsymbol{\theta})\boldsymbol{a}$. Set

$$U(\nabla l_N^W(\boldsymbol{\theta})\boldsymbol{a} \mid \mathbf{Y}_N)$$
$$= \sum_{k=0}^{K} \Big\{ \big(\frac{m_{+k}}{n_R}\big)^3 \frac{1}{(n_k + m_{+k})^2} E\Big(\Big| \nabla \log(\rho(k, \mathbf{X}; \boldsymbol{\theta}))\boldsymbol{a} - \frac{1}{\pi^*_{+k}} \mathbf{A}_k(\boldsymbol{\theta})\boldsymbol{a} \Big|^3 \mid Y = k \Big) \Big\}.$$

As $N \to \infty$, we have

$$N\mathrm{Var}\,(\nabla l_N^W(\boldsymbol{\theta}) \mid \mathbf{Y}_N) = \mathbf{V}_b(\boldsymbol{\theta}) - \sum_{k=0}^{K} \Big\{ \frac{1}{\lambda_k + \lambda_R \pi^*_{+k}} \mathbf{A}_k(\boldsymbol{\theta})^{\times 2} \Big\} + O(1),$$

$$N^2 U(\nabla l_N^W(\boldsymbol{\theta})\boldsymbol{a} \mid \mathbf{Y}_N) = \sum_{k=0}^{K} \Big\{ \frac{(\pi^*_{+k})^3}{(\lambda_k + \lambda_R \pi^*_{+k})^2} E\Big(\Big| \nabla \log(\rho(k, \mathbf{X}; \boldsymbol{\theta}))\boldsymbol{a} - \frac{1}{\pi^*_{+k}} \mathbf{A}_k(\boldsymbol{\theta})\boldsymbol{a} \Big|^3 \mid Y = k \Big) \Big\} + O(1).$$

According to Liapounov's CLT (Lehmann (1999, Thm. 2.7.2)) and Slutsky's Theorem,

$$\sqrt{N} \Big\{ \nabla l_N^W(\boldsymbol{\theta})\boldsymbol{a} - E(\nabla l_N^W(\boldsymbol{\theta})\boldsymbol{a} \mid \mathbf{Y}_N) \Big\} \mid \mathbf{Y}_N$$
$$\xrightarrow{D} \mathrm{Normal}\Big(\mathbf{0}^q, \boldsymbol{a}' \Big[\mathbf{V}_b(\boldsymbol{\theta}) - \sum_{k=0}^{K} \Big\{ \frac{1}{\lambda_k + \lambda_R \pi^*_{+k}} \mathbf{A}_k(\boldsymbol{\theta})^{\times 2} \Big\} \Big] \boldsymbol{a} \Big).$$

The Dominated Convergence Theorem allows us to remove the conditioning on $\mathbf{Y}_N$. An application of the Cramer-Wold theorem (Lehmann (1999, Thm. 5.1.8)) then results in

$$\sqrt{N} \Big\{ \nabla l_N^W(\boldsymbol{\theta}) - E(\nabla l_N^W(\boldsymbol{\theta}) \mid \mathbf{Y}_N) \Big\}$$
$$\xrightarrow{D} \mathrm{Normal}\Big(\mathbf{0}^q, \mathbf{V}_b(\boldsymbol{\theta}) - \sum_{k=0}^{K} \Big\{ \frac{1}{\lambda_k + \lambda_R \pi^*_{+k}} \mathbf{A}_k(\boldsymbol{\theta})^{\times 2} \Big\} \Big). \tag{3.2}$$

Condition (C3) and the fact that $\Pr(Y = k \mid \mathbf{X} = \mathbf{x}) = \rho(k, \mathbf{x}; \boldsymbol{\theta}^*)$ imply that, at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$,

$$\sum_{k=0}^{K} \mathbf{A}_k(\boldsymbol{\theta}^*) = \mathbf{0}^q,$$
$$\sum_{\substack{j,k=0; \\ j \neq k}}^{K} \mathbf{A}_k(\boldsymbol{\theta}^*)' \mathbf{A}_j(\boldsymbol{\theta}^*) = -\sum_{k=0}^{K} \mathbf{A}_k(\boldsymbol{\theta}^*)^{\times 2}.$$

Because $E(\nabla l_N^W(\boldsymbol{\theta}) \mid \mathbf{Y}_N)$ is a linear combination of the multinomial vector $\mathbf{M}_N$, we have

$$
\begin{aligned}
&E(E(\nabla l_N^W(\boldsymbol{\theta}^*) \mid \mathbf{Y}_N)) = \sum_{k=0}^{K} \mathbf{A}_k(\boldsymbol{\theta}^*) = \mathbf{0}^q,\\
&\operatorname{Var}\left(E(\nabla l_N^W(\boldsymbol{\theta}^*) \mid \mathbf{Y}_N)\right)\\
&\quad = \sum_{k=0}^{K} \Big\{ \frac{\operatorname{Var}(m_{+k})}{(n_R \pi_{+k}^*)^2} \mathbf{A}_k(\boldsymbol{\theta}^*)^{\times 2} \Big\} + \sum_{\substack{j,k=0;\\ j\neq k}}^{K} \Big\{ \frac{\operatorname{Cov}(m_{+k}, m_{+j})}{(n_R)^2 \pi_{+k}^* \pi_{+j}^*} \mathbf{A}_k(\boldsymbol{\theta}^*)' \mathbf{A}_j(\boldsymbol{\theta}^*) \Big\}\\
&\quad = \frac{1}{n_R} \sum_{k=0}^{K} \Big\{ \frac{1-\pi_{+k}^*}{\pi_{+k}^*} \mathbf{A}_k(\boldsymbol{\theta}^*)^{\times 2} \Big\} - \frac{1}{n_R} \sum_{\substack{j,k=0;\\ j\neq k}}^{K} \Big\{ \mathbf{A}_k(\boldsymbol{\theta}^*)' \mathbf{A}_j(\boldsymbol{\theta}^*) \Big\}\\
&\quad = \frac{1}{n_R} \sum_{k=0}^{K} \Big\{ \frac{1}{\pi_{+k}^*} \mathbf{A}_k(\boldsymbol{\theta}^*)^{\times 2} \Big\}
\end{aligned}
$$

and, most importantly,

$$
\sqrt{N} E(\nabla l_N^W(\boldsymbol{\theta}^*) \mid \mathbf{Y}_N) \xrightarrow{D} N\Big(\mathbf{0}^q, \frac{1}{\lambda_R} \sum_{k=0}^{K} \Big\{ \frac{1}{\pi_{+k}^*} \mathbf{A}_k(\boldsymbol{\theta})^{\times 2} \Big\} \Big). \tag{3.3}
$$

Note that $E(\nabla l_N^W(\boldsymbol{\theta}^*) \mid \mathbf{Y}_N)$ is an orthogonal projection of $\nabla l_N^W(\boldsymbol{\theta}^*)$ onto the sample space of $\mathbf{Y}_N$. It then follows from (3.2) and (3.3) that

$$
\sqrt{N} \nabla l_N^W(\boldsymbol{\theta}^*) \xrightarrow{D} \text{Normal}\Big(\mathbf{0}^q, \mathbf{V}_a(\boldsymbol{\theta}^*) + \mathbf{V}_b(\boldsymbol{\theta}^*) \Big).
$$

Regarding $\nabla^2 l_N^W(\boldsymbol{\theta})$, the uniform convergence of (2.3) to the joint distribution of $\mathbf{X}$ and Y (Vardi (1985)) implies that, at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$,

$$
\nabla^2 l_N^W(\boldsymbol{\theta}^*) \xrightarrow{a.e.} \sum_{k=0}^{K} \int_{\mathscr{X}} \nabla^2 \log(\rho(k, \mathbf{x}; \boldsymbol{\theta}^*)) f(\mathbf{x}, k) d\mathbf{x}.
$$

Because $f(\mathbf{x}, k) = \rho(k, \mathbf{x}; \boldsymbol{\theta}^*) g(\mathbf{x})$, the right-hand side of the above is

$$
\begin{aligned}
&\sum_{k=0}^{K} \int_{\mathscr{X}} \nabla^2 \rho(k, \mathbf{x}; \boldsymbol{\theta}^*) g(\mathbf{x}) d\mathbf{x} - \sum_{k=0}^{K} \int_{\mathscr{X}} \{\nabla \log(\rho(k, \mathbf{x}; \boldsymbol{\theta}^*))\}^{\times 2} f(\mathbf{x}, k) d\mathbf{x}\\
&\quad = -\sum_{k=0}^{K} \mathbf{B}_k(\boldsymbol{\theta}^*) \equiv \mathbf{H}(\boldsymbol{\theta}^*).
\end{aligned}
$$

This completes the proof.

Remark 1. Recall that π^*_{+k}, the marginal probability of Y, is unknown. From the above proof, we observe that its model-free estimator $\tilde{\pi}_{+k} = m_{+k}/n_R$ contributes to the variance of $\sqrt{N}\nabla l_N^W(\boldsymbol{\theta}^*)$. Morgenthaler and Vardi (1986) overlooked the randomness of $\tilde{\pi}_{+k}$ and incorrectly derived the asymptotic variance of $\sqrt{N}\nabla l_N^W(\boldsymbol{\theta}^*)$ to be $\mathbf{V}_b(\boldsymbol{\theta}^*)$. Moreover, even if one could replace π^*_{+k} with $\tilde{\pi}_{+k}$, $\mathrm{Var}\,(\sqrt{N}\nabla l_N^W(\boldsymbol{\theta}^*))$ should be

$$\mathbf{V}_b(\boldsymbol{\theta}^*) - \sum_{k=0}^{K} \Big\{ \frac{1}{\lambda_k + \lambda_R \pi^*_{+k}} \mathbf{A}_k(\boldsymbol{\theta}^*)^{\times 2} \Big\} + o(1).$$

Since $\{\mathbf{A}_k(\boldsymbol{\theta}^*)'\mathbf{A}_k(\boldsymbol{\theta}^*)\}_{k=0}^K$ are all positive-semi-definite matrices, Morgenthaler and Vardi's formula underestimates the asymptotic variance of $\sqrt{N}\nabla l_N^W(\boldsymbol{\theta}^*)$.

When $\lambda_R = 1$, the ODE sample becomes a random sample. In this case, $\mathbf{V}_a(\boldsymbol{\theta}^*)$ is a square matrix of zeros and $\mathbf{H}(\boldsymbol{\theta}^*) = -\mathbf{V}_b(\boldsymbol{\theta}^*)$. The standard likelihood approach thus requires only $\mathbf{H}(\boldsymbol{\theta}^*)$ be full-rank (Rao (1973, Sec. 5.e.2)). To accommodate ODE sampling, we impose the stronger condition (C5) so that each $\mathbf{B}_k(\boldsymbol{\theta}^*)$, $k = 0, \ldots, K$, is positive-definite. This allows $\mathbf{H}(\boldsymbol{\theta}^*)$ to be negative-definite and hence, invertible. Application of Lemma 1 and Rao's general likelihood theory (Rao (1973, Sec. 5.e.2)) yields the following theorem.

Theorem 1. *Suppose that* $f(\mathbf{x}, y) = \rho(y, \mathbf{x}; \boldsymbol{\theta}^*)g(\mathbf{x})$ *and* $\boldsymbol{\theta}^*$ *is in the interior of the parameter space* Θ. *As* $N \to \infty$, $n_g/N \to \lambda_R$, $\lambda_R \neq 0$, *and* $n_k/N \to \lambda_k$, $k = 0, \ldots, K$. *If* (C1)−(C5) *are satisfied, the MLE at* (2.4), $\hat{\boldsymbol{\theta}}^W$, *satisfies*

$$\sqrt{N}(\hat{\boldsymbol{\theta}}^W - \boldsymbol{\theta}^*) \xrightarrow{D} \mathrm{Normal}(\mathbf{0}^q, \boldsymbol{\Sigma}(\boldsymbol{\theta}^*)),$$

where $\boldsymbol{\Sigma}(\boldsymbol{\theta}^*) = \mathbf{H}(\boldsymbol{\theta}^*)^{-1}\{\mathbf{V}_a(\boldsymbol{\theta}^*) + \mathbf{V}_b(\boldsymbol{\theta}^*)\}\mathbf{H}(\boldsymbol{\theta}^*)^{-1}$ *with* $\mathbf{H}(\boldsymbol{\theta}^*)$, $\mathbf{V}_a(\boldsymbol{\theta}^*)$, *and* $\mathbf{V}_b(\boldsymbol{\theta}^*)$ *as in Lemma* 1.

Proof. Since Theorem 1 follows from Lemma 1 along fairly well-known lines, we only sketch out the steps involved in its proof. Our derivation relies heavily on the fact that $\tilde{f}(\mathbf{x}, y)$ at (2.3) converges uniformly to $f(\mathbf{x}, y)$ (Vardi (1985)). Consequently,

$$l_N^W(\boldsymbol{\theta}) \overset{uniform}{\longrightarrow} \sum_{k=0}^{K} \int_{\mathscr{X}} \log(\rho(k, \mathbf{x}; \boldsymbol{\theta})) f(\mathbf{x}, k) d\mathbf{x} \equiv l^W(\boldsymbol{\theta}).$$

It is easy to verify that under (C1) and (C3), $\nabla l^W(\boldsymbol{\theta}^*) = \mathbf{0}^q$ and $\nabla^2 l^W(\boldsymbol{\theta}^*) = \mathbf{H}(\boldsymbol{\theta}^*)$. Condition (C5) implies that $\mathbf{H}(\boldsymbol{\theta}^*)$ is negative-definite. Thus, $\boldsymbol{\theta}^*$ must maximize $l^W(\boldsymbol{\theta})$. Conditions (C1) and (C4) imply that $\hat{\boldsymbol{\theta}}^W$, the estimator that maximizes $l_N^W(\boldsymbol{\theta})$, exists a.e. and converges a.e. to $\boldsymbol{\theta}^*$. A Taylor series expansion of $\nabla l_N^W(\hat{\boldsymbol{\theta}}^W)$ at $\boldsymbol{\theta}^*$ yields

$$\nabla l_N^W(\hat{\boldsymbol{\theta}}^W) = \nabla l_N^W(\boldsymbol{\theta}^*) + (\hat{\boldsymbol{\theta}}^W - \boldsymbol{\theta}^*)\nabla^2 l_N^W(\breve{\boldsymbol{\theta}}), \tag{3.4}$$

where $\breve{\boldsymbol{\theta}} = \gamma\boldsymbol{\theta}^* + (1-\gamma)\hat{\boldsymbol{\theta}}^W$ for some $\gamma \in [0,1]$. Obviously, $\breve{\boldsymbol{\theta}} \xrightarrow{a.e.} \boldsymbol{\theta}^*$. Under (C2), we have $\nabla^2 l_N^W(\breve{\boldsymbol{\theta}}) \xrightarrow{a.e.} \mathbf{H}(\boldsymbol{\theta}^*)$. Since $\hat{\boldsymbol{\theta}}^W$ maximizes $\nabla l_N^W(\boldsymbol{\theta})$, $\nabla l_N^W(\hat{\boldsymbol{\theta}}^W) = \mathbf{0}$. Then (3.4) implies that $\hat{\boldsymbol{\theta}}^W - \boldsymbol{\theta}^* = -\nabla l_N^W(\boldsymbol{\theta}^*)\{\nabla^2 l_N^W(\breve{\boldsymbol{\theta}})\}^{-1}$. An application of Lemma 1 and Slutsky's Theorem completes the proof.

Remark 2. Morgenthaler and Vardi (1986) incorrectly gave the asymptotic variance of $\sqrt{N}(\hat{\boldsymbol{\theta}}^W)$ as $\mathbf{H}(\boldsymbol{\theta}^*)^{-1}\mathbf{V}_b(\boldsymbol{\theta}^*)\mathbf{H}(\boldsymbol{\theta}^*)^{-1}$. The case study presented in Section 4 shows that this severely underestimates the variance of the intercept estimator in generalized linear models. However, if one of the row/column vectors in $\mathbf{H}(\boldsymbol{\theta}^*)^{-1}$ is orthogonal to all the $\mathbf{A}_k(\boldsymbol{\theta}^*)$ terms, some of the diagonal elements in $\boldsymbol{\Sigma}(\boldsymbol{\theta}^*)$, including the variance of the slope estimator in generalized linear models, are unaffected. Also, when the ODE sample contains only a random sample ($\lambda_R = 1$), both formulas lead to $\boldsymbol{\Sigma}(\boldsymbol{\theta}^*) = -\mathbf{H}(\boldsymbol{\theta}^*)^{-1}$.

Remark 3. The analysis of a TSOD sample requires that the second-stage sample contains units from each level of Y (Breslow, McNeney and Wellner (2003)). From the proofs of Lemma 1 and Theorem 1, we can see that this restriction is unnecessary for ODE sampling; as long as the random sample component is present, the model parameter is theoretically estimable, even if some or all of the Y-strata are missing.

For a realized ODE sample, $\boldsymbol{\Sigma}(\boldsymbol{\theta}^*)$ can be approximated by its plug-in estimator, which requires integrations in (3.1) with respect to an empirical distribution of $f(\mathbf{x}, y)$. The ODE sample offers three nonparametric estimates of $f(\mathbf{x}, y)$: the first utilizes its random sample component alone; the second estimates $\Pr(\mathbf{X} \mid Y = k)$ from the $Y = k$ stratum and then multiplies it by $\tilde{\pi}_{+k}$; the third is given by Vardi's NMPLE of $f(\mathbf{x}, y)$. We prefer Vardi's estimator, because it is constructed from the entire ODE sample. Accordingly, replace π^*_{+k} in (3.1) with $\tilde{\pi}_{+k}$. The variance of $\hat{\boldsymbol{\theta}}^W$ is approximated by $N^{-1}\hat{\boldsymbol{\Sigma}} = N^{-1}\{\nabla^2 l_N^W(\hat{\boldsymbol{\theta}}^W)\}^{-1}(\hat{\mathbf{V}}_a + \hat{\mathbf{V}}_b)\{\nabla^2 l_N^W(\hat{\boldsymbol{\theta}}^W)\}^{-1}$, where

$$\hat{\mathbf{V}}_a + \hat{\mathbf{V}}_b = \sum_{k=0}^{K}\Big\{\frac{Nn_k}{m_{+k}(n_k+m_{+k})}\hat{\mathbf{A}}_k^{\times 2}\Big\} + \sum_{k=0}^{K}\Big\{\frac{Nm_{+k}}{n_R(n_k+m_{+k})}\hat{\mathbf{B}}_k\Big\},$$

$$\hat{\mathbf{A}}_k = \frac{m_{+k}}{n_R(n_k+m_{+k})}\sum_{i=1}^{N}\{I(Y_i = k)\nabla\log(\rho(k,\mathbf{X};\hat{\boldsymbol{\theta}}^W))\},$$

$$\hat{\mathbf{B}}_k = \frac{m_{+k}}{n_R(n_k+m_{+k})}\sum_{i=1}^{N}[I(Y_i = k)\{\nabla\log(\rho(k,\mathbf{X};\hat{\boldsymbol{\theta}}^W))\}^{\times 2}].$$

The asymptotic variance of the PL estimator also involves integrating over $f(\mathbf{x}, y)$. We prefer estimating this variance using (2.3) as well (details are available from the authors).

4. Case Study Under Properly Specified Models

This section implements the WL and PL methods under properly specified generalized linear models.

4.1. Models using the canonical link

For a categorical outcome that is measured only on a nominal scale, the canonical link is commonly used to describe the effect of linear combinations of $\mathbf{X}$ on Y. The model is

$$\begin{aligned} &\log\Big(\frac{\rho(k,\mathbf{x};\boldsymbol{\theta})}{\rho(0,\mathbf{x};\boldsymbol{\theta})}\Big) = \alpha_k + \boldsymbol{\beta}_k\mathbf{x}', \qquad k = 1,\dots,K, \\ &\sum_{k=0}^{K}\rho(k,\mathbf{x};\boldsymbol{\theta}) = 1. \end{aligned} \tag{4.1}$$

The $1 \times q$ parameter vector $\boldsymbol{\theta}$ here consists of intercept terms, $\alpha_1,\dots,\alpha_k$, and slope terms, $\boldsymbol{\beta}_1,\dots,\boldsymbol{\beta}_k$, i.e., $\boldsymbol{\theta} = (\alpha_1\cdots\alpha_k\boldsymbol{\beta}_1\cdots\boldsymbol{\beta}_k)$. When $K = 1$, (4.1) reduces to the binary logit model, where the term "logit" refers to the link function $\text{logit}(p) = \log(p/(1-p))$. When $K \geq 2$, (4.1) is known as the baseline-category logit model (Sec. 7.1.1 of Agresti (2002)). Under this model,

$$\begin{aligned} \mathbf{A}_k(\boldsymbol{\theta}^*)\mathbf{H}(\boldsymbol{\theta}^*)^{-1} &= \Big(-\mathbf{e}_k \quad \mathbf{0}^{q-K}\Big), \quad k = 1,\dots,K, \\ \mathbf{A}_0(\boldsymbol{\theta}^*)\mathbf{H}(\boldsymbol{\theta}^*)^{-1} &= \Big(\sum_{k=1}^{K}\mathbf{e}_k \quad \mathbf{0}^{q-K}\Big), \end{aligned}$$

where $\mathbf{e}_k$ represents the $1\times K$ unit vector whose kth element is one. It is thus seen that omitting $\mathbf{V}_a(\boldsymbol{\theta}^*)$ causes the variance of the WL intercept estimator of (4.1) to be underestimated while the variance of the slope estimator is not affected. Also note that without the random sample component ($n_R = 0$), the PL objective function (2.2) can be viewed as the likelihood function from a random sample with parameters $\alpha_k + \log(\pi_{+0}n_k/(\pi_{+k}n_0))$ and $\boldsymbol{\beta}_k$, $k = 1,\dots,K$. As a result, only $\boldsymbol{\beta}_1,\dots,\boldsymbol{\beta}_K$ can be estimated. The values $\alpha_1,\dots,\alpha_K$ and $\pi_{+1},\dots,\pi_{+K}$ cannot be identified separately from a Y-stratified sample. In this situation, the PL method is equivalent to the method developed specifically for case-control studies (Prentice and Pyke (1979)). As mentioned in Section 2, existence of Vardi's NPMLE requires the presence of the random sample component. The WL method is thus inapplicable for analyzing a Y-stratified sample. In fact, plugging $n_R = m_0 = \cdots = m_K = 0$ into (2.3) would lead to an undefined term of zero divided by zero.

Consider the case in which Y is binary, $\mathbf{X} = (X_1\ X_2)$, X_1 and X_2 are independent, $X_1 \sim \text{Normal}(0,1)$, and $X_2 \sim \text{Bernoulli}(0.5)$. ODE samples with

$N = 500$ and $n_0 = n_1$ were simulated according to the following logit model at $\alpha^* = -2.5$, $\beta_1^* = 0.2$, $\beta_2^* = 0.6$ ($\pi_{+1}^* \approx 0.1$):

$$\text{logit}(\rho(1, (x_1\ x_2); \boldsymbol{\theta})) = \alpha + \beta_1 x_1 + \beta_2 x_2. \tag{4.2}$$

Table 1.a results from fitting (4.2) to one thousand simulated ODE samples at $n_R/N = 1$, 0.8, 0.5, 0.2, or 0. All likelihood maximizations were successful. At $n_R/N = 0$, the WL method is infeasible and the PL method estimates only β_1 and β_2. As n_R/N increases, the SE of the intercept estimator decreases while the SE of the slope estimator increases. From the coverage rate (CR) of confidence intervals constructed with a 95% nominal level, we observe that Morgenthaler and Vardi's formula underestimates the SE of the WL estimator for α and has no effect on the SE of the WL estimators for β_1 and β_2 (see the CR labeled with "$*$").

4.2. Models using non-canonical links

For a binary outcome, alternatives to the logit link include the probit and complimentary-log-log links. When $K \geq 2$ and levels of Y are ordered, cumulative logit/probit/complimentary-log-log links are generally preferred over the baseline-category logit link. If one chooses these links, the diagonal elements in $\mathbf{V}_a(\boldsymbol{\theta}^*)$ are positive. Nevertheless, results of our simulation show that omitting $\mathbf{V}_a(\boldsymbol{\theta}^*)$ has a great impact on estimating the variance of the WL intercept estimator and, to a lesser degree, on that of the WL slope estimator. For continuous or partially continuous $\mathbf{X}$, the intercept terms in these models can be identified, in theory, from a Y-stratified sample using the PL method. However, our simulation shows that without the random sample component, the likelihood maximization either fails to converge or carries large estimation bias for the intercept terms.

Let Y be a trinary outcome, i.e., $K = 2$, and let $\mathbf{X}$ be a univariate standard normal random variable. One thousand ODE samples with $N = 600$ and $n_0 = n_1 = n_2$ were simulated according to the following cumulative logit model at $\alpha_1^* = -2$, $\alpha_2^* = -1$, $\beta^* = 0.2$ ($\pi_{+0}^* \approx 0.12$, $\pi_{+1}^* \approx 0.15$):

$$\begin{aligned} \text{logit}(\rho(0, x; \boldsymbol{\theta})) &= \alpha_1 + \beta x, \\ \text{logit}(\rho(0, x; \boldsymbol{\theta}) + \rho(1, x; \boldsymbol{\theta})) &= \alpha_2 + \beta x. \end{aligned} \tag{4.3}$$

Results in Table 1.b are obtained by fitting the simulated samples with Model (4.3). Convergence of the likelihood maximization took place for all the random and ODE samples ($n_R/N = 1$, 0.85, 0.5, 0.15), whereas maximization of the PL function failed in 94% of the Y-stratified samples ($n_R/N = 0$). Again, omitting $\mathbf{V}_a(\boldsymbol{\theta}^*)$ causes the SE of the intercept estimator to be underestimated. As n_R/N

Table 1. Simulation results for properly specified models. Values labeled with "$*$" correspond to the coverage rates (CRs) of 95% confidence intervals constructed using Morgenthaler and Vardi's formula. Results posted here are based on samples with converged likelihood maximization.

a. Model (4.2).

		$\hat{\alpha}$			$\hat{\beta}_1$			$\hat{\beta}_2$		
n_R/N		Bias	SE	CR	Bias	SE	CR	Bias	SE	CR
1	PL/WL	-0.039	0.251	0.952	0.002	0.154	0.938	0.013	0.315	0.953
0.8	PL	-0.037	0.224	0.949	0.008	0.122	0.952	0.015	0.235	0.950
	WL	-0.037	0.225	0.949	0.008	0.122	0.954	0.016	0.235	0.952
				0.897*			0.954*			0.952*
0.5	PL	-0.036	0.248	0.949	0.004	0.101	0.947	0.006	0.201	0.953
	WL	-0.037	0.248	0.95	0.005	0.102	0.949	0.006	0.202	0.952
				0.787*			0.949*			0.952*
0.2	PL	-0.040	0.372	0.957	0.000	0.097	0.943	0.011	0.194	0.940
	WL	-0.041	0.372	0.957	0.001	0.098	0.953	0.012	0.194	0.941
				0.553*			0.953*			0.941*
0	PL	NA	NA	NA	0.005	0.092	0.955	0.008	0.184	0.948

b. Model (4.3).

		$\hat{\alpha}_1$			$\hat{\alpha}_2$			$\hat{\beta}$		
n_R/N		Bias	SE	CR	Bias	SE	CR	Bias	SE	CR
1	PL/WL	-0.003	0.120	0.961	0.000	0.091	0.955	0.007	0.093	0.939
0.85	PL	-0.010	0.137	0.962	-0.005	0.103	0.932	0.009	0.087	0.950
	WL	-0.010	0.137	0.962	-0.005	0.103	0.932	0.009	0.087	0.955
				0.898*			0.907*			0.954*
0.5	PL	-0.015	0.182	0.944	-0.009	0.135	0.950	0.001	0.082	0.941
	WL	-0.015	0.182	0.943	-0.009	0.135	0.952	0.001	0.083	0.945
				0.729*			0.797*			0.944*
0.15	PL	-0.045	0.330	0.963	-0.018	0.240	0.959	0.006	0.081	0.950
	WL	-0.045	0.330	0.964	-0.017	0.240	0.960	0.006	0.083	0.955
				0.387*			0.546*			0.954*
0	PL	0.541	0.824	1.000	0.844	0.452	1.000	0.024	0.058	1.000

increases, we gain precision in estimating the intercept, but lose precision in estimating the slope.

4.3. Comparison of the PL and WL methods

For an ODE sample when the assumed working model is correct, Table 1

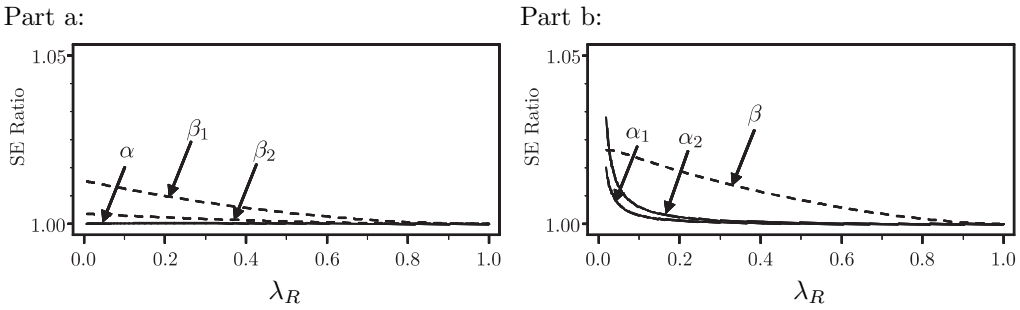

Figure 1. Plots of the asymptotic SE ratio of the WL vs. PL estimator at various values of . Proportions of different stratified components are held equal here. Part a corresponds to Model (4.2). Part b corresponds to Model (4.3). Solid lines represent SE ratio for the intercept terms in the generalized linear model. Dashed lines represent SE ratio for the slope terms.

demonstrates that the PL and WL methods perform similarly: (i) they both have low estimation bias; (ii) the SE of the WL estimator is only slightly higher than that of the PL estimator; (iii) the CRs of their confidence intervals both approach the nominal level. This similarity is also observed from our simulations using the binary probit/complementary-log-log models, and the trinary baseline-category logit models (results not shown).

Recall that the PL estimator is asymptotically most efficient. If so, when does the PL method outperform the WL method? To answer this question, we computed the asymptotic SE of both estimators using the same parameter settings as in Table 1. Figure 1 plots the ratio of their asymptotic SE against an array of λ_R values (proportions of different strata are held equal). The fact that the curves are always above one confirms that the PL method is asymptotically more efficient than the WL method. Nevertheless, the difference between these two methods is negligible unless λ_R is very close to zero. A heuristic explanation for this is that estimation of π_{+1} at small λ_R relies more on the model assumption than on the random sample component. Consequently, the PL method works better than the WL method. This advantage, however, may be lost in practice; if some of the outcome levels are rare, one needs a sufficiently large random sample to observe enough $\mathbf{X}$ at each level of Y. The advantage of collecting a Y-stratified sample is that it estimates the odds ratio between Y and $\mathbf{X}$ without requiring a large sample size. In reality, the size of the Y-stratified sample is not many folds larger than that of the random sample. For example, the size of Doll and Hill's (1950) case-control data is around fourteen hundred and Doll's cohort data has over twenty-five thousand participates (Doll et al. (2004)). This corresponds to a λ_R value around 0.95.

Profiling and weighting are two different strategies for constructing objective functions from complex survey data. In the analysis of a TSOD sample, profiling involves exploiting the model assumption to patch up the missing $\mathbf{X}$ values in the first-stage sample with what is observed in the second-stage sample (Scott and Wild (1997)), whereas weighting refers to using the sampling weight to adjust the likelihood function assembled from the second-stage sample (Kalbfleisch and Lawless (1988)). Breslow and Chatterjee (1999) observed that these two strategies generate comparable mean-squared errors when stratification in the second stage is exclusively on Y. Nevertheless finer stratification, such as stratification on both Y and a partially measured $\mathbf{X}$, manifests the high efficiency associated with likelihood profiling (Breslow and Chatterjee (1999); Lawless, Kalbfleisch and Wild (1999)). It is important to stress that the ODE sampling considered here involves only stratification on Y. Parameter estimation under more intricate enriched sampling plans calls for reformulation of the WL/PL objective functions. More works need to be done in this area.

5. Robustness

This section investigates the robustness of the WL and PL methods when the assumed working model is incorrect. Because there is no perfect model, Scott and Wild (1986) suggested that a meaningful model parameter value maximizes the expectation of the sampling-weight-adjusted log-likelihood function. For a binary outcome, Manski and Thompson (1989) considered the loss function $-\log(1 - |Y - \rho(Y, \mathbf{X}; \boldsymbol{\theta})|)$ and stated that among all models of the form $\rho(Y\mathbf{X}; \boldsymbol{\theta})$, the best is at $\boldsymbol{\theta}^*$, where

$$\boldsymbol{\theta}^* \in \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}} \Big\{ \sum_{k=0}^{1} \int_{\mathscr{X}} \log(1 - |k - \rho(k, \mathbf{X}; \boldsymbol{\theta})|) f(\mathbf{x}, k) d\mathbf{x} \Big\}. \tag{5.1}$$

Both ideologies require that $\boldsymbol{\theta}^*$ maximizes the integral of $\log(\rho(Y, \mathbf{X}; \boldsymbol{\theta}))$ over the joint distribution of $\mathbf{X}$ and Y which, in ODE sampling, corresponds to the asymptote of the WL objective function $l_N^W(\boldsymbol{\theta})$ at (2.4). The uniform convergence of Vardi's NPMLE to $f(\mathbf{x}, y)$ suggests that the WL method offers a consistent estimator of $\boldsymbol{\theta}^*$ even when the model is misspecified. In contrast, the PL method relies heavily on the validity of the model. Consistency of its estimator does not hold if the model is incorrect.

Chatterjee, Chen, and Breslow (2003) defined $\boldsymbol{\theta}^*$ according to (5.1) and examined the bias of their pseudoscore estimator from a TSOD sample. Our simulation study adopted their setting and generated ODE samples with binary outcomes under the logit model with a quadratic term: $\operatorname{logit}(\Pr(Y = 1 \mid X = x)) = -2 + x + \delta x^2$, $X \sim \text{Normal}(0, 1)$. The model used to fit the ODE sample

Table 2. Simulation results under misspecified models. The true model is $\text{logit}(\Pr(Y = 1 \mid X = x)) = -2 + x + \delta x^2$. The working model is $\text{logit}(\rho(1, x; \boldsymbol{\theta})) = \alpha + \beta x$. All likelihood maximizations converged.

Part a: $\delta = 0.3$. $(\alpha^* = -1.808, \beta^* = 1.205)$

		$\hat{\alpha}$			$\hat{\beta}$		
N		Bias	SE	CR	Bias	SE	CR
500	PL	0.007	0.180	0.955	-0.054	0.138	0.915
	WL	-0.015	0.184	0.961	0.014	0.141	0.954
1000	PL	0.003	0.127	0.956	-0.060	0.100	0.882
	WL	-0.019	0.130	0.959	0.007	0.102	0.945
2000	PL	0.021	0.092	0.937	-0.069	0.067	0.831
	WL	-0.001	0.094	0.949	-0.002	0.069	0.945

Part b: $\delta = 0.6$. $(\alpha^* = -1.484, \beta^* = 1.115)$

		$\hat{\alpha}$			$\hat{\beta}$		
N		Bias	SE	CR	Bias	SE	CR
500	PL	0.013	0.173	0.948	-0.094	0.148	0.869
	WL	-0.013	0.179	0.951	0.014	0.154	0.947
1000	PL	0.016	0.120	0.951	-0.098	0.101	0.821
	WL	-0.010	0.124	0.956	0.011	0.106	0.952
2000	PL	0.021	0.086	0.939	-0.103	0.070	0.692
	WL	-0.004	0.089	0.941	0.005	0.074	0.958

lacks the quadric term: $\text{logit}(\rho(1, x; \boldsymbol{\theta})) = \alpha + \beta x$. We set δ to be either 0.3 or 0.6. This represents the situation where the model is either moderately or severely misspecified. According to (5.1), $\boldsymbol{\theta}^* = (\alpha\beta)$ was (-1.808 1.205) at $\delta = 0.3$, and (-1.484 1.115) at $\delta + 0.6$. Results in Table 2 are based on one thousand ODE samples with $n_R/N = 0.5$ and $n_0/N = n_1/N = 0.25$. As we anticipated, the bias of the WL estimator got closer to zero when the size of the ODE sample increased, while that of the PL estimator drifted away from zero. Table 2 also demonstrates that the SE of the PL estimator was smaller than that of the WL estimator, but the difference was quite small. The WL method had CRs close to the desired 95% nominal level in all cases we examined. For the PL method, its CR of the confidence interval for β^* dropped to as low as 69.2% when the model was severely misspecified ($\delta = 0.6$) and the sample size was large ($N = 2,000$). Whittemore (1997) and Chatterjee, Chen, and Breslow (2003) also recognized this lack of robustness associated with heavily model-based methods while analyzing other types of complex survey data.

In practice, there is a sense of balance between controlling the bias and SE in the analysis of an ODE sample. If one is certain about the model assumption, the PL method has the best efficiency and, of course, should be employed. On the other hand, if the fitness of the presumed model is suspect and the sample size is large, eliminating the estimation bias should be the main concern; then we recommend the WL method.

6. An Example

Here, we illustrate the WL and PL methods using an ODE sample drawn from a national survey. Between 1988 and 1994, the National Center for Health Statistics conducted a survey (also known as NHANES III) to obtain health and nutrition information on the US population. For convenience, we only consider the 16,971 adults whose age, race, blood pressure and body weight are all non-missing (see Hosmer and Lemeshow (2000)). Each individual in the survey dataset carried a sampling weight to represent the number of people in the US possessing the same health and demographic characteristics. Values of this sampling weight ranged from 226 to 139,745. Treat the NHANES III data as a mirror image of the US population. A random sample from the US population is therefore equivalent to a Probability-Proportional-to-Size-With-Replacement (PPSWR) sample from the NHANES III data, where "Size" refers to the sampling weight. PPSWR sampling was executed via SAS PROC SURVEYSELECTR®with option METHOD=PPS_WR (SAS Institute (2004)). Let $Y = 1$ be the event that an adult had high blood pressure (systolic blood pressure over 140 mmHg); $Y = 0$, otherwise. The Y-stratified sample was collected by taking independent PPSWR samples from the high and regular blood pressure subpopulations. The resulting data was fit with the binary logit linear model where age, race and body weight served as predictors. Table 3 shows that combining the random sample with the Y-stratified sample improved the estimation precision not only for the slope terms, but also for the intercept term. For example, the SE estimate of age was 0.0146 from the random sample, and it was 0.0117 from the Y-stratified sample. Analyzing the combined sample cut SE estimates to 0.0076 (PL method) and 0.0079 (WL method). For the intercept term, the SE estimate was reduced by almost 50% from 1.39 (random sample) to 0.70 (PL method) and 0.72 (WL method). SE estimates computed using Morgenthaler and Vardi's formula (values labeled with '$*$' in Table 3) were all numerically smaller than those computed using our proposed formula. Rounding to two significant digits, nevertheless, camouflages this underestimation except for the intercept term. Even though the WL and PL methods provided slightly different parameter estimates, both indicate that senior, black, and obese individuals are among the high-risk group for hypertension.

Table 3. Analysis of an ODE sample taken from a national survey. Values labeled with "$*$" correspond to the SEs given by Morgenthaler and Vardi's formula.

		Intercept		Age (yr.)		Race (Black vs. Other)		Weight (lb.)	
Sample	Method	Est.	SE	Est.	SE	Est.	SE	Est.	SE
Random $n_R = 250$	PL/WL	-6.39	1.39	0.081	0.0146	0.60	0.62	0.0011	0.0053
Y-stratified $n_0 = n_1 = 125$	PL	NA	NA	0.096	0.0117	1.17	0.57	0.0114	0.0040
ODE	PL	-7.65	0.70	0.083	0.0076	0.69	0.37	0.0083	0.0026
$n_R = 250$	WL	-7.52	0.72	0.082	0.0079	0.65	0.37	0.0078	0.0026
$n_0 = n_1 = 125$			0.70*		0.0079*		0.37*		0.0026*

7. Discussion

Studies that collect outcome-stratified samples often match units at different levels of Y according to some other discrete variable, say W. Wang and Zhou (2006) applied the PL method to an ODE sample where the stratification is on both Y and W. Their approach treats W as auxiliary in the sense that $\Pr(Y = k \mid \mathbf{X} = \mathbf{x}, W) = \Pr(Y = k \mid \mathbf{X} = \mathbf{x}) = \rho(k, \mathbf{x}; \boldsymbol{\theta})$, $k = 0, \ldots, K$. Moreover, the joint distribution of $\mathbf{X}$ and W is assumed not to depend on $\boldsymbol{\theta}$. In adopting Wang and Zhou's formulation, the WL method calls for only small alterations to (2.4). To see this, suppose W has L levels. According to Vardi (1985), the WL method should maximize

$$l_N^W(\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{l=1,k=0}^{L,K} \Big\{ \frac{m_{+lk}}{n_R} \frac{I(W_i = l, Y_i = k)}{n_{lk} + m_{+lk}} \log(\rho(Y_i, \mathbf{X}_i; \boldsymbol{\theta})) \Big\},$$

where n_{lk} is the size of the $W = l$, $Y = k$ stratum, and m_{+lk} is the count of units in the random component with $W = l$ and $Y = k$, $l = 1, \ldots, L$, $k = 0, \ldots, K$. So far, it is not clear how this type of stratification affects the precision of parameter estimation in ODE sampling. It would be interesting to compare the WL and PL methods under this finer stratification.

In summary, this paper explores semiparametric methods for the analysis of an ODE sample when the conditional probability of the outcome variable given the predictor is specified up to certain unknown parameters but the marginal distributions of the outcome and the predictor are unknown. Under reasonable regularity conditions, the estimator that maximizes the WL objective function has an asymptotic normal distribution with mean the true model parameter value. Although the PL method is asymptotically most efficient, it involves estimating the marginal probability of the outcome based on the model. In contrast,

the WL objective function has fewer arguments over which to maximize. Simulation and an asymptotic comparison indicate that performance of the WL method is often comparable to the PL method under properly specified models. On the other hand, the PL method is vulnerable to model misspecification, whereas the WL method still offers a meaningful parameter estimate. The robustness of the WL method leads us to promote its use, especially for the situation where the fitness of the assumed working model is uncertain and the sample size is large. Future research will concentrate on developing likelihood-ratio tests for model goodness-of-fit and nested models, and on applying these methodologies to data.

Acknowledgements

This research was completed while the first and third authors were assistant professors at North Dakota State University, Fargo. We would like to thank the Editor, an Associate Editor, and the anonymous referees for their helpful comments on this paper.

References

Agresti, A. A. (2002). *Categorical Data Analysis*. Wiley-Interscience, Hoboken, NJ.

Breslow, N. E. and Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Appl. Statist.* **48**, 457-468.

Breslow, N., McNeney, B. and Wellner, J. A. (2003). Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. *Ann. Statist.* **31**, 1110-1139.

Chatterjee, N., Chen, H. Y. and Breslow, N. E. (2003). A pseudoscore estimator for regression problems with two-phase sampling. *J. Amer. Statist. Assoc.* **98**, 158-168.

Chen, H. Y. (2003). A note on the prospective analysis of outcome-dependent samples. *J. Roy. Statist. Soc. Ser. B* **65**, 575-584.

Cosslett, S. R. (1981a). Efficient estimation of discrete-choice models. In *Structural Analysis of Discrete Data With Econometric Applications* (Edited by C. Manski and D. McFadden), 51-111. The MIT Press, Cambridge, MA.

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

Cosslett, S. R. (1993). Estimation from endogenously stratified samples. In *Handbook of Statistics* (Vol.11): Econometric (Edited by C. Maddala, C. Rao and H. Vinod), 1-44. Elsevier/North-Holland, New York /Amsterdam.

Doll, R. and Hill, A. B. (1950). Smoking and carcinoma of the lung. *Br. Med. J.* **221**, 739-748.

Doll, R., Peto, R., Boreham, J. and Sutherland, I. (2004). Mortality in relation to smoking: 40 years' observations on male British doctors. *Br. Med. J.* **328**, 1519-1527.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*. John Wiley, New York.

Kalbfleisch, J. D. and Lawless, J. F. (1988). Likelihood analysis of multi-state models for disease incidence and mortality. *Statist. Med.* **7**, 149-160.

Kullback, S. (1997). *Information Theory and Statistics*. Dover Publications, New York.

Lawless, J. F., Kalbfleisch, J. D. and Wild, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *J. Roy. Statist. Soc. Ser. B* **61**, 413-438.

Lehmann, E. L. (1999). *Elements of Large-Sample Theory*. Springer, New York.

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

Manski, C. F. and Thompson, T. S. (1989). Estimation of best predictors of binary response. *J. Econometrics* **40**, 97-123.

Morgenthaler, S. and Vardi, Y. (1986). Choice-based samples: a nonparametric approach. *J. Econometrics* **32**, 109-125.

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

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. John Wiley, New York.

SAS Institute. (2004). *SAS/STAT* 9.1 *User's Guide*. SAS Publishing, Cary, NC.

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

Scott, A. and Wild, C. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* **84**, 57-71.

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

Wang, X. F. and Zhou, H. B. (2006). A semiparametric empirical likelihood method for biased sampling schemes with auxiliary covariates. *Biometrics* **62**, 1149-1160.

Whittemore, A. S. (1997). Multistage sampling designs and estimating equations. *J. Roy. Statist. Soc. Ser. B* **59**, 589-602.

Zhou, H. B., Weaver, M. A., Qin, J., Longnecker, M. P. and Wang, M. C. (2002). A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. *Biometrics* **58**, 413-421.

7169 Maple Bluff Place, Indianapolis, IN 46236, U.S.A.

E-mail: kangqing@hotmail.com

Department of Statistics, Kansas State University, Manhattan, KS 66506, U.S.A.

E-mail: nels@ksu.edu

Elanco Animal Health, Greenfield, IN 46140, U.S.A.

E-mail: vahlch@lilly.com

(Received November 2008; accepted May 2009)