The International Journal of Biostatistics

Volume 8, *Issue* 1 2012 *Article* 16

Maximum Likelihood Estimation for Semiparametric Density Ratio Model

Guoqing Diao, *George Mason University*
Jing Ning, *University of Texas M.D. Anderson Cancer Center*
jing qin, *biostatistics research brance, NIAID, NIH*

Recommended Citation:
Diao, Guoqing; Ning, Jing; and qin, jing (2012) "Maximum Likelihood Estimation for Semiparametric Density Ratio Model," *The International Journal of Biostatistics*: Vol. 8: Iss. 1, Article 16.
DOI: 10.1515/1557-4679.1372

Maximum Likelihood Estimation for Semiparametric Density Ratio Model

Guoqing Diao, Jing Ning, and jing qin

Abstract

In the statistical literature, the conditional density model specification is commonly used to study regression effects. One attractive model is the semiparametric density ratio model, under which the conditional density function is the product of an unknown baseline density function and a known parametric function containing the covariate information. This model has a natural connection with generalized linear models and is closely related to biased sampling problems. Despite the attractive features and importance of this model, most existing methods are too restrictive since they are based on multi-sample data or conditional likelihood functions. The conditional likelihood approach can eliminate the unknown baseline density but cannot estimate it. We propose efficient estimation procedures based on the nonparametric likelihood. The nonparametric likelihood approach allows for general forms of covariates and estimates the regression parameters and the baseline density simultaneously. Therefore, the nonparametric likelihood approach is more versatile than the conditional likelihood approach especially when estimation of the conditional mean or other quantities of the outcome is of interest. We show that the nonparametric maximum likelihood estimators are consistent, asymptotically normal, and asymptotically efficient. Simulation studies demonstrate that the proposed methods perform well in practical settings. A real example is used for illustration.

KEYWORDS: semiparametric density ratio model, conditional likelihood, generalized linear model, nonparametric maximum likelihood estimation, semiparametric efficiency

Author Notes: The authors are grateful to the Associate Editor and the two anonymous referees for their careful reading of the earlier version of the manuscript, and providing valuable comments and suggestions. The first author's research was supported in part by National Cancer Institute grant CA150698.

1 Introduction

Consider a multinomial model

$$P(Y=k|X)=\frac{\exp(\alpha_k+\beta_k X)}{\sum_{j=1}^{K}\exp(\alpha_j+\beta_j X)},\ \ k=1,2,...,K$$

where $\alpha_1=\beta_1=0$ and the distribution of X has an unspecified density $f(x)$. In typical case and control studies (Anderson and Philips, 1981), one may collect covariate information X in each category $Y=k,k=1,2,\ldots,K$. In contrast to prospective studies, sample size in the kth group $n_k,k=1,2...,K$ are fixed in case and control studies. Using Bayes's formula, one can easily observe that for $k=2,....,K$,

$$\frac{f(x|Y=k)}{f(x|Y=1)}=\left[\frac{P(Y=k|x)f(x)}{P(Y=k)}\right]\left[\frac{P(Y=1|x)f(x)}{P(Y=1)}\right]^{-1}=\exp(\alpha_k^*+\beta_k x),$$

where $f(x|Y=k)$ is the conditional density of X given $Y=k$ and $\alpha_k^*=\alpha_k-\log[P(Y=k)/P(Y=1)]$. In other words $f(x|Y=k)$ and $f(x|Y=1)$ are linked by the exponential tilting functions $\exp(\alpha_k^*+\beta_k x),k=2,\ldots,K$. Note that α_k^* is a normalizing constant satisfying

$$1=\int f(x|Y=k)dx=\int \exp(\alpha_k^*+\beta_k x)f(x|Y=1)dx$$

or

$$\exp(\alpha_k^*)=\frac{1}{\int \exp(\beta_k x)f(x|Y=1)dx}.$$

Immediately, we have

$$f(x|Y=k)=\frac{\exp(\beta_k x)f(x|Y=1)}{\int \exp(\beta_k z)f(z|Y=1)dz},k=2,\ldots,K.$$

The above model is also called the density ratio model, where $f(x|Y=1)$ is an unspecified baseline density. The density ratio model has a close connection with the Cox proportional hazards model (Cox, 1972), where the hazard functions satisfy

$$\lambda_k(t)=\lambda(t)\exp(\gamma_k),k=2,\ldots,K,$$

in which $\lambda(t)$ is an unspecified baseline hazard. However, in contrast to the Cox proportional hazards model, a normalizing constant of $\int \exp(\beta x) f(x|Y=1)dx$ is needed in the density ratio model.

In statistical literature the density ratio model can be viewed as a special case of biased sampling problems, where $w_k(x,\beta) \equiv \exp(\beta_k x), k = 2,\ldots,K$ can be treated as the weighting functions. When the target population cannot be observed directly, then one has a biased sampling problem. Biased sampling problems were described long ago. In particular, Cox (1969) and Vardi (1982) discussed length-biased sampling problems and their many important applications. Biased sampling problems have also been discussed extensively in areas such as case and control studies, missing data problems, and casual inference. When the weighting functions do not depend on any unknown parameters, Multiple biased sampling problems for weighting functions that do not depend on any unknown paraters have also been systematically discussed previously, by Vardi (1985) and Gill, Vardi, and Wellner (1988). By contrast, for situations in which weighting functions do depend on some unknown parameters, Anderson (1979), Qin and Zhang (1997), and Gilbert, Lele, and Vardi (1999) studied the semiparametric maximum likelihood estimator. Chen (2001) studied the nonparametric maximum likelihood estimator under the choice-based sampling problem. Liang and Qin (2000) used a pairwise conditional likelihood approach to eliminate the baseline density function. More recently, Bondell (2005) and Bondell (2008) proposed robust estimation procedures incorporating minimum distance approaches based on the cumulative distribution functions and the characteristic functions, respectively.

With the exception of Qin and Liang (1999) and Liang and Qin (2000), the aforementioned methods are concerned with multi-sample data. It is straightforward to generalize the discrete density ratio model to a density ratio model with a vector of covariates $\mathbf{X}$

$$f(y|\mathbf{X}) = \frac{f(y)\exp(y\beta^T\mathbf{X})}{\int_{\mathscr{Y}} f(z)\exp(z\beta^T\mathbf{X})d\mu(z)}. \tag{1}$$

In the above model, for a continuous outcome variable, $\mu(\cdot)$ is the dominating Lebesgue measure n $\mathscr{Y}$ and $f(y)$ is the unspecified baseline density function; whereas for a discrete outcome variable, $\mu(\cdot)$ is the dominating counting measure in $\mathscr{Y}$ and $f(\cdot)$ is the unspecified baseline probability mass function. The vector of covariates $\mathbf{X}$ may include both discrete and continuous covariates. In statistical literature $f(y)$ is also called the "carrier density". Clearly if the form of $f(y)$ is known, then

$$f(y|\mathbf{X}) = f(y)\exp\{\alpha(\beta^T\mathbf{X}) + y\beta^T\mathbf{X}\},\ \exp\{\alpha(\beta^T\mathbf{X})\} = \frac{1}{\int_{\mathscr{Y}} f(y)\exp(y\beta^T\mathbf{X})dy},$$

where $\alpha(\beta^T\mathbf{X})$ is a known function of $\beta^T\mathbf{X}$. This density ratio model becomes the well-known exponential family model. The multinomial model can also be shown by Bayes's rule to be a special case of the density ratio model (1) with a discrete covariate. However, if $f(y)$ is unknown, then $\alpha(\beta^T\mathbf{X})$ is also unknown. The conventional parametric likelihood cannot be used.

The density ratio model (1) also has a close connection with the generalized linear models. Consider a generalized linear model with conditional density function

$$f(y|\mathbf{X};\theta,\phi) = \exp[a(\phi)\{y\theta - b(\theta)\} + c(y,\phi)], \tag{2}$$

where $\theta = g(\gamma^T\mathbf{X})$, g is a known link function, ϕ is the dispersion parameter, and functions $a(\phi), b(\theta)$, and $c(y,\phi)$ are known. This model was proposed by Nelder and Wedderburn (1972). When g is the identity function, the density ratio model includes the generalized linear model (2) as a special case, in which, $\beta = a(\phi)\gamma$ and $f(y) = \exp[-a(\phi)b(\theta) + c(y,\phi)]$. The generalized linear models with the identify function of g include the linear regression, logistic regression, and Poisson regression models.

In the density ratio model (1), for two individuals with outcome and covariates values of $(y,\mathbf{x})$ and $(y^*,\mathbf{x}^*)$, respectively,

$$\frac{f(y|\mathbf{x})/f(y^*|\mathbf{x})}{f(y|\mathbf{x}^*)/f(y^*|\mathbf{x}^*)} = \exp\{(y^*-y)\beta^T(\mathbf{x}^*-\mathbf{x})\}.$$

Thus β characterizes the effect of $\mathbf{x}$ on the "odds" of y through its probability (density) function. It is therefore also called the generalized "odds ratio" model (Qin and Liang, 1999, Liang and Qin, 2000). This is similar to the Cox proportional hazards model (Cox, 1972) in which the hazards ratio has a known parametric form and the McCullagh (1980) proportional odds ratio model for ordinal categorical response data.

In the Cox proportional hazards model, Cox (1972, 1975) used the product of conditional likelihood approach at each failure time point to eliminate the baseline hazard $\lambda(t)$. In the density ratio model (1), Kalbfleisch (1978) was able to eliminate the baseline density $f(y)$ and also the unknown normalizing constant function $\alpha(\beta^T\mathbf{X})$ by conditioning on the order statistics $y_{(1)} < ... < y_{(n)}$. However, this method is not practical since it involves a permutation with $n!$ terms. For moderate sample size, say $n = 30$, there are at least 10^{30} permutation terms. To reduce computational burden, Qin and Liang (1999) and Liang and Qin (2000) proposed a pair-wise conditional likelihood method. The pair-wise likelihood can eliminate the unknown baseline density and the normalizing function as well. Moreover, it has a nice simple form. However, the conditional likelihood approach cannot estimate the entire distribution. In many applications, the estimation of the conditional mean or other quantities of Y is of interest. The conditional likelihood approach is not applicable under such situations. Recently, Rathouz and Gao (2009) considered the maximum likelihood estimation for model (1) with discrete outcome variables, which is essentially a parametric approach.

To overcome the limitations of existing methods, in Section 2, we develop efficient nonparametric likelihood-based estimation procedures which allow for general forms of covariates and simultaneously estimate the regression parameters and the baseline density function. The proposed nonparametric maximum likelihood estimators (NPMLEs) are shown to be consistent and asymptotically normal. In addition, the NPMLEs of the regression parameters are asymptotically efficient, that is, the limiting covariance matrix achieves the semiparametric efficiency bound defined in Chapter 3 of Bickel, Klaassen, Ritov, and Wellner (1993). Simulation studies in Section 3 demonstrate that the proposed estimators perform well in practical situations. An application to a well-known acute leukemia survival data set (Feigl and Zelen, 1965) is provided in Section 4. We offer a discussion of our methods in Section 5. Technical details are relegated to the Appendix.

2 Main Results

Let $(Y, \mathbf{X})$ represent a pair of a response and a $q \times 1$ vector of covariates in the population. We consider the density ratio model in (1). Let $F(y) = \int_{-\infty}^{y} f(z)dz$ denote the baseline cumulative distribution function (CDF). The density ratio model (1) is considered a semiparametric model, in the sense that $F(\cdot)$ is an infinite-dimensional parameter. Given a random sample of n observations $\{(Y_i, \mathbf{X}_i), i = 1, \ldots, n\}$, our goal is to make inferences about the unknown parameters (β, F).

In this section, we develop efficient estimation and inference procedures about (β, F) based on the nonparametric likelihood approach. The likelihood function from a random sample of n observations has the form

$$L_n(\beta, F) = \prod_{i=1}^{n} \frac{f(Y_i)\exp(Y_i \beta^T \mathbf{X}_i)}{\int_{\mathscr{Y}} f(z)\exp(z\beta^T \mathbf{X}_i)dz}.$$

Ideally, we want to obtain the estimators of (β, F) by maximizing the above likelihood function. The maximum of $L_n(\beta, F)$, however, does not exist, because we can always let $f(Y_i)$ go to infinity for some i. Therefore, we allow F to be a right-continuous function and replace $f(y)$ with $F\{y\}$, the jump size of $F(\cdot)$ at y. The same technique has also been used in the derivation of the empirical cumulative distribution function and in the estimation of the cumulative hazard function in the Cox proportional hazards model. We show that the NPMLE of F has jumps only at the observed data points $Y_1,\ldots,Y_n$. By contrast, if F does not jump at any one of the Y_i's $(i=1,\ldots,n)$, then $F\{Y_i\}=0$. This leads to a zero value for the likelihood function. On the other hand, if F has an additional mass outside Y_i's, say y^*, then the likelihood can be written as

$$\begin{aligned} L_n^*(\beta,F) &= \prod_{i=1}^{n} \frac{F\{Y_i\}\exp(Y_i\beta^T\mathbf{X}_i)}{F\{y^*\}\exp(y^*\beta^T\mathbf{X}_i)+\sum_{k=1}^{n}F\{Y_k\}\exp(Y_k\beta^T\mathbf{X}_i)} \\ &\leq \prod_{i=1}^{n} \frac{F\{Y_i\}\exp(Y_i\beta^T\mathbf{X}_i)}{\sum_{k=1}^{n}F\{Y_k\}\exp(Y_k\beta^T\mathbf{X}_i)} \\ &= \prod_{i=1}^{n} \frac{\{F\{Y_i\}/c\}\exp(Y_i\beta^T\mathbf{X}_i)}{\sum_{k=1}^{n}\{F\{Y_k\}/c\}\exp(Y_k\beta^T\mathbf{X}_i)}, \end{aligned}$$

where

$$c=\sum_{i=1}^{n}F\{Y_i\}.$$

Therefore we have a larger likelihood by using a new distribution function with jumps $F\{Y_i\}/c$ at $Y_i, i=1,\ldots,n$. This shows that the NPMLE of F has positive jumps at and only at the observed data points. Therefore the nonparametric likelihood function, still denoted by $L_n(\beta,F)$ for simplicity, takes the form

$$L_n(\beta,F)=\prod_{i=1}^{n}\frac{F\{Y_i\}\exp(Y_i\beta^T\mathbf{X}_i)}{\sum_{k=1}^{n}F\{Y_k\}\exp(Y_k\beta^T\mathbf{X}_i)}. \quad (3)$$

We maximize $l_n(\beta,F)\equiv\log L_n(\beta,F)$ subject to the constraint $\sum_{k=1}^{n}F\{Y_k\}=1$ to obtain the NPMLEs of (β,F), denoted by $(\widehat{\beta}_n,\widehat{F}_n)$, where $\widehat{F}_n(y)=\sum_{k=1}^{n}I(Y_k\leq y)\widehat{F}_n\{Y_k\}$ and $I(\cdot)$ is an indicator function. It is easy to see that $(\widehat{\beta}_n,\widehat{F}_n)$ exist since the nonparametric likelihood in (3) is bounded from above by one.

To maximize $l_n(\beta, F)$ subject to the constraint $\sum_{k=1}^n F\{Y_k\} = 1$, we consider the Lagrangian

$$H_n(\beta, F, \lambda) = l_n(\beta, F) - \lambda \left(\sum_{k=1}^n F\{Y_k\} - 1 \right),$$

where λ is the Lagrange multiplier. We take the derivative of H_n with respect to $F\{Y_i\}$ and set it equal to 0,

$$\frac{\partial H_n(\beta, F, \lambda)}{\partial F\{Y_i\}} = \frac{1}{F\{Y_i\}} - \sum_{j=1}^n \frac{\exp(Y_i \beta^T \mathbf{X}_j)}{\sum_{k=1}^n F\{Y_k\} \exp(Y_k \beta^T \mathbf{X}_j)} - \lambda = 0.$$

Multiplying both sides by $F\{Y_i\}$, summing over i, and taking the constraint into account, we obtain

$$\sum_{i=1}^n \sum_{j=1}^n \frac{F\{Y_i\} \exp(Y_i \beta^T \mathbf{X}_j)}{\sum_{k=1}^n F\{Y_k\} \exp(Y_k \beta^T \mathbf{X}_j)} = n - \lambda.$$

By exchanging the summation indices i and j on the left hand side, it is easy to verify that $\lambda = 0$. Therefore, the NPMLE $\widehat{F}_n\{Y_i\}$ satisfies

$$\widehat{F}_n\{Y_i\} = \left\{ \sum_{j=1}^n \frac{\exp(Y_i \widehat{\beta}_n^T \mathbf{X}_j)}{\sum_{k=1}^n \widehat{F}_n\{Y_k\} \exp(Y_k \widehat{\beta}_n^T \mathbf{X}_j)} \right\}^{-1}. \tag{4}$$

Based on this result, we use an iterative algorithm to compute the NPMLEs.

Step 1. Start with initial estimates $\beta^{(0)}$ and $F^{(0)}$.

Step 2. Insert $\beta^{(0)}$ and $F^{(0)}$ into the right-hand side of equation (4) to obtain $F^{(1)}$.

Step 3. Insert $F^{(1)}$ into $l_n(\beta, F)$ and maximize the parametric likelihood at fixed $F^{(1)}$ by solving the score equation

$$\frac{\partial l_n(\beta, F^{(1)})}{\partial \beta} = \mathbf{0}.$$

Step 4. Repeat steps 2 and 3 until convergence.

Alternatively, we use the following reparameterization to reduce the constrained optimization problem to an unconstrained optimization problem

$$F\{Y_i\} = \frac{\exp(\zeta_i)}{\sum_{j=1}^{n} \exp(\zeta_j)},$$

where $\zeta_n = 0$. Note that this reparameterization has a nice connection with multinomial logit model and ζ_i can be treated as the intercept for the ith category. We then use the quasi-Newton algorithm described in Chapter 10 of Press, Teukolsky, Vetterling, and Flannery (1992) to maximize the nonparametric likelihood over $(\beta, \zeta_1, ..., \zeta_{n-1})$ simultaneously. This quasi-Newton algorithm requires only the input of the nonparametric log-likelihood function and the first partial derivatives of the nonparametric log-likelihood with respect to $(\beta, \zeta_1, ..., \zeta_{n-1})$. The quasi-Newton algorithm has been implemented in standard software such as SAS and R and has been successfully applied in the statistical literature. When there is only one binary covariate, it can be shown that the NPMLE of β in the density ratio model (1) is exactly the same as the MLE of the log-odds ratio in the logistic regression model fitting the probability of X given Y. In this case, we observed that the NPMLE of β obtained from the quasi-Newton algorithm is the same as the MLE of the log-odds ratio in the logistic regression model obtained from standard statistical software such as SAS and R, which provides an empirical validation of the quasi-Newton algorithm. The quasi-Newton algorithm is computationally efficient and converges very quickly. It takes about 0.1 second to analyze a data set with 100 observations in the simulation studies using the quasi-Newton algorithm on a Dell PowerEdge 2900 server. In our experience, the iterative algorithm also works well and in general yields the same parameter estimates as the quasi-Newton algorithm. However, the iterative algorithm may converge slowly or fail to converge especially when the absolute values of the true regression parameters are large. Moreover, the quasi-Newton algorithm is computationally more efficient than the iterative algorithm. Therefore, throughout the simulation studies and real data analysis in this paper, we use the quasi-Newton algorithm to compute the NPMLEs. An efficient computer program implemented in C language is available upon request.

We next establish the asymptotic results of $(\widehat{\beta}_n, \widehat{F_n})$. First we impose the following assumptions:

(C1) The covariates $\mathbf{X}$ are bounded with probability one, and if there exists a constant vector $\mathbf{c}$ such that $\mathbf{c}^T\mathbf{X} = \mathbf{0}$ with probability one, then $\mathbf{c} = \mathbf{0}$. Furthermore, the support of $\mathbf{X}$ contains $\mathbf{0}$ and the covariance matrix of $\mathbf{X}$ is positive definite.

(C2) The true parameters (β_0, F_0) belong to the following space

$$\begin{aligned}\mathscr{F} = \{(\beta, F) : &\beta \in \mathscr{B}_0 \text{ and } F \text{ is a distribution function with density} \\ &f \text{ on } \mathscr{Y} \subset R\},\end{aligned}$$

where $\mathscr{B}_0$ is a known compact set in R^q. Furthermore, there exist positive constants g_1 and g_2 such that with probability one

$$g_1 \le \int_{\mathscr{Y}} \exp(z\beta_0^T \mathbf{X}) dF_0(z) \le g_2, \quad \left| \int_{\mathscr{Y}} z \exp(z\beta_0^T \mathbf{X}) dF_0(z) \right| \le g_2$$
$$\text{and } g_1 \le \int_{\mathscr{Y}} z^2 \exp(z\beta_0^T \mathbf{X}) dF_0(z) \le g_2.$$

Remark: The first part of assumption (C1) is the standard assumption for regression models. The last condition in (C1) ensures that **X** does not include a constant term. In fact, if **X** contains a constant term, the corresponding regression coefficient can be absorbed into the unspecified baseline density function. The first part of assumption (C2) is common in semiparametric models and the second part ensures that the conditional density $f(y|\mathbf{X})$ exists as well as the first and second derivatives of conditional density with respect to β.

Under assumptions (C1) and (C2), we show that parameters (β, F) are identifiable. Suppose that two sets of parameters, (β, F) and $(\widetilde{\beta}, \widetilde{F})$ in $\mathscr{F}$ give the same likelihood function for the observed data, that is,

$$\frac{F'(Y)\exp(Y\beta^T\mathbf{X})}{\int_{\mathscr{Y}} \exp(z\beta^T\mathbf{X})dF(z)} = \frac{\widetilde{F}'(Y)\exp(Y\widetilde{\beta}^T\mathbf{X})}{\int_{\mathscr{Y}} \exp(z\widetilde{\beta}^T\mathbf{X})d\widetilde{F}(z)}, \tag{5}$$

where $F'(\cdot)$ and $\widetilde{F}'(\cdot)$ are the first derivatives of F and $\widetilde{F}$, respectively. Because the equality (5) holds for any $\mathbf{X}$, by letting $\mathbf{X} = \mathbf{0}$, immediately we obtain $F(y) = \widetilde{F}(y)$ for any y. It follows that

$$\frac{\exp(Y\beta^T\mathbf{X})}{\int_{\mathscr{Y}} \exp(z\beta^T\mathbf{X})dF(z)} = \frac{\exp(Y\widetilde{\beta}^T\mathbf{X})}{\int_{\mathscr{Y}} \exp(z\widetilde{\beta}^T\mathbf{X})dF(z)}.$$

This equality holds for any y in $\mathscr{Y}$. Choose two arbitrary unequal constants y_1 and y_2 in $\mathscr{Y}$. With some simple algebra, we obtain

$$(y_1 - y_2)\beta^T\mathbf{X} = (y_1 - y_2)\widetilde{\beta}^T\mathbf{X}.$$

Therefore, condition $(C1)$ gives $\beta = \widetilde{\beta}$. The identifiability of the parameters (β, F) is established.

With the identifiability results, we can establish the following consistency result.

Theorem 1. Under conditions (C1) and (C2), $||\widehat{\beta}_n - \beta_0|| + \sup_{\mathscr{Y}} |\widehat{F}_n - F_0| \to 0$, almost surely, where $||\cdot||$ is the Euclidean norm.

The proofs of this theorem and Theorem 2 below are given in the Appendix.

The asymptotic normality of the NPMLEs $(\widehat{\beta}_n, \widehat{F}_n)$ is stated as follows.

Theorem 2. Under conditions (C1) and (C2), the random element $\sqrt{n}(\widehat{\beta}_n - \beta_0, \widehat{F}_n - F_0)$ converges weakly to a tight, zero-mean Gaussian process in the metric space $l^\infty(\mathscr{H})$, where

$$\mathscr{H} = \{(\mathbf{h}_1, h_2) : \mathbf{h}_1 \in R^q, h_2 \text{ is a function on } \mathscr{Y}; ||\mathbf{h}_1|| \le 1, |h_2|_V \le 1\},$$

and $|h_2|_V$ denotes the total variation of h_2 in $\mathscr{Y}$. Furthermore, $\widehat{\beta}_n$ is asymptotically efficient.

Denote the random element in the limiting distribution by $\psi \in l^\infty(\mathscr{H})$. Theorem 2 implies that for any $(\mathbf{h}_1, h_2) \in \mathscr{H}$, $\sqrt{n}(\widehat{\beta}_n - \beta_0)^T \mathbf{h}_1 + \sqrt{n} \int_{\mathscr{Y}} h_2(t) d(\widehat{F}_n - F_0)$ is asymptotically normal with mean zero and variance $\text{Var}(\psi[\mathbf{h}_1, h_2])$, and this normal approximation is uniform in $(\mathbf{h}_1, h_2)$. To estimate the variance of $(\widehat{\beta}_n, \widehat{F}_n)$, we view (3) as a parametric likelihood with $F\{Y_i\} (i = 1, ..., n-1)$ and β the parameters. Note that $F\{Y_n\} = 1 - \sum_{i=1}^{n-1} F\{Y_i\}$. Then we can estimate the asymptotic covariance matrix for these parameters by the inverse of the observed information matrix according to the parametric likelihood theory. Alternatively, one can consider to estimate the asymptotic covariance matrix of $\widehat{\beta}_n$ with the profile log-likelihood function (Murphy and van der Vaart, 2000). This approach requires a numerical approximation of the second derivatives of the profile log-likelihood function at $\widehat{\beta}_n$, since the profile log-likelihood function does not have an analytic form. Another limitation is that this profile likelihood approach does not provide a variance estimation for $\widehat{F}_n$, which may be of interest in practice.

We can construct the Wald test statistic for testing $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$ and show that it is asymptotically chi-square distributed with the degrees of freedom being the dimension of $\mathbf{X}$. Alternatively, we can construct the likelihood ratio test statistic

$$LRT = -2[\log l_n(\beta_0, \widetilde{F}_n) - \log l_n(\widehat{\beta}_n, \widehat{F}_n)],$$

where $\widetilde{F}_n$ is the NPMLE of F under the null hypothesis. By using the results of Murphy and van der Vaart (1997, 2000), we can show that the likelihood ratio test statistic is also asymptotically chi-squared distributed with the degrees of freedom being the dimension of $\mathbf{X}$.

3 Numerical results

We conducted simulation studies to evaluate the performances of the proposed nonparametric likelihood approaches. For comparison, we also computed the maximum pair-wise conditional likelihood estimators (MPCLEs) of Liang and Qin (2000) and the maximum triple-wise conditional likelihood estimators (MTCLEs) by maximizing

$$\begin{aligned} L_{PC}(\beta) &\propto \prod_{i<k} \frac{f(Y_i|\mathbf{X}_i)f(Y_k|\mathbf{X}_k)}{f(Y_i|\mathbf{X}_i)f(Y_k|\mathbf{X}_k)+f(Y_i|\mathbf{X}_k)f(Y_k|\mathbf{X}_i)} \\ &= \prod_{i<k} \frac{1}{1+\exp\{-(Y_i-Y_k)\beta^T(\mathbf{X}_i-\mathbf{X}_k)\}}. \end{aligned}$$

and

$$\begin{aligned} L_{TC}(\beta) &\propto \prod_{i<j<k} \frac{f(Y_i|\mathbf{X}_i)f(Y_j|\mathbf{X}_j)f(Y_k|\mathbf{X}_k)}{\sum_{(i_1,j_1,k_1)\in\mathscr{C}} f(Y_i|\mathbf{X}_{i_1})f(Y_j|\mathbf{X}_{j_1})f(Y_k|\mathbf{X}_{k_1})} \\ &= \prod_{i<j<k} \frac{\exp(Y_i\beta^T\mathbf{X}_i+Y_j\beta^T\mathbf{X}_j+Y_k\beta^T\mathbf{X}_k)}{\sum_{(i_1,j_1,k_1)\in\mathscr{C}}\exp(Y_i\beta^T\mathbf{X}_{i_1}+Y_j\beta^T\mathbf{X}_{j_1}+Y_k\beta^T\mathbf{X}_{k_1})}, \end{aligned}$$

respectively, where $\mathscr{C}$ is a set of all possible permutations of (i,j,k). The MTCLEs are expected to be more efficient than the MPCLEs. However, the triple-wise conditional likelihood approach is computationally more intensive than the pair-wise conditional likelihood approach, because one needs to calculate likelihood contributions from each triplet, leading to the calculation of $\binom{n}{3}$ terms. Throughout the simulation studies and real data analysis, we use the quasi-Newton algorithm for maximizing $L_{PC}(\beta)$ and $L_{TC}(\beta)$. We have double-checked the first derivatives of $L_{PC}(\beta)$ and $L_{TC}(\beta)$ with respect to β at the parameter estimates and have found that they are all very close to 0.

In the first set of simulations, we generate Y from a normal distribution with mean $\beta_1X_1+\beta_2X_2$ and variance 1, where $X_1 \sim Ber(0.5)$ and $X_2 \sim N(0,1)$. In the second set of simulations, we generate Y from an exponential distribution with mean $1/(1-\beta_1X_1-\beta_2X_2)$, where $X_1 \sim Ber(0.5)$ and X_2 is a uniform random variable in $[0,1]$. In all simulations, we fix $\beta_1=-0.5$ and $\beta_2=0.5$. We considered sample sizes of 100 and 200. The results based on 1,000 replicates are summarized in Tables 1 and 2 for normal baseline density and exponential baseline density, respectively.

Under all simulation settings, the NPMLEs of (β,F) appear to be unbiased; the standard error estimate reflects accurately the true variation. As the sample size increases, the biases and standard deviations of the NPMLEs decrease. We considered two approaches to construct the 95% confidence intervals: one is based on the normal approximation and the other one is based on the chi-square approximation of the likelihood ratio statistics. Both confidence intervals have proper coverage probabilities close to the nominal level. As expected, the MPCLEs and MTCLEs have small biases, but they are less efficient than the NPMLEs. Compared to the NPMLEs, the efficiency loss of the MPCLEs ranged from 3% to 13% whereas the efficiency loss of the MTCLEs ranged from 1% to 8%.

We also compared the computation time of each method. The average computation time for analyzing one data set with 100 subjects was 0.125 second, 0.13 second, and 4.05 seconds for the nonparametric likelihood approach, the pair-wise and triple-wise conditional likelihood approaches, respectively. The corresponding computation times increased to 0.52 second, 0.55 second, and 49.0 seconds, respectively, when the sample size increased to 200. The nonparametric likelihood approach is as computationally efficient as the pair-wise likelihood approach while estimating the regression parameters and the baseline cumulative distribution function simultaneously. The triple-wise conditional likelihood approach is computationally intensive and the computational burden increases rapidly as the sample size increases.

Figure 1 displays the average of the baseline CDF estimates based on 1,000 replicates with a sample size of 100 and the true CDF curves. There is very little bias of the NPMLE of the baseline CDF under both simulation settings.

In addition, we conducted simulation studies to assess the size and power of the likelihood ratio test described in Section 2. As expected, the likelihood ratio test accurately controls the type I error rate at the nominal significance level under the null hypothesis and yields reasonable power under the alternative hypothesis.

4 An application

As an illustration, we used a well-known acute myelogenous leukemia survival data set previously analyzed by Feigl and Zelen (1965). This data set consists of two groups of patients who died of acute myelogenous leukemia. Among a total of 33 patients, 17 patients termed AG positive were identified by the presence of Auer rods and/or significant granulature of the leukemic cells in the bone marrow at diagnosis. For the remaining 16 AG-negative patients, these factors were absent. The response variable was survival time (in weeks) from the date of diagnosis. The survival times ranged from 1 to 156 weeks (median, 56 weeks) in the AG-positive group and 2 to 65 weeks (median, 7.5 weeks) in the AG-negative group. It is known that in acute myelogenous leukemia, the white blood count (WBC) relates directly to the disease severity; the higher the WBC, the more severe the disease. The log WBCs at the time of diagnosis ranged from 2.875 to 5 (median, 4.021). There were no censored data.

We first fitted the density ratio model to the leukemia survival data set. Two covariates were included in the model: the log WBC and the AG status which takes value 1 for the AG-positive group and value 2 for the AG-negative group. Table 3 summarizes the estimation results based on the pair-wise and triple-wise conditional likelihoods as well as the nonparametric likelihood. The standard error and 95% confidence interval estimates with the conditional likelihood approach were obtained from 500 bootstrap samples. Both WBC and AG status appear to have significant effects on the survival time, with p-values 0.012 and 0.013, respectively, based on nonparametric likelihood approach. Three approaches yielded similar parameter estimates; however, the nonparametric likelihood approach was more efficient than the conditional likelihood approaches, with smaller standard error estimates and narrower 95% confidence interval estimates.

To illustrate the goodness-of-fit of the density ratio model, we present the empirical and model-fitted survival curves for each AG group in Figure 2. The model-fitted survival function is calculated as the empirical average of the predicted survival functions within each AG group. The model-fitted survival curves agree very well with the empirical survival curves indicating a good fit of the density ratio model with the leukemia data. For comparison, we also fitted the exponential regression model described in Feigl and Zelen (1965) and the Cox proportional hazards model. Figure 3 presents the empirical and corresponding model-fitted survival curves. The parametric exponential regression model fits the data poorly. Among the three models, the density ratio model was found to fit the leukemia data the best.

5 Discussion

In this paper we have studied a semiparametric density ratio model in which the baseline carrier density is not specified. This model has a natural connection with generalized linear models and is closely related to biased sampling problems. We expect that this model will have wide applicability in a broad area of statistical problems.

When there are tied values in the outcome observations, we can rewrite the nonparametric likelihood as

$$L_n(\beta, F) = \prod_{i=1}^{n} \frac{F\{Y_i\}\exp(Y_i \beta^T \mathbf{X}_i)}{\sum_{k=1}^{m} F\{Y_{(k)}\}\exp(Y_{(k)}\beta^T \mathbf{X}_i)},$$

where $Y_{(1)}, \ldots, Y_{(m)}$ are the distinct observed outcome data points in an ascending order and m is the total number of the distinct outcome data points. The same estimation and inference procedures can be applied.

We have described two algorithms to obtain the NPMLEs of the unknown parameters in the density ratio model. One is an intuitive iterative algorithm based on the result in equation (5) and the other one is the quasi-Newton algorithm. For fixed β, Vardi (1985) and Gilbert et al. (1999) showed that the algorithm is convergent for the baseline distribution function estimation. However, they were not able to prove that the log-likelihood function is concave over all parameters. It is not clear whether the concave property of the log-likelihood function actually holds. Instead, we can prove the local convergence. We use $\widehat{\beta}_{PC}$, the maximum pair-wise conditional likelihood estimator of β as an initial estimate of β and

$$\frac{\sum_{i=1}^{n} I(Y_i \leq y) \exp(-Y_i \widehat{\beta}_{PC}^T \mathbf{X}_i)}{\sum_{i=1}^{n} \exp(-Y_i \widehat{\beta}_{PC}^T \mathbf{X}_i)}$$

as an initial estimate of $F(y)$. We can then achieve local convergence. Our empirical experience suggests that the quasi-Newton algorithm is efficient and reliable. First, when there is only one binary covariate the NPMLE of the regression parameter from the quasi-Newton algorithm is exactly the same as the MLE of the log-odds ratio in the logistic regression model obtained from standard statistical packages such as SAS and R. Second, although it is less intuitive than the iterative algorithm, the quasi-Newton algorithm obtains the same parameter estimates as the iterative algorithm when the latter works. Furthermore, the quasi-Newton algorithm is computationally more efficient than the iterative algorithm and always converges in the analysis of thousands of generated data sets in the simulation studies. Finally, we calculated the first derivatives of the log-likelihood function with respect to the unknown parameters at the NPMLEs obtained from the quasi-Newton algorithm in the simulation studies and the real data analysis. We found that they were all very close to 0 suggesting that the quasi-Newton algorithm indeed solves the score equations for estimating the unknown parameters.

In the simulation studies, we observed a small gain in efficiency of the NPMLE over the MPCLE and MTCLE for estimating β_1. This phenomenon may occur because the covariate X_1 is a Bernoulli random variable rather than a continuous random variable, such as X_2. Similar findings were reported in Farewell (1979), which compared the conditional and unconditional likelihood approaches for the estimation of logistic models based on retrospective data. Nevertheless, the proposed NPMLEs are still more efficient than the MPCLEs and MTCLEs. Furthermore, the nonparametric likelihood approach estimates the regression parameters and the baseline density simultaneously. Therefore, unlike conditional likelihood approaches, the nonparametric likelihood approach can be used to estimate the conditional means and the conditional cumulative distribution functions given covariate values.

We have illustrated the proposed methods by applying them to a leukemia survival data set. As an alternative approach, the density ratio model appears to fit the data better than the commonly used Cox proportional hazards model. However, the proposed methods are not tailored for censored data. It would be interesting to explore the extension of the proposed approach to censored data. One challenging issue is how to derive the denominator term in model (1) when the last observation is right censored. Future research in this direction is warranted.

In the real data analysis, we used an ad-hoc graphical approach to assess the fit of the density ratio model to the leukemia data. It would be interesting to develop a formal lack-of-fit test to assess the adequacy of the density ratio model (1). Bondell (2007) constructed a goodness-of-fit test statistic for two-sample data via a discrepancy between two competing kernel density estimators of the underlying conditional distributions. We are currently investigating the extension of this approach to the goodness-of-fit test for the semiparametric density ratio model with general forms of covariates.

Similar to the stratified Cox proportional hazards model, we can also use a stratified version of the density ratio model

$$f_k(y|\mathbf{X}) = \frac{f_k(y)\exp(y\beta^T\mathbf{X})}{\int_{\mathscr{Y}} f_k(z)\exp(z\beta^T\mathbf{X})dz}, \quad k = 1, 2, \ldots, K$$

where $f_k(y)$ is the baseline density for the kth stratum. If the sample size from each stratum is relatively small, the nonparametric likelihood method may not work as well as the pair-wise or triple-wise conditional likelihood method since there is not enough information for estimating the baseline densities $f_k(y)$'s. This phenomenon was also observed in standard stratified logistic regression models (Farewell, 1979, Lubin, 1981).

The proposed method can also be thought as a semiparametric generalization of the analysis of covariance model (ANCOVA), where

$$f(y|X,Z) = \frac{f(y)\exp(y\beta X + y\gamma Z + y\xi XZ)}{\int_{\mathscr{Y}} f(t)\exp(t\beta X + t\gamma Z + t\xi XZ)dt}.$$

In this example, X is the continuous covariate and Z is the discrete covariate. The term $y\xi XZ$ is the interaction between X and Z. If the underlying carrier density

$f(y)$ is a normal density, then this is the standard ANCOVA model. However, if the form of $f(y)$ is unknown, then we have a semiparametric model. One may be interested in testing $\xi = 0$, that is, there is no interaction between X and Z. Based on the density ratio model we also have tested the interaction effect between WBC and AG status on the survival time in the leukemia data set , and the resulting p-value was 0.405, implying no strong evidence of an interaction effect. When both X and Z are discrete variables, Fokianos, Kedem, Qin, and Short (2001) discussed the generalization of the traditional ANOVA model to a semiparametric ANOVA model, which is a special case of the density ratio model in (1).

In the Introduction section, we showed that the density ratio model can be considered a semiparametric version of the generalized linear models with g being the identity function in model (2). When g is not the identity function, there is no direct connection between the two models. One can modify the density ratio such that

$$f(y|\mathbf{X}) = \frac{f(y)\exp(yg(\beta^T\mathbf{X})}{\int_{\mathscr{Y}} f(z)\exp(zg(\beta^T\mathbf{X}))dz}.$$

The investigation of such a model is beyond the scope of this paper but certainly warrants future research.

Appendix: Proof of Asymptotic Results

We introduce some notations that will be used throughout the proof. Let $\mathbf{O}_i$ denote the observations for the ith subject consisting of $(Y_i, \mathbf{X}_i)$. Let $\mathbf{P}_n$ and $\mathbf{P}$ be the empirical measure and the expectation of n i.i.d. observations $\mathbf{O}_1, \ldots, \mathbf{O}_n$. That is, for any measurable function $g(\mathbf{O})$,

$$\mathbf{P}_n[g(\mathbf{O})] = \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{O}_i), \quad \mathbf{P}[g(\mathbf{O})] = E[g(\mathbf{O})].$$

A. Proof of Theorem 1

Since $\widehat{F}_n$ is bounded in $\mathscr{Y}$, by Helly's selection theorem, a subsequence of $\widehat{F}_n$, still indexed by $\{n\}$, can be found to converge point-wise to a distribution function F^* in $\mathscr{Y}$ and the same subsequence of $\widehat{\beta}_n$ converges to some β^*. We shall prove that $\beta^* = \beta_0$ and $F^* = F_0$.

Recall from (4) that $\widehat{F}_n\{Y_i\}$ satisfies

$$\widehat{F}_n\{Y_i\} = \frac{1}{n\mathbf{P}_n[Q(y, \mathbf{O}; \widehat{\beta}_n, \widehat{F}_n)]}\Bigg|_{y=Y_i}$$

where

$$Q(y,\mathbf{O};\beta,F)=\frac{\exp(y\beta^T\mathbf{X})}{\int_{\mathscr{Y}}\exp(z\beta^T\mathbf{X})dF(z)}.$$

In view of (4), we construct another step function $\widetilde{F}_n$ with jumps only at the observed Y_i and the jump sizes satisfy

$$\widetilde{F}_n\{Y_i\}=\left.\frac{1}{n\mathbf{P}_n[Q(y,\mathbf{O};\beta_0,F_0)]}\right|_{y=Y_i}.$$

We verify that $\widetilde{F}_n$ converges to F_0 in $\mathscr{Y}$ with probability one. Since both $\{\beta^T\mathbf{X}:\beta\in\mathscr{B}_0\}$ and $\{F(Y):F$ is a distribution function in $\mathscr{Y}\}$ are P-Donsker classes, the preservation of the Donsker property based on Theorem 2.10.6 of van der Vaart and Wellner (1996) implies that the class

$$\mathscr{F}_1=\{Q(y,\mathbf{O};\beta,F):y\in\mathscr{Y},\beta\in\mathscr{B}_0,F\text{ is a distribution function in }\mathscr{Y}\}$$

is a bounded P-Donsker class. Since a P-Donsker class is also a Glivenko-Cantelli class, by the Glivenko-Cantelli theorem in var der Vaart and Wellner (1996), $\widetilde{F}_n(t)$ uniformly converges to $E(I(Y\le t)/\mu(Y)]$, where $\mu(y)=E[Q(y,\mathbf{O};\beta_0,F_0)]$. It can be shown that

$$\begin{aligned}E\left[\frac{I(Y\le t)}{\mu(Y)}\right]&=E\left[\int_{\mathscr{Y}}\frac{I(y\le t)\exp(y\beta_0^T\mathbf{X})}{\mu(y)\int_{\mathscr{Y}}\exp(z\beta_0^T\mathbf{X})dF_0(z)}dF_0(y)\right]\\&=\int_{\mathscr{Y}}I(y\le t)dF_0(y)=F_0(t).\end{aligned}$$

Consequently, we conclude that $\widetilde{F}_n$ uniformly converges to F_0 in $\mathscr{Y}$ with probability one.

Since $(\widehat{\beta}_n,\widehat{F}_n)$ maximizes the log-likelihood function $l_n(\beta,F)$, we have

$$n^{-1}l_n(\widehat{\beta}_n,\widehat{F}_n)-n^{-1}l_n(\beta_0,\widetilde{F}_n)\ge 0.$$

By taking limits on both sides, we can show that

$$-K((\beta^*,F^*),(\beta_0,F_0))\ge 0$$

where $K(\cdot,\cdot)$ denotes the Kullback-Leibler information of (β^*, F^*) with respect to the true parameters. From the identifiability result proved in Section 2, we immediately obtain $\beta^* = \beta_0$ and $F^* = F_0$. This establishes consistency of $(\widehat{\beta}_n, \widehat{F}_n)$.

B. Proof of Theorem 2

We prove Theorem 2 by verifying the four conditions in Theorem 3.3.1 of van der Vaart and Wellner (1996). We first define a neighborhood of the true parameters (β_0, F_0), denoted by

$$\mathscr{U} = \{(\beta, F) : ||\beta - \beta_0|| + \sup_{y \in \mathscr{Y}} |F(y) - F_0(y)| < \varepsilon_0\}$$

for a small constant ε_0. It is obvious that $\mathscr{U}$ is a Donsker class. Based on the consistency results, $(\widehat{\beta}_n, \widehat{F}_n)$ belongs to $\mathscr{U}$ with probability close to one when sample size n is large enough.

For any one-dimensional submodel given as $\{\beta + \varepsilon \mathbf{h}_1, F + \varepsilon \int_{\mathscr{Y}} Q_F[h_2] dF\}$, $(\beta, F) \in \mathscr{U}, \mathbf{H} \equiv (\mathbf{h}_1, h_2) \in \mathscr{H}$, and $Q_F[h_2](y) = h_2(y) - \int_{\mathscr{Y}} h_2 dF$, the score function for a single observation $\mathbf{O}$ takes the form

$$\begin{aligned} W(\mathbf{O}; \beta, F)[\mathbf{H}] = & \left\{ Y - \frac{\int_{\mathscr{Y}} z \exp(z\beta^T \mathbf{X}) dF(z)}{\int_{\mathscr{Y}} \exp(z\beta^T \mathbf{X}) dF(z)} \right\} \mathbf{h}_1^T \mathbf{X} \\ & + Q_F[h_2](Y) - \frac{\int_{\mathscr{Y}} Q_F[h_2](z) \exp(z\beta^T \mathbf{X}) dF(z)}{\int_{\mathscr{Y}} \exp(z\beta^T \mathbf{X}) dF(z)}. \end{aligned} \tag{6}$$

We define

$$U_n(\beta, F)[\mathbf{H}] = \mathbf{P}_n\{W(\mathbf{O}; \beta, F)[\mathbf{H}]\}$$

and

$$U(\beta, F)[\mathbf{H}] = \mathbf{P}\{W(\mathbf{O}; \beta, F)[\mathbf{H}]\}.$$

Thus, it follows that both $U_n(\beta, F)[\mathbf{H}]$ and $U(\beta, F)[\mathbf{H}]$ are maps from $\mathscr{U}$ to $l^\infty(\mathscr{H})$, $\sqrt{n}\{U_n(\beta, F) - U(\beta, F)\}$ is an empirical process in the space $l^\infty(\mathscr{H})$, $U_n(\widehat{\beta}_n, \widehat{F}_n) = 0$, and $U(\beta_0, F_0) = 0$.

We shall now prove the first property stated in Theorem 3.3.1 of van der Vaart and Wellner (1996)

$$\begin{aligned}&\sqrt{n}(U_n-U)(\widehat{\beta}_n,\widehat{F_n})-\sqrt{n}(U_n-U)(\beta_0,F_0)\\&=o_P(1+\sqrt{n}||\widehat{\beta}_n-\beta_0||+\sqrt{n}\sup_{\mathscr{Y}}|\widehat{F_n}-F_0|).\end{aligned}\tag{7}$$

Based on the explicit expression in (6), $W(\mathbf{O};\beta,F)[\mathbf{H}]$ is continuously differentiable with respect to β, and assumption (C2) implies that there exists a positive constant g_3, such that

$$\left\|\frac{dW(\mathbf{O};\beta,F)}{d\beta}\right\|\leq g_3,$$

with probability one. Furthermore,

$$\begin{aligned}|W(\mathbf{O};\beta,F_1)-W(\mathbf{O};\beta,F_2)|\leq g_4\Bigg\{&\left|\int_{\mathscr{Y}}z\exp(z\beta^T\mathbf{X})d(F_1(z)-F_2(z))\right|\\&+\left|\int_{\mathscr{Y}}h_2(z)d(F_1(z)-F_2(z))\right|\\&+\left|\int_{\mathscr{Y}}h_2(z)\exp(z\beta^T\mathbf{X})d(F_1(z)-F_2(z))\right|\\&+\left|\int_{\mathscr{Y}}\exp(z\beta^T\mathbf{X})d(F_1(z)-F_2(z))\right|\end{aligned}$$

for some positive constant g_4. Therefore,

$$\sup_{\mathbf{H}\in\mathscr{H}}E\left[\left\{W(\mathbf{O};\beta,F)[\mathbf{H}]-W(\mathbf{O};\beta_0,F_0)[\mathbf{H}]\right\}^2\right]$$

converges to zero if $||\beta-\beta_0||+\sup_{\mathscr{Y}}|F-F_0|\to 0$. In addition, by the same arguments for $\mathscr{F}_1$, the class

$$\mathscr{F}_2=\{W(\mathbf{O};\beta,F)[\mathbf{H}]-W(\mathbf{O};\beta_0,F_0)[\mathbf{H}]:(\beta,F)\in\mathscr{U},\mathbf{H}\in\mathscr{H}\}$$

is P-Donsker. According to Lemma 3.3.5 of van der Vaart and Wellner (1996), property (7) holds.

Since the following class

$$\{W(\mathbf{O};\beta_0,F_0)[\mathbf{H}]:\mathbf{H}\in\mathscr{H}\}$$

is P-Donsker, by the Donsker theorem, $\sqrt{n}(U_n - U)(\beta_0, F_0)$ converges to a tight random element ξ, a zero mean Gaussian process indexed by $\mathbf{H} \in \mathscr{H}$ and the covariance between $\xi(\mathbf{H}_1)$ and $\xi(\mathbf{H}_2)$ is equal to

$$E\left[W(\mathbf{O};\beta_0, F_0)[\mathbf{H}_1] \times W(\mathbf{O};\beta_0, F_0)[\mathbf{H}_2]\right].$$

The second property of Theorem 3.3.1 of van der Vaart and Wellner (1996) is verified.

By using the smoothness of $U(\beta, F)$, we can immediately prove that $U(\beta, F)$ is Frechet-differentiable at (β_0, F_0). The derivative of $U(\beta, F)$ at (β_0, F_0), denoted by $U'(\beta_0, F_0)$ is a map from the space

$$\{(\beta - \beta_0, F - F_0) : (\beta, F) \in \mathscr{U}\}$$

to $l^\infty(\mathscr{H})$.

It remains to verify that U' is continuously invertible at (β_0, F_0). According to the argument in the Appendix of Zeng and Lin (2007), it suffices to prove that for any one-dimensional submodel given as $\{\beta_0 + \varepsilon\mathbf{h}_1, F_0 + \varepsilon \int Q_{F_0}[h_2]dF_0\}$, $\mathbf{H} \in \mathscr{H}$, the Fisher information along the submodel is non-singular. If the Fisher information along this submodel is singular, the score function along this submodel is zero with probability one. We will show that $W(\mathbf{O};\beta_0, F_0)[\mathbf{H}] = 0$ yields that $\mathbf{h}_1 = \mathbf{0}$ and $Q_{F_0}[h_2] = 0$. We follow the ideas of proving the identifiability of the model. Let $\mathbf{X} = \mathbf{0}$, we obtain $Q_{F_0}[h_2](Y) - \int_{\mathscr{Y}} Q_{F_0}[h_2](z)dF_0(z) = 0$. Therefore $Q_{F_0}[h_2](y) = 0$ for any $y \in \mathscr{Y}$. Let y_1 and y_2 be two different values in $\mathscr{Y}$. We have

$$\left\{ y_1 - \frac{\int_{\mathscr{Y}} z\exp(z\beta^T\mathbf{X})dF(z)}{\int_{\mathscr{Y}} \exp(z\beta^T\mathbf{X})dF(z)} \right\} \mathbf{h}_1^T\mathbf{X} = 0$$

and

$$\left\{ y_2 - \frac{\int_{\mathscr{Y}} z\exp(z\beta^T\mathbf{X})dF(z)}{\int_{\mathscr{Y}} \exp(z\beta^T\mathbf{X})dF(z)} \right\} \mathbf{h}_1^T\mathbf{X} = 0.$$

With simple algebra, we obtain $(y_1 - y_2)\mathbf{h}_1^T\mathbf{X} = 0$. It follows from assumption (C1) that $\mathbf{h}_1 = \mathbf{0}$. We have thus proved the nonsigularity of the Fisher information matrix along any nontrivial submodel.

We now have verified all four properties of Theorem 3.3.1 of van der Vaart and Wellner (1996). We conclude that $\sqrt{n}(\widehat{\beta}_n - \beta_0, \widehat{F}_n - F_0)$ weakly converges to a tight Gaussian random element $-U'^{-1}\xi$ in $l^\infty(\mathscr{H})$. Furthermore, it can be shown that $\widehat{\beta}_n$ is an asymptotic linear estimator for β_0 and that the corresponding influence functions are on the space spanned by the score functions. The semiparametric efficiency theory in Chapter 3 of Bickel et al. (1993) implies that $\widehat{\beta}_n$ is semiparametrically efficient.

References

Anderson, J. A. (1979): "Robust inference using logistic models," *Bulletin of the International Statistical Institute*, 2, 35–53.

Anderson, J. A. and P. R. Philips (1981): "Regression, discrimination and measurement models for ordered categorical variables," *Applied Statistics*, 30, 22–31.

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

Bondell, H. D. (2005): "Minimum distance estimation for the logistic regression model," *Biometrika*, 92, 724–731.

Bondell, H. D. (2007): "Testing goodness-of-fit in logistic case-control studies," *Biometrika*, 94, 487–495.

Bondell, H. D. (2008): "A characteristic function approach to the biased sampling model, with application to robust logistic regression," *Journal of Statistical Planning and Inference*, 138, 742–755.

Chen, K. (2001): "Parametric models for response-biased sampling," *Journal of the Royal Statistical Society: Series B*, 63, 775–789.

Cox, D. R. (1969): "Some sampling problems in technology," in N. L. Johnson and H. J. Smith, eds., *New Developments in Survey Sampling*, New York: Wiley-Interscience, 506–527.

Cox, D. R. (1972): "Regression model and life-tables (with Discussion)," *Journal of the Royal Statistical Society, Series B*, 34, 187–220.

Cox, D. R. (1975): "Partial likelihood," *Biometrika*, 62, 269–276.

Farewell, V. (1979): "Some results on the estimation of logistic models based on retrospective data," *Biometrika*, 66, 27–32.

Feigl, P. and M. Zelen (1965): "Estimation of exponential survival probabilities with concomitant information," *Biometrics*, 21, 826–838.

Fokianos, K., B. Kedem, J. Qin, and D. A. Short (2001): "A semiparametric approach to the one-way layout," *Technometrics*, 43, 56–65.

Gilbert, P., S. Lele, and Y. Vardi (1999): "Maximum likelihood estimation in semiparametric selection bias models with application to aids vaccine trials," *Biometrika*, 86, 27–43.

Gill, R. D., Y. Vardi, and J. A. Wellner (1988): “Large sample theory of empirical distributions in biased sampling models,” *Annals of Statistics*, 16, 1069–1112.

Liang, K. Y. and J. Qin (2000): “Regression analysis under non-standard situations: A pairwise pseudo-likelihood approach,” *Journal of the Royal Statistical Society: Series B*, 62, 773–786.

Lubin, J. H. (1981): “An empirical evaluation of the use of conditional and unconditional likelihoods for case-control data,” *Biometrika*, 68, 567–571.

McCullagh, P. (1980): “Regression models for ordinal data (with discussion),” *Journal of the Royal Statistical Society: Series B*, 42, 109–142.

Murphy, S. A. and A. W. van der Vaart (1997): “Semiparametric likelihood ratio inference,” *Annals of Statistics*, 25, 1471–1509.

Murphy, S. A. and A. W. van der Vaart (2000): “On the profile likelihood,” *Journal of the American Statistical Association*, 95, 449–465.

Nelder, J. A. and R. M. W. Wedderburn (1972): “Generalized linear models,” *Journal of the Royal Statistical Society: Series A*, 135, 370–384.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992): *Numerical Recipes in C: The Art of Scientific Computing, Second Edition*, Cambridge: Cambridge University Press.

Qin, J. and K. Y. Liang (1999): “Generalized odds ratio model and pairwise conditional likelihood,” *Technical Report*.

Qin, J. and B. Zhang (1997): “A goodness of fit test for logistic regression models based on case-control data,” *Biometrika*, 84, 609–618.

Rathouz, P. and L. Gao (2009): “Generalized linear models with unspecified reference distribution,” *Biostatistics*, 10, 205–218.

van der Vaart, A. and J. Wellner (1996): *Weak Convergence and Empirical Processes: With Applications to Statistics*, New York: Springer-Verlag.

Vardi, Y. (1982): “Nonparametric estimation in presence of length bias,” *Annals of Statistics*, 10, 616–620.

Vardi, Y. (1985): “Empirical distribution in selection bias models,” *Annals of Statistics*, 13, 178–203.

Zeng, D. and D. Y. Lin (2007): “Maximum likelihood estimation in semiparametric regression models with censored data (with discussion),” *Journal of the Royal Statistical Society: Series B*, 69, 507–564.

Table 1. Simulation results for normal baseline density

Setup			NPMLE						MPCLE		MTCLE	
Par	n	True	Bias	SE	SEE	CP	CP*	MSE	MSE	RE	MSE	RE
β_1	100	-0.500	-0.024	0.223	0.221	0.956	0.950	0.050	0.053	1.046	0.052	1.024
β_2		0.500	0.028	0.135	0.129	0.955	0.951	0.019	0.020	1.060	0.019	1.028
$F(-0.5)$		0.309	-0.001	0.060	0.061	0.942	-	0.004	-	-	-	-
$F(0.0)$		0.500	0.002	0.068	0.067	0.941	-	0.005	-	-	-	-
$F(0.5)$		0.691	0.002	0.063	0.061	0.936	-	0.004	-	-	-	-
β_1	200	-0.500	-0.013	0.152	0.153	0.952	0.948	0.023	0.024	1.028	0.024	1.012
β_2		0.500	0.015	0.089	0.089	0.960	0.955	0.008	0.009	1.093	0.009	1.058
$F(-0.5)$		0.309	-0.001	0.043	0.043	0.941	-	0.002	-	-	-	-
$F(0)$		0.500	0.001	0.048	0.048	0.949	-	0.002	-	-	-	-
$F(0.5)$		0.691	0.002	0.044	0.043	0.941	-	0.002	-	-	-	-

Par, parameter; Bias, difference between the average of parameter estimates and true parameter value; SE, empirical standard deviation of the parameter estimates; SEE, average of standard error estimates; CP: coverage probability of the 95% confidence interval based on the normal approximation of the NPMLE; CP*: coverage probability of the 95% confidence interval based on the chi-square approxiamtion of the likelihood ratio statistic; MPCLE, maximum pair-wise conditional likelihood estimator; MTCLE, maximum triple-wise conditional likelihood estimator; MSE: mean squared error; RE: relative efficiency of the NPMLE.

Table 2. Simulation results for exponential baseline density

Setup			NPMLE						MPCLE		MTCLE	
Par	n	True	Bias	SE	SEE	CP	CP*	MSE	MSE	RE	MSE	RE
β_1	100	-0.500	-0.050	0.248	0.235	0.943	0.938	0.064	0.066	1.031	0.065	1.011
β_2		0.500	0.041	0.351	0.354	0.971	0.961	0.125	0.140	1.122	0.133	1.066
$F(0.5)$		0.393	-0.004	0.075	0.077	0.935	-	0.006	-	-	-	-
$F(1.0)$		0.632	-0.004	0.085	0.086	0.932	-	0.007	-	-	-	-
$F(1.5)$		0.777	-0.004	0.076	0.077	0.940	-	0.006	-	-	-	-
β_1	200	-0.500	-0.022	0.162	0.160	0.953	0.954	0.027	0.027	1.028	0.027	1.014
β_2		0.500	0.025	0.242	0.239	0.958	0.948	0.059	0.067	1.131	0.064	1.078
$F(0.5)$		0.393	-0.001	0.053	0.054	0.953	-	0.003	-	-	-	-
$F(1.0)$		0.632	-0.002	0.061	0.06	0.935	-	0.004	-	-	-	-
$F(1.5)$		0.777	-0.002	0.055	0.054	0.926	-	0.003	-	-	-	-

Par, parameter; Bias, difference between the average of parameter estimates and true parameter value; SE, empirical standard deviation of the parameter estimates; SEE, average of standard error estimates; CP: coverage probability of the 95% confidence interval based on the normal approximation of the NPMLE; CP*: coverage probability of the 95% confidence interval based on the chi-square approxiamtion of the likelihood ratio statistic; MPCLE, maximum pair-wise conditional likelihood estimator; MTCLE, maximum triple-wise conditional likelihood estimator; MSE: mean squared error; RE: relative efficiency of the NPMLE.

Table 3. Estimation results for the leukemia study

Parameter	Estimate	SE	95% Confidence Interval
		NPMLE	
$\log_{10}$(WBC)	-0.0355	0.0141	(-0.0632,-0.0078)
AG status	-0.0403	0.0162	(-0.0721,-0.0085)
		MPCLE	
$\log_{10}$(WBC)	-0.0324	0.0217	(-0.0938,-0.0127)
AG status	-0.0381	0.0183	(-0.0893,-0.0167)
		MTCLE	
$\log_{10}$(WBC)	-0.0324	0.0212	(-0.0896,-0.0135)
AG status	-0.0381	0.0176	(-0.0863,-0.0177)

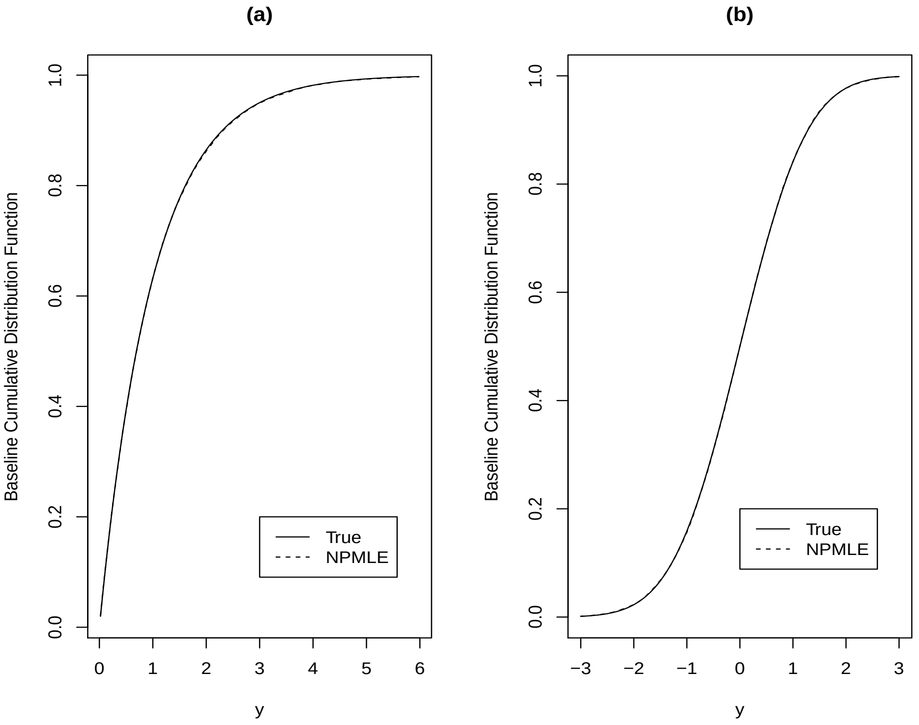

Figure 1: True and average of the NPMLE of the baseline cumulative distribution function based on 1,000 replicates: (a) standard normal; (b) exponential with mean of one.

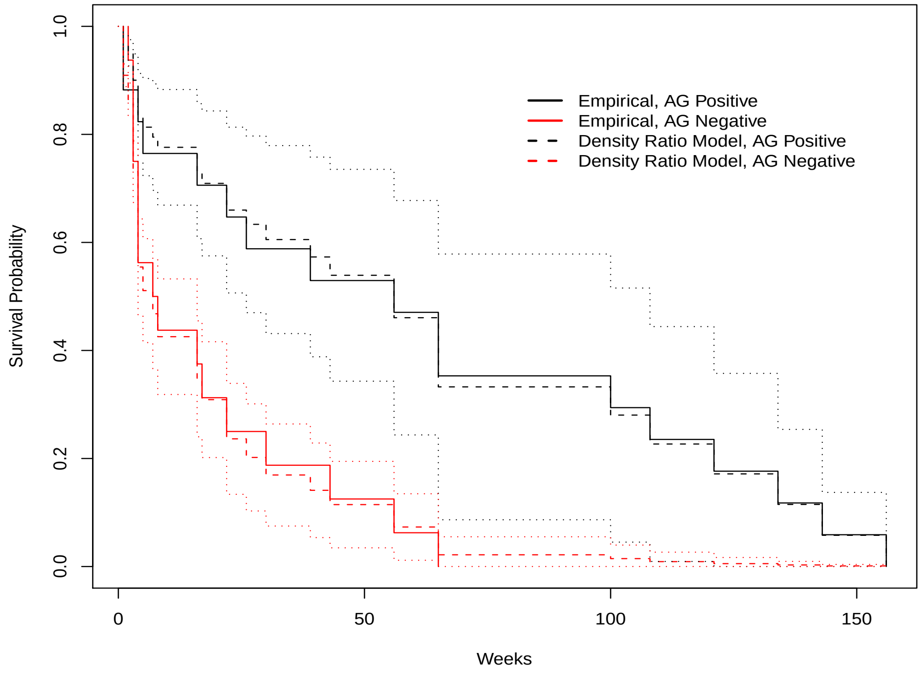

Figure 2: Empirical and model-fitted survival probabilities under the density ratio model. The black and red dotted curves correspond to the 95% point-wise confidence bands of the survival functions based on the density ratio model for the AG-positive group and AG-negative group, respectively.

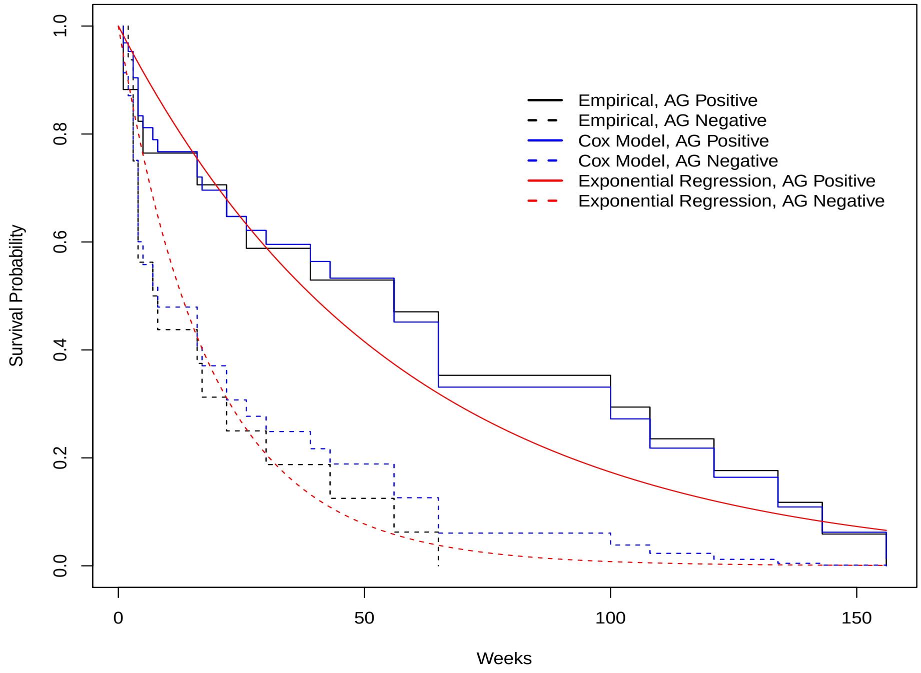

Figure 3: Empirical and model-fitted survival probabilities under the exponential regression model of Feigl and Zelen (1965) and the Cox proportional hazards model.