Some Problems in Statistical Inference Under Order Restrictions

by

Zhiguo Li

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2008

Doctoral Committee:

Associate Professor Bin Nan, Co-Chair
Professor Jeremy Taylor, Co-Chair
Associate Professor George Michailidis
Associate Professor Moulinath Banerjee
Associate Professor Douglas Schaubel

To my son Jerry

ACKNOWLEDGEMENTS

Firstly, my deepest appreciation goes to my dissertation chair, Dr. Bin Nan , for his guidance, encouragement, as well as his financial assistance. His sharp insight into statistics, broad knowledge, keen interest in research and his strategy of attacking problems benefited me a lot.

I would also like to express my sincere gratitude to my advisor Dr. Jeremy Taylor. I always learned a lot from the valuable discussions with him. This work would not have been finished without his constructive suggestions as well as his consistent financial support.

I also owe thanks to Dr. George Michailidis, Dr. Douglas Schaubel and Dr. Moulinath Banerjee for serving in my committee. Special thanks go to Dr. Timothy Johnson, who is always ready to help, and his great assistance in R programming and Latex was really invaluable to me.

Last, but not the least, I thank my parents, my brothers and sisters, and my wife, for their love, support and their confidence in me.

My son Jerry came to this world just before I finished my dissertation. He kept me so busy and made my final step of the long journey slower, but I feel much more joyful than before his arrival.

PREFACE

In many statistical problems, it is necessary to take into account the order restrictions of the unknown parameters of interest. Sometimes, it is reasonable to assume order restrictions on the parameters. For a simple example, consider the comparison of the treatment effects of a drug and a placebo. Most of the time it would be reasonable to assert that the drug has a larger effect than the placebo in treating the disease. A similar situation arises when several doses of a drug are applied. The treatment effect would be higher for a higher dose. These are among the simplest examples of isotonic regression analysis, which has a long history and has been extensively studied in the literature (see, for example, Barlow et al., 2002 and Robertson et al. 1988). In these kind of situations, the purpose of using estimators that take the order restrictions into consideration is to gain efficiency. If the true parameters indeed satisfy the order restrictions, then the estimators that take this into account are more efficient than the estimators that ignore the order restriction. In some other situations, the statistical model itself enforces order restrictions on the parameters. A common example is the estimation of the nonparametric distribution function or the cumulative hazard function of a random time to failure, or the estimation of the baseline cumulative hazard function in a Cox regression model. In these problems, one has to take into account the monotonic nature of the distribution function or the cumulative hazard function. In this dissertation, I consider three different statistical problems in which an order restriction on the unknown parameters is either

natural or reasonable, and discuss methods of estimation and inference of the unknown parameters under the restrictions. The order restriction of the parameters is the major concern for the problem in Chapter IV. In Chapters II and III, the main focus is missing covariates in Cox regression models, while the order restrictions of the parameters need to be taken care of in parameter estimation.

In the second chapter, we consider the Cox regression model with grouped survival data coming from case-cohort studies. The order restriction arises since the baseline cumulative hazard function is an increasing function. The problem with fully observed data has been studied in the literature. Here what we are interested in is the case in which the data come from case-cohort studies or more general two phase stratified sampling. In case-cohort studies, the covariate of interest is observed for all the cases and a subsample of controls. This results in missing covariates, but the probability that the covariate is observed is known for every subject. In this situation, we propose using the weighted likelihood method, in which one maximizes the inverse selection probability weighted likelihood function, to fit Cox models. The weighted likelihood estimator can be easily calculated via the Newton-Raphson iteration. The asymptotic properties of the estimator are studied. It is shown that, when the weights (or the probabilities that the covariates are observed) are reasonably estimated, the weighted likelihood estimator is more efficient than the weighted likelihood estimator when the true weights are used. This study is motivated by an HIV vaccine efficacy trial, and the proposed weighted likelihood method is applied to analyze the data coming from that trial.

In fact, treating the data from the vaccine trial as grouped survival data is only an approximation, and the true structure of the data is actually current status data. Therefore, in Chapter III we further develop the method to fit proportional hazards

models with current status data with missing covariates. We still assume the probability that the covariate is observed is a known function of a variable that is observed for every subject. We study the weighted likelihood estimator for the estimation of the unknown cumulative baseline hazard function and the logarithm of the hazard ratios. We propose an adapted version of the "iterative convex minorant algorithm" to compute the weighted likelihood estimator and establish the asymptotic properties of the estimator with true weights and with estimated weights, and show that the estimator with estimated weights is more efficient than the estimator with true weights. Since in this model the baseline cumulative hazard function is only estimable at the $n^{1/3}$ rate, and thus no existing theory is available to prove the asymptotic normality of the estimator of the log hazard ratios when the weights are estimated, we establish a general theorem for this purpose. For the estimation of variance of the estimator of the log hazard ratios, we investigate the weighted bootstrap method. It turns out that it works well for the estimator with true weights. For the estimator with estimated weights, the bootstrap does not work, so the asymptotic variance is estimated by estimating components of the variance formula separately, using nonparametric smoothing. We also did simulation studies and analyzed the vaccine trial data to illustrate the weighted likelihood method.

In Chapter IV, the problem is concerned with the inference of ordered probabilities of binomial random variables. The case in which some of the adjacent cell probabilities are equal or close to each other is of particular interest, since difficulties arise in this case and it is not well studied in the literature. We suppose that there is a one-way or two-way table, in each cell of the table, there is a binomial trial with a certain probability of "success", and the probabilities are ordered along either way of the table. This is a convenient way of describing the situation in which there

are a binary response variable and one or two categorical covariates, and the conditional probability of response given levels of the categorical covariate(s) are ordered according to levels of either covariate. The order restrictions on the probabilities are the belief of the investigator based on commonsense or scientific knowledge. For simplicity, we suppose that there are two binomial random variables and the probabilities of "success" are p_1 and p_2 and they satisfy $p_1 \leq p_2$. The maximum likelihood estimator under the order restriction (restricted MLE) of the probabilities is a natural estimator, which guarantees the order restriction and is more efficient than the estimator which ignores the order restriction. It can also be easily calculated using the "pool adjacent violators" algorithm. However, the usual normal approximation to the distribution of the estimator is not appropriate when $p_1 = p_2$ or under a local alternative type of assumption. Hence, the confidence intervals constructed by using this approximation do not have correct coverage rates when $p_1 = p_2$ or when they are very close to each other. In an attempt to resolve this problem, we find the correct asymptotic distribution of the restricted MLE when some of the two adjacent probabilities are equal or satisfy a local alternative type assumption. Confidence intervals are constructed based on these asymptotic distributions. The coverage rates of the confidence intervals are improved, especially when the true adjacent cell probabilities are equal. But when the adjacent probabilities are not equal but close to each other, these confidence intervals still do not perform well. Further, we propose using bootstrap methods to construct confidence intervals. Several types of bootstrap confidence intervals and two types of confidence intervals based on the asymptotic distribution are compared in a simulation study and the bootstrap percentile confidence interval is found to have good performance and outperforms the other types of intervals. In addition, the bootstrap procedure can be applied to problems with

parameters of higher dimension without any difficulty.

TABLE OF CONTENTS

LIST OF TABLES

Table

CHAPTER I

Weighted Likelihood Method for Grouped Survival Data in Case-Cohort Studies

1.1 Introduction

Interval censored data arise often in HIV studies where times to HIV infection are not exactly observed, but instead the two time points within which the infection happens are observed. The time points may be, for instance, the times of clinic visits. This type of data are commonly seen in practice, for example patients in clinical trials may be monitored for clinical response at a set of visit times. A special case of interval censored failure times occurs when the visit times are fixed in advance and are the same for all subjects. In this case the failure times are grouped into a discrete set of time intervals. For such a data structure, Kalbfleisch and Prentice (1973) and Prentice and Gloeckler (1978), among others, proposed and developed methods for maximum likelihood estimation of the relative risks and survival function in the proportional hazards model (Cox, 1972; Cox, 1975).

The case-cohort design was proposed by Prentice (1986) for large cohort studies (e.g., prevention trials) for which the covariates of interest are expensive to collect. In such a design, the covariate values are collected only for those subjects who experience the failure event during the follow-up period and for a subcohort that is randomly sampled from the study cohort. For right censored data, Self and Prentice (1988)

derived the asymptotic theory for a pseudo likelihood estimator of the parameters in a general relative risk model, including the proportional hazards model as a special case.

Gilbert et al. (2005) employed the Self-Prentice method to analyze data from the first randomized placebo-controlled Phase 3 trial of a preventive HIV vaccine (Flynn et al., 2005). Forthal et al. (2007) also analyzed these data, using an alternative psuedo likelihood estimator for the Cox model with case-cohort sampling (Estimator II of Borgan et al., 2000). These analyses addressed the objective to evaluate in vaccine recipients the association between anti-HIV antibody levels generated by the vaccine and subsequent HIV infection. Trial participants were immunized with vaccine or placebo at months 0, 1, 6, 12, 18, 24 and 30. Volunteers testing negative for HIV infection at month 0 were enrolled, and HIV infection tests were administered at each immunization visit and at the final follow-up visit at month 36. Serum and plasma samples were obtained from all volunteers at the immunization visits as well as at visits 2 weeks after the immunization visits, scheduled for measuring peak immunologic response values. The assays were performed for all vaccine recipients who became HIV infected and for a stratified random sample of the uninfected vaccine recipients, selected after the trial. Covariates measured on everyone include demographic variables, geographic region, race, and baseline behavioral risk score (taking integer values from 0 to 7).

For study participants who acquired HIV infection during the study, the infection time can only be determined to be between the dates of the last negative and first positive HIV tests. In both Gilbert et al.'s (2005) and Forthal et al.'s (2007) Cox model analyses of the case-cohort data, the time to infection was approximated by the midpoint of the dates of the last negative and first positive tests. Approximat-

ing interval censoring to right censoring, however, may introduce bias in parameter estimation. It is desirable to develop a more general method that takes the interval censoring nature of the failure times into account.

We propose a weighted likelihood approach to fit a proportional hazards model with grouped survival data and stratified case-cohort covariate sampling, and apply the method to evaluate the association between the newest antibody measurement described in Forthal et al. (2007) and HIV infection . The method maximizes the inverse selection probability weighted log likelihood function (or log partial likelihood function). The weighted likelihood approach has been used in other missing data problems; see Breslow and Wellner (2007) and references cited therein. In our case, we consider both true weights and estimated weights, where the true weights are calculated by using the true selection probabilities determined by design and the estimated weights are calculated by using sample fractions within strata. Both methods lead to consistent and asymptotically normal estimators of the parameters, and the variances of the estimators can be consistently estimated. As pointed out by many authors including Breslow and Wellner (2007), the method with estimated weights is more efficient. The numerical calculations can be readily carried out via Newton-Raphson iteration. We apply multiple imputation to handle missing immunological responses in the subcohort. We present the proposed methods and asymptotic results in Section 1.2 and report a simulation study in Section 1.3. In Section 1.4 we apply the proposed method to the vaccine trial study and make concluding remarks in Section 1.5. We provide detailed technical derivations and proofs of the asymptotic properties in the Appendices.

1.2 The Weighted Likelihood Method

Consider the general setting of grouped survival data. Let T be the underlying time to the event of interest, and C be the underlying censoring time. Let X be a p-dimensional covariate (process). Assume noninformative censoring and C is independent of T given X. In the HIV vaccine trial study, however, neither T nor C is completely observed. Instead, T is either known to be in one of the m fixed time intervals: $(t_0, t_1], (t_1, t_2], \ldots, (t_{m-1}, t_m)$, where $0 = t_0 < t_1 < \ldots < t_{m-1} < t_m = +\infty$, or right censored at a visit time t_j, $1 \leq j \leq m-1$. In either case, X will be observed up to the last observed visit time. The two cases coincide when $j = m - 1$.

Suppose we only observe data in the first R_i intervals for subject i, where $1 \leq R_i \leq m - 1$; then the subject either experiences an event in the R_ith interval or is right censored at t_{R_i}. Let $\Delta_{ij} = 1$ if the event for the ith subject falls into the jth interval and $\Delta_{ij} = 0$ otherwise, $1 \leq j \leq R_i$, and denote $\Delta_{i,R_i+1} = 1 - \sum_{j=1}^{R_i} \Delta_{ij}$ and $\Delta_i = (\Delta_{i1}, \ldots, \Delta_{i,R_i+1})^T$. In fact $\Delta_{ij} = 0$ for all $j < R_i$, but we keep the vector notation Δ_i for ease of technical derivation. Note that R_i is a random variable and the length of Δ_i varies with R_i. Let the covariate be componentwise constant in each of the R_i observed time intervals and denote $X_i = (X_{i1}, \ldots, X_{i,R_i})^T$, where X_{ij} is the p-dimensional covariate vector for the ith subject in the jth interval. Assume that in a full cohort, we would have n i.i.d. observations (Δ_i, R_i, X_i), $1 \leq i \leq n$, which is equivalent to observing i.i.d. observations $(\Delta_{i,R_i+1}, R_i, X_i)$, $1 \leq i \leq n$.

Suppose T follows a Cox regression model, i.e., the hazard function can be written as

$$\lambda(t|X(t)) = \lambda(t) \exp(X(t)^T \beta), \tag{1.1}$$

where $X(t)$ is the p-dimensional covariate vector at time t and $\beta = (\beta_1, \ldots, \beta_p)^T$. Let

$\Lambda(t)$ be the baseline cumulative hazard function, and denote $\alpha_k = \Lambda(t_k) - \Lambda(t_{k-1})$ and $\gamma_k = \log \alpha_k, k = 1, 2, \ldots, m$, where α_m and γ_m are equal to $+\infty$. Then the conditional probability of the event for the ith subject falling into the jth interval given X_i is

$$P(\Delta_{ij} = 1|X_i) = e^{-\sum_{k=1}^{j-1} e^{\gamma_k + X_{ik}^T \beta}} \left(1 - e^{-e^{\gamma_j + X_{ij}^T \beta}}\right), 1 \leq j \leq m.$$

Here for notational convenience we assume that $\sum_{k=1}^{0} e^{\gamma_k + X_{ik}^T \beta} = 0$. Note that the above expression only involves covariates observed up to time t_j for a fixed j.

By the conditional independence of T_i and C_i given X_i, the conditional probability mass function of (Δ_i, R_i) given X_i can be written as

$$\begin{aligned} P(\Delta_i &= \delta_i, R_i = j|X_i) \\ &= \prod_{\ell=1}^{j+1} \left(e^{-\sum_{k=1}^{\ell-1} e^{\gamma_k + X_{ik}^T \beta}}\right)^{\delta_{i\ell}} \left(1 - e^{-e^{\gamma_j + X_{ij}^T \beta}}\right)^{\delta_{ij}} f(\delta_i, j|X_i) \\ &\equiv L(\theta|\Delta_i = \delta_i, R_i = j) f(\delta_i, j|X_i), \qquad 1 \leq j \leq m-1, \end{aligned}$$

where $f(\delta_i, j|X_i)$ does not contain any information about θ and hence can be dropped when constructing the likelihood function for θ. Detailed derivation is given in Appendix A. Note that $L_i(\theta) \equiv L(\theta|\Delta_i, R_i)$ above is more complicated than necessary for numerical evaluation. But its current form will be very helpful in deriving asymptotic properties for the proposed estimator, which will be easily seen in the Appendices C and D. Also note that $L_i(\theta)$ reduces to the likelihood contribution of the ith subject in Prentice and Gloeckler (1978).

1.2.1 Estimation with True Weights

In case-cohort studies, the covariates are not observed for all subjects. Here we consider the Bernoulli sampling scheme (Manski and Lerman, 1977) for selecting the subcohort. Each subject is examined for a covariate V_i (which can either be

part of X_i or be an ancillary variable(s)) that is measured in all subjects (i.e., at phase one), and is then independently selected at phase two into the subcohort with probability $P(i \in SC|V_i) = \pi(V_i)$, where "$SC$" stands for subcohort and $\pi(\cdot)$ is a known function. The covariate X is assembled only for subjects in the subcohort and for those who experience the failure event during follow-up. The data resulting from this sampling scheme preserve an i.i.d. structure and satisfy the missing at random (MAR) assumption (Little and Rubin, 2002), because the probability that the covariate X is missing depends only on V and $\Delta_{i,R_{i+1}}$, which are always observed.

Kulich and Lin (2004) distinguished between "N-estimation" and "D-estimation" for right censored data in case-cohort sampling designs, where N-estimation uses weights that are independent of failure status while D-estimation uses weights that depend on failure status. The main reason for distinguishing these approaches is that the martingale theory applies for N-estimation, but not for D-estimation. This distinction is irrelevant for our methodology for grouped failure time data because it does not have any difficulty in handling failure status dependent weights.

For the observed data in a case-cohort study, we propose the following weighted likelihood function for making inferences on θ:

$$L_{w,n}(\theta) = \prod_{i=1}^{n} \Big\{ L_i(\theta) \Big\}^{w_i},$$

where

$$w_i = (1 - \Delta_{i,R_i+1}) + \frac{I(i \in SC)}{\pi(V_i)} \Delta_{i,R_i+1}, \quad 1 \leq i \leq n.$$

Clearly the weight w_i depends on the failure status of subject i. It is easily seen that only subjects with completely observed covariates contribute to the weighted likelihood function, and w_i is the inverse of the probability that subject i is selected from the original cohort to have covariate X_i measured. The logarithm of the weighted

likelihood function is

$$\begin{aligned}\ell_{w,n}(\theta) &= \sum_{i=1}^{n} w_i \ell_i(\theta) \\ &= \sum_{i=1}^{n} w_i \left\{ - \sum_{j=1}^{R_i+1} \left(\Delta_{ij} \sum_{k=1}^{j-1} e^{\gamma_k + X_{ik}^T \beta} \right) + \Delta_{iR_i} \log \left(1 - e^{-e^{\gamma_{R_i} + X_{iR_i}^T \beta}} \right) \right\}. \qquad (1.2)\end{aligned}$$

We call the maximizer of $\ell_{w,n}(\theta)$ the weighted likelihood estimator of θ, denoted by $\hat{\theta}_n$, which can be obtained by solving the following weighted log likelihood estimating equation for θ:

$$\frac{\partial}{\partial \theta} \ell_{w,n}(\theta) = \sum_{i=1}^{n} w_i \frac{\partial}{\partial \theta} \ell_i(\theta) = 0. \qquad (1.3)$$

The Newton-Raphson method can be employed to solve the above estimating equation. Note that the covariates after the R_ith interval do not contribute to the log likelihood function and its derivatives. Define the matrix of the second derivatives as

$$I_n = \begin{pmatrix} I_{\gamma\gamma,n} & I_{\gamma\beta,n} \\ I_{\gamma\beta,n}^T & I_{\beta\beta,n} \end{pmatrix} = \begin{pmatrix} -\partial^2 \ell_{w,n}(\theta)/\partial\gamma\partial\gamma^T & -\partial^2 \ell_{w,n}(\theta)/\partial\gamma\partial\beta^T \\ -\partial^2 \ell_{w,n}(\theta)/\partial\beta\partial\gamma^T & -\partial^2 \ell_{w,n}(\theta)/\partial\beta\partial\beta^T \end{pmatrix}.$$

The numerical inversion of I_n is necessary in Newton-Raphson iteration, which may be difficult if there are many intervals (m is large). Following the idea of Prentice and Gloeckler (1978) and Finkelstein (1986), however, the inversion can be simplified by using the following equality

$$I_n^{-1} = \begin{pmatrix} I_{\gamma\gamma,n}^{-1} + AB^{-1}A^T & -AB^{-1} \\ -B^{-1}A^T & B^{-1} \end{pmatrix},$$

where $A = I_{\gamma\gamma,n}^{-1} I_{\gamma\beta,n}, B = I_{\beta\beta,n} - I_{\gamma\beta,n}^T I_{\gamma\gamma,n}^{-1} I_{\gamma\beta,n}$, which only involves inverting the p-dimensional matrix B since $I_{\gamma\gamma,n}$ is diagonal (see Appendix B for explicit forms of the derivatives of the weighted log likelihood). Then the Newton-Raphson method

updates values of $\theta = (\gamma^T, \beta^T)^T$ iteratively via

$$\begin{pmatrix} \gamma^{(k)} \\ \beta^{(k)} \end{pmatrix} = \begin{pmatrix} \gamma^{(k-1)} \\ \beta^{(k-1)} \end{pmatrix} + \left\{ I_n^{-1} \frac{\partial \ell_{w,n}(\theta)}{\partial \theta} \right\}_{\theta=\theta^{(k-1)}}$$

until the algorithm converges; here the superscript (k) represents values in the kth iteration. Note that when the sample size is small, or some time intervals are narrow, there may be no observed events in an interval, in which case the Newton-Raphson procedure will fail. A simple remedy is to combine such an interval with its neighbor to make the number of events in the combined interval greater than zero.

The dependency of the sampling probabilities on covariates and outcome makes the case-cohort design a biased sampling design. The inverse selection probability weighted estimating equation (1.3) corrects the bias, however, because by MAR we have

$$E(w_i|\Delta_i, R_i, X_i, V_i) = (1 - \Delta_{i,R_i+1}) + \Delta_{i,R_i+1} \frac{P(i \in SC|V_i)}{\pi(V_i)} = 1, \tag{1.4}$$

and hence

$$\begin{aligned} E\left\{ w_i \frac{\partial \ell_i(\theta)}{\partial \theta} \right\} &= EE\left\{ w_i \frac{\partial \ell_i(\theta)}{\partial \theta} \middle| \Delta_i, R_i, X_i, V_i \right\} \\ &= E\left\{ \frac{\partial \ell_i(\theta)}{\partial \theta} E(w_i|\Delta_i, R_i, X_i, V_i) \right\} \\ &= E\left\{ \frac{\partial \ell_i(\theta)}{\partial \theta} \right\} = 0. \end{aligned}$$

A naive approach to the analysis would simply put $w_i = 1$ for all subjects with covariates completely observed and $w_i = 0$ otherwise. We call the corresponding estimator the naive estimator. Since the equality (1.4) does not hold for all i, in general the naive estimator will be asymptotically biased, which is verified by the simulation study in Section 1.3.

For full cohort data, Prentice and Gloeckler (1978) provided an intuitive discussion on the asymptotic properties of the maximum likelihood estimator for grouped survival data. We give a set of mild regularity conditions in the following theorem that formally establishes both consistency and asymptotic normality of the weighted likelihood estimator with true weights that are usually known for a case-cohort design, which includes the maximum likelihood estimator of Prentice and Gloeckler (1978) as a special case. The proof is given in Appendix C.

Theorem I.1. *Suppose the parameter space Θ is compact and the true parameter θ_0 is an interior point of Θ. Assume the following conditions hold:*

(i) The covariate X has bounded support.

(ii) The variance matrix of X_{ij} is positive definite for all $1 \leq j \leq m-1$.

(iii) $\pi(V_i) \geq \delta > 0$ for all i and some $\delta > 0$.

(iv) $P(C_i \geq t_{m-1}|X_i) > 0$ with probability 1.

If the maximizer $\hat{\theta}_n$ of $\ell_{w,n}(\theta)$ does not occur on the boundary of Θ, then as $n \to \infty$, $\hat{\theta}_n$ converges to θ_0 in probability, and $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution to a Gaussian random variable with mean zero and variance $\Sigma(\theta_0) = I^{-1}(\theta_0)D(\theta_0)I^{-1}(\theta_0)$, where $I(\theta) = E_{\theta_0}\{\partial^2 \ell_i(\theta)/\partial\theta\partial\theta^T\}$ and $D(\theta) = E_{\theta_0}[\{w_i\partial\ell_i(\theta)/\partial\theta\}\{w_i\partial\ell_i(\theta)/\partial\theta\}^T]$.

Note that the compactness of Θ and the boundedness of X guarantee that the probability of observing an event in each of the m intervals is strictly bounded between 0 and 1. The asymptotic variance $\Sigma(\theta_0)$ can be consistently estimated by the sandwich estimator

$$\hat{\Sigma}_n(\hat{\theta}_n) = \hat{I}_n^{-1}(\hat{\theta}_n)\hat{D}_n(\hat{\theta}_n)\hat{I}_n^{-1}(\hat{\theta}_n),$$

where $\hat{I}_n(\theta) = n^{-1}\sum_{i=1}^n w_i\{\partial^2\ell_i(\theta)/\partial\theta\partial\theta^T\}$, $\hat{D}_n(\theta) = n^{-1}\sum_{i=1}^n w_i^2\{\partial\ell_i(\theta)/\partial\theta\}\{\partial\ell_i(\theta)/\partial\theta\}^T$.

1.2.2 Estimation with Estimated Weights

Although the sampling probabilities $\pi(V_i)$ are known, using estimated weights in which $\pi(V_i)$ is replaced by its estimator can improve the efficiency of the weighted likelihood estimator (Robins et al., 1994; Breslow and Wellner, 2007). Suppose that all censored subjects are divided into S strata by the variable $V \in \mathcal{V} \equiv \{\nu_1, \ldots, \nu_S\}$, and in this subsection, we denote the true sampling probabilities by $\pi(\nu_s) = p_{0s}$, $1 \le s \le S$. Suppose that there are n_s subjects in stratum s, out of whom n_s^* are selected into the subcohort by the independent Bernoulli sampling. We assume that when $n \to \infty$, $n_s/n \to \alpha_s > 0$, $1 \le s \le S$. Instead of using the true sampling probabilities $p_0 = (p_{01}, \ldots, p_{0S})^T$ in the weight function w, we now replace each p_{0s} with the sampling fraction $\hat{p}_s = n_s^*/n_s$, $1 \le s \le S$, and set $\hat{\pi}(V_i) = \hat{p}_s$ if $V_i = \nu_s$, $1 \le s \le S$. Now the estimated weight function becomes

$$w_i(\hat{p}) = (1 - \Delta_{i,R_{i+1}}) + \frac{I(i \in SC)}{\hat{\pi}(V_i)} \Delta_{i,R_{i+1}}, \qquad 1 \le i \le n.$$

Denote the maximizer of $\sum_{i=1}^n w_i(\hat{p})\ell(\theta; X_i)$ by $\tilde{\theta}_n$. The following theorem establishes the consistency and asymptotic normality of $\tilde{\theta}_n$, but with a different asymptotic variance matrix to that of $\hat{\theta}_n$ given in Theorem I.1. A detailed proof is given in Appendix D.

Theorem I.2. *Under the same conditions in Theorem II.1, $\tilde{\theta}_n$ is consistent and $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ converges in distribution to a Gaussian random variable with mean zero and variance*

$$\Sigma(\theta_0) - I^{-1}(\theta_0) B(\theta_0, p_0) G_{22} B^T(\theta_0, p_0) I^{-1}(\theta_0)$$

as $n \to \infty$, *where*

$$\begin{aligned} B(\theta, p) &= E_{\theta_0}[\{\partial \ell(\theta; X_i)/\partial \theta\}\{\partial w_i(p)/\partial p\}^T], \\ G_{22} &= \text{diag}\{p_{01}(1-p_{01})/\alpha_1, \ldots, p_{0S}(1-p_{0S})/\alpha_S\}, \end{aligned}$$

which can be consistently estimated by

$$\begin{aligned} \hat{B}(\theta, p) &= \frac{1}{n}\sum_{i=1}^{n}\{\partial \ell(\theta; X_i)/\partial \theta\}\{\partial w_i(p)/\partial p\}^T, \\ \hat{G}_{22} &= \text{diag}\{n\hat{p}_1(1-\hat{p}_1)/n_1, \cdots, n\hat{p}_S(1-\hat{p}_S)/n_S\}. \end{aligned}$$

1.2.3 Multiple Imputation Approach to Handling Missing Covariate Data

Due to the expense of measuring the antibody responses in the HIV vaccine trial, the antibody level for vaccine recipients who failed was only measured at the beginning of the first interval (at the month 6.5 visit) and at the visit immediately preceding the failure visit, and for censored vaccine recipients it was only measured at month 6.5 and at a randomly selected visit month after month 6.5. Since the missing elements of X for subject i are missing by design, depending only on $\Delta_{i,R_{i+1}}$, the missing mechanism is MAR (Little and Rubin, 2002). To handle this type of missing data, we propose using multiple imputation to fill in the missing components of X.

Specifically, suppose only X_2 can be missing. For each time interval 2 through $m-1$ (excluding the last interval), we impute the missing values of X_2 by random draws from a linear regression model with the covariate in the first interval as the predictor, which is fitted separately for cases and non-cases. For example, to impute missing covariate values in the second interval for cases, we first fit a linear model $X_{22} = c_0 + c_1 X_{21} + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$, using all the cases with complete data for X_{22}. After obtaining estimates $\hat{c} = (\hat{c}_0, \hat{c}_1)^T$ and $\hat{\sigma}^2$, we then take a random draw of σ^{*2} from $\hat{\sigma}^2 \chi_{n+1}$, where n is the number of subjects included in the linear regression,

and c^* and ε^* are random draws from $N(\hat{c}, \sigma^{*2}(A^TA)^{-1})$ and $N(0, \sigma^{*2})$, respectively, where A is the design matrix of the linear regression. Finally, we fill in the missing value X_{22} by $\hat{X}_{22} = c_1^* + c_2^* X_{21} + \varepsilon^*$. We construct 10 complete data sets following this procedure. For each imputed data set, we calculate the weighted likelihood estimator of β and its variance estimate, and then combine the 10 sets of results using the method of Little and Rubin (2002) to obtain the final estimate and its variance estimate. Confidence intervals for β are calculated using the t distribution following Little and Rubin (2002).

1.3 Simulation Studies

We conducted simulations to assess the performance of the weighted likelihood estimator by comparing the bias, efficiency and coverage properties to other estimators including the maximum likelihood estimator for full cohort data, the naive estimator for case-cohort data, and the Self-Prentice (1988) pseudo likelihood estimator for case-cohort data. The pseudo likelihood estimation is based on approximating interval censoring by right censoring, whereby event times are defined by the midpoint of the left- and right-censoring intervals.

We consider two covariates (X_1, X_2), where the corresponding regression coefficients are $(1, -1)^T$. Note that the subscript of X here denotes covariate component, not an index for study subject as in Section 1.2. To match the HIV vaccine trial (Flynn et al., 2005), we set the time origin as 6.5 months post-entry (the time by which the study subjects are "fully immunized") and use six time intervals ($m = 6$) with fixed visit times at months 12, 18, 24, 30, and 36. The covariate X_1 is set to be discrete and time-independent, which takes values 1 and 2 with equal probability. The covariate $X_2 = (X_{21}, X_{22}, X_{23}, X_{24}, X_{25})^T$ is specified as a 5-variate random vector corresponding to the five post-immunization visits at months 6.5, 12.5, 18.5,

24.5, 30.5, where X_{2j} is the covariate value of X_2 in the jth interval. The conditional distribution of X_2 given X_1 is normal, i.e., $X_2|X_1 = k \sim N(\mu_k, \Sigma)$, $k = 1, 2$, with $\mu_1 = (0.1, 0.2, 0.3, 0.4, 0.5)^T$, $\mu_2 = (0, 0.1, 0.2, 0.3, 0.4)^T$, and

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix},$$

where $\rho = 0.7$. With this set-up the covariates X_{2j}, $j = 1, \ldots, 5$, are positively correlated following an AR(1) model, and X_1 and X_2 are also correlated.

We choose the cohort size n as 200, 500 or 3000. When $n = 200$, the probability of selecting censored subjects into the subcohort is 0.333 and the baseline hazard is a constant value 0.015; When $n = 500$, the probability of selecting censored subjects into the subcohort is 0.25 and the baseline hazard is a constant value 0.02; when $n = 3000$, the selection probability is 0.085 for censored subjects and the baseline hazard is a constant value 0.005. With these settings there are approximately 90 completely observed subjects when $n = 200$, among whom about 40 are failures, approximately 200 completely observed subjects when $n = 500$, among whom about half are failures, and approximately 400 completely observed subjects when $n = 3000$, among whom about 150 are failures. The last situation resembles the HIV vaccine trial data that will be analyzed in the next section. The survival times are generated from a piecewise exponential distribution specified by model (1.1) (with $\lambda_0(t) \equiv c$ specified above). Censoring times are generated from a discrete uniform subdistribution at months $(12, 18, 24, 30)$ combined with a truncation at month 36 to yield about 25 early dropouts (prior month 36), similar to what was observed in

the HIV study. One thousand simulation runs are conducted under each simulation setting.

For each simulation run, parameter estimates are obtained by solving equation (1.3) with estimated weights using the Newton-Raphson method. The initial value of β is set to be zero, and the initial value of γ is obtained from the Kaplan-Meier curve $S^{(0)}(\cdot)$, calculated by pushing the failure time to the right end point of the interval in which an event occurs, via $\gamma_j^{(0)} = \log[\log\{S^{(0)}(t_j)\} - \log\{S^{(0)}(t_{j+1})\}]$, $1 \leq j \leq m-1$. Then the variance estimator is calculated from the expressions given in Theorems II.1 and II.2, and the 95% Wald confidence interval for each parameter is obtained based on the asymptotic normality. Bias, coverage percentage, the average of the estimated standard deviations, and the empirical standard deviation are calculated from the 1000 simulation runs. Since the parameter of interest is β, only the bias for estimating γ is reported. The relative efficiency of the weighted likelihood estimator of β versus the maximum likelihood estimator (MLE) computed from the full data is calculated by the ratio of empirical variances.

In addition to evaluating the different methods with no missing components in X, we evaluate the weighted likelihood method with multiple imputation, by coarsening the simulated X_2 covariates to have missing components in the pattern described in SubSection 2.2.3. Tables 1.1 and 1.2 summarize the simulation results, where weights are estimated by sampling fractions. From Table 1.1 we see that the weighted likelihood estimators have reasonably small biases. The standard deviation estimators for $\hat{\beta}$ are accurate, which lead to accurate coverage percentages. The multiple imputation method works well. It is not surprising that the weighted likelihood method for case-cohort data is less efficient than the maximum likelihood estimator for the full cohort data. However, under case-cohort sampling the weighted likelihood method

is much more efficient than the naive method that uses simple random sampling. In addition, by ignoring the biased sampling nature of the case-cohort sampled data, the naive estimator is clearly biased. The pseudo likelihood method of Self and Prentice (1988) that uses approximated right censored data is also more biased than the weighted likelihood method for grouped survival data. From Table 1.1 we see that the bias of $\hat{\gamma}$ is severe for both the naive method and the pseudo likelihood method, whereas it is very small for the weighted likelihood method.

To better illustrate the efficiency gain of the weighted likelihood estimator with estimated weights compared to the estimator with true weights, we generate an auxiliary variable V that is a coarsening of X. Particularly, $V = 1$ if the average of X_2 over the five intervals is less than 1 and $X_1 = 1$; $V = 2$ if the average of X_2 is less than 1 and $X_1 = 2$; $V = 3$ if the average of X_2 is greater than 1 and $X_1 = 1$; and $V = 4$ if the average of X_2 over the five intervals is greater than 1 and $X_1 = 2$; The subcohort is selected by stratified Bernoulli sampling from the 4 strata defined by V. When $n = 200$, the subcohort sampling probabilities are 0.4, 0.4, 0.7, and 0.7 for the 4 strata. When $n = 500$, the sampling probabilities are 0.2, 0.2, 0.7, and 0.7. When $n = 3000$, the sampling probabilities are 0.05, 0.05, 0.25, and 0.25. The probabilities are determined such that the numbers of failures and controls selected into the subcohort are approximately the same as in the previous simulation. Results are give in Table 1.3, which clearly show the advantage of using estimated weights.

1.4 Analysis of the HIV Vaccine Trial Data

We now analyze the HIV vaccine trial data using the weighted likelihood method to investigate the association between antibody levels and HIV infection. We investigate the newest antibody measurement described in Forthal et al. (2007), which quantitates the degree to which the serum of a vaccine recipient reduces (relative

to control serum) the avidity of the binding of soluble CD4 to the GNE8 strain of HIV. We refer to this antibody variable as the GNE8 CD4 avidity level. We focus on measurements taken at month 6.5, 12.5, 18.5, 24.5, and 30.5 to evaluate the relationship between peak GNE8 CD4 avidity levels and the rate of HIV infection. Because this antibody variable was only obtained from vaccine recipients who tested HIV negative at month 6, and the main scientific goal is to evaluate the association in vaccine recipients after they received the third immunization at month 6.5, the time intervals for analysis are $[6.5, 12)$, $[12.5, 18)$, $[18, 24)$, $[24, 30)$, $[30, 36)$, and $[36, \infty)$, where month 36 is the time of the final study visit.

The GNE8 CD4 avidity level was measured for all infected vaccine recipients and for a stratified random sample of uninfected vaccine recipients. Placebo recipients are not used in the analysis because their GNE8 CD4 avidity levels all equal 0. We only consider men in the analysis since only 4 women were included in the case-cohort sample. The stratification variable is defined by four demographic subgroups: white low risk men, nonwhite low risk men, white higher risk men, and nonwhite higher risk men, with sampling fractions 0.047, 0.176, 0.208, and 0.450, respectively. Here low (higher) risk subjects are those who had baseline behavioral risk score (defined in Flynn et al., 2005) below or equal to (greater than) 2. The entire cohort size of vaccine recipients at the time origin month 6.5 is 3370, of whom 131 became HIV infected by month 36. Among uninfected vaccine recipients, 115, 73, 71, and 18 were sampled from the four strata for measuring the GNE8 CD4 avidity level. Among the 277 sampled uninfected vaccine recipients, 254 were right censored at month 36, and 23 were right censored at an earlier visit time.

In addition to the primary covariate of interest peak GNE8 CD4 avidity level, other covariates included in the Cox model analysis are race (white or nonwhite)

and baseline behavioral risk score. The baseline risk score is categorized into three groups: low (< 2), medium (2 or 3), and high (> 3). The peak antibody level is time-dependent, but is assumed constant between two adjacent vaccine shots. It is measured at time-points described at the beginning of SubSection 1.2.3.

To handle the missing covariate data we use the multiple imputation approach described in SubSection 1.2.3. During the data exploration we found that the contribution of the antibody level in model (1.1) is monotone, but not linear, with faster increase at lower antibody levels. By trying out a few power transformations of the antibody level, we found the one fifth power transformation seemed to provide an estimated linear effect. Hence we implemented this transformation in the final analysis.

The results are presented in Table 1.4. We first investigated interactions between antibody level and the other covariates, and none are statistically significant. On main effects, the race effect is not statistically significant, while baseline risk group is highly significant. Compared to the low risk group, the estimated relative hazard of HIV infection for the medium or high risk groups is approximately tripled, controlling for antibody level and race. The GNE8 CD4 avidity levels are significantly inversely associated with HIV infection rate. Note that on their original scale the antibody levels range from 0 to about 0.75, and their transformed values range from 0 to about 0.95. From Table 1.3 we see that the estimated log relative hazard of infection for every 0.1 unit increase in the one fifth power of antibody level is -0.120 with 95% confidence interval of $(-0.203, -0.034)$, controlling for race and baseline risk score. Transformed back to the original scale, the strength of association is larger at lower values of the antibody level. For example, an antibody level of 0.25 compared to 0 reduces the hazard of HIV infection by about 59.8%; an antibody level of

0.5 compared to 0.25 reduces the hazard by 12.7%; and the antibody level of 0.75 compared to 0.5 reduces the hazard by 8.5%, controlling for race and baseline risk score.

1.5 Discussion

It should also be noted that, although the weighted likelihood estimator provides an intuitively reasonable method that can be easily carried out numerically, it is not the most efficient estimator. Efficient estimation will in general involve the joint distribution of covariates and high-dimensional integration, and hence is much more complicated, especially when some covariates are continuous. When covariates are discrete, a simpler derivation is possible, but not pursued here.

Appendix A: Derivation of $P(\Delta_i = \delta_i, R_i = j|X_i)$

Clearly the pair of random variables (Δ_i, R_i), or equivalently (Δ_{i,R_i+1}, R_i), is completely determined by (T_i, C_i). In particular, the set $\{\Delta_{i,R_i+1} = 0, R_i = j\}$ is equivalent to observing the event in $(t_{j-1}, t_j]$, which in turn is equivalent to the set $\{T_i \in (t_{j-1}, t_j], C_i \geq t_j\}$; and the set $\{\Delta_{i,R_i+1} = 1, R_i = j\}$ is equivalent to censoring the event at time t_j, which in turn is equivalent to the set $\{T_i \geq t_j, C_i \in (t_{j-1}, t_j]\}$. Let δ_i denote the realized vector values of Δ_i. Then by the conditional independence of T_i and C_i given X_i, the conditional probability mass function of (Δ_i, R_i) given X_i

can be written as

$$
\begin{aligned}
&P(\Delta_i = \delta_i, R_i = j|X_i) \\
&\quad = P\left\{T_i \in (t_{j-1}, t_j], C_i \geq t_j \middle| X_i\right\}^{1-\delta_{i,j+1}} P\left\{T_i \geq t_j, C_i \in (t_{j-1}, t_j] \middle| X_i\right\}^{\delta_{i,j+1}} \\
&\quad = \left\{e^{-\sum_{k=1}^{j-1} e^{\gamma_k + X_{ik}^T\beta}}\left(1 - e^{-e^{\gamma_j + X_{ij}^T\beta}}\right)\right\}^{1-\delta_{i,j+1}} \left\{e^{-\sum_{k=1}^{j} e^{\gamma_k + X_{ik}^T\beta}}\right\}^{\delta_{i,j+1}} f(\delta_i, j|X_i) \\
&\quad = \prod_{\ell=1}^{j}\left\{e^{-\sum_{k=1}^{\ell-1} e^{\gamma_k + X_{ik}^T\beta}}\left(1 - e^{-e^{\gamma_\ell + X_{i\ell}^T\beta}}\right)\right\}^{\delta_{i\ell}} \left\{e^{-\sum_{k=1}^{j} e^{\gamma_k + X_{ik}^T\beta}}\right\}^{\delta_{i,j+1}} f(\delta_i, j|X_i) \\
&\quad = \prod_{\ell=1}^{j+1}\left(e^{-\sum_{k=1}^{\ell-1} e^{\gamma_k + X_{ik}^T\beta}}\right)^{\delta_{i\ell}} \left(1 - e^{-e^{\gamma_j + X_{ij}^T\beta}}\right)^{\delta_{ij}} f(\delta_i, j|X_i) \\
&\quad \equiv L(\theta|\Delta_i = \delta_i, R_i = j) f(\delta_i, j|X_i), \qquad 1 \leq j \leq m-1,
\end{aligned}
$$

where $f(\delta_i, j|X_i) = \{P(C_i \geq t_j|X_i)\}^{1-\delta_{i,j+1}}\{P(t_j < C_i \leq t_{j+1}|X_i)\}^{\delta_{i,j+1}}$.

Appendix B: Derivatives of the Weighted Log Likelihood

Denote $h_{ij} = e^{\gamma_j + X_{ij}^T\beta}$, $1 \leq i \leq n$, $1 \leq j \leq m-1$. The first order derivatives of the weighted likelihood function are $\partial\ell_{w,n}(\theta)/\partial\theta = \sum_{i=1}^{n} w_i \partial\ell_i(\theta)/\partial\theta$, where

$$
\begin{aligned}
\frac{\partial\ell_i(\theta)}{\partial\beta} &= -\sum_{j=1}^{R_i+1}\left(\Delta_{ij}\sum_{k=1}^{j-1} h_{ik}X_{ik}\right) + \Delta_{iR_i}\frac{h_{iR_i}e^{-h_{iR_i}}}{1 - e^{-h_{iR_i}}}X_{iR_i}, \\
\frac{\partial\ell_i(\theta)}{\partial\gamma_s} &= -\sum_{j=s+1}^{R_i+1}\{\Delta_{ij}h_{is}I(R_i \geq s)\} + \Delta_{is}\frac{h_{is}e^{-h_{is}}}{1 - e^{-h_{is}}}I(R_i = s), \quad 1 \leq s \leq m-1.
\end{aligned}
$$

Let

$$
b_{ij} = \frac{h_{ij}e^{-h_{ij}}}{1 - e^{-h_{ij}}}\left(1 - \frac{h_{ij}}{1 - e^{-h_{ij}}}\right), \quad 1 \leq i \leq n,\ 1 \leq j \leq m-1.
$$

Then the second order derivatives are $\partial^2 \ell_{w,n}(\theta)/\partial\theta\partial\theta^T = \sum_{i=1}^n w_i \partial^2 \ell_i(\theta)/\partial\theta\partial\theta^T$, where

$$\frac{\partial^2 \ell_i(\theta)}{\partial\beta\partial\beta^T} = -\sum_{j=1}^{R_i+1}\left(\Delta_{ij}\sum_{k=1}^{j-1} h_{ik}X_{ik}X_{ik}^T\right) + \Delta_{iR_i}b_{iR_i}X_{iR_i}X_{iR_i}^T,$$

$$\frac{\partial^2 \ell_i(\theta)}{\partial\gamma_s^2} = -\sum_{j=s+1}^{R_i+1}\{\Delta_{ij}h_{is}I(R_i \geq s)\} + \Delta_{is}b_{is}I(R_i = s), \quad 1 \leq s \leq m-1,$$

$$\frac{\partial^2 \ell_i(\theta)}{\partial\beta\partial\gamma_s} = -\sum_{j=s+1}^{R_i+1}\{\Delta_{ij}h_{is}X_{is}I(R_i \geq s)\} + \Delta_{is}b_{is}X_{is}I(R_i = s),$$

$$\frac{\partial^2 \ell_{w,n}(\theta)}{\partial\gamma_s\partial\gamma_t} = 0, \quad s \neq t.$$

Appendix C: Proof of Theorem II.1

The proof of consistency of $\hat{\theta}_n$ is based on Theorem 5.7 of van der Vaart (1998), which can be reduced to the following Lemma 1 that is more relevant to our problem. In the following we omit the word "outer" from outer probability and outer integral, and refer the detailed arguments to van der Vaart and Wellner (1996), Chapter 1.

LEMMA 1: *For i.i.d. observations $Z_1, \cdots, Z_n$, let $M_n(\theta) = n^{-1}\sum_{i=1}^n m_\theta(Z_i)$ and $M(\theta) = Em_\theta(Z)$, where $\theta \in \Theta \subset R^d$. Assume that Θ is compact, $M(\theta)$ is continuous and has a unique maximizer at θ_0, and the measurable function $\theta \mapsto m_\theta(Z)$ is continuous for every Z and dominated by an integrable function. Then any sequence of estimators $\hat{\theta}_n$ satisfying $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_p(1)$ converges in probability to θ_0 as $n \to \infty$.*

Proof: Since Θ is compact and the function $\theta \mapsto m_\theta(Z)$ is continuous for every Z and dominated by an integrable function, the class of functions $\{m_\theta : \theta \in \Theta\}$ is Glivenko-Cantelli (see example 19.8 in van der Vaart, 1998). Hence we have the uniform convergence of $M_n(\theta)$, i.e., $\sup_{\theta\in\Theta}|M_n(\theta) - M(\theta)| \to 0$ in probability as

$n \to \infty$. On the other hand, by the compactness of Θ and the fact that the function $M(\theta)$ has a unique maximizer at θ_0, we have $\sup_{\|\theta-\theta_0\|\geq\varepsilon} M(\theta) < M(\theta_0)$ for every $\varepsilon > 0$. Hence the conditions of Theorem 5.7 of van der Vaart (1998) are satisfied, and it follows that $\hat{\theta}_n \to \theta_0$ in probability. $\square$

We now apply Lemma 1 to prove the consistency of $\hat{\theta}_n$ in Theorem II.1. By Lemma 1, it suffices to show that the class of functions $\{w\ell(\theta) : \theta \in \Theta\}$ are continuous and bounded by an integrable function, and $\mu(\theta) = E_{\theta_0}\{w\ell(\theta)\}$ is continuous and has a unique maximizer at θ_0, where $\ell(\theta)$ is the log likelihood function for one subject with the subscript i suppressed. From equation (2) we see that $\ell(\theta)$ is continuous and bounded by a constant since γ, β and X_j are all bounded. In addition, w is bounded by Condition (iii). Thus the function $\theta \mapsto w\ell(\theta)$ is uniformly bounded by an integrable function. Then the continuity of $\mu(\theta)$ follows from the dominated convergence theorem. It remains to show that $\mu(\theta)$ has a unique maximizer at θ_0.

Let $\mu^*(\theta) = \mu(\theta) - \mu(\theta_0)$. Denote the joint density of (Δ, R, X) by p_θ. Then for any $\theta \in \Theta$ we have

$$\begin{aligned}
\mu^*(\theta) &= E_{\theta_0}\{w\ell(\theta) - w\ell(\theta_0)\} \\
&= E_{\theta_0}\left\{w \log \frac{p_\theta(\Delta, R, X)}{p_{\theta_0}(\Delta, R, X)}\right\} \\
&= E_{\theta_0}\left\{\log \frac{p_\theta(\Delta, R, X)}{p_{\theta_0}(\Delta, R, X)} E_{\theta_0}(w|\Delta, R, X, V)\right\} \\
&= E_{\theta_0}\left\{\log \frac{p_\theta(\Delta, R, X)}{p_{\theta_0}(\Delta, R, X)}\right\} \qquad \text{by (1.4)} \\
&\leq \log E_{\theta_0}\left\{\frac{p_\theta(\Delta, R, X)}{p_{\theta_0}(\Delta, R, X)}\right\} \qquad \text{by the Jensen's inequality} \\
&= \log 1 = 0.
\end{aligned}$$

Hence $\mu(\theta)$ is maximized at θ_0. Note that the above calculation shows that $\mu^*(\theta)$ is equivalent to the negative Kullback-Leibler divergence and thus less than or equal

to 0. Furthermore, since the equality in (1.5) holds if and only if $p_{\theta_0}(\Delta, R, X) = p_\theta(\Delta, R, X)$ with probability 1, we have that $\mu(\theta) = \mu(\theta_0)$ if and only if $p_{\theta_0}(\Delta, R, X) = p_\theta(\Delta, R, X)$ with probability 1. Denote $\theta_0 = (\gamma_{1,0}, \cdots, \gamma_{m-1,0}, \beta_0^T)^T$. Then by (2.1) we have $\gamma_k + X_k^T\beta = \gamma_{k,0} + X_k^T\beta_0$, or equivalently $X_k^T(\beta - \beta_0) = \gamma_{k,0} - \gamma_k$, with probability 1, for all k. Since $\mathrm{Var}(X_k) > 0$, we must have $\beta = \beta_0$ and $\gamma_k = \gamma_{k,0}$ for all k, i.e., $\theta = \theta_0$. Therefore, $\mu(\theta)$ has a unique maximizer at θ_0. Thus the consistency of $\hat{\theta}_n$ follows from Lemma 1. □

The proof of asymptotic normality of $\hat{\theta}_n$ in Theorem I.1 can be done by applying Theorem 5.23 of van der Vaart (1998), which is listed as Lemma 2 in the following for ease of reference.

LEMMA 2: *Let $Z_1, \ldots, Z_n$ be a random sample from some distribution P. For each θ in an open subset of Euclidean space, let $z \mapsto m_\theta(z)$ be a measurable function such that $\theta \mapsto m_\theta(z)$ is differentiable at θ_0 for $P-$almost every z with derivative $\dot{m}_{\theta_0}(z)$ and such that, for every θ_1 and θ_2 in a neighborhood of θ_0 and a measurable function $\dot{m}$ with $E_{\theta_0}\dot{m}^2 < \infty$,*

$$|m_{\theta_1}(z) - m_{\theta_2}(z)| \leq \dot{m}(z)\|\theta_1 - \theta_2\|.$$

Furthermore, assume that the map $\theta \mapsto E_{\theta_0} m_\theta$ admits a second order Taylor expansion at a point of maximum θ_0 with nonsingular symmetric second derivative matrix V_{θ_0}. If $\sum_{i=1}^n m_{\hat{\theta}_n}(Z_i) \geq \sup_\theta \sum_{i=1}^n m_\theta(Z_i) - o_p(1)$ and $\hat{\theta}_n \to_p \theta_0$, then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{m}_{\theta_0}(Z_i) + o_p(1).$$

Proof: See van der Vaart (1998), page 54. □

We introduce some additional notation before proving the asymptotic normality of $\hat{\theta}_n$. We still suppress the subscript i for subject i because we have i.i.d. observations. For a single observation, let $D = 1$ if the subject either has a failure observed or is right censored at t_{m-1} (the last visit time), and $D = 0$ if the subject is right censored at a time earlier than t_{m-1}. We also extend the length of Δ to m if an event is observed (it is m when the failure time is censored at t_{m-1}) by adding $m - r$ zeros to the remaining intervals after the interval that contains the event. Then the likelihood function for the subject can be decomposed as

$$\begin{aligned} L(\theta) &= \left[\prod_{j=1}^{m}\left\{e^{-\sum_{k=1}^{j-1}e^{\gamma_k+X_k^T\beta}}\left(1-e^{-e^{\gamma_j+X_j^T\beta}}\right)\right\}^{\Delta_j}\right]^{D}\left\{e^{-\sum_{j=1}^{r}e^{\gamma_j+X_j^T\beta}}\right\}^{1-D} \\ &\equiv \{L^{(1)}(\theta)\}^{D}\{L^{(2)}(\theta)\}^{1-D}. \end{aligned}$$

Likewise, the log likelihood function can be written as

$$\ell(\theta) = D\ell^{(1)}(\theta) + (1-D)\ell^{(2)}(\theta), \tag{1.5}$$

where $\ell^{(1)}(\theta) = \log L^{(1)}(\theta)$, and $\ell^{(2)}(\theta) = \log L^{(2)}(\theta)$.

We are now in a position to prove the asymptotic normality of $\hat{\theta}_n$ by checking the conditions of Lemma 2. Identify Z and $m_\theta(Z)$ in the lemma with (Δ, R, X) and $w\ell(\theta)$. Obviously the map $z \mapsto m_\theta(z)$ is measurable and $\theta \mapsto m_\theta(z)$ is differentiable at any θ in Θ for every z. By (1.5) and the boundedness of w, θ and (Δ, R, X), every element of $\dot{m}_\theta(z) = \partial m_\theta(z)/\partial\theta$ is bounded in both θ and z by a common constant, say, C. By the mean value theorem and the Cauchy-Schwartz inequality we have

$$\begin{aligned} &|m_{\theta_1}(z) - m_{\theta_2}(z)| = |\dot{m}_{\theta^*}(z)^T(\theta_1 - \theta_2)| \\ \leq\ &\|\dot{m}_{\theta^*}(z)\| \cdot \|\theta_1 - \theta_2\| \leq (p + m - 1)C\|\theta_1 - \theta_2\|, \end{aligned}$$

where θ^* lies on the line segment between θ_1 and θ_2. Hence we can take $\dot{m}(z)$ in Lemma 2 to be $(m + p - 1)C$ and the condition $E_{\theta_0}\dot{m}^2(Z) < \infty$ is automatically

satisfied. Since elements in both $\partial m_\theta(z)/\partial\theta$ and $\partial^2 m_\theta(z)/\partial\theta\partial\theta^T$ are bounded by integrable functions, by the dominated convergence theorem we can exchange the second order derivative and the expectation. Hence the map $\theta \mapsto E_{\theta_0} m_\theta$ admits a second-order Taylor expansion. Now we only need to show that V_{θ_0} in Lemma 2 is nonsingular.

By (1.4) we have $E_{\theta_0} m_\theta = E_{\theta_0}\{w\ell(\theta)\} = E_{\theta_0}\ell(\theta)$. Hence $V_{\theta_0} = E_{\theta_0}\{\partial^2\ell(\theta)/\partial\theta\partial\theta^T\}_{\theta=\theta_0} = -I(\theta_0)$. Since

$$I(\theta_0) = -E_{\theta_0}\left\{\frac{\partial^2\ell(\theta)}{\partial\theta\partial\theta^T}\right\}_{\theta=\theta_0} = E_{\theta_0}\left\{\frac{\partial\ell(\theta)}{\partial\theta}\left(\frac{\partial\ell(\theta)}{\partial\theta}\right)^T\right\}_{\theta=\theta_0},$$

if $I(\theta_0)$ singular, then there must exist a nonzero constant real vector α such that $\alpha^T I(\theta_0)\alpha = 0$, which implies by (1.5) that

$$E_{\theta_0}\left\{\alpha^T\frac{\partial\ell(\theta)}{\partial\theta}\right\}^2_{\theta=\theta_0} = E_{\theta_0}\left\{D\left(\alpha^T\frac{\partial\ell^{(1)}(\theta)}{\partial\theta}\right)^2 + (1-D)\left(\alpha^T\frac{\partial\ell^{(2)}(\theta)}{\partial\theta}\right)^2\right\}_{\theta=\theta_0} = 0.$$

Hence $E_{\theta_0}[D\{\alpha^T\partial\ell^{(1)}(\theta)/\partial\theta\}^2]_{\theta=\theta_0} = 0$. Again by (1.5) we have,

$$\begin{aligned}
\left.\frac{\partial\ell^{(1)}(\theta)}{\partial\gamma_s}\right|_{\theta=\theta_0} &= -\sum_{j=s+1}^{m}\Delta_j h_s^0 + \Delta_s\frac{h_s^0 e^{-h_s^0}}{1-e^{-h_s^0}}, \qquad s = 1, 2, \cdots, m-1, \\
\left.\frac{\partial\ell^{(1)}(\theta)}{\partial\beta}\right|_{\theta=\theta_0} &= \sum_{j=1}^{m}\Delta_j\left(-\sum_{k=1}^{j-1}h_k^0 X_k + \frac{h_j^0 e^{-h_j^0}}{1-e^{-h_j^0}}X_j\right),
\end{aligned}$$

where $h_s^0 = e^{\gamma_s + X_s^T\beta_s}|_{\theta=\theta_0}, 1 \le s \le m-1$. Hence we have

$$\begin{aligned}
E_{\theta_0}\left\{D\left(\alpha^T\frac{\partial\ell^{(1)}(\theta)}{\partial\theta}\right)^2\right\}_{\theta=\theta_0} &= E_{\theta_0}\left\{\sum_{j=1}^{m}D\Delta_j f_j(X)\right\}^2_{\theta=\theta_0} \\
&= E_{\theta_0}\left\{\sum_{j=1}^{m}D\Delta_j f_j^2(X)\right\}_{\theta=\theta_0} \\
&= \sum_{j=1}^{m}E_{\theta_0}\left\{P(\Delta_j = D = 1|X)f_j^2(X)\right\} = 0 \qquad (1.6)
\end{aligned}$$

for some function f_j. Now by (2.1), (1.5) and Assumption (iv), we obtain

$$P(\Delta_j = D = 1|X) = e^{-\sum_{k=1}^{j-1}h_k^0}\left(1-e^{-h_j^0}\right)P(C \ge t_j|X) > 0, \quad j < m,$$

and

$$P(\Delta_m = D = 1|X) = e^{-\sum_{k=1}^{j-1} h_k^0} \left(1 - e^{-h_j^0}\right) P(C \geq t_{m-1}|X) > 0.$$

Hence (1.6) holds if and only if $f_j(X) = 0$ with probability 1 for all j. Denoting $\alpha = (c_1, \cdots, c_{m-1}, \bar{\alpha}^T)^T$, then we can write

$$\begin{aligned}
\alpha^T \frac{\partial \ell^{(1)}(\theta)}{\partial \theta}\bigg|_{\theta=\theta_0} &= \left\{ \sum_{s=1}^{m-1} c_s \frac{\partial \ell^{(1)}(\theta)}{\partial \gamma_s} + \bar{\alpha}^T \frac{\partial \ell^{(1)}(\theta)}{\partial \beta} \right\}_{\theta=\theta_0} \\
&= \sum_{s=1}^{m-1} c_s \left\{ -\sum_{j=s+1}^{m} \Delta_j + \frac{\Delta_s e^{-h_s^0}}{1 - e^{-h_s^0}} \right\} h_s^0 \\
&\quad + \bar{\alpha}^T \sum_{j=1}^{m} \Delta_j \left\{ -\sum_{k=1}^{j-1} h_k^0 X_k + \frac{h_j^0 e^{-h_j^0}}{1 - e^{-h_j^0}} X_j \right\}.
\end{aligned}$$

Therefore the coefficient of Δ_1 is $f_1(X) = (c_1 + \bar{\alpha}^T X_1) h_1^0 e^{-h_1^0}/(1 - e^{-h_1^0})$. By setting $f_1(X)$ to be 0, we obtain $c_1 + \bar{\alpha}^T X_1 = 0$ with probability 1. Since $\mathrm{Var}(X_1) > 0$, this implies $\bar{\alpha} = 0$ and then it follows that $c_1 = 0$. Now $f_2(X)$ becomes $f_2(X) = c_2 h_2^0 e^{-h_2^0}/(1 - e^{-h_2^0})$, so we have $c_2 = 0$. By continuing this procedure we conclude that $c_3 = \cdots = c_{m-1} = 0$. Therefore, we obtain $\alpha = 0$, which contradicts the assumption of nonzero α. This shows that $I(\theta_0)$ must be nonsingular. Then by Lemma 2 and the consistency of $\hat{\theta}_n$ that we have already shown, we obtain

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = I^{-1}(\theta_0) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_i \left. \frac{\partial \ell_i(\theta)}{\partial \theta} \right|_{\theta=\theta_0} + o_p(1),$$

and asymptotic normality is guaranteed by the central limit theorem since $w_i\{\partial \ell_i(\theta)/\partial \theta\}$ is bounded and thus square integrable. □

Appendix D: Proof of Theorem II.2

Similar to the proof of consistency of $\hat{\theta}_n$ in Theorem II.1, the consistency of $\tilde{\theta}_n$ follows directly from Theorem 5.7 of van der Vaart (1998), in which the random

objective function $M_n(\theta)$ is more general and contains estimated weights in our case. Based on the proof of Theorem II.1, we only need to show

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\{w_i(\hat{p})-w_i(p_0)\}\ell(\theta;X_i)\right|\to 0$$

in probability as $n\to\infty$. This follows easily by the boundness of $\ell(\theta,X_i)$:

$$\begin{aligned}\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\{w_i(\hat{p})-w_i(p_0)\}\ell(\theta;X_i)\right| &\le \frac{C}{n}\sum_{i=1}^{n}|w_i(\hat{p})-w_i(p_0)| \\ &\le C\sum_{s=1}^{S}\left|\frac{1}{\hat{p}_s}-\frac{1}{p_{0s}}\right|\to 0 \qquad (1.7)\end{aligned}$$

in probability as $n\to\infty$ for some constant C.

Since the vector of sample fractions $\hat{p}$ is an asymptotic linear estimator of p_0, together with the asymptotic linearity of $\hat{\theta}_n$ established in Theorem II.1, we have the joint asymptotic normality of $(\hat{\theta}_n,\hat{p})$:

$$\sqrt{n}\begin{pmatrix}\hat{\theta}_n-\theta_0\\ \hat{p}-p_0\end{pmatrix}\to_d N\begin{pmatrix}\Sigma(\theta_0) & G_{12}\\ G_{21} & G_{22}\end{pmatrix}, \qquad (1.8)$$

where $G_{22}=\mathrm{diag}\{p_{01}(1-p_{01})/\alpha_1,\cdots,p_{0S}(1-p_{0S})/\alpha_S\}$ is the asymptotic variance matrix of $\hat{p}$, and $G_{12}=G_{21}^T$ is the asymptotic covariance matrix between $\hat{\theta}_n$ and $\hat{p}$.

Based on the differentiability of the weighted log likelihood function $\sum_{i=1}^{n}w_i(\hat{p})\ell(\theta;X_i)$ to θ and that the maximizer $\tilde{\theta}_n$ does not occur on the boundary of the parameter space, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}w_i(\hat{p})\left.\frac{\partial\ell(\theta;X_i)}{\partial\theta}\right|_{\theta=\tilde{\theta}_n}=0.$$

By the Taylor expansion of the left hand side of the above equation around $\theta=\theta_0$, it follows that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}w_i(\hat{p})\left.\frac{\partial\ell(\theta;X_i)}{\partial\theta}\right|_{\theta=\theta_0}+A_n\sqrt{n}(\tilde{\theta}_n-\theta_0)=0, \qquad (1.9)$$

where

$$A_n = \frac{1}{n}\sum_{i=1}^{n} w_i(\hat{p})\ \frac{\partial^2 \ell(\theta; X_i)}{\partial\theta\partial\theta^T}\bigg|_{\theta=\theta_0} + R_n,$$

and R_n is $o_p(1)$ by the boundness of the third derivative of $\ell(\theta; X_i)$ to θ. Similar to (1.7) by the boundness of the second derivative of $\ell(\theta; X_i)$ to θ, further by the weak law of large numbers and equation (4) we have

$$A_n = \frac{1}{n}\sum_{i=1}^{n} w_i(p_0)\ \frac{\partial^2 \ell(\theta; X_i)}{\partial\theta\partial\theta^T}\bigg|_{\theta=\theta_0} + o_p(1) = I(\theta_0) + o_p(1).$$

By the Taylor expansion again to the first term of the left hand side of (1.9) around p_0 and the boundness of the derivatives of ℓ and w, we can write

$$\begin{aligned} A_n\sqrt{n}(\tilde{\theta}_n - \theta_0) &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_i(p_0)\ \frac{\partial \ell(\theta; X_i)}{\partial\theta}\bigg|_{\theta=\theta_0} \\ &\quad - \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell(\theta; X_i)}{\partial\theta}\left(\frac{\partial w(p)}{\partial p}\right)^T\bigg|_{\theta=\theta_0, p=p_0} \sqrt{n}(\hat{p} - p_0) + o_p(1) \\ &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_i(p_0)\ \frac{\partial \ell(\theta; X_i)}{\partial\theta}\bigg|_{\theta=\theta_0} - B(\theta_0, p_0)\sqrt{n}(\hat{p} - p_0) + o_p(1), \end{aligned}$$

where

$$B(\theta, p) = E\left\{\frac{\partial \ell(\theta; X_i)}{\partial\theta}\left(\frac{\partial w(p)}{\partial p}\right)^T\right\}.$$

Therefore, by the nonsingularity of $I(\theta_0)$ proved in Theorem II.1, we conclude that

$$\begin{aligned} \sqrt{n}(\tilde{\theta}_n - \theta_0) &= -I^{-1}(\theta_0)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_i(p_0)\ \frac{\partial \ell(\theta; X_i)}{\partial\theta}\bigg|_{\theta=\theta_0} \\ &\quad -I^{-1}(\theta_0)B(\theta_0, p_0)\sqrt{n}(\hat{p} - p_0) + o_p(1). \end{aligned} \tag{1.10}$$

In view of (1.8) and (1.10), it now follows from the result in Pierce (1982) that

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) \to_d N(0, \Sigma(\theta_0) - I^{-1}(\theta_0)B(\theta_0, p_0)G_{22}B^T(\theta_0, p_0)I^{-1}(\theta_0),$$

as $n \to \infty$. □

Table 1.1: Summary statistics of simulations, with true parameter values $\beta_1 = 1$ and $\beta_2 = -1$.

Method	Parameter	Bias	Coverage Percentage	Average SD	Empirical SD	Relative efficiency (from empirical variances)
$n = 200$. Mean sample size of completely observed subjects in the case-cohort sample is 90, in which the mean number of censored subjects selected in the subcohort is 50.						
Weighted	β_1	-0.007	0.963	0.440	0.435	0.636
likelihood	β_2	0.044	0.942	0.203	0.211	0.720
Full data	β_1	-0.007	0.968	0.093	0.347	1
MLE	β_2	0.014	0.956	0.173	0.179	1
Naive	β_1	0.172	0.923	0.372	0.362	–
estimator	β_2	-0.080	0.907	0.175	0.177	–
Pseudo	β_1	-0.349	0.813	0.131	0.146	–
likelihood	β_2	0.360	0.722	0.262	0.293	–
Multiple	β_1	0.008	0.970	0.481	0.457	–
imputation	β_2	0.074	0.924	0.223	0.230	–
$n = 500$. Mean sample size of completely observed subjects in the case-cohort sample is 200, in which the mean number of censored subjects selected in the subcohort is 100.						
Weighted	β_1	-0.022	0.942	0.295	0.302	0.580
likelihood	β_2	0.026	0.931	0.133	0.136	0.607
Full data	β_1	-0.026	0.955	0.230	0.230	1
MLE	β_2	0.010	0.954	0.108	0.106	1
Naive	β_1	0.218	0.824	0.233	0.239	–
estimator	β_2	-0.128	0.761	0.108	0.108	–
Pseudo	β_1	-0.261	0.780	0.131	0.146	–
likelihood	β_2	0.249	0.675	0.262	0.293	–
Multiple	β_1	0.030	0.964	0.301	0.287	–
imputation	β_2	0.011	0.959	0.145	0.147	–
$n = 3000$. Mean sample size of completely observed subjects in the case-cohort sample is 400, in which the mean number of censored subjects selected in the subcohort is 250.						
Weighted	β_1	-0.003	0.945	0.208	0.215	0.561
likelihood	β_2	0.016	0.935	0.096	0.106	0.412
Full data	β_1	-0.018	0.948	0.066	0.161	1
MLE	β_2	-0.002	0.940	0.067	0.068	1
Naive	β_1	0.275	0.562	0.156	0.160	–
estimator	β_2	-0.183	0.229	0.067	0.068	–
Pseudo	β_1	-0.090	0.863	0.102	0.118	–
likelihood	β_2	0.099	0.774	0.203	0.234	–
Multiple	β_1	0.028	0.935	0.215	0.227	–
imputation	β_2	0.019	0.920	0.098	0.110	–

Table 1.2: Biases for estimation of the γ_i's in the simulations.

	Weighted likelihood	Full data MLE	Naive estimator	Pseudo likelihood	Multiple imputation
			$n = 200, \quad \gamma_i = -2.41$		
γ_1	-0.13	-0.10	0.45	0.28	0.13
γ_2	-0.07	-0.04	0.56	0.31	0.04
γ_3	-0.02	-0.01	0.68	0.42	0.07
γ_4	-0.04	-0.01	0.77	0.33	0.02
γ_5	-0.06	-0.05	0.85	0.24	-0.03
			$n = 500, \quad \gamma_i = -2.12$		
γ_1	0.01	-0.02	0.57	0.53	0.06
γ_2	-0.01	-0.01	0.64	0.29	0.03
γ_3	-0.02	-0.03	0.72	0.30	0.09
γ_4	-0.00	-0.03	0.85	0.24	0.03
γ_5	-0.02	-0.02	1.04	0.31	0.02
			$n = 3000, \quad \text{true } \gamma_i \equiv -3.51$		
γ_1	-0.01	-0.01	1.55	1.60	0.04
γ_2	-0.01	-0.00	1.63	1.23	0.04
γ_3	-0.00	-0.01	1.76	1.28	0.03
γ_4	-0.00	-0.00	1.93	1.28	0.04
γ_5	-0.01	-0.00	2.11	1.28	0.01

Table 1.3: Comparing the weighted likelihood methods using true weights and estimated weights.

	$\beta_1 = 1$				$\beta_2 = -1$			
	bias	SE1	SE2	coverage	bias	SE1	SE2	coverage
				$n = 200$				
true weights	0.046	0.195	0.212	0.938	-0.033	0.456	0.446	0.959
estimated weights	0.037	0.185	0.181	0.917	-0.019	0.397	0.390	0.959
				$n = 500$				
true weights	0.020	0.129	0.121	0.939	0.001	0.288	0.278	0.940
estimated weights	0.014	0.122	0.117	0.939	0.004	0.255	0.243	0.935
				$n = 3000$				
true weights	0.018	0.095	0.087	0.932	0.013	0.203	0.207	0.948
estimated weights	0.018	0.085	0.080	0.937	0.007	0.158	0.166	0.955

SE1: empirical standard deviation
SE2: average of estimated standard deviations

Table 1.4: Estimated log relative hazards (RHs) of HIV infection in the vaccine trial.

	$(\text{Antibody})^{1/5}$	White	Medium risk score	High risk score
log(RH)	-1.204	-0.191	1.249	1.109
95% CI	(-2.027, -0.342)	(-0.736, 0.354)	(0.728, 1.771)	(0.489, 1.728)
P value	0.009	0.492	<0.001	<0.001

White: 1 for white, 0 for nonwhite
Medium risk group: risk score is equal to 2 or 3
High risk group: risk score is greater than 3

References

Borgan, O., Langholz, B., Samuelsen, S. O., Goldstein, L. and Pogoda, J. (2000). Exposure stratified case-cohort designs. *Lifetime Data Analysis* **6,** 39-58.

Breslow, N. E. and Wellner, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics* , to appear.

Cox, D. R. (1972). Regression models 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.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42,** 845-854.

Flynn, N. M., Forthal, D. N., Harro, C. D., Judson, F. N., Mayer, K. H., Para, M. F., and the rgp 120 HIV Vaccine Study Group (2005). Placebo-controlled phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. *Journal of Infectious Diseases* **191,** 654-665.

Forthal, D. N., Gilbert, P. B., Landucci, G. and Phan, T. (2007). Recombinant gp120 vaccine-induced antibodies inhibit clinical strains of HIV-1 in the presence of Fc receptor-bearing effector cells and correlate inversely with HIV infection rate. *Journal of Immunology* **178,** 6596-6603.

Gilbert, P. B., Peterson, M. L., Follmann, D., Hudgens, M. G., Francis, D. P., Gurwith, M., Heyward, W. L., Jobes, D. V. , Popovic, V., Self, S. G., Sinangil, F., Burke, D. and Berman, P. W. (2005). Correlation between immunologic responses to a recombinant glycoprotein 120 Vaccine and incidence of HIV-1 infection in a phase 3 HIV-1 preventive vaccine trial. *Journal of Infectious Diseases* **191,** 666-677.

Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal likelihoods based on Cox's

regression and life model. *Biometrika* **60**, 267-278.

Kulich, M. and Lin, D. Y. (2004). Improving efficiency of relative-risk estimation in case-cohort studies. *Journal of the American Statistical Association* **99,** 832-844.

Little, R. J. A. and Rubin, D. B. (2000). *Statistical Analysis with Missing Data*. John Wiley, New York.

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

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73,** 1-11.

Prentice, R. L. and Gloeckler, L. A. (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics* **34,** 57-67.

Robins, J. M., Rotnitzky, A. and Zhao L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89,** 846-866.

Self, S. G. and Prentice, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Annals of Statistics* **16,** 64-81.

CHAPTER II

Missing Covariates in Cox Regression with Current Status Data

2.1 Introduction

The Cox proportional hazards model (Cox, 1972) with right censored data has been extensively studied in the literature (see, for example, Andersen et al., 1993, Fleming and Harrington, 1991, or Kalbfleisch and Prentice, 2002). Besides right censoring, other types of censoring also arise in practice. For example, failure time data may be interval censored, i.e., we only know that the failure time for a subject falls into some random interval. In the so-called "case 1" interval censoring, we only know whether the failure event has occurred prior to a random observation time Y or not, and hence the observed data are (Δ, Y), where $\Delta = I_{(T \leq Y)}$ and T is the time to event. This type of data are also called current status data. Groeneboom and Wellner (1992) studied the nonparametric maximum likelihood estimator (NPMLE) of the distribution function of T with current status data. They developed an efficient iterative algorithm to compute the nonparametric likelihood estimate of the underlying distribution function of T based on n independent and identically distributed (i.i.d.) observations of (Δ, Y), which converges faster than the EM algorithm. They proved the consistency of the NPMLE and showed its $n^{1/3}$ rate of convergence and derived its (non-normal) asymptotic distribution. In the regression setting, instead

of the marginal distribution of T, we are interested in the association between the failure time T and a vector of covariates Z. Therefore, the observed data consist of i.i.d. copies of (Δ, Y, Z), where Z is a d-dimensional covariate vector. Huang (1996) studied the maximum likelihood estimation for both the baseline cumulative hazard function and the log hazard ratio parameter in a proportional hazards model with current status data, and developed an algorithm based on Groeneboom and Wellner (1992) for computing the maximum likelihood estimates. Other work on this topic includes Murphy and Van Der Vaart (2000) and Van Der Vaart (2002), and the former treated this problem as an example of their profile likelihood approach. The maximum likelihood estimators of the finite dimensional vector of log hazard ratios and the infinite dimensional baseline cumulate hazard function are proved to be consistent and asymptotically efficient, and the former converges to a normal random variable at $n^{1/2}$ rate while the latter converges at $n^{1/3}$ rate, the same rate as in Groeneboom and Wellner (1992).

In this chapter we consider the problem of fitting the proportional hazards model with current status data when covariates are not always observed. We assume that covariates are missing at random (MAR, a terminology of Little and Rubin Little and Rubin (2002)), and the probability of observing those covariates is known or can be reasonably estimated. This kind of missing data problem can arise in many situations. One example is two-phase stratified sampling, which was originally proposed by Neyman (1938). It was proposed to estimate the population mean of a target variable that is costly or hard to measure. At phase one an auxiliary variable, which is correlated to the target variable and easy to measure, is measured on a relatively large sample. Then at phase two, the target variable is measured in a random subsample drawn by random sampling stratified by the variable measured at phase one.

The case-cohort design, proposed by Prentice (1986), is a special type of two-phase stratified sampling design, in which the strata are defined by the outcome (failure status) and possibly other auxiliary variables that are measured for every subject.

For statistical inference in such type of missing data problems, the usual likelihood approach can be very difficult if not impossible. The weighted likelihood method, however, can be easily applied, in which one maximizes the inverse probability weighted version of the log likelihood function (see e.g. Kalbfleisch and Lawless, 1988 and Skinner et al., 1989), or solves a weighted version of the score equation (see e.g. Manski and Lerman, 1977) to estimate the parameter of interest. When the weighted likelihood approach is applied to parametric models, the asymptotic properties of the regular estimators with $n^{1/2}$ rate follow readily from the results for M-estimation (see e.g. Van Der Vaart, 1998). For example, Li et al. (2008) employed the weighted likelihood method in a proportional hazards model for grouped survival data coming from an HIV vaccine study, where the covariates of interest are obtained from a case-cohort design. In a recent work on semiparametric models for two-phase sampling designs in which the infinite dimensional nuisance parameter can be estimated at $n^{1/2}$ rate, Breslow and Wellner (2007) considered the weighted likelihood method and derived asymptotic results for both Bernoulli sampling and finite population stratified sampling in selecting the phase two sample. They also derived asymptotic results for the weighted likelihood estimator using estimated sampling probabilities in the weight function for the Bernoulli sampling, showing the efficiency gain comparing to the estimator using true sampling probabilities.

When we fit the proportional hazards model to current status data with missing covariates using the weighted likelihood, however, we may expect a slower than $n^{1/2}$ convergence rate for the baseline cumulative hazard function based on the result of

Huang (1996) for the full data. Hence the theory developed by Breslow and Wellner (2007) does not apply. In Section 2.7, we construct a general theorem that generalizes Theorem 6.1 in Wellner and Zhang (2007), which was developed for their pseudo likelihood method, and apply the theorem to show that our proposed estimators for the log hazard ratio are $n^{1/2}$-consistent and asymptotically normal, and that using estimated weights improves efficiency as shown for the case in Breslow and Wellner (2007).

The construction of this chapter is as follows. In Section 2.2 we provide an algorithm that is modified from the one given in Huang (1996) for computing the weighted likelihood estimator. In Section 2.3, we first establish the asymptotic properties of the weighted likelihood estimators using true weights, then show that similar asymptotic results hold for the weighted likelihood estimator using estimated weights and that such an estimator is more efficient than the one obtained by using true weights. We discuss variance estimation in Section 2.4, and conduct simulations and analyze the data from a case-cohort HIV vaccine study in Section 2.5. A brief discussion is given in Section 2.6. In Section 2.7, we introduce a general theorem that is useful for the proof of asymptotic properties of the weighted likelihood estimator using estimated weights. All the major proofs are given in Section 2.8. We adopt the empirical process notation of Van Der Vaart and Wellner (1996) throughout the article by denoting Pf as the integral of f with respect to the probability measure P, $\mathbb{P}_n f$ as the integral of f with respect to the empirical measure $\mathbb{P}_n$, which is the sample average of f for i.i.d. data, and $\mathbb{G}_n f = n^{1/2}(\mathbb{P}_n - P)f$.

2.2 The Weighted Likelihood Estimator

Suppose failure time T and observation time Y are independent given covariate Z, and T follows a proportional hazards model:

$$\Lambda(t|Z) = \Lambda(t)e^{\beta^T Z},$$

where $\Lambda(t|Z)$ is the conditional cumulative hazard function given Z, and $\Lambda(t)$ is the baseline cumulative hazard function. We consider the case that covariate Z can be missing. The probability of observing Z is denoted as $\pi_\alpha(\Delta, V)$, which may depend on a parameter α, the failure status Δ, and an auxiliary variable V that is observed for everyone. For example, in a case-cohort design with stratified sampling of the subcohort, the probability of observing covariate Z is $\pi_\alpha(\Delta, V) = \Delta + (1 - \Delta)\sum_{j=1}^{J} p_j I_{(V\in\mathcal{V}_j)}$, where $\mathcal{V}_1, \cdots, \mathcal{V}_J$ are J strata determined by the value of the auxiliary variable V, $\alpha = (p_1, \cdots, p_J)^T$, and p_j is the probability that a subject is sampled into the subcohort from stratum $j, 1 \le j \le J$. The parameter α may or may not be known. In Section 2.4 we shall discuss the effect of estimating α from observed data. It is possible that V is part of Z. The density of a single observation $X \equiv (\Delta, Y, Z, V)$ at $x \equiv (\delta, y, z, v)$ can be written as

$$p_{\beta,\Lambda}(x) = \left(1 - e^{-\Lambda(y)e^{\beta^T z}}\right)^{\delta} \left(e^{-\Lambda(y)e^{\beta^T z}}\right)^{1-\delta} f(y, z, v), \tag{2.1}$$

where $f(y, z, v)$ is the joint density of (Y, Z, V). The parameter of interest is the log hazard ratio β, and $\Lambda(\cdot)$ is a nuisance parameter.

Let $X_1, \cdots, X_n$ be n i.i.d. copies of X. The log likelihood function, up to an additive constant, is

$$l(\beta, \Lambda) = \sum_{i=1}^{n} \Delta_i \log\left(1 - e^{-\Lambda(Y_i)e^{\beta^T Z_i}}\right) - (1 - \Delta_i)\Lambda(Y_i)e^{\beta^T Z_i}.$$

Because Z_i's are only observed for a subsample, the NPMLE can be too complicated to be useful. However, we can use the following weighted version of the log likelihood function

$$l_n^w(\beta,\Lambda) = \sum_{i=1}^{n} w_i \left\{ \Delta_i \log\left(1 - e^{-\Lambda(Y_i)e^{\beta^T Z_i}}\right) - (1-\Delta_i)\Lambda(Y_i)e^{\beta^T Z_i} \right\}, \tag{2.2}$$

where $w_i = \xi_i/\pi_\alpha(\Delta_i, V_i)$ with $\xi_i = 1$ if Z_i is observed and 0 otherwise, $1 \le i \le n$. For simplicity, here and in the sequel we suppress the dependence of w on α, Δ and V except in Subsections 2.3.2 and 2.4.2, where we discuss the weighted likelihood estimator with estimated weights. The weighted likelihood estimator of the true parameter $(\beta_0, \Lambda_0(\cdot))$ is defined as the maximizer of the weighted log likelihood function (2.2) with discretized Λ at observed time points and denoted by

$$(\hat\beta_n, \hat\Lambda_n) = \text{argmax}\mathbb{P}_n wl(\beta,\Lambda;X).$$

Let $(Y_{(1)},\cdots,Y_{(n)})$ be the order statistics of $(Y_1,\cdots,Y_n)$. Let $\Delta_{(i)}, Z_{(i)}$ and $w_{(i)}$ be the values of Δ, Z and w associated with $Y_{(i)}$, $1 \le i \le n$. Following Huang (1996) we assume that $\Delta_{(1)} = 1$, $\Delta_{(n)} = 0$, and the estimator $\hat\Lambda_n(\cdot)$ is a right-continuous step function on $[0, Y_{(n)}]$ with jumps at $Y_{(i)}$'s and $\hat\Lambda_n(0) = 0$. Replacing Λ by its estimator $\hat\Lambda_n$, we obtain the following score equation for β by differentiating the objective function (2.2) with respect to β and setting it to 0:

$$\sum_{i=1}^{n} w_{(i)} \left\{ \Delta_{(i)} \frac{e^{-\hat\Lambda_n(Y_{(i)})e^{\hat\beta_n^T Z_{(i)}}}}{1 - e^{-\hat\Lambda_n(Y_{(i)})e^{\hat\beta_n^T Z_{(i)}}}} - (1-\Delta_{(i)}) \right\} \hat\Lambda_n(Y_{(i)}) e^{\hat\beta_n^T Z_{(i)}} Z_i = 0. \tag{2.3}$$

Due to the monotonicity constraint on $\hat\Lambda_n$, there is no such a simple score equation for $\hat\Lambda_n$. However, analogous to Groeneboom and Wellner (1992) and Huang (1996), $\hat\Lambda_n$ can be characterized by a set of inequalities and an equality, which are given in

the following Theorem II.1. In the development in this section, we assume that there are $k_n(\leq n)$ distinct observation times $Y_1^* < Y_2^* < \cdots < Y_{k_n}^*$. Both Theorems II.1 and II.2 below allow multiple observations at the same time point. Although Y is assumed to be a continuous random variable, in practice two or more observation times may be tied due to rounding errors.

Theorem II.1. *The weighted likelihood estimator* $(\hat{\beta}_n, \hat{\Lambda}_n)$ *satisfies the score equation (2.3) and*

$$\sum_{Y_j \geq Y_i^*} w_j e^{\hat{\beta}_n^T Z_j} \left\{ \Delta_j \frac{e^{-\hat{\Lambda}_n(Y_j) e^{\hat{\beta}_n^T Z_j}}}{1 - e^{-\hat{\Lambda}(Y_j) e^{\hat{\beta}_n^T Z_j}}} - (1 - \Delta_j) \right\} \leq 0, \tag{2.4}$$

for $i = 1, 2, \cdots, k_n$, *and*

$$\sum_{i=1}^{n} w_i e^{\hat{\beta}_n^T Z_i} \left\{ \Delta_i \frac{e^{-\hat{\Lambda}_n(Y_i) e^{\hat{\beta}_n^T Z_i}}}{1 - e^{-\hat{\Lambda}_n(Y_i) e^{\hat{\beta}_n^T Z_i}}} - (1 - \Delta_i) \right\} = 0. \tag{2.5}$$

Remark 2.1. By the pool adjacent violators algorithm, see Robertson et al. (1988), for example, the estimates $\hat{\Lambda}_n(Y_{(1)}), \cdots, \hat{\Lambda}_n(Y_{(n)})$ form a number of constant blocks (called level blocks), say, $\hat{\Lambda}_n(Y_{(1)}) = \cdots = \hat{\Lambda}_n(Y_{(b_1)}) < \hat{\Lambda}_n(Y_{(b_1+1)}) = \cdots = \hat{\Lambda}_n(Y_{(b_2)}) < \cdots < \hat{\Lambda}_n(Y_{(b_m+1)}) = \cdots = \hat{\Lambda}_n(Y_{(n)})$, for some $1 \leq b_1 < b_2 < \cdots < b_m \leq n-1$. By maximizing the sum of weighted log likelihood functions in each level block, we get a stronger result than (2.5):

$$\sum_{j=b_i+1}^{b_{i+1}} w_{(j)} e^{\hat{\beta}_n^T Z_{(j)}} \left\{ \Delta_{(j)} \frac{e^{-\hat{\Lambda}_n(Y_{(j)}) e^{\hat{\beta}_n^T Z_{(j)}}}}{1 - e^{-\hat{\Lambda}_n(Y_{(j)}) e^{\hat{\beta}_n^T Z_{(j)}}}} - (1 - \Delta_{(j)}) \right\} = 0, \tag{2.6}$$

for each $i = 0, 1, \cdots, m$, where $b_0 = 0$ and $b_{m+1} = n$. When there is no missing covariates, a result parallel to (2.6) is implicit in Huang (1996) for the MLE of (β, Λ) where no w_i is involved.

Our next result yields an iterative algorithm to compute $\hat{\Lambda}_n(\cdot, \beta)$ for any fixed β. This algorithm is more efficient than the pool adjacent violators algorithm that can

also be applied to calculate $\hat{\Lambda}_n(\cdot, \beta)$ for a fixed β. Following Huang (1996) we define

$$
\begin{aligned}
W_\Lambda(Y_i^*) &= \sum_{Y_j \le Y_i^*} w_j e^{\beta^T Z_j} \left\{ \Delta_j \frac{e^{-\Lambda(Y_j)e^{\beta^T Z_j}}}{1 - e^{-\Lambda(Y_j)e^{\beta^T Z_j}}} - (1 - \Delta_j) \right\}, \\
G_\Lambda(Y_i^*) &= \sum_{j=1}^{i} \Delta G_\Lambda(Y_j^*) \quad \text{with} \\
\Delta G_\Lambda(Y_j^*) &= \sum_{Y_k = Y_j^*} w_k e^{\beta^T Z_k} \left\{ \Delta_k \frac{e^{\beta^T Z_k} e^{-\Lambda(Y_k)e^{\beta^T Z_k}}}{\left(1 - e^{-\Lambda(Y_k)e^{\beta^T Z_k}}\right)^2} + \frac{1 - \Delta_k}{\Lambda(Y_k)} \right\}, \qquad (2.7) \\
V_\Lambda(Y_i^*) &= W_\Lambda(Y_i^*) + \sum_{Y_j^* \le Y_i^*} \Lambda(Y_j^*) \Delta G_\Lambda(Y_j^*),
\end{aligned}
$$

where we have added the quantity $w_k e^{\beta^T Z_k}(1-\Delta_k)/\Lambda(Y_k)$ in the original definition of $\Delta G_\Lambda(\cdot)$ on page 545 of Huang (1996) to make $\Delta G_\Lambda(Y_j^*) \equiv G_\Lambda(Y_j^*) - G_\Lambda(Y_{j-1}^*) > 0$ with $G_\Lambda(Y_0^*) \equiv 0$, $1 \le j \le k_n$, a required condition for the algorithm. See the following Remark 2.2 for an explanation.

Theorem II.2. *For any fixed β, $\hat{\Lambda}_n(\cdot; \beta)$ maximizes $l_n^w(\beta, \Lambda)$ if and only if $\hat{\Lambda}_n(\cdot; \beta)$ is the left derivative of the greatest convex minorant of the "self-induced" cumulative sum diagram defined by the points $(0, 0)$ and*

$$
\left(G_{\hat{\Lambda}_n(\cdot,\beta)}(Y_i^*), V_{\hat{\Lambda}_n(\cdot,\beta)}(Y_i^*) \right), \quad 1 \le i \le k_n.
$$

REMARK 2.2. The function $G_\Lambda(\cdot)$ in Theorem II.2 can be chosen arbitrarily, as long as $\Delta G_\Lambda(Y_i^*) > 0$, $1 \le i \le k_n$, and the constructed $V_\Lambda(\cdot)$ is nondecreasing. The point is clearly seen in the proof of Proposition 1.4 and Remark 1.4 of Groeneboom and Wellner (1992). The choices in both Groeneboom and Wellner (1992) and Huang (1996) are based on a second order Taylor expansion of the log likelihood function. It works well for the nonparametric estimation of the marginal distribution function of T, but numerical issue arises in the semiparametric regression case since such chosen $G_\Lambda(\cdot)$ (determined by the first of two terms in the summands in (2.7)) has

zero increments at all observation times for censored subjects. This problem can be resolved by adding a positive quantity to the increments of $G_\Lambda(\cdot)$ at those time points as what we have done in (2.7). Such added quantity also makes $V_\Lambda(\cdot)$ nondecreasing.

Since for a fixed β there is no closed form for $\Lambda(\cdot, \beta)$, we follow Huang (1996) to iterate between β and Λ to find the weighted likelihood estimator $(\hat{\beta}_n, \hat{\Lambda}_n)$. For a fixed β, Λ is obtained iteratively by the algorithm given in Theorem II.2. Then for the updated Λ, β is updated by solving score equation (2.3) using the Newton-Raphson method. This procedure is repeated until convergence. Simulation studies show that the algorithm converges very quickly.

2.3 Asymptotic Properties

2.3.1 The Weighted Likelihood Estimator with True Weights

Asymptotic properties of the estimator are based on the following assumptions.

(A) The parameter space for β, $\mathcal{B} \subset R^d$, is compact, and the true parameter β_0 is an interior point of $\mathcal{B}$.

(B) The observation time Y possesses a Lebesgue density that is continuous and positive on an interval $[\sigma, \tau]$ with $\sigma > 0$ and vanishes outside this interval, and the joint distribution $F(y, z)$ of (Y, Z) has bounded second order partial derivative with respect to y.

(C) The cumulative hazard function Λ satisfies $1/M \leq \Lambda \leq M$ on $[\sigma, \tau]$ for some positive constant M. The true parameter Λ_0 satisfies $0 < \Lambda_0(\sigma-) < \Lambda_0(\tau) < M$ and is continuously differentiable with positive derivative on $[\sigma, \tau]$.

(D) The covariate vector Z is bounded and $E[\mathrm{var}(Z|Y)] > 0$.

(E) There exists a constant ϵ such that $\pi_\alpha(\Delta, V) \geq \epsilon > 0$ for all α in a neighborhood of the true parameter α_0.

Denote the parameter space for Λ defined in (C) by Φ and the parameter space for (β, Λ) by Θ. The above Assumptions (A) to (D) are basically the same as those in Huang (1996) and Van Der Vaart (2002) for the full data NPMLE for the Cox model with current status data. They are imposed mainly for technical reasons, but also make practical sense. For instance, τ can be viewed as the time of the end of study. Assumption (D) ensures the identifiability of β as well as the positive definiteness of the efficient information matrix for β (see proofs of Theorems II.3 and II.6). Assumption (E) is a common assumption for missing data problems.

Let $|\cdot|$ be the Euclidian norm, and $\|\Lambda\|_2 = \{\int \Lambda^2(y) dQ_Y(y)\}^{1/2}$ for every $\Lambda \in \Phi$, where $Q_Y(y)$ is the probability measure of the censoring variable Y. Define the distance in $R^d \times \Phi$ as $d((\beta_1, \Lambda_1), (\beta_2, \Lambda_2)) = |\beta_1 - \beta_2| + \|\Lambda_1 - \Lambda_2\|_2$. The consistency of the weighted likelihood estimator $(\hat{\beta}_n, \hat{\Lambda}_n)$ can be proved by applying Theorem 5.8 and Lemma 5.9 in Van Der Vaart (2002), in which Theorem 5.8 gives a sufficient condition for the consistency of an M-estimator of a parameter in a general metric space, while Lemma 5.9 provides a way of checking the conditions in Theorem 5.8. We drop the word "outer" and its corresponding notation for the outer measure throughout the article, which we believe would not cause any confusion.

Theorem II.3. *Under Assumptions (A) to (E), we have* $\hat{\beta}_n \to_p \beta_0$ *and* $\hat{\Lambda}_n(t) \to_p \Lambda_0(t)$ *for every* $t \in (\sigma, \tau)$*, as* $n \to \infty$*.*

In fact, the above convergence also holds almost surely. Convergence in probability suffices for our purpose. To derive the rate of convergence of $(\hat{\beta}_n, \hat{\Lambda}_n)$, we need to calculate the bracketing entropy number of the class of functions $\{m(\beta, \Lambda; X) : (\beta, \Lambda) \in \Theta\}$, where $m(\beta, \Lambda; X) = w\ell(\beta, \Lambda; X)$ and $\ell(\beta, \Lambda; X) = \log\{(p_{\beta,\Lambda} + p_{\beta_0,\Lambda_0})/2\}$. The function ℓ was introduced by Van Der Vaart (2002) for technical convenience (see the proof of Theorem II.3). For a probability measure P and a class of functions

$\mathcal{F}$ in $L_2(P)$, denote the ε-bracketing number of $\mathcal{F}$ by $N_{[]}(\varepsilon, \mathcal{F}, L_2(P))$. The logarithm of the bracketing number is called the entropy number of $\mathcal{F}$. The entropy number of $\{m(\beta, \Lambda; X) : (\beta, \Lambda) \in \Theta\}$ turns out to be of the same order as that of $\{\ell(\beta, \Lambda; X) : (\beta, \Lambda) \in \Theta\}$ (see the proof of Theorem II.4), and the latter is $O(1/\varepsilon)$ by Lemma 8.6 of Van Der Vaart (2002).

Theorem II.4. *Under Assumptions (A) to (E), there exists a constant C such that for every $\varepsilon > 0$,*

$$\log N_{[]}(\varepsilon, \{m(\beta, \Lambda; X) : (\beta, \Lambda) \in \Theta\}, L_2(P)) \leq C/\varepsilon.$$

Based on the above results, the rate of convergence of $(\hat{\beta}_n, \hat{\Lambda}_n)$ can be derived by Theorem 3.2.5 in Van Der Vaart and Wellner (1996), which states that the rate of convergence is determined by the smoothness of the model and the modulus of continuity of the objective function. The following theorem shows exactly the same result as the full data case, see Theorem 3.3 of Huang (1996), also see Lemma 8.5 of Van Der Vaart (2002).

Theorem II.5. *Under Assumptions (A) to (E), we have $|\hat{\beta}_n - \beta_0| + \|\hat{\Lambda}_n - \Lambda_0\|_2 = O_p(n^{-1/3})$.*

Similar to the full data case, the overall rate of convergence is dominated by $\hat{\Lambda}_n$ that has a cubic root-n rate, while the rate of convergence for $\hat{\beta}_n$ is root-n that will be given in the following Theorem II.6. To prove the asymptotic normality of $\hat{\beta}_n$, we can apply Theorem 6.1 of Wellner and Zhang (2007), which generalizes Theorem 6.1 of Huang (1996). The theorem provides a set of sufficient conditions for the asymptotic normality of the M-estimator of the finite dimensional parameter in a semiparametric model. We further generalize it in Section 2.7 to the case where an additional parameter in the objective function is estimated a priori, which takes

care of the weighted likelihood estimation using estimated weights. If the additional parameter is known, then our general theorem, Theorem II.10 in Section 2.7, reduces to Theorem 6.1 of Wellner and Zhang (2007). For ease of reference, we apply II.10 to show the asymptotic normality of $\hat{\beta}_n$ in Section 2.8. Now we introduce some necessary notation.

Suppose that Λ_η is a parametric submodel in Φ passing through Λ at $\eta = 0$. Let $H = \{h : h = \partial\Lambda_\eta/\partial\eta|_{\eta=0}\}$ be the collection of all directions to approach Λ. The efficient score function for β is defined as the projection of the ordinary score function for β onto the orthogonal complement of the tangent space for Λ, which is the closure of the linear space spanned by H (see Bickel et al., 1993 for detailed discussions). When there is no missing data, let $l_1(\beta, \Lambda; X)$ be the score function for β, i.e., $l_1(\beta, \Lambda; X) = \partial l(\beta, \Lambda; X)/\partial\beta = e^{\beta^T Z}\Lambda(Y)Q(X)Z$, where

$$Q(X) = \Delta\frac{e^{-e^{\beta^T Z}\Lambda(Y)}}{1 - e^{-e^{\beta^T Z}\Lambda(Y)}} - (1 - \Delta).$$

Denote the score for Λ as $l_2(\beta, \Lambda; X)[h] = \partial l(\beta, \Lambda_\eta; X)/\partial\eta = e^{\beta^T Z}Q(X)h(Y)$ for every $h \in H$, and $l_2(\beta, \Lambda; X)[\mathbf{h}] = (l_2(\beta, \Lambda; X)[h_1], \cdots, l_2(\beta, \Lambda; X)[h_d])^T$ for any $\mathbf{h} = (h_1, \cdots, h_d)^T$, where $h_k \in H, 1 \le k \le d$. Then by Huang, (1996) (or Murphy and Van Der Vaart, 2000) and Van Der Vaart (2002) for a more precise argument on the calculation based on an approximated least favorable submodel) for the full data, the efficient score function for β is

$$\begin{aligned} \tilde{l}(\beta, \Lambda; X) &= l_1(\beta, \Lambda; X) - l_2(\beta, \Lambda; X)[\mathbf{h}^*] \\ &= e^{\beta^T Z}\Lambda(Y)Q(X)\left[Z - \frac{E\left(Ze^{2\beta^T Z}r(Y, Z; \beta, \Lambda)|Y\right)}{E\left(e^{2\beta^T Z}r(Y, Z; \beta, \Lambda)|Y\right)}\right], \end{aligned} \tag{2.8}$$

where

$$r(Y, Z; \beta, \Lambda) = \frac{e^{-\Lambda(Y)e^{\beta^T Z}}}{1 - e^{-\Lambda(Y)e^{\beta^T Z}}},$$

and

$$\mathbf{h}^*(y) = \Lambda_0(y)\frac{E(Ze^{2\beta_0^T Z}Q^2(X)|Y=y)}{E(e^{2\beta_0^T Z}Q^2(X)|Y=y)}. \tag{2.9}$$

The information matrix for β is then given by

$$\begin{aligned} I(\beta) &= E\{\tilde{l}(\beta,\Lambda;X)\}^{\otimes 2} \\ &= E\left\{e^{2\beta^T Z}\Lambda^2(Y)Q^2(X)\left[Z-\frac{E\left(Ze^{2\beta^T Z}r(Y,Z;\beta,\Lambda)|Y\right)}{E\left(e^{2\beta^T Z}r(Y,Z;\beta,\Lambda)|Y\right)}\right]^{\otimes 2}\right\}. \end{aligned} \tag{2.10}$$

We then have the following theorem of asymptotic normality for the weighted likelihood estimator $\hat{\beta}_n$ obtained by using true weights. We can see that the asymptotic variance matrix is the full data asymptotic variance plus an additional nonnegative definite matrix that reflects the loss of efficiency due to missing data.

Theorem II.6. *Under Assumptions (A) to (E) and that $\alpha = \alpha_0$, we have*

$$n^{1/2}(\hat{\beta}_n - \beta_0) = I^{-1}(\beta_0)n^{1/2}\,\mathbb{P}_n w\tilde{l}(\beta_0,\Lambda_0;X) + o_p(1) \to_d N(0,\Sigma),$$

as $n \to \infty$, where $\Sigma = I^{-1}(\beta_0) + I^{-1}(\beta_0)DI^{-1}(\beta_0)$, and

$$D = E\left[\frac{1-\pi_\alpha(\Delta,V)}{\pi_\alpha(\Delta,V)}\left\{\tilde{l}(\beta_0,\Lambda_0;X)\right\}^{\otimes 2}\right].$$

2.3.2 The Weighted Likelihood Estimator with Estimated Weights

Here we denote the weight by $w(\alpha)$, where $\alpha = (\alpha_1,\cdots,\alpha_J)^T$ with true value $\alpha_0 = (\alpha_{01},\cdots,\alpha_{0J})^T$. No matter α_0 is known or not, it may be replaced by a good estimator $\hat{\alpha}_n = (\hat{\alpha}_{n1},\cdots,\hat{\alpha}_{nJ})^T$, then the estimated weight $w(\hat{\alpha}_n)$ is used in the weighted likelihood function. Let

$$(\tilde{\beta}_n,\tilde{\Lambda}_n) = \operatorname{argmax}\mathbb{P}_n w(\hat{\alpha}_n)l(\beta,\Lambda;X)$$

be the weighted likelihood estimator of (β_0,Λ_0) obtained by using estimated weights. When the nuisance parameter can be estimated at root-n rate, the efficiency gain

of the estimator $\tilde{\beta}_n$ comparing to $\hat{\beta}_n$ that is obtained using true weights has been discussed by many authors, see e.g. Pierce (1982), Robinson et al. (1994), Breslow and Wellner (2007), and Li et al (2008), among many others. It turns out that for the current setting in which the nuisance parameter can only be estimated at a slower than root-n rate, such an efficiency gain for the estimation of the parameter of interest also holds under mild conditions. The detail follows.

Theorem II.7. *Suppose $\hat{\alpha}_n \to_p \alpha_0$ and $w(\alpha)$ is differentiable with uniformly bounded first order derivative $\dot{w}(\alpha)$ in a neighborhood of α_0. Then under Assumptions (A) to (E), we have $\tilde{\beta}_n \to_p \beta_0$ and $\tilde{\Lambda}_n(t) \to_p \Lambda_0(t)$ for every $t \in (\sigma, \tau)$.*

Theorem II.8. *Suppose $En^{1/2}|\hat{\alpha}_n - \alpha_0|$ is bounded, and $w(\alpha)$ is twice differentiable with uniformly bounded first and second order derivatives $\dot{w}(\alpha)$ and $\ddot{w}(\alpha)$ in a neighborhood of α_0. Then under Assumptions (A) to (E), we have $|\tilde{\beta}_n - \beta_0| + \|\tilde{\Lambda}_n - \Lambda_0\|_2 = O_p(n^{-1/3})$.*

Remark 3.1. The uniform boundedness of $\dot{w}(\alpha)$ and $\ddot{w}(\alpha)$ is not too restrictive. For example, for a case-cohort design with a stratified Bernoulli sampled subcohort, we have $\pi_\alpha(\Delta, V) = \Delta + (1-\Delta)\sum_{j=1}^{J} p_j I_{(V \in \mathcal{V}_j)}$, and the above conditions are satisfied as long as all the stratified selection probabilities p_j's are bounded away from 0. The same is true for a two-phase design in which the second stage sample is selected by a stratified Bernoulli sampling. More generally, if $\pi_\alpha(\Delta, V)$ follows a logistic model, say, $\operatorname{logit} \pi_\alpha(\Delta, V) = \alpha_0 + \alpha_1^T V + \alpha_2 \Delta$, then the conditions are still satisfied given that V is bounded. The boundedness of $En^{1/2}|\hat{\alpha}_n - \alpha_0|$ is a little more restrictive. The asymptotic normality of $n^{1/2}(\hat{\alpha}_n - \alpha_0)$ is neither sufficient nor necessary for this to hold, while the condition that $En^{1/2}|\hat{\alpha}_n - \alpha_0|$ converges to a finite limit as $n \to \infty$ is stronger than necessary. Nevertheless, in the two most important cases: case-cohort

sampling and two-phase stratified sampling, $\hat{p}_j$ is the proportion of subjects selected from stratum j, $1 \le j \le J$. Then it is easy to show that $En^{1/2}|\hat{p}_j - p_{0j}|$ converges to a finite limit as $n \to \infty$, and hence the sequence is bounded.

The following theorem shows the asymptotic normality of $\tilde{\beta}_n$ as well as the efficiency gain of $\tilde{\beta}_n$ comparing to $\hat{\beta}_n$, which can be proved by applying II.10 that will be introduced in Section 2.7.

Theorem II.9. *Under the same conditions in Theorem II.8, we have*

$$n^{1/2}(\tilde{\beta}_n - \beta_0) = I^{-1}(\beta_0)n^{1/2}\,\mathbb{P}_n w\tilde{l}(\beta_0, \Lambda_0; X) - Cn^{1/2}(\hat{\alpha}_n - \alpha_0) + o_p(1)$$

as $n \to \infty$, where $C = I^{-1}(\beta_0)P\{\tilde{l}(\beta_0, \Lambda_0; X)\dot{w}^T(\alpha_0)\}$. Furthermore, if $\hat{\alpha}_n$ is asymptotically efficient with influence function ℓ^α, then

$$\begin{aligned} n^{1/2}(\tilde{\beta}_n - \beta_0) &= I^{-1}(\beta_0)n^{1/2}\,\mathbb{P}_n w\tilde{l}(\beta_0, \Lambda_0; X) - Cn^{1/2}\,\mathbb{P}_n\ell^\alpha + o_p(1) \\ &\to_d N(0, \Sigma - C\Sigma_\alpha C^T), \end{aligned}$$

where Σ was defined in Theorem II.6 and $\Sigma_\alpha = E(\ell^{\alpha\otimes 2})$.

2.4 Variance Estimation

2.4.1 Using True Weights

When α_0 is given and $w(\alpha_0)$ is used in the estimation of β, the asymptotic variance given in Theorem II.6 can be used to obtain the variance estimator of the weighted likelihood estimator $\hat{\beta}_n$. However, as discussed in Huang (1996), smoothing is inevitable in such calculation.

Without smoothing, the weighted bootstrap with i.i.d. weights, also called the "wild boostrap" (see e.g. Van Der Vaart and Wellner (1996)), turns out to be an effective and robust approach in variance estimation for the weighted likelihood estimator with true weights. See Ma and Kosorok (2005) for a detailed argument of

using the weighted bootstrap method to the general M-estimation in a semiparametric model.

Suppose that $u_1, \cdots, u_n$ are n i.i.d. nonnegative and bounded random weights, independent of $X_1, \cdots, X_n$ and $w_1, \cdots, w_n$, and satisfying $E(u_i) = 1$ and $\text{var}(u_i) = \delta_0 < \infty$ for a constant δ_0. Denote the estimator of β obtained by maximizing the objective function $\mathbb{P}_n uwl(\beta, \Lambda; X)$ by $\hat{\beta}_n^*$. Randomly generate $(u_1, \cdots, u_n)$ repeatedly, say, B times, and obtain corresponding $\hat{\beta}_n^*$ that are denoted by $\hat{\beta}_{n1}^*, \cdots$, and $\hat{\beta}_{nB}^*$. A variance estimator of $\hat{\beta}_n$ is then obtained from the empirical variance of $\hat{\beta}_{n1}^*, \cdots, \hat{\beta}_{nB}^*$ rescaled by δ_0. Analogous to the case in Ma and Kosorok (2005), this weighted bootstrap estimation of variance can be justified in the following way.

Since u is bounded with mean 1 and independent of the X_i's and w_i's, we have $E\{uwl(\beta, \Lambda; X)\} = E\{wl(\beta, \Lambda; X)\}$. By Theorem II.6 we have

$$n^{1/2}(\hat{\beta}_n^* - \beta_0) = I^{-1}(\beta_0) n^{1/2} \mathbb{P}_n^* w\tilde{l}(\beta_0, \Lambda_0; X) + o_p(1),$$

where $\mathbb{P}_n^* w\tilde{l}(\beta_0, \Lambda_0; X) = \mathbb{P}_n uw\tilde{l}(\beta_0, \Lambda_0; X)$. Hence

$$n^{1/2}(\hat{\beta}_n^* - \hat{\beta}_n) = I^{-1}(\beta_0) n^{1/2} (\mathbb{P}_n^* - \mathbb{P}_n) w\tilde{l}(\beta_0, \Lambda_0; X) + o_p(1).$$

By Theorem 2 of Ma and Kosorok (2005) we know that, conditional on $(X_1, w_1), \cdots, (X_n, w_n)$, $(n/\delta_0)^{1/2}(\hat{\beta}_n^* - \hat{\beta}_n)$ has the same asymptotic distribution as that of $n^{1/2}(\hat{\beta}_n - \beta_0)$ unconditionally.

2.4.2 Using Estimated Weights

Unfortunately, the above weighted bootstrap method does not work for the weighted likelihood estimator $\tilde{\beta}_n$ with estimated weights. To see this, we assume $\tilde{\beta}_{n1}^*, \cdots$, and $\tilde{\beta}_{nB}^*$ are the B bootstrap estimates of β_0 based on estimated weights. Then by Theorem II.9,

$$n^{1/2}(\tilde{\beta}_n^* - \beta_0) = I^{-1}(\beta_0) n^{1/2} \mathbb{P}_n^* w\tilde{l}(\beta_0, \Lambda_0; X) + n^{1/2} C(\hat{\alpha}_n - \alpha_0) + o_p(1),$$

hence

$$n^{1/2}(\tilde{\beta}_n^* - \tilde{\beta}_n) = I^{-1}(\beta_0)n^{1/2}\,(\mathbb{P}_n^* - \mathbb{P}_n)w\tilde{l}(\beta_0, \Lambda_0; X) + o_p(1).$$

This implies that the asymptotic distribution of $(n/\delta_0)^{1/2}(\tilde{\beta}_n^* - \hat{\beta}_n)$ conditional on all the observed data is the same as that of $n^{1/2}(\hat{\beta}_n - \beta_0)$ unconditionally. Therefore, the empirical variance of $\tilde{\beta}_{n1}^*, \cdots, \tilde{\beta}_{nB}^*$ actually estimates the asymptotic variance of $\hat{\beta}_n$ (after rescaling), not the asymptotic variance of $\tilde{\beta}_n$ that is of interest.

We propose using the smoothing technique to calculate the asymptotic variance given in Theorem II.9, which involves estimating $I(\beta_0)$ as well as the matrices $E(w^2\tilde{l}^{\otimes 2}(\beta_0, \Lambda_0; X))$ and C. The full data information matrix $I(\beta_0)$ can be estimated by $\sum_{i=1}^n w_i\tilde{l}_n^{\otimes 2}(\tilde{\beta}_n, \tilde{\Lambda}_n; X_i)/n$, where $\tilde{l}_n$ is $\tilde{l}$ with conditional expectations replaced by their estimates obtained from nonparametric smoothing. Similarly, in estimating $E(w^2\tilde{l}^{\otimes 2}(\beta_0, \Lambda_0; X))$, we use

$$\frac{1}{n}\sum_{i=1}^n w_i^2(\hat{\alpha}_n)\tilde{l}_n(\tilde{\beta}_n, \tilde{\Lambda}_n; X_i) = \frac{1}{n}\sum_{i\in\mathcal{C}} \frac{1}{\pi_{\hat{\alpha}_n}^2(\Delta_i, V_i)}\tilde{l}_n^{\otimes 2}(\tilde{\beta}_n, \tilde{\Lambda}_n; X_i),$$

and in estimating the matrix C, we use

$$\frac{1}{n}\sum_{i=1}^n \tilde{l}_n(\tilde{\beta}_n, \tilde{\Lambda}_n; X_i)\dot{w}_i^T(\hat{\alpha}_n) = -\frac{1}{n}\sum_{i\in\mathcal{C}} \tilde{l}_n(\tilde{\beta}_n, \tilde{\Lambda}_n; X_i)\frac{\dot{\pi}_{\hat{\alpha}_n}^T(\Delta_i, V_i)}{\pi_{\hat{\alpha}_n}^2(\Delta_i, V_i)},$$

here $\mathcal{C}$ denotes the set of indices of all subjects with complete data and $\dot{\pi}_\alpha(\Delta, V) = \partial\pi_\alpha(\Delta, V)/\partial\alpha$. Finally, the matrix Σ_α needs to be estimated. The estimator of Σ_α depends on the model used for estimating α_0. For a two-phase stratified sampling, for example, $\alpha = (p_1, \cdots, p_J)^T$ and p_j is estimated by the sampling proportion $\hat{p}_j$ in stratum j, $1 \le j \le J$, then we have $\hat{\Sigma}_\alpha = \mathrm{diag}(n\hat{p}_1(1-\hat{p}_1)/n_1, \cdots, n\hat{p}_J(1-\hat{p}_J)/n_J)$, where n_j is the number of subjects in stratum j, $1 \le j \le J$, among n subjects. Note that the vector of sampling proportions $\hat{p}_j$'s is the maximum likelihood estimator of α and hence most efficient.

2.5 Numerical Results

2.5.1 Simulations

A simulation study is conducted to explore the performance of the proposed weighted likelihood estimators. We assume the unobserved time to failure T follows a proportional hazards model given covariate Z with a constant baseline hazard function $\lambda(t) \equiv c$, which implies that the failure time has an exponential distribution. The censoring time Y is assumed to be uniformly distributed in the interval between 0.5 and 8.5. The covariate Z has two components Z_1 and Z_2, where $Z_1 \sim N(0,1)$, and Z_2 is a categorical covariate with $Pr(Z_2 = 0) = Pr(Z_2 = 1) = 0.5$. The true parameter for β is $\beta_0 = (1, -1)^T$. We consider two sample sizes, $n = 500$ and $n = 3000$. When $n = 500$, we take $c = 0.03$; when $n = 3000$, we take $c = 0.01$. We first generate n i.i.d. samples of (Δ, Y, Z) and then generate missing covariates. The missing covariates are generated via a case-cohort sampling method. We assume that Z_1 can be missing while Z_2 is always observed. The probability of missing Z_1 is 0 for a subject with a failure event, and depends on an auxiliary variable V for a censored subject. The auxiliary variable V is associated with the covariates of interest in the following way: $V = 1$ when $Z_1 < 1$ and $Z_2 = 0$, $V = 2$ when $Z_1 < 1$ and $Z_2 = 1$, $V = 3$ when $Z_1 >= 1$ and $Z_2 = 0$, and $V = 4$ when $Z_1 >= 1$ and $Z_2 = 1$. When $n = 500$, the probability of missing covariate Z_1 is $p = 0.2$ if $V = 1$ or 2, and $p = 0.7$ if $V = 3$ or 4. When $n = 3000$, $p = 0.05$ if $V = 1$ or 2, and $p = 0.15$ if $V = 3$ or 4. Under these circumstances, when sample size $n = 500$, there are about 170 subjects with covariates fully observed, among whom about 100 are observed to have a failure event; and when $n = 3000$, there are about 400 subjects with fully observed covariates, among whom 250 are failures. The setting for $n = 3000$ here mimics the setting for the HIV case-cohort study in the next subsection.

We then calculate the weighted likelihood estimator $(\hat{\beta}_n, \hat{\Lambda}_n)$ using the iterative algorithm in Section 2.2 for each generated data set. We choose $(0, 0)$ as the initial value of $\hat{\beta}_n$, and then iterate between $\hat{\beta}_n$ and $\hat{\Lambda}_n$ until convergence. The same procedure is executed to obtain $(\tilde{\beta}_n, \tilde{\Lambda}_n)$. We run 500 replications for the simulation, and then obtain point estimates and biases of the estimators of β_0. Variance estimates of $\hat{\beta}_n$ are obtained by the weighted bootstrap procedure and that of $\tilde{\beta}_n$ are obtained by using smoothing splines. To apply the weighted bootstrap method, we generate independent weight u from a uniform distribution on $(0, 2)$, and use 100 bootstrap samples to estimate variance for each simulated data set. Smoothing splines can be used for the variance estimation for both $\hat{\beta}_n$ and $\tilde{\beta}_n$ in evaluating the quantities $E(\cdot|Y = y)$. The actual calculation is implemented in R. To be specific, for a function $h(Y, Z_1, Z_2)$, we have

$$\begin{aligned} E[h(Y, Z_1, Z_2)|Y] &= E[h(Y, Z_1, 1)|Y, Z_2 = 1]Pr(Z_2 = 1|Y) \\ &\quad + E[h(Y, Z_1, 0)|Y, Z_2 = 0]Pr(Z_2 = 0|Y), \end{aligned}$$

where $E[h(Y, Z_1, 1)|Y, Z_2 = 1]$, $E[h(Y, Z_1, 0)|Y, Z_2 = 0]$ and $Pr(Z_2 = 1|Y)$ are calculated separately using the weighted generalized additive models (function "gam") with cubic smoothing splines to Y and Gaussian (or logit) link function. Default smoothing parameter values are used.

Biases, means of estimated variances, empirical variances, and coverage proportions (CP) of 95% confidence intervals for the estimators of coefficients of Z_1 and Z_2 are presented in Table 2.1. The biases are reasonably small across the board, particularly for the larger sample size. Variance estimators, obtained either by weighted bootstrap or smoothing, are very close to corresponding empirical variances and yield reasonably good coverage proportions. Comparing empirical variances of the

weighted likelihood estimator with true weights and those with estimated weights, the efficiency gain of the latter is clear, supporting our theoretical results in Section 2.2.

2.5.2 A Case-Cohort Study from An HIV Vaccine Trial

We illustrate our method here by analyzing a case-cohort study from one of the largest phase 3 HIV-1 vaccine efficacy trials in the world (see Flynn et al., 2005 and Gilbert et al., 2005). The trial demonstrated lack of efficacy of the vaccine, but Gilbert et al. Gilbert et al. (2005) undertook a secondary objective, which was to determine whether antibody responses are correlated with the incidence of HIV-1 infection among vaccine recipients. The trial was designed to have multiple visits and either vaccine or placebo was administered at each visit. For simplicity, we only consider the infection status at the last visit and thus have the current status data to work with. The original trial consists of 5095 men and 308 women who received the study vaccine or placebo at a $2:1$ ratio. Gilbert et al. (2005) designed a case-cohort study that consisted of all 241 infected subjects and 167, a fraction of 5%, uninfected subjects, all were selected from vaccine recipients. They found that the peak antibody levels reached a high level at month 6.5 (after the second vaccine shot) and became relatively stable afterwards. We consider the only functional assay, the MN neutralization titer, among all antibody responses and use its peak level at month 6.5 (hence infections prior month 6.5 are excluded) as the covariate of interest in our analysis. This antibody in principle should be most relevant for HIV protection. Cubic-root power transformation of this variable is used to achieve a better linear effect in the Cox model. Several demographic variables are also considered, but only the baseline behavioral risk score is significant. Since only the sample fraction of 5% for uninfected subjects was provided by Gilbert et al. (2005), we use the weighted

likelihood method with estimated weights in our analysis. The final result is given in Table 2. We can see that the antibody MN neutralization titer has a protection effect against HIV infection, which is consistent with the finding in Gilbert et al. (2005) where an analysis for approximated right censored data was conducted.

2.6 A General Theorem

In this section, we provide a general theorem that generalizes Theorem 6.1 of Wellner and Zhang (2007) by replacing one of the nuisance parameters by its estimator in the objective function that will be maximized with respect to all other parameters. We will follow their notation closely.

Given i.i.d. observations $X_1, \cdots, X_n$, suppose that the estimates $(\tilde{\beta}_n, \tilde{\Lambda}_n)$ of unknown parameters (β, Λ) are set to be the maximizer of the objective function $\mathbb{P}_n m(\beta, \Lambda, \hat{\alpha}_n; X)$, where $\hat{\alpha}_n$ is an estimator of the true parameter α_0, $\beta \in R^d$, and $\Lambda \in \mathcal{F}$, an infinite dimensional Banach space. Here we assume α_0 to be finite dimensional, though it can be more general. Suppose that Λ_η is a parametric submodel in $\mathcal{F}$ passing through Λ, that is, $\Lambda_\eta \in \mathcal{F}$ and $\Lambda_{\eta=0} = \Lambda$. Let $H = \{h : h = \partial\Lambda_\eta/\partial\eta|_{\eta=0}\}$

be the collection of all directions to approach Λ. For any $h \in H$, we define

$$\begin{aligned}
m_1(\beta,\Lambda,\alpha;x) &= \left(\frac{\partial m(\beta,\Lambda,\alpha;x)}{\partial\beta_1},\cdots,\frac{\partial m(\beta,\Lambda,\alpha;x)}{\partial\beta_d}\right)^T,\\
m_2(\beta,\Lambda,\alpha;x)[h] &= \left.\frac{\partial m(\beta,\Lambda_\eta,\alpha;x)}{\partial\eta}\right|_{\eta=0},\\
m_3(\beta,\Lambda,\alpha;x) &= \frac{\partial m(\beta,\Lambda,\alpha;x)}{\partial\alpha},\\
m_{11}(\beta,\Lambda,\alpha;x) &= \frac{\partial^2 m(\beta,\Lambda,\alpha;x)}{\partial\beta\partial\beta^T},\\
m_{12}(\beta,\Lambda,\alpha;x)[h] &= \left.\frac{\partial m_1(\beta,\Lambda_\eta,\alpha;x)}{\partial\eta}\right|_{\eta=0},\\
m_{13}(\beta,\Lambda,\alpha;x) &= \frac{\partial^2 m(\beta,\Lambda,\alpha;x)}{\partial\beta\partial\alpha^T},\\
m_{21}(\beta,\Lambda,\alpha;x)[h] &= \frac{\partial m_2(\beta,\Lambda,\alpha;x)[h]}{\partial\beta},\\
m_{22}(\beta,\Lambda,\alpha;x)[h_1,h_2] &= \left.\frac{\partial m_2(\beta,\Lambda_{\eta_2},\alpha;x)[h_1]}{\partial\eta_2}\right|_{\eta_2=0},\\
m_{23}(\beta,\Lambda,\alpha;x)[h] &= \frac{\partial m_2(\beta,\Lambda,\alpha;x)[h]}{\partial\alpha}.
\end{aligned}$$

We also define

$$\begin{aligned}
S_1(\beta,\Lambda,\alpha) &= Pm_1(\beta,\Lambda,\alpha;X),\\
S_2(\beta,\Lambda,\alpha)[h] &= Pm_2(\beta,\Lambda,\alpha;X)[h],\\
S_3(\beta,\Lambda,\alpha) &= Pm_3(\beta,\Lambda,\alpha;X)\\
S_{1n}(\beta,\Lambda,\alpha) &= \mathbb{P}_n m_1(\beta,\Lambda,\alpha;X),\\
S_{2n}(\beta,\Lambda,\alpha)[h] &= \mathbb{P}_n m_2(\beta,\Lambda,\alpha;X)[h],\\
\dot{S}_{11}(\beta,\Lambda,\alpha) &= Pm_{11}(\beta,\Lambda,\alpha;X),\\
\dot{S}_{12}(\beta,\Lambda,\alpha)[h] &= \dot{S}_{21}^T(\beta,\Lambda,\alpha)[h] = Pm_{12}(\beta,\Lambda,\alpha;X)[h],\\
\dot{S}_{13}(\beta,\Lambda,\alpha) &= Pm_{13}(\beta,\Lambda,\alpha;X),\\
\dot{S}_{22}(\beta,\Lambda,\alpha)[h_1,h_2] &= Pm_{22}(\beta,\Lambda,\alpha;X)[h_1,h_2],\\
\dot{S}_{23}(\beta,\Lambda,\alpha)[h] &= Pm_{23}(\beta,\Lambda,\alpha;X)[h].
\end{aligned}$$

Furthermore, for $\mathbf{h} = (h_1, \cdots, h_d)^T \in H^d$, where $h_j \in H$ for $1 \leq j \leq d$, we denote

$$
\begin{aligned}
m_2(\beta, \Lambda, \alpha; x)[\mathbf{h}] &= (m_2(\beta, \Lambda, \alpha; X[h_1], \cdots, m_2(\beta, \Lambda, \alpha; X[h_d])^T, \\
m_{12}(\beta, \Lambda, \alpha; x)[\mathbf{h}] &= (m_{12}(\beta, \Lambda, \alpha; X[h_1], \cdots, m_{12}(\beta, \Lambda, \alpha; X[h_d]), \\
m_{21}(\beta, \Lambda, \alpha; x)[\mathbf{h}] &= (m_{21}(\beta, \Lambda, \alpha; X[h_1], \cdots, m_{21}(\beta, \Lambda, \alpha; X[h_d])^T, \\
m_{22}(\beta, \Lambda, \alpha; x)[\mathbf{h}, h] &= (m_{22}(\beta, \Lambda, \alpha; X[h_1, h], \cdots, m_{22}(\beta, \Lambda, \alpha; X[h_d, h])^T, \\
m_{23}(\beta, \Lambda, \alpha; x)[\mathbf{h}] &= (m_{23}(\beta, \Lambda, \alpha; X[h_1], \cdots, m_{23}(\beta, \Lambda, \alpha; X[h_d])^T, \\
S_2(\beta, \Lambda, \alpha)[\mathbf{h}] &= Pm_2(\beta, \Lambda, \alpha; X)[\mathbf{h}], \\
S_{2n}(\beta, \Lambda, \alpha)[\mathbf{h}] &= \mathbb{P}_n m_2(\beta, \Lambda, \alpha; X)[\mathbf{h}], \\
\dot{S}_{12}(\beta, \Lambda, \alpha)[\mathbf{h}] &= Pm_{12}(\beta, \Lambda, \alpha; X)[\mathbf{h}], \\
\dot{S}_{21}(\beta, \Lambda, \alpha)[\mathbf{h}] &= Pm_{21}(\beta, \Lambda, \alpha; X)[\mathbf{h}], \\
\dot{S}_{22}(\beta, \Lambda, \alpha)[\mathbf{h}, h] &= Pm_{22}(\beta, \Lambda, \alpha; X)[\mathbf{h}, h], \\
\dot{S}_{23}(\beta, \Lambda, \alpha)[\mathbf{h}] &= Pm_{23}(\beta, \Lambda, \alpha; X)[\mathbf{h}],
\end{aligned}
$$

The following conditions are parallel to those in Theorem 6.1 of Wellner and Zhang (2007), but here they are adapted to accommodate a more general setting.

A1. $|\hat{\alpha}_n - \alpha_0| = o_p(1)$, $|\tilde{\beta}_n - \beta_0| = o_p(1)$, and $\|\tilde{\Lambda}_n - \Lambda_0\| = O_p(n^{-\gamma})$ for some $\gamma > 0$ and some norm $\|\cdot\|$.

A2. There exists an $\mathbf{h}^* = (h_1^*, \cdots, h_d^*)^T$, where $h_j^* \in L_2(P)$, $j = 1, 2, \cdots, d$, such that

$$\dot{S}_{12}(\beta_0, \Lambda_0, \alpha_0)[h] - \dot{S}_{22}(\beta_0, \Lambda_0, \alpha_0)[\mathbf{h}^*, h] = 0,$$

for all $h \in H$. Moreover, the matrix

$$
\begin{aligned}
A &= -\dot{S}_{11}(\beta_0, \Lambda_0, \alpha_0) + \dot{S}_{21}(\beta_0, \Lambda_0, \alpha_0)[\mathbf{h}^*] \\
&= -P\{m_{11}(\beta_0, \Lambda_0, \alpha_0; X) - m_{21}(\beta_0, \Lambda_0, \alpha_0; X)[\mathbf{h}^*]\}
\end{aligned}
$$

is non-singular.

A3. $S_1(\beta_0, \Lambda_0, \alpha_0) = 0$ and $S_2(\beta_0, \Lambda_0, \alpha_0)[\mathbf{h}^*] = 0$.

A4. The estimator $(\tilde{\beta}_n, \tilde{\Lambda}_n)$ satisfies

$$S_{1n}(\tilde{\beta}_n, \tilde{\Lambda}_n, \hat{\alpha}_n) = o_p(n^{-1/2}) \quad \text{and} \quad S_{2n}(\tilde{\beta}_n, \tilde{\Lambda}_n, \hat{\alpha}_n)[\mathbf{h}^*] = o_p(n^{-1/2}).$$

A5. For any $\delta_n \downarrow 0$ and $C > 0$, let

$$\Theta_n = \{(\beta, \Lambda, \alpha) : |(\beta^T, \alpha^T) - (\beta_0^T, \alpha_0^T)| \leq \delta_n, \|\Lambda - \Lambda_0\|_2 \leq Cn^{-\gamma}\}.$$

We have

$$\begin{aligned} &\sup_{(\beta,\Lambda,\alpha)\in\Theta_n} |n^{1/2}(S_{1n} - S_1)(\beta, \Lambda, \alpha) - n^{1/2}(S_{1n} - S_1)(\beta_0, \Lambda_0, \alpha_0)| \\ &\quad = \; o_p(1), \end{aligned}$$

and

$$\begin{aligned} \sup_{(\beta,\Lambda,\alpha)\in\Theta_n} &|n^{1/2}(S_{2n} - S_2)(\beta, \Lambda, \alpha)[\mathbf{h}^*] \\ &- n^{1/2}(S_{2n} - S_2)(\beta_0, \Lambda_0, \alpha_0)[\mathbf{h}^*]| = o_p(1). \end{aligned}$$

A6. For some $\mu > 1$ satisfying $\mu\gamma > 1/2$, and for $(\beta, \Lambda, \alpha) \in \Theta_n$,

$$\begin{aligned} &\Big|S_1(\beta, \Lambda, \alpha) - S_1(\beta_0, \Lambda_0, \alpha_0) - \dot{S}_{11}(\beta_0, \Lambda_0, \alpha_0)(\beta - \beta_0) \\ &\qquad - \dot{S}_{12}(\beta_0, \Lambda_0, \alpha_0)[\Lambda - \Lambda_0] - \dot{S}_{13}(\beta_0, \Lambda_0, \alpha_0)(\alpha - \alpha_0)\Big| \\ &\quad = \; o(|\beta - \beta_0|) + o(|\alpha - \alpha_0|) + O(\|\Lambda - \Lambda_0\|^\mu), \end{aligned}$$

and

$$\begin{aligned}|S_2(\beta,\Lambda,p)[\mathbf{h}^*] &- S_2(\beta_0,\Lambda_0,\alpha_0)[\mathbf{h}^*] \\ &- \dot{S}_{21}(\beta_0,\Lambda_0,\alpha_0)[\mathbf{h}^*](\beta-\beta_0) \\ &- \dot{S}_{22}(\beta_0,\Lambda_0,\alpha_0)[\mathbf{h}^*,\Lambda-\Lambda_0] \\ &- \dot{S}_{23}(\beta_0,\Lambda_0,\alpha_0)[\mathbf{h}^*](\alpha-\alpha_0)| \\ = \; & o(|\beta-\beta_0|)+o(|\alpha-\alpha_0|)+O(\|\Lambda-\Lambda_0\|^{\mu}).\end{aligned}$$

Theorem II.10. *Suppose that conditions A1 to A6 hold. Then we have*

$$n^{1/2}(\tilde{\beta}_n-\beta_0)=A^{-1}n^{1/2}\mathbb{P}_n m^*(\beta_0,\Lambda_0,\alpha_0;X)-Cn^{1/2}(\hat{\alpha}_n-\alpha_0)+o_{p^*}(1),$$

where

$$m^*(\beta_0,\Lambda_0,\alpha_0;X)=m_1(\beta_0,\Lambda_0,\alpha_0;X)-m_2(\beta_0,\Lambda_0,\alpha_0;X)[\mathbf{h}^*],$$

and

$$C=A^{-1}(\dot{S}_{13}(\beta_0,\Lambda_0,\alpha_0)-\dot{S}_{23}(\beta_0,\Lambda_0,\alpha_0)[\mathbf{h}^*]).$$

If $n^{1/2}(\hat{\alpha}_n-\alpha_0)$ is asymptotically normal with influence function ℓ^{α}, then $n^{1/2}(\tilde{\beta}_n-\beta_0)$ is asymptotically normal. Furthermore, if $\hat{\alpha}_n$ is asymptotically efficient, then $n^{1/2}(\tilde{\beta}_n-\beta_0)\to_d N(0,\Omega)$ with

$$\Omega=A^{-1}E[m^*(\beta_0,\Lambda_0;X)^{\otimes 2}](A^{-1})^T-CE(\ell^{\alpha\otimes 2})C^T.$$

Proof: By A1, A3 and A5,

$$S_{1n}(\tilde{\beta}_n,\tilde{\Lambda}_n,\hat{\alpha}_n)-S_1(\tilde{\beta}_n,\tilde{\Lambda}_n,\hat{\alpha}_n)-S_{1n}(\beta_0,\Lambda_0,\alpha_0)=o_p(n^{-1/2}).$$

In view of A4, this reduces to

$$S_{1n}(\beta_0,\Lambda_0,\alpha_0)+S_1(\tilde{\beta}_n,\tilde{\Lambda}_n,\hat{\alpha}_n)=o_p(n^{-1/2}).$$

Then by A6, it follows that

$$\begin{aligned}
&\dot{S}_{11}(\beta_0,\Lambda_0,\alpha_0)(\tilde{\beta}_n-\beta_0)+\dot{S}_{12}(\beta_0,\Lambda_0,\alpha_0)[\tilde{\Lambda}_n-\Lambda_0] \\
&\qquad +\dot{S}_{13}(\beta_0,\Lambda_0,\alpha_0)(\hat{\alpha}_n-\alpha_0)+S_{1n}(\beta_0,\Lambda_0,\alpha_0) \\
&= o(|\tilde{\beta}_n-\beta_0|)+o(|\hat{\alpha}_n-\alpha_0|)+O(\|\tilde{\Lambda}_n-\Lambda_0\|^2) \qquad (2.11)\\
&= o_p(n^{-1/2}),
\end{aligned}$$

In a similar way, we obtain

$$S_{2n}(\beta_0,\Lambda_0,\alpha_0)[\mathbf{h}^*]+S_2(\tilde{\beta}_n,\tilde{\Lambda}_n,\hat{\alpha}_n)[\mathbf{h}^*]=o_p(n^{-1/2}),$$

and then

$$\begin{aligned}
&\dot{S}_{21}(\beta_0,\Lambda_0,\alpha_0)[\mathbf{h}^*](\tilde{\beta}_n-\beta_0)+\dot{S}_{22}(\beta_0,\Lambda_0,\alpha_0)[\mathbf{h}^*,\tilde{\Lambda}_n-\Lambda_0] \\
&\qquad +\dot{S}_{23}(\beta_0,\Lambda_0,\alpha_0)[\mathbf{h}^*](\hat{\alpha}_n-\alpha_0)+S_{2n}(\beta_0,\Lambda_0,\alpha_0)[\mathbf{h}^*] \\
&= o(|\tilde{\beta}_n-\beta_0|)+o(|\hat{\alpha}_n-\alpha_0|)+O(\|\tilde{\Lambda}_n-\Lambda_0\|^2) \qquad (2.12)\\
&= o_p(n^{-1/2}).
\end{aligned}$$

Subtracting (2.12) from (2.11) and rearranging terms, by A2 we obtain

$$\begin{aligned}
n^{1/2}(\tilde{\beta}_n-\beta_0) &= n^{1/2}A^{-1}\mathbb{P}_n m^*(\beta_0,\Lambda_0,\alpha_0;X)-Cn^{1/2}(\hat{\alpha}_n-\alpha_0) \qquad (2.13)\\
&\quad +o_p(1).
\end{aligned}$$

When $n^{1/2}(\hat{\alpha}_n-\alpha_0)$ is asymptotically normal with influence function ℓ^α, the right hand side of the above equation converges to a zero mean normal random variable by the classical central limit theorem. Furthermore, when $\hat{\alpha}_n$ is efficient, $n^{1/2}(\tilde{\beta}_n-\beta_0)\to_d N(0,\Omega)$ follows from (2.13) and the result in Pierce (1982), with Ω being stated in the theorem. $\square$

2.7 Proofs of Theoretical Results in Section 3

2.7.1 Proof of Theorem II.1

This proof of this theorem follows from the same idea as the proof of Proposition 1.1 in Groeneboom and Wellner (1992).

Let $\tilde{x} = (x_1, \cdots, x_{r_n})$. Define

$$\phi_i(x) = \sum_{Y_j = Y_i^*} w_j \left\{ \Delta_j \log(1 - e^{-xe^{\hat{\beta}_n^T Z_j}}) - (1 - \Delta_j) x e^{\hat{\beta}_n^T Z_j} \right\},$$

$1 \leq i \leq r_n$, and

$$\phi(\tilde{x}) = \sum_{i=1}^{r_n} \phi_i(x_i),$$

where $\tilde{x}$ satisfies the constraint

$$0 \leq x_1 \leq x_2 \cdots \leq x_{r_n}. \tag{2.14}$$

Suppose $\tilde{a} = (a_1, \cdots, a_{r_n})$ maximizes $\phi(\tilde{x})$ under constraint (2.14). Then the vector $\tilde{a} + \varepsilon \tilde{1}_i$ satisfies constraint (2.14), for any $\varepsilon > 0$, and $1 \leq i \leq r_n$, where $\tilde{1}_i$ is the r_n dimensional vector with the first $r_n - i$ components 0 and the last i components 1. Since $\tilde{a}$ maximizes $\phi(\tilde{x})$, we have

$$\begin{aligned}
\lim_{\varepsilon \downarrow 0} \frac{\phi(\tilde{a} + \varepsilon \tilde{1}_i) - \phi(\tilde{a})}{\varepsilon} &= \sum_{Y_j \geq Y_i^*} \phi_j'(a_j) \\
&= \sum_{Y_j \geq Y_i^*} w_j e^{\hat{\beta}_n^T Z_j} \left\{ \Delta_j \frac{e^{-\hat{\Lambda}_n(Y_j) e^{\hat{\beta}_n^T Z_j}}}{1 - e^{-\hat{\Lambda}(Y_j) e^{\hat{\beta}_n^T Z_j}}} - (1 - \Delta_j) \right\} \\
&\leq 0,
\end{aligned}$$

for any $i = 1, \cdots, r_n$. Moreover, since ϕ is differentiable and attains a maximum at

$\tilde{a}$, we obtain

$$\begin{aligned}\lim_{h\to 0}\frac{\phi(\tilde{a}+h\tilde{a})-\phi(\tilde{a})}{h} &= \sum_{i=1}^{r_n} a_i\phi_i'(a_i)\\ &= \sum_{i=1}^{n} w_i\hat{\Lambda}_n(Y_i)e^{\hat{\beta}_n^T Z_i}\left\{\Delta_i\frac{e^{\hat{\beta}_n^T Z_i}e^{-\hat{\Lambda}_n(Y_i)e^{\hat{\beta}_n^T Z_i}}}{1-e^{-\hat{\Lambda}_n(Y_i)e^{\hat{\beta}_n^T Z_i}}}-(1-\Delta_i)\right\}\\ &= 0.\end{aligned}$$

Conversely, suppose that $\tilde{a}$ satisfies (2.4), (2.5) and (2.14), with $\hat{\Lambda}_n(Y_i)$ replaced by a_i. Since ϕ is concave, if $\tilde{x}$ satisfies (2.14), then

$$\phi(\tilde{x})-\phi(\tilde{a})\le\langle\bigtriangledown\phi(\tilde{a}),\tilde{x}-\tilde{a}\rangle. \tag{2.15}$$

By (2.5), $\langle\bigtriangledown\phi(\tilde{a}),\tilde{a}\rangle=0$, thus (2.15) becomes $\phi(\tilde{x})-\phi(\tilde{a})\le\langle\bigtriangledown\phi(\tilde{a}),\tilde{x}\rangle$. Now write $\tilde{x}$ in form $\tilde{x}=\sum_{i=1}^{r_n}\alpha_i\tilde{1}_i$, where $\alpha_i=x_{r_n-i+1}-x_{r_n-i}\ge 0$, and $\alpha_0=0$. Then for any $\tilde{x}$ satisfying (2.14),

$$\begin{aligned}\phi(\tilde{x})-\phi(\tilde{a}) &\le \langle\bigtriangledown\phi(\tilde{a}),\tilde{x}\rangle\\ &= \sum_{i=1}^{r_n}\alpha_i\sum_{Y_j\ge Y_i}\phi_j'(a_j)\\ &\le 0,\end{aligned}$$

and hence $\tilde{a}$ maximizes $\phi(\tilde{x})$. □

2.7.2 Proof of Theorem II.2

The proof of this theorem follows along the same lines as Proposition 1.4 in Groeneboom and Wellner (1992).

For simplicity of notation, we write $\hat{\Lambda}_n$ instead of $\hat{\Lambda}_n(\cdot;\beta)$. By definition, the left derivative of the convex minorant of the cumulative sum diagram is given by

$$h_n(\tau_i)=\frac{V_{\hat{\Lambda}_n}(\tau_i)-V_{\hat{\Lambda}_n}(\tau_{i-1})}{G_{\hat{\Lambda}_n}(\tau_i)-G_{\hat{\Lambda}_n}(\tau_{i-1})},$$

at the successive locations τ_i of the vertices of the greatest convex minorant of the cumulative sum diagram, and $\tau_0 = 0$. Moreover, defining $Y_0^* = 0, \Delta V_{n,i} = V_{\hat{\Lambda}_n}(Y_i^*) - V_{\hat{\Lambda}_n}(Y_{i-1}^*), \Delta G_{n,i} = G_{\hat{\Lambda}_n}(Y_i^*) - G_{\hat{\Lambda}_n}(Y_{i-1}^*)$, and $\Delta W_{n,i} = W_{\hat{\Lambda}_n}(Y_i^*) - W_{\hat{\Lambda}_n}(Y_{i-1}^*), 1 \leq i \leq r_n$, we have that h_n maximizes

$$\sum_{i=1}^{r_n} \left\{ h(Y_i^*) - \frac{\Delta V_{n,i}}{\Delta G_{n,i}} \right\}^2 \Delta G_{n,i},$$

over all nondecreasing functions h such that $h(0) = 0$. This means by Theorem 1.3.2 in Robertson, Wright and Dykstra (1988) that

$$\sum_{i=1}^{r_n} \left\{ \frac{\Delta V_{n,i}}{\Delta G_{n,i}} - h_n(Y_i^*) \right\} h(Y_i^*) \Delta G_{n,i} \leq 0, \tag{2.16}$$

$$\sum_{i=1}^{r_n} \left\{ \frac{\Delta V_{n,i}}{\Delta G_{n,i}} - h_n(Y_i^*) \right\} h_n(Y_i^*) \Delta G_{n,i} = 0, \tag{2.17}$$

for all nondecreasing h such that $h(0) = 0$. We now show that (2.16) implies (2.4) and (2.17) implies (2.5), with h_n replaced by $\hat{\Lambda}_n$.

For a fixed i, let $h(t) = 1_{(t \geq Y_i^*)}$, then (2.16) reduces to

$$\sum_{j \geq i} \left\{ \Delta V_{n,j} - h_n(Y_j^*) \Delta G_{n,j} \right\} \leq 0,$$

or

$$\sum_{j \geq i} \Delta W_{n,j} = \sum_{Y_j \geq Y_i^*} w_j e^{\beta^T Z_j} \left\{ \Delta_j \frac{e^{-h_n(Y_j^*) e^{\beta^T Z_j}}}{1 - e^{-h_n(Y_j^*) e^{\beta^T Z_j}}} - (1 - \Delta_j) \right\} \leq 0.$$

Similarly, by (2.17), we get

$$\begin{aligned} & \sum_{i=1}^{r_n} h_n(Y_i^*) \{ \Delta V_{n,i} - h_n(Y_i^*) \Delta G_{n,i} \} = \sum_i h_n(Y_i^*) \Delta W_{h_n}(Y_i^*) \\ = & \sum_{i=1}^{n} h_n(Y_i^*) w_i e^{\beta^T Z_i} \left\{ \Delta_i \frac{e^{-h_n(Y_i^*) e^{\beta^T Z_i}}}{1 - e^{-h_n(Y_i^*) e^{\beta^T Z_i}}} - (1 - \Delta_i) \right\} \\ = & \ 0. \end{aligned}$$

It now follows from Theorem II.1 that h_n maximizes (2.2) for the fixed β.

Now suppose $\hat{\Lambda}_n$ maximizes (2.2) for a fixed β, then by Theorem II.1, (2.4) and (2.5) hold. For any nondecreasing function h such that $h(0) = 0$, define $\alpha_i = h(Y^*_{(r_n-i+1)}) - h(Y^*_{(r_n-i)}), 1 \le i \le r_n$, and $\Delta V_{n,i}, \Delta W_{n,i}$ are defined as before, except that $\hat{\Lambda}_n$ is now replaced by h_n, then $(h(Y^*_1), \cdots, h(Y^*_{r_n})) = \sum_{i=1}^n \alpha_i \tilde{1}_i$, and

$$\begin{aligned}\sum_{i=1}^{r_n} \left\{ \frac{\Delta V_{n,i}}{\Delta G_{n,i}} - \hat{\Lambda}_n(Y^*_i) \right\} h(Y^*_i) \Delta G_{n,i} &= \sum_{i=1}^{r_n} \Delta W_{n,i} h(Y^*_i) = \langle \Delta \bar{W}, \sum_{i=1}^{r_n} \alpha_i \tilde{1}_i \rangle \\ &= \sum_{i=1}^{n} \alpha_i \langle \Delta \bar{W}, \tilde{1}_i \rangle \le 0,\end{aligned}$$

by (2.4), where $\Delta \bar{W} = (\Delta W_{n,1}, \cdots, \Delta W_{n,r_n})$. In addition, (2.5) is equivalent to (2.17) with $h_n(Y^*_i)$ replaced by $\hat{\Lambda}_n(Y^*_i)$. Again, by Robertson, Wright and Dykstra (1988), (2.16) and (2.17) both imply that $\hat{\Lambda}_n$ minimizes $\sum_{i=1}^{r_n} \{h(Y^*_i) - \Delta V_{n,i}/\Delta G_{n,i}\}^2 \Delta G_{n,i}$ over all nondecreasing functions h such that $h(0) = 0$. By the pool adjacent violators algorithm, this implies that

$$\hat{\Lambda}_n(\tau_i) = \frac{V_{\hat{\Lambda}_n}(\tau_i) - V_{\hat{\Lambda}_n}(\tau_{i-1})}{G_{\hat{\Lambda}_n}(\tau_i) - G_{\hat{\Lambda}_n}(\tau_{i-1})},$$

which is the left derivative of the greatest convex minorant of the cumulative sum diagram consisting of the points $P_j, j = 0, 1, \cdots, r_n$. □

2.7.3 Proof of Theorem II.3

Following Van Der Vaart (2002), we introduce the functions $\ell(\beta, \Lambda; X) = \log\{(p_{\beta,\Lambda} + p_0)/2\}$ and $m(\beta, \Lambda; X) = w\ell(\beta, \Lambda; X)$, where $p_0 = p_{\beta_0,\Lambda_0}$. Although $\mathbb{P}_n m(\beta, \Lambda; X)$ is not maximized at $(\hat{\beta}_n, \hat{\Lambda}_n)$, it is true that $\mathbb{P}_n m(\hat{\beta}_n, \hat{\Lambda}_n; X) \ge \mathbb{P}_n m(\beta_0, \Lambda_0; X)$. Only this less restrictive condition is needed by Theorem 5.8 in Van Der Vaart (2002). Note that, under our assumptions, p_0 is bounded and bounded away from 0, so it follows that $m(\beta, \Lambda; X)$ is uniformly bounded. Then by Theorem 5.8 and Lemma 5.9 in Van Der Vaart (2002), to prove the consistency of $(\hat{\beta}_n, \hat{\Lambda}_n)$, it suffices to show that

the parameter space for (β, Λ) is compact, the map $(\beta, \Lambda) \mapsto p_{\beta,\Lambda}(x)$ is continuous for every x, and the map $(\beta, \Lambda) \mapsto Pm(\beta, \Lambda; X)$ achieves a unique maximum at (β_0, Λ_0).

The compactness of the parameter space $\mathcal{B}$ of β is from Assumption (A). By the theorem on page 239 of Billingsley (1999), the parameter space Φ of Λ is compact if Φ is closed, and for each sequence $\{\Lambda_n, n \geq 1\}$ in Φ, there exists a subsequence $\{\Lambda_{n'}\}$ and $\Lambda_0 \in \Phi$, such that $\|\Lambda_{n'} - \Lambda_0\|_2 \to 0$, as $n' \to \infty$. By the same diagonal argument used to prove Helly's selection theorem (see e.g. Shorack, 2000), for any sequence $\{\Lambda_n, n \geq 1\}$ in Φ, there exists a subsequence $\{\Lambda_{n'}\}$ and Λ_0, such that $|\Lambda_{n'}(y) - \Lambda_0(y)| \to 0$, for every continuity point of Λ_0. But this implies, by the dominated convergence theorem, that $\|\Lambda_{n'} - \Lambda_0\|_2 \to 0$, since the density of Y is bounded above and bounded away from 0. In addition, Φ is clearly closed. Therefore, Φ is also compact. The continuity of the map $(\beta, \Lambda) \mapsto p_{\beta,\Lambda}(x)$ for every x is clearly seen from equation (2.1).

We now show that the map $(\beta, \Lambda) \mapsto Pm(\beta, \Lambda; X)$ achieves a unique maximum at (β_0, Λ_0). By the fact that $E(w|\Delta, V) = 1$, we have $P\{m(\beta, \Lambda; X) - m(\beta_0, \Lambda_0; X)\} = P\{\ell(\beta, \Lambda; X) - \ell(\beta_0, \Lambda_0; X)\}$ that is negative Kullback-Leibler divergence and hence is always less than or equal to 0. It is 0 if and only if $p_{\beta,\Lambda} = p_0$ with probability 1, or equivalently, $e^{\beta^T Z}\Lambda(Y) = e^{\beta_0^T Z}\Lambda_0(Y)$ with probability 1. Denoting $\bar{\beta} = \beta - \beta_0$, this is equivalent to $\bar{\beta}^T Z = -\log \Lambda(Y) + \log \Lambda_0(Y)$, with probability 1, and hence $\bar{\beta}^T E\mathrm{var}(Z|Y)\bar{\beta} = 0$. But since $E\mathrm{var}(Z|Y) > 0$ by Assumption (D), this implies that $\beta = \beta_0$, and then $\Lambda(t) = \Lambda_0(t)$ follows. By Theorem 5.8 and Lemma 5.9 in Van Der Vaart (2002), we conclude that $\hat{\beta}_n \to \beta_0$ and $\|\hat{\Lambda}_n - \Lambda_0\|_2 \to 0$ in probability (almost surely), as $n \to \infty$. By the fact that the density of Y is bounded away from 0, the latter is equivalent to $\int_\sigma^\tau (\hat{\Lambda}_n(t) - \Lambda_0(t))^2 dt \to 0$ in probability (almost surely). Since $\Lambda_0(\cdot)$ is continuous and strictly monotone, it further implies that $\hat{\Lambda}_n(t) \to \Lambda_0(t)$ in

probability (almost surely) for every $t \in (\sigma, \tau)$. $\square$

2.7.4 Proof of Theorem II.4

Denote the ε-bracketing number of the set of functions $\{\ell(\beta, \Lambda; X) : (\beta, \Lambda) \in \Theta\}$ by $n(\varepsilon)$. Then there exist ε-brackets $[\ell^L_{\beta_i,\Lambda_i}, \ell^U_{\beta_i,\Lambda_i}]$, $1 \le i \le n(\varepsilon)$, such that for any $\ell(\beta, \Lambda; X)$, we have $\ell^L_{\beta_i,\Lambda_i} \le \ell(\beta, \Lambda; X) \le \ell^U_{\beta_i,\Lambda_i}$ for some i, which translates to $m^L_{\beta_i,\Lambda_i} \equiv w\ell^L_{\beta_i,\Lambda_i} \le m(\beta, \Lambda; X) \le w\ell^U_{\beta_i,\Lambda_i} \equiv m^U_{\beta_i,\Lambda_i}$. By Assumption (E) we know that $|w| < K$ for some constant $K < \infty$, hence

$$\|m^L_{\beta_i,\Lambda_i} - m^U_{\beta_i,\Lambda_i}\|^2 = E[w^2(\ell^L_{\beta_i,\Lambda_i} - \ell^U_{\beta_i,\Lambda_i})^2] \le K^2 E(\ell^L_{\beta_i,\Lambda_i} - \ell^U_{\beta_i,\Lambda_i})^2 \le K^2\varepsilon^2.$$

This shows that every $[m^L_{\beta_i,\Lambda_i}, m^U_{\beta_i,\Lambda_i}]$ is a $K\varepsilon$-bracket for the set of functions $\{m(\beta, \Lambda; X) : (\beta, \Lambda) \in \Theta\}$, and the brackets $[m^L_{\beta_i,\Lambda_i}, m^U_{\beta_i,\Lambda_i}]$, $1 \le i \le n(\varepsilon)$, cover $\{m(\beta, \Lambda; X) : (\beta, \Lambda) \in \Theta\}$. Form Lemma 8.6 of Van Der Vaart (2002) we know that the ε-bracketing number of $\{\ell(\beta, \Lambda; X) : (\beta, \Lambda) \in \Theta\}$ is of the order $e^{c/\varepsilon}/\varepsilon^d$ for some positive constant c. Hence the ε-bracketing number of $\{m(\beta, \Lambda; X) : (\beta, \Lambda) \in \Theta\}$ is of the order $e^{c'/\varepsilon}/\varepsilon^d$, for some constant c', which yields the desirable result. $\square$

2.7.5 Proof of Theorem II.5

It is well known that for any pair of probability densities p and q, we have $E_p(\log q/p) \le -\int (p^{1/2} - q^{1/2})^2 d\mu$, where μ is the dominating measure for the densities (see e.g. equation (8.1) in Van Der Vaart (2002)). Then by Lemma 8.7 of Van Der Vaart (2002), we have

$$\begin{aligned} & P\{m(\beta, \Lambda; X) - m(\beta_0, \Lambda_0; X)\} \\ &\quad = P\{\ell(\beta, \Lambda; X) - \ell(\beta_0, \Lambda_0; X)\} \\ &\quad \le -\int (p^{1/2}_{\beta,\Lambda} - p^{1/2}_0)^2 d\mu \\ &\quad \le -C\int_\sigma^\tau (\Lambda(t) - \Lambda_0(t))^2 dt - C|\beta - \beta_0|^2, \end{aligned}$$

for some constant C. The rest of the proof follows exactly the same as the proof of Lemma 8.5 in Van Der Vaart (2002). We still provide it here because it will be helpful to the proof of Theorem II.8. Based on the above calculation and Assumption (B), to apply Theorem 3.2.5 of Van Der Vart and Wellner (1996) we can choose

$$d^2((\beta,\Lambda),(\beta_0,\Lambda_0)) = \|\Lambda(t)-\Lambda_0(t)\|_2^2 + |\beta-\beta_0|^2.$$

Let $\mathcal{F} = \{m(\beta,\Lambda;X) : (\beta,\Lambda)\in\Theta\}$. By Theorem II.4, for a sufficiently small δ we have

$$J_{[]}(\delta,\mathcal{F},L_2(P)) = \int_0^\delta \left\{1+\log N_{[]}(\varepsilon,\mathcal{F},L_2(P))\right\}^{1/2} d\varepsilon \le C_1\delta^{1/2}$$

for some constant C_1. Hence by Lemma 3.4.2 of Van Der Vart and Wellner (1996), we obtain

$$\mathrm{E}\sup_{d((\beta,\Lambda),(\beta_0,\Lambda_0))<\delta} |\mathbb{G}_n\{m(\beta,\Lambda;X)-m(\beta_0,\Lambda_0;X)\}| \le C_2\phi_n(\delta) \tag{2.18}$$

for some constant C_2, where

$$\phi_n(\delta) = \delta^{1/2}\left(1+M\frac{\delta^{1/2}}{\delta^2 n^{1/2}}\right),$$

and M is a constant satisfying $\sup|m(\beta,\Lambda;x)-m(\beta_0,\Lambda_0;x)| \le M$. Thus by the consistency of $(\hat\beta_n,\hat\Lambda_n)$ provided in Theorem II.3 and Theorem 3.2.5 of Van Der Vart and Wellner (1996), it follows that $d((\hat\beta_n,\hat\Lambda_n),(\beta_0,\Lambda_0)) = O_p(n^{-1/3})$. □

2.7.6 Proof of Theorem II.6

The proof proceeds along the same lines as the proof of Theorem 3.4 in Huang (1996) by verifying Conditions A1-A6 in II.10 with $\hat\alpha_n$ fixed at α_0 and $m(\beta,\Lambda,\alpha_0;X) = w(\alpha_0)l(\beta,\Lambda;X) = w(\alpha_0)[\Delta\log(1-e^{-\Lambda(Y)e^{\beta^TZ}}) - (1-\Delta)\Lambda(Y)e^{\beta^TZ}]$. Here we denote the estimators of (β,Λ) as $(\hat\beta_n,\hat\Lambda_n)$ instead of $(\tilde\beta_n,\tilde\Lambda_n)$, the notation used in Theorem II.10. For notational simplicity, we drop α_0 wherever it appears in this proof. For instance, we write $m(\beta,\Lambda;X)$ instead of $m(\beta,\Lambda,\alpha_0;X)$.

By Theorem II.5, Condition A1 in II.10 is satisfied with $\gamma = 1/3$ and $\|\cdot\|$ being the L_2 norm with respect to the probability measure of Y. In order to verify A2, we first need to find an $\mathbf{h}^* \in L_2(P)$ such that $\dot{S}_{12}(\beta_0, \Lambda_0; X)[h] - \dot{S}_{22}(\beta_0, \Lambda_0; X)[\mathbf{h}^*, h] = 0$ for all $h \in H$. Because $E(w|X) = 1$, such a condition reduces to the exact same condition for the full data where $w \equiv 1$, hence holds with the $\mathbf{h}^*$ given in (2.9), which is the least favorable direction for the full data. See Huang (1996), Murphy-Van Der Vaart (2000) or Van Der Vaart (2002) for details. Furthermore, A is the information matrix for β for the full data, and its non-singularity is guaranteed by Assumption (D). We thus have verified Condition A2. Condition A3 holds automatically because, by $E(w|X) = 1$, S_1 and S_2 are equal to the expectations of full data scores for β and Λ, and hence equal to 0 at (β_0, Λ_0).

The first part of Condition A4 is trivial because $\hat{\beta}_n$ is obtained from equation $S_{1n}(\hat{\beta}_n, \hat{\Lambda}_n) = 0$. Due to the monotonicity constraint on Λ, however, we may not exactly have $S_{2n}(\hat{\beta}_n, \hat{\Lambda}_n)[\mathbf{h}^*] = 0$. We now verify that $S_{2n}(\hat{\beta}_n, \hat{\Lambda}_n)[\mathbf{h}^*] = o_p(n^{-1/2})$. Similar to Huang (1996), we define $\xi_0 = \mathbf{h}^* \circ \Lambda_0^{-1}$. The characterization of $\hat{\Lambda}_n$ given by (2.6) yields that

$$\sum_{j=k_i+1}^{k_{i+1}} w_{(j)} e^{\hat{\beta}_n^T Z_{(j)}} \left(\Delta_{(j)} \frac{e^{-e^{\hat{\beta}_n^T Z_{(j)}} \hat{\Lambda}_n(Y_{(j)})}}{1 - e^{-e^{\hat{\beta}_n^T Z_{(j)}} \hat{\Lambda}_n(Y_{(j)})}} - (1 - \Delta_{(j)}) \right) = 0,$$

for each $i = 0, 1, \cdots, n$, and thus

$$\sum_{j=1}^{n} w_{(j)} \xi_0(\hat{\Lambda}_n(Y_{(j)})) e^{\hat{\beta}_n^T Z_{(j)}} \left(\Delta_{(j)} \frac{e^{-e^{\hat{\beta}_n^T Z_{(j)}} \hat{\Lambda}_n(Y_{(j)})}}{1 - e^{-e^{\hat{\beta}_n^T Z_{(j)}} \hat{\Lambda}_n(Y_{(j)})}} - (1 - \Delta_{(j)}) \right) = 0.$$

Therefore, noting that $\mathbf{h}^* = \mathbf{h}^* \circ \Lambda_0^{-1} \circ \Lambda_0 = \xi_0 \circ \Lambda_0$, we can write

$$\begin{aligned} S_{2n}(\hat{\beta}_n, \hat{\Lambda}_n)[\mathbf{h}^*] &= \mathbb{P}_n\Big\{we^{\hat{\beta}_n^T Z}\mathbf{h}^*(Y)(\Delta r(Y, Z; \hat{\beta}_n, \hat{\Lambda}_n) - (1-\Delta))\Big\} \\ &= \mathbb{P}_n\Big\{we^{\hat{\beta}_n^T Z}(\xi_0 \circ \Lambda_0(Y) - \xi_0 \circ \hat{\Lambda}_n(Y))(\Delta r(Y, Z; \hat{\beta}_n, \hat{\Lambda}_n) \\ &\qquad - (1-\Delta))\Big\} \\ &= I_1 + I_2, \end{aligned}$$

where

$$\begin{aligned} I_1 &= (\mathbb{P}_n - P)\Big\{we^{\hat{\beta}_n^T Z}(\xi_0 \circ \Lambda_0(Y) - \xi_0 \circ \hat{\Lambda}_n(Y))(\Delta r(Y, Z; \hat{\beta}_n, \hat{\Lambda}_n) \\ &\qquad - (1-\Delta))\Big\}, \\ I_2 &= P\Big\{we^{\hat{\beta}_n^T Z}(\xi_0 \circ \Lambda_0(Y) - \xi_0 \circ \hat{\Lambda}_n(Y))(\Delta r(Y, Z; \hat{\beta}_n, \hat{\Lambda}_n) \\ &\qquad - (1-\Delta))\Big\}. \end{aligned}$$

We want to show that both I_1 and I_2 are of order $o_p(n^{-1/2})$. Let

$$\psi(x; \beta, \Lambda) = we^{\beta^T z}(\xi_0 \circ \Lambda_0(y) - \xi_0 \circ \Lambda(y))(\delta r(y, z; \beta, \Lambda) - (1-\delta)).$$

For any $\eta > 0$, it will be verified in Lemma II.11, given at the end of this proof, that the entropy number of the class of functions

$$\Psi_0(\eta) = \{\psi(x; \beta, \Lambda) : |\beta - \beta_0| + \|\Lambda - \Lambda_0\|_2 \leq \eta, \beta \in \mathcal{B}, \Lambda \in \Phi\}$$

is of order $1/\varepsilon$, and hence $\Psi_0(\eta)$ is a Donsker class. By Assumptions (B), (C), (D) and equation (2.9) we see that function $\psi(X; \beta, \Lambda)$ converges to $\psi(X; \beta_0, \Lambda_0) = 0$ in quadratic mean as $d((\beta, \Lambda), (\beta_0, \Lambda_0)) \to 0$. Then by Corollary 2.3.12 of Van Der Vart and Wellner (1996), we have

$$\sup_{\psi \in \Psi_0(Cn^{-1/3})} (\mathbb{P}_n - P)\psi(X; \beta, \Lambda) = o_p(n^{-1/2}),$$

which shows that I_1 is of order $o_p(n^{-1/2})$. On the other hand, since $\mathbf{h}^*$ has bounded derivative by Assumption (C), ξ_0 also has bounded derivative. Applying the Cauchy-Schwartz inequality and Theorem II.5 together with the fact that $E(w|X) = 1$, we obtain

$$\begin{aligned}
I_2 &= P\Big\{e^{\hat{\beta}_n^T Z}(\xi_0 \circ \Lambda_0(Y) - \xi_0 \circ \hat{\Lambda}_n(Y))(\Delta r(Y, Z; \hat{\beta}_n, \hat{\Lambda}_n) - (1 - \Delta))\Big\} \\
&= P\left\{e^{\hat{\beta}_n^T Z}(\xi_0 \circ \Lambda_0(Y) - \xi_0 \circ \hat{\Lambda}_n(Y))\frac{e^{-e^{\hat{\beta}_n^T Z \hat{\Lambda}_n(Y)}} - e^{-e^{\beta_0^T Z \Lambda_0(Y)}}}{1 - e^{-e^{\hat{\beta}_n^T Z \hat{\Lambda}_n(Y)}}}\right\} \\
&\leq C\Big\{P(\hat{\Lambda}_n(Y) - \Lambda_0(Y))^2\Big\}^{1/2} \times \Big\{P\Big(e^{\hat{\beta}_n^T Z}\hat{\Lambda}_n(Y) - e^{\beta_0' Z}\Lambda_0(Y)\Big)^2\Big\}^{1/2} \\
&= O_p(n^{-2/3}).
\end{aligned}$$

Hence I_2 is also $o_p(n^{1/2})$. The second equality above is obtained by an iterated conditional expectation argument in which the inner conditional expectation is calculated given (Y, Z) and $(\hat{\beta}_n, \hat{\Lambda}_n)$ is treated as fixed.

To verify condition A5, we consider the following classes of functions

$$\begin{aligned}
\Psi_1(\eta) = \Big\{&wl_1(\beta, \Lambda; x) - wl_1(\beta_0, \Lambda_0; x) : \\
&\qquad |\beta - \beta_0| + \|\Lambda - \Lambda_0\|_2 \leq \eta, \beta \in \mathcal{B}, \Lambda \in \Phi\Big\}, \\
\Psi_2(\eta) = \Big\{&wl_2(\beta, \Lambda; x)[\mathbf{h}^*] - wl_2(\beta_0, \Lambda_0; x)[\mathbf{h}^*] : \\
&\qquad |\beta - \beta_0| + \|\Lambda - \Lambda_0\|_2 \leq \eta, \beta \in \mathcal{B}, \Lambda \in \Phi\Big\},
\end{aligned}$$

for $\eta > 0$, where l_1 and l_2 are the scores for β and Λ, respectively. Given by Lemma II.11 that the entropy numbers of $\Psi_1(\eta)$ and $\Psi_2(\eta)$ are both of order $1/\eta$, we know that both $\Psi_1(\eta)$ and $\Psi_2(\eta)$ are Donsker, and hence condition A5 is satisfied.

Finally, by a Taylor expansion of $S_1(\beta, \Lambda)$ and $S_2(\beta, \Lambda)[\mathbf{h}^*]$ at (β_0, Λ_0), it is easy to see that condition A6 is satisfied with $\mu = 2$, and we have $\mu\gamma = 2 \times (1/3) > 1/2$.

Then by Theorem II.10 we have,

$$n^{1/2}(\hat{\beta}_n - \beta_0) = I^{-1}(\beta_0)n^{1/2}\mathbb{P}_n w\tilde{l}(\beta_0, \Lambda_0; X) + o_{p^*}(1) \to_d N(0, \Sigma)$$

as $n \to \infty$, where $\Sigma = I^{-1}(\beta_0)BI^{-1}(\beta_0)$ with

$$\begin{aligned} B &= E\left[w^2\{\tilde{l}(\beta_0, \Lambda_0; X)\}^{\otimes 2}\right] \\ &= E\left[E(w^2|X)\{\tilde{l}(\beta_0, \Lambda_0; X)\}^{\otimes 2}\right] \\ &= E\left[\left\{1 + \frac{1 - \pi_\alpha(\Delta, V)}{\pi_\alpha(\Delta, V)}\right\}\left\{\tilde{l}(\beta_0, \Lambda_0; X)\right\}^{\otimes 2}\right] \\ &\equiv I(\beta_0) + D, \end{aligned}$$

hence $\Sigma = I^{-1}(\beta_0) + I^{-1}(\beta_0)DI^{-1}(\beta_0)$.

The following is the lemma that has been used in the above proof. Its proof follows similarly to the proof of Lemma 7.1 in Huang (1996) with the uniform boundedness of w, Λ, Z and the derivative of ξ_0, hence is omitted here.

Lemma II.11. *For the above classes of functions $\Psi_0(\eta), \Psi_1(\eta)$ and $\Psi_2(\eta)$, we denote their L_2 covering numbers as $N_0(\varepsilon, \Psi_0, L_2(Q)), N_1(\varepsilon, \Psi_1, L_2(Q))$ and $N_2(\varepsilon, \Psi_2, L_2(Q))$, respectively. Then under Assumptions (A) to (E),*

$$\sup_Q N_i(\varepsilon, \Psi_i, L_2(Q)) \leq C_{1i}/\varepsilon^d e^{1/\varepsilon}, \quad i = 0, 1, 2,$$

hence for sufficiently small ε, the entropy numbers satisfy

$$\sup_Q \log N_i(\varepsilon, \Psi_i, L_2(Q)) \leq C_{2i}/\varepsilon, \quad i = 0, 1, 2,$$

where C_{1i} and C_{2i}, $i \in \{0, 1, 2\}$, are constants and Q runs through all probability measures.

2.7.7 Proof of Theorem II.7

The proof follows the same idea used in the proof of Theorem 5.8 in Van Der Vaart (2002). Define $m(\beta, \Lambda, \alpha; X) = w(\alpha)\log\{(p_{\beta,\Lambda} + p_{\beta_0,\Lambda_0})\}/2$. In the proof of Theorem II.3 we have showed that $(\beta_0, \Lambda_0, \alpha_0)$ is the unique maximizer of $Pm(\beta, \Lambda, \alpha_0; X)$.

Hence,

$$\sup_{(\beta,\Lambda):\, d((\beta,\Lambda),(\beta_0,\Lambda_0))>\delta} Pm(\beta,\Lambda,\alpha_0;X) < Pm(\beta_0,\Lambda_0,\alpha_0;X) \tag{2.19}$$

holds for every $\delta > 0$. By the definition of $(\tilde{\beta}_n, \tilde{\Lambda}_n)$, we have

$$\begin{aligned} \mathbb{P}_n m(\tilde{\beta}_n, \tilde{\Lambda}_n, \hat{\alpha}_n; X) &\geq \mathbb{P}_n m(\beta_0, \Lambda_0, \hat{\alpha}_n; X) \\ &= \mathbb{P}_n m(\beta_0, \Lambda_0, \alpha_0; X) + o_p(1), \end{aligned} \tag{2.20}$$

where the equality is obtained by Taylor expansion and the uniform boundedness of $\dot{w}(\alpha)$. From Theorem II.4 we know that the class of functions $\{m(\beta, \Lambda, \alpha_0; X) : (\beta, \Lambda) \in \Theta\}$ is Donsker and hence Glivenko-Cantelli. Thus from (2.19) and (2.20) we have

$$\begin{aligned} 0 &\leq Pm(\beta_0, \Lambda_0, \alpha_0; X) - Pm(\tilde{\beta}_n, \tilde{\Lambda}_n, \alpha_0; X) \\ &= \mathbb{P}_n m(\beta_0, \Lambda_0, \alpha_0; X) - \mathbb{P}_n m(\tilde{\beta}_n, \tilde{\Lambda}_n, \alpha_0; X) + o_p(1) \\ &\leq \mathbb{P}_n m(\tilde{\beta}_n, \tilde{\Lambda}_n, \hat{\alpha}_n; X) - \mathbb{P}_n m(\tilde{\beta}_n, \tilde{\Lambda}_n, \alpha_0; X) + o_p(1) \\ &= o_p(1), \end{aligned} \tag{2.21}$$

where the last step is again obtained by Taylor expansion and the uniform boundedness of $\dot{w}(\alpha)$. By inequality (2.19), for every $\delta > 0$ we have

$$\Big\{ d((\tilde{\beta}_n, \tilde{\Lambda}_n), (\beta_0, \Lambda_0)) \geq \delta \Big\} \subset \Big\{ Pm(\tilde{\beta}_n, \tilde{\Lambda}_n, \alpha_0; X) < Pm(\beta_0, \Lambda_0, \alpha_0; X) \Big\}$$

with the sequence of the events on the right going to a null event in view of (2.21), which yields the almost sure (thus in probability) convergence of $(\tilde{\beta}_n, \tilde{\Lambda}_n)$. $\square$

2.7.8 Proof of Theorem II.8

Let $\ell(\beta, \Lambda; X) = \log\{(p_{\beta,\Lambda} + p_{\beta_0,\Lambda_0})/2\}$ as before and $S_n(\beta, \Lambda) = \mathbb{P}_n w(\hat{\alpha}_n)\ell(\beta, \Lambda; X)$. Clearly $S_n(\tilde{\beta}_n, \tilde{\Lambda}_n) \geq S_n(\beta_0, \Lambda_0)$. A Taylor expansion on α at α_0 yields

$$S_n(\beta, \Lambda) = \mathbb{P}_n w(\alpha_0)\ell(\beta, \Lambda; X) + \mathbb{P}_n \dot{w}^T(\alpha_0)\ell(\beta, \Lambda; X)(\hat{\alpha}_n - \alpha_0) + (\hat{\alpha}_n - \alpha_0)^T \mathbb{P}_n \ddot{w}(\alpha_n^*)\ell(\beta, \Lambda; X)(\hat{\alpha}_n - \alpha_0), \tag{2.22}$$

where α_n^* is a point between α_0 and $\hat{\alpha}_n$. To apply Theorem 3.2.5 of Van Der Vart and Wellner (1996), we define $\mathbb{M}_n^0(\beta, \Lambda) = \mathbb{P}_n w(\alpha_0)\ell(\beta, \Lambda; X)$, $\mathbb{M}(\beta, \Lambda) = Pw(\alpha_0)\ell(\beta, \Lambda; X)$, and $\mathbb{M}_n(\beta, \Lambda) = M_n^0(\beta, \Lambda) + P\dot{w}^T(\alpha_0)\ell(\beta, \Lambda; X)(\hat{\alpha}_n - \alpha_0)$. Then by the uniform boundedness of $\ddot{w}$, it is easy to see that the third term on the right hand side of equality (2.22) is $O_p(n^{-1})$. Thus (2.22) becomes

$$S_n(\beta, \Lambda) = \mathbb{M}_n(\beta, \Lambda) + n^{-1/2}\Big\{\mathbb{G}_n \dot{w}^T(\alpha_0)\ell(\beta, \Lambda; X)\Big\}(\hat{\alpha}_n - \alpha_0) + O_p(n^{-1}).$$

Applying Theorem II.4 with w replaced by $\dot{w}^{(j)}$, $1 \leq j \leq J$, where J is the dimension of α, we know that the classes of functions $\{\dot{w}(\alpha_0)^{(j)}\ell(\beta, \Lambda; X) : \beta \in \mathcal{B}, \Lambda \in \Phi\}$, $1 \leq j \leq J$, are Donsker. Hence

$$\sup_{\beta,\Lambda} |\mathbb{G}_n \dot{w}^{(j)}(\alpha_0)\ell(\beta, \Lambda; X)| = O_p(1), \ 1 \leq j \leq J,$$

and we have $S_n(\beta, \Lambda) = \mathbb{M}_n(\beta, \Lambda) + O_p(n^{-1})$. The inequality $S_n(\tilde{\beta}_n, \tilde{\Lambda}_n) \geq S_n(\beta_0, \Lambda_0)$ then implies that $\mathbb{M}_n(\tilde{\beta}_n, \tilde{\Lambda}_n) \geq \mathbb{M}_n(\beta_0, \Lambda_0) - |O_p(n^{-1})|$, which further implies $\mathbb{M}_n(\tilde{\beta}_n, \tilde{\Lambda}_n)$ $\geq \mathbb{M}_n(\beta_0, \Lambda_0) - |O_p(r_n^{-2})|$ with $r_n = n^{1/3}$.

By the triangle inequality and the calculation in the proof of Theorem II.5, we

obtain

$$
\begin{aligned}
& E \sup_{d((\beta,\Lambda),(\beta_0,\Lambda_0))<\delta} \left| n^{1/2}(\mathbb{M}_n - \mathbb{M})(\beta,\Lambda) - n^{1/2}(\mathbb{M}_n - \mathbb{M})(\beta_0,\Lambda_0) \right| \\
& \quad \le E \sup_{d((\beta,\Lambda),(\beta_0,\Lambda_0))<\delta} \left| n^{1/2}(\mathbb{M}_n^0 - \mathbb{M})(\beta,\Lambda) - n^{1/2}(\mathbb{M}_n^0 - \mathbb{M})(\beta_0,\Lambda_0) \right| \\
& \qquad + E \sup_{d((\beta,\Lambda),(\beta_0,\Lambda_0))<\delta} \left| n^{1/2}(\mathbb{M}_n - \mathbb{M}_n^0)(\beta,\Lambda) - n^{1/2}(\mathbb{M}_n - \mathbb{M}_n^0)(\beta_0,\Lambda_0) \right| \\
& \quad \le C\delta^{1/2} \left(1 + M \frac{\delta^{1/2}}{\delta^2 n^{1/2}} \right) \\
& \qquad + \sum_{j=1}^{J} \sup_{d((\beta,\Lambda),(\beta_0,\Lambda_0))<\delta} |A^{(j)}(\beta,\Lambda) - A^{(j)}(\beta_0,\Lambda_0)| E n^{1/2} |\hat{\alpha}_{nj} - \alpha_{0j}|, \qquad (2.23)
\end{aligned}
$$

where $A^{(j)}$ is the jth component of $P\dot{w}(\alpha_0)\ell(\cdot\,,\cdot\,;X)$. Based on the assumptions on model (2.2) and the uniform boundedness of $\dot{w}(\alpha_0)$, we know that for $1 \le j \le J$,

$$
\begin{aligned}
& |A^{(j)}(\beta,\Lambda) - A^{(j)}(\beta_0,\Lambda_0)| \\
& \quad = |P\dot{w}^{(j)}(\alpha_0)\{\ell(\beta,\Lambda;X) - \ell(\beta_0,\Lambda_0;X)\}| \\
& \quad \le C_j \Big[|\beta - \beta_0| + \{P(\Lambda(Y) - \Lambda_0(Y))^2\}^{1/2} \Big] \\
& \quad = C_j d((\beta,\Lambda),(\beta_0,\Lambda_0)) \\
& \quad \le C_j \delta
\end{aligned}
$$

for some constant C_j. Together with the boundedness of $En^{1/2}|\hat{\alpha}_{nj} - \alpha_{0j}|$, the above inequality implies that the term (2.23) is bounded by $K\delta \le K\delta^{1/2}(1 + M\delta^{1/2}/(\delta^2 n^{1/2}))$ for a constant K and sufficiently small δ. Hence,

$$
\begin{aligned}
& E \sup_{d((\beta,\Lambda),(\beta_0,\Lambda_0))<\delta} \left| n^{1/2}(\mathbb{M}_n - \mathbb{M})(\beta,\Lambda) - n^{1/2}(\mathbb{M}_n - \mathbb{M})(\beta_0,\Lambda_0) \right| \\
& \qquad \le C^* \delta^{1/2} \left(1 + M \frac{\delta^{1/2}}{\delta^2 n^{1/2}} \right)
\end{aligned}
$$

for a constant C^*.

Finally, the inequality $\mathbb{M}(\beta,\Lambda) - \mathbb{M}(\beta_0,\Lambda_0) \le -Cd^2((\beta,\Lambda),(\beta,\Lambda)_0)$ has already been established in the proof of Theorem II.5. Thus the conditions of Theorem 3.2.5

of Van Der Vart and Wellner (1996) are all satisfied with the same function $\phi_n(\delta)$ as that derived in the proof of Theorem II.5. Hence, $(\tilde{\beta}_n, \tilde{\Lambda}_n)$ converges at the same rate as $(\hat{\beta}_n, \hat{\Lambda}_n)$, which is $n^{1/3}$. $\square$

2.7.9 Proof of Theorem II.9

We prove the theorem by checking Conditions A1 to A6 in II.10 with $m(\beta, \Lambda, \alpha; X) = w(\alpha)l(\beta, \Lambda; X)$. Condition A1 holds with $\gamma = 1/3$ by Theorem II.8. Conditions A2 and A3 have been verified in the proof of Theorem II.6. We now verify Condition A4.

The first part of A4 holds automatically since we have $S_{1n}(\tilde{\beta}_n, \tilde{\Lambda}_n, \hat{\alpha}_n) = 0$. For the second part, we also define $\xi_0 = \mathbf{h}^* \circ \Lambda_0^{-1}$ with $\mathbf{h}^*$ given in (2.9). Using the same argument as that in the proof of Theorem II.6 and taking a Taylor expansion with respect to α at α_0, we obtain

$$S_{2n}(\tilde{\beta}_n, \tilde{\Lambda}_n, \hat{\alpha}_n) = J_1 + (\hat{\alpha}_n - \alpha_0)^T J_2 + (\hat{\alpha}_n - \alpha_0)^T J_3 (\hat{\alpha}_n - \alpha_0),$$

where

$$J_1 = \mathbb{P}_n \Big\{ w(\alpha_0) e^{\tilde{\beta}_n^T Z} (\xi_0 \circ \Lambda_0(Y) - \xi_0 \circ \hat{\Lambda}_n(Y)) (\Delta r(Y, Z; \tilde{\beta}_n, \tilde{\Lambda}_n) - (1 - \Delta)) \Big\},$$

$$J_2 = \mathbb{P}_n \Big\{ \dot{w}(\alpha_0) e^{\tilde{\beta}_n^T Z} (\xi_0 \circ \Lambda_0(Y) - \xi_0 \circ \hat{\Lambda}_n(Y)) (\Delta r(Y, Z; \tilde{\beta}_n, \tilde{\Lambda}_n) - (1 - \Delta)) \Big\},$$

and

$$J_3 = \mathbb{P}_n \Big\{ \ddot{w}(\alpha_n^*) e^{\tilde{\beta}_n^T Z} (\xi_0 \circ \Lambda_0(Y) - \xi_0 \circ \hat{\Lambda}_n(Y)) (\Delta r(Y, Z; \tilde{\beta}_n, \tilde{\Lambda}_n) - (1 - \Delta)) \Big\}$$

for some α_n^* lying between α_0 and α_n. In the proof of Theorem II.6, we have shown that $J_1 = o_p(n^{-1/2})$. It is easy to see that $J_3 = O_p(1)$ by the boundedness assumptions, hence $(\hat{\alpha}_n - \alpha_0)^T J_3 (\hat{\alpha}_n - \alpha_0) = o_p(n^{-1/2})$ because $|\hat{\alpha}_n - \alpha_0| = O_p(n^{-1/2})$. We now show that $J_2 = o_p(1)$.

Let $J_2 = K_1 + K_2$, where

$$K_1 = (\mathbb{P}_n - P)\Big\{\dot{w}(\alpha_0)e^{\tilde{\beta}_n^T Z}(\xi_0 \circ \Lambda_0(Y) - \xi_0 \circ \hat{\Lambda}_n(Y))(\Delta r(Y, Z; \tilde{\beta}_n, \tilde{\Lambda}_n) - (1-\Delta)]\Big\},$$

$$K_2 = P\Big\{\dot{w}(\alpha_0)e^{\tilde{\beta}_n^T Z}(\xi_0 \circ \Lambda_0(Y) - \xi_0 \circ \hat{\Lambda}_n(Y))(\Delta r(Y, Z; \tilde{\beta}_n, \tilde{\Lambda}_n) - (1-\Delta))\Big\}.$$

Replacing w by $\dot{w}$, which is uniformly bounded, in the definition of function $\psi(x; \beta, \Lambda)$ and following the same calculation as that for Ψ_0, I_1 and I_2 in the proof of Theorem II.6, we obtain that both K_1 and K_2 are $o_p(1)$. Thus we have verified Condition A4.

To verify A5, it suffices to show that the classes of functions

$$\Psi_1^*(\eta) = \Big\{w(\alpha)l_1(\beta, \Lambda; x) - w(\alpha_0)l_1(\beta_0, \Lambda_0; x) : |\alpha - \alpha_0| + |\beta - \beta_0| + \|\Lambda - \Lambda_0\|_2 \leq \eta, \alpha \in R^J, \beta \in \mathcal{B}, \Lambda \in \Phi\Big\},$$

$$\Psi_2^*(\eta) = \Big\{w(\alpha)l_2(\beta, \Lambda; x)[\mathbf{h}^*] - w(\alpha_0)l_2(\beta_0, \Lambda_0; x)[\mathbf{h}^*] : |\alpha - \alpha_0| + |\beta - \beta_0| + \|\Lambda - \Lambda_0\|_2 \leq \eta, \alpha \in R^J, \beta \in \mathcal{B}, \Lambda \in \Phi\Big\}$$

are Donsker. This follows in a similar way as that in Lemma II.11.

Finally, A6 is verified by Taylor expansions of functions $S_1(\beta, \Lambda, \alpha)$ and $S_2(\beta, \Lambda, \alpha)[\mathbf{h}^*]$ at $(\beta_0, \Lambda_0, \alpha_0)$. We also have $\mu = 2$ and $\mu\gamma > 1/2$. When $\hat{\alpha}_n$ is efficient with influence function ℓ^α, then the last part of the Theorem follows from the result of Pierce (1982).

A geometric interpretation of the efficiency gain using estimated weights for the missing data problem is given in the following. Let $\dot{\mathcal{P}}_{\Lambda,\alpha}^{\perp}$ be the orthogonal complement of the tangent space of (Λ, α) in $L_2(P)$. Then the influence function of the regular asymptotic linear estimator $\tilde{\beta}_n$ is in $\dot{\mathcal{P}}_{\Lambda,\alpha}^{\perp}$. Since the score function (or equivalently the influence function) of $\hat{\alpha}_n$ for data missing at random is orthogonal to

$\dot{\mathcal{P}}^{\perp}_{\Lambda,\alpha}$, we know that $\hat{\alpha}_n$ is asymptotically independent of $\tilde{\beta}_n$, which yields the result given by Pierce (1982). For technical details of this simple interpretation, we refer to Bickel et al. (1993), Robinson et al. (1994) and Yu and Nan (2006). $\square$

Table 2.1: Summary statistics of simulations, with true parameter values $\beta_1 = 1$ and $\beta_2 = -1$. Scenario 1: $n = 500$, which yields about 170 completely observed subjects including about 100 failures. Scenario 2: $n = 3000$, which yields about 400 completely observed subjects including about 250 failures.

Method	Full Data MLE		True Weights		Estimated Weights	
Parameter	β_1	β_2	β_1	β_2	β_1	β_2
Scenario 1						
Bias	-0.022	0.016	-0.028	0.072	-0.033	0.074
Bootstrap Variance	0.020	0.021	0.033	0.036	–	–
Smoothing Variance	0.019	0.020	0.031	0.034	0.028	0.030
Empirical Variance	0.021	0.022	0.033	0.037	0.030	0.033
Bootstrap CP	0.946	0.946	0.953	0.946	–	–
Smoothing CP	0.940	0.945	0.941	0.923	0.941	0.925
Scenario 2						
Bias	0.013	-0.020	0.013	0.029	0.009	0.030
Bootstrap Variance	0.005	0.007	0.018	0.019	–	–
Smoothing Variance	0.006	0.007	0.020	0.018	0.015	0.012
Empirical Variance	0.006	0.007	0.019	0.020	0.013	0.014
Bootstrap CP	0.940	0.945	0.940	0.955	–	–
Smoothing CP	0.960	0.940	0.946	0.928	0.948	0.932

Table 2.2: Estimates of log hazard ratios for MN neutralizing titer (MN) and the baseline behavioral risk score.

Variable	MN	Medium Risk	High Risk
Estimate	-0.6544	0.8976	2.3854
Variance	0.1051	0.0628	0.2941
P-value	0.0435	0.0003	< 0.0001

Reference: the group with risk scores equal to 0
Medium Risk: the group with risk scores from 1 to 3
High Risk: the group with risk scores greater than 3

References

Anderson, P. K., Borgan Ø., Gill R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer, New York.

Banerjee, M. and Wellner J.A. (2001). Likelihood Ratio Tests for Monotone Functions. *Annals of Statistics* **29**, 1699-1731.

Banerjee, M., Biswas, P. and Ghosh, D. (2006). A Semiparametric Binary Regression Model Involving Monotonicity Constraints. *Scandinavian Journal of Statistics* **33,** 673-697.

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

Billingsley, P. (1999). *Convergence of Probability measures.* Wiley, New York.

Borgan, Ø., Langholz, B., Samuelsen, S. O., Doldstein, L., and Pogoda, J. (2000). Exposure stratified case-cohort designs. *Lifetime Data Analysis.* **6,** 39-58.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34,** 187-220.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis.* Jon Wiley, New York.

Gilbert, P. B., Peterson, M. L., Follmann, D., Hudgens, M. G., Francis, D. P., Gurwith, M., Heyward, W. L., Jobes, D. V. , Popovic, V., Self, S. G., Sinangil, F., Burke, D. and Berman, P. W. (2005). Correlation between immunologic responses to a recombinant glycoprotein 120 Vaccine and incidence of HIV-1 infection in a phase 3 HIV-1 preventive vaccine trial. *Journal of Infectious Diseases* **191,** 666-677.

Groeneboom, P. and Wellner, J.A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation.* Birkhäuser, Basel.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censored data. *Annals of Statistics* **24,** 540-568.

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

Kalbfleisch, J. D. and Prentice R. L., (2002). *The Statistical Analysis of Failure Time Data.* John Wiley, New York.

Little, R. J. A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data.* Jon Wiley, New York.

Li, Z., Gilbert, P. and Nan, B. (2007). Weighted likelihood method for grouped survival data in case-cohort studies with application to HIV vaccine trials. *Technical report, Dept. of Biostatistics, University of Michigan,* **70**

Lin, D. Y.. (2000). On fitting Cox's proportional hazards models to survey data. *Biometrika* **87,** 37-47.

Ma, S. and Kosorok, M. R. (2005). Robust semiparametric M-estimation and the weighted bootstrap. *Journal of Multivariate Analysis* **96,** 190-217.

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

Murphy, S. A. and Van Der Vaart, A.W. (2000). On profile likelihood. *Journal of the American Statistical Association* **95,** 449-465.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association* **33,** 101-116.

Pierce, D. A. (1982). The asymptotic effect of substituting estimators for parameters in certain types of statistics. *Annals of Statistics* **10,** 475-478.

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73,** 1-11.

Robertson, T., Wright, F. T. and Dykstra R. L., (1988). *Order-Restricted Statistical Inference*. Wiley, New York.

Robins, J. M., Rotnitzky, A. and Zhao L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of American Statistical Association.* **89,** 846-866.

Skinner, C. J., Holt D. and Smith, T. M. F. (eds.) (1989). *Analysis of Complex Surveys*. John Wiley & Sons, New York.

Van Der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Van Der Vaart, A.W. (2002). Semiparametric Statistics. *In Lectures on Probability and Statistics, Ecole d'ete de Saint-Flour XXIX - 1999, Lecture Notes in Mathematics* **1781,** 330-457.

Van Der Vaart, A.W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.

Wellner, J. A. and Breslow N. E. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics* **34,** 86-102.

Wellner, J. A. and Zhang Y. (2007). Two likelihood-based semiparametric estimation methods for panel count data with covariates. *Annals of Statistics* **,** in press.

CHAPTER III

Inference for Ordered Binomial Probabilities when the Truth Can Be on the Boundary

3.1 Introduction

In this chapter, we consider the inference of ordered binomial probabilities. We suppose that there is a binary outcome variable for which the probability of response depends on one or more ordered categorical covariates. In biomedical studies, sometimes it is reasonable or natural to assume that the probabilities are ordered according to the categorical covariates. In other words, they change monotonically as each of the categorical variables changes level. The problem can be conveniently described as inference of cell probabilities in a one-way (if there is one covariate) or two-way contingency table. There is a binomial trial in each of the cells and the probabilities of "success" are ordered in either way of the table. The goal is to estimate the probability of the response under each combination of levels of the covariates, while incorporating the order restriction of the parameters. The statistical problems with restricted parameters have a long history and a vast literature (see, for example, Barlow et al., 2002 and Robertson et al., 1988). The purpose of incorporating the order restriction is to gain efficiency of the estimator compared with the estimator ignoring the restriction. Taylor et al. (2007) considered this problem with two categorical covariates and compared various methods via simulations. The methods include an

"empirical" estimator which is the usual maximum likelihood estimator ignoring the order restriction (unrestricted MLE), the isotonic regression estimator, which is the maximum likelihood estimator under the order restriction (restricted MLE), and a Bayesian estimator in which the ordering is introduced through prior distributions. They found that utilizing Bayesian isotonic regression can improve efficiency and minimize bias and guarantee order restriction in a wide variety of scenarios. However, when the true parameters in two adjacent cells are close to each other, the biases become large and the variances are not correctly estimated by using the asymptotic variance formula of the estimator which ignores the order restriction. In the setting of testing hypotheses with ordered alternatives, Agresti and Coull (1996) presented two likelihood ratio tests for comparison of binomial proportions; Agresti and Coull (1998) presented the likelihood ratio test and Nair (1987) examined the properties of the so called cumulative chi-squared-type tests of such alternatives in contingency tables. See Agresti and Coull (2002) for a survey of ways of taking order restrictions into account in the analysis of contingency tables.

Since the restricted MLE is a very natural estimator, which guarantees that the order restriction is always satisfied, and improves efficiency, we are particularly interested in the inference based on this estimator. Due to the difficulty in inference when some of the true adjacent cell probabilities are equal or close to each other, we will focus on this particular situation. We attempt to construct confidence intervals of the cell probabilities that have robust performance whether the parameters in adjacent cells are close to each other or well separated, though our main goal is to handle the case in which some of the adjacent probabilities are equal or close. At first, we find that the difficulty arises because, when two adjacent probabilities are equal or close, the distribution of the estimator cannot be well approximated by

a normal distribution in the usual way. Actually, the asymptotic distributions of the estimator in these situations are even not normal. In the following section, we will derive the asymptotic distributions of the estimator and construct confidence intervals based on these asymptotic distributions. In Section 3.3, we consider several types of bootstrap confidence intervals, which can improve the performance of the confidence intervals based on the asymptotic distributions of the estimators.

3.2 Inference Based on Asymptotic Distributions of the Estimator

3.2.1 Asymptotic Distributions

At first, we assume that the categorical covariate is one dimensional and has two levels. Denote the binary outcome variable by Y, and the covariate by V. Assume that $p_1 = P(Y = 1|V = 1) \leq p_2 = P(Y = 1|V = 2)$. Suppose that there are n_i subjects and d_i events in the group with $V = i$, $i = 1, 2$. The restricted MLE of p_1 and p_2 are $\tilde{p}_{1n} = \min\left(d_1/n_1, (d_1 + d_2)/(n_1 + n_2)\right)$ and $\tilde{p}_{2n} = \max\left(d_2/n_2, (d_1 + d_2)/(n_1 + n_2)\right)$, respectively. This can be easily seen as follows. If $d_1/n_1 \leq d_2/n_2$, then $(d_1/n_1, d_2/n_2)$ maximizes the likelihood function in the restricted region of (p_1, p_2). If $d_1/n_1 > d_2/n_2$, then the maximizer of the likelihood function is on the boundary and hence $\tilde{p}_{1n} = \tilde{p}_{2n} = (d_1 + d_2)/(n_1 + n_2)$. When $p_1 = p_2$, the asymptotic distributions of $\tilde{p}_{1n}$ and $\tilde{p}_{2n}$ are not normal, as stated in the following theorem.

Theorem III.1. *Suppose that* $p_1 = p_2$ *and* $\lim_{n\to\infty} n_2/n_1 = c$, *then*

$$\sqrt{n_1}(\tilde{p}_{1n} - p_1) \to_d \min\left[W_1, \frac{1}{1+c}W_1 + \frac{\sqrt{c}}{1+c}W_2\right],$$

and

$$\sqrt{n_2}(\tilde{p}_{2n} - p_2) \to_d \max\left[W_2, \frac{\sqrt{c}}{1+c}W_1 + \frac{c}{1+c}W_2\right],$$

as $n \to \infty$, *where* W_1, W_2 *are independent and* $W_i \sim N(0, p_1(1-p_1)), i = 1, 2$.

Proof: By the continuous mapping theorem (Billingsley, 1999), and the fact that $\sqrt{n_1}(d_1/n_1 - p_1) \to_d W_1$ and $\sqrt{n_2}(d_2/n_2 - p_2) \to_d W_2$, it follows that

$$\begin{aligned}
\sqrt{n_1}(\tilde{p}_{1n} - p_1) &= \sqrt{n_1}\left[\min\left(\frac{d_1}{n_1}, \frac{d_1+d_2}{n_1+n_2}\right) - p_1\right] \\
&= \min\left[\sqrt{n_1}\left(\frac{d_1}{n_1} - p_1\right), \frac{n_1}{n_1+n_2}\sqrt{n_1}\left(\frac{d_1}{n_1} - p_1\right)\right. \\
&\qquad \left. + \frac{n_2}{n_1+n_2}\sqrt{\frac{n_1}{n_2}}\sqrt{n_2}\left(\frac{d_2}{n_2} - p_2\right)\right] \\
&\to_d \min\left[W_1, \frac{1}{1+c}W_1 + \frac{\sqrt{c}}{1+c}W_2\right].
\end{aligned}$$

Similarly,

$$\begin{aligned}
\sqrt{n_1}(\tilde{p}_{2n} - p_2) &= \sqrt{n_2}\left[\max\left(\frac{d_2}{n_2}, \frac{d_1+d_2}{n_1+n_2}\right) - p_2\right] \\
&= \max\left[\sqrt{n_2}\left(\frac{d_2}{n_2} - p_2\right), \frac{n_1}{n_1+n_2}\sqrt{\frac{n_2}{n_1}}\sqrt{n_1}\left(\frac{d_1}{n_1} - p_1\right)\right. \\
&\qquad \left. + \frac{n_2}{n_1+n_2}\sqrt{n_2}\left(\frac{d_2}{n_2} - p_2\right)\right] \\
&\to_d \max\left[W_2, \frac{\sqrt{c}}{1+c}W_1 + \frac{c}{1+c}W_2\right]. \quad \square
\end{aligned}$$

The theorem can be extended to higher dimensions. For example, if there are 3 ordered cell probabilities, say, p_1, p_2 and p_3, and $p_1 = p_2 = p_3$, the asymptotic distributions of the ordered MLEs $\tilde{p}_{1n}, \tilde{p}_{2n}, \tilde{p}_{3n}$ of p_1, p_2, p_3 follow in a similar way as the above theorem. As an example, we give the result for $\tilde{p}_{1n}$. The restricted MLE of p_1 is

$$\tilde{p}_{1n} = \begin{cases} \frac{d_1}{n_1}, & if \frac{d_1}{n_1} \le \frac{d_2}{n_2} \le \frac{d_3}{n_3}, or \frac{d_1}{n_1} \le \frac{d_2+d_3}{n_2+n_3}, \frac{d_2}{n_2} > \frac{d_3}{n_3}, \\ \frac{d_1+d_2}{n_1+n_2}, & if \frac{d_1}{n_1} > \frac{d_2}{n_2}, \frac{d_1+d_2}{n_1+n_2} \le \frac{d_3}{n_3}, \\ \frac{d_1+d_2+d_3}{n_1+n_2+n_3}, & if \frac{d_1}{n_1} > \frac{d_2}{n_2}, \frac{d_1+d_2}{n_1+n_2} > \frac{d_3}{n_3}, or \frac{d_2}{n_2} > \frac{d_3}{n_3}, \frac{d_1}{n_1} > \frac{d_2+d_3}{n_2+n_3}. \end{cases} \tag{3.1}$$

Suppose that W_1, W_2, W_3 are independent and $W_i \sim N(0, p_i(1-p_i))$, $i = 1, 2, 3$. Denote $c_2 = \lim_{n\to\infty} n_2/n_1$ and $c_3 = \lim_{n\to\infty} n_3/n_1$. The continuous mapping theorem and the fact that $\sqrt{n_i}(d_i/n_i - p_i) \to_d N(0, p_1(1-p_1))$, for $i = 1, 2, 3$ yield

$\sqrt{n_1}(\tilde{p}_{1n} - p_1) \rightarrow_d W$, where $W = W_1$, if $W_1 \leq W_2/\sqrt{c_2} \leq W_3/\sqrt{c_3}$, or $W_1 \leq (\sqrt{c_2}W_2 + \sqrt{c_3}W_3)/(c_2 + c_3)$ and $W_2/\sqrt{c_2} > W_3/\sqrt{c_3}$; $W = (W_1 + \sqrt{c_2}W_2)/\sqrt{1 + c_2}$, if $W_1 > W_2/\sqrt{c_2}, (\sqrt{c_3}W_1 + \sqrt{c_2 c_3}W_2)/(1 + c_2) \leq W_3$; $W = \left(W_1 + \sqrt{c_2}W_2 + \sqrt{c_3}W_3\right)/(1 + c_2 + c_3)$, if $W_1 > \sqrt{c_2}W_2, (\sqrt{c_3}W_1 + \sqrt{c_2 c_3}W_2)/(1 + c_2) > W_3$, or $W_3/\sqrt{c_3} > W_2/\sqrt{c_2}$ and $W_1 > (c_2 W_2 + \sqrt{c_3}W_3)/(c_2 + c_3)$. In principle, the asymptotic distribution of the restricted MLE in even higher dimensions can be derived analogously. We can still write out the explicit formula of the estimator by the pool adjacent violators algorithm and then write out the asymptotic distributions accordingly, but it becomes much more complicated with more parameters. We assume that there are $m (m \geq 4)$ ordered probabilities $p_1, p_2, \cdots, p_m$. In the special case where all $n_j, 1 \leq j \leq m$ are equal, or more generally, $\lim_{n\to\infty} n_j/n_1 = 1, 1 \leq j \leq m$, the asymptotic distribution of $(\tilde{p}_{1n}, \cdots, \tilde{p}_{mn})^T$ can be expressed in a simple form. Denote $T(\cdot)$ to be the function which transforms the unrestricted MLE $(\hat{p}_{1n}, \cdots, \hat{p}_{mn})^T$ to the restricted MLE $(\tilde{p}_{1n}, \cdots, \tilde{p}_{mn})^T$, that is, $(\tilde{p}_{1n}, \cdots, \tilde{p}_{mn})^T = T(\hat{p}_{1n}, \cdots, \hat{p}_{mn})$. Then it is easy to see that, when $p_1 = p_2 = \cdots = p_m$,

$$\sqrt{n_1}\begin{pmatrix} \tilde{p}_{1n} - p_1 \\ \vdots \\ \tilde{p}_{mn} - p_1 \end{pmatrix} \rightarrow_d T(W_1, \cdots, W_m),$$

as $n \to \infty$, where $W_1, \cdots, W_m$ are independent and $W_i \sim N(0, p_1(1-p_1)), 1 \leq i \leq m$.

However, the above results have limited application, since in practice it is rarely the case that $p_1 = p_2$, and we never know whether it is true. In order to approximate the distribution of the restricted MLE of the cell probabilities in a wider range of situations, we use a more general assumption than that $p_1 = p_2$, that is, we assume that $p_2 = p_1 + \Delta/\sqrt{n_1}$, where Δ is an unknown constant which controls the difference between p_1 and p_2. Under this assumption, it is easy to obtain the

asymptotic distributions of the restricted MLEs of p_1 and p_2. Towards this end, we first establish the following result. In the following theorem, we consider a series of binomial random variables with probabilities of "success" going to a constant p with a $\sqrt{n}$ rate. Note that for a particular n, there is only one binomial random variable, and here n has no connection with n_1 and n_2 mentioned above.

Theorem III.2. *Suppose that* $d_n \sim B(n, p_n), n \geq 1$, *where* $p_n = p + \frac{\Delta}{\sqrt{n}}$, p *and* Δ *are constants,* $0 \leq p \leq 1$, *and* $\Delta \geq 0$. *Under these assumptions, we have*

$$\sqrt{n}\left(\frac{d_n}{n} - p_n\right) \rightarrow_d N(0, p(1-p)),$$

as $n \rightarrow \infty$.

Proof: Suppose $S_i = \{$"success" in the ith Bernoulli trial$\}$, $1 \leq i \leq n$, and $d_n = \sum_{i=1}^{n} S_i$. Then we can write

$$\xi_n = \sqrt{n}\left(\frac{d_n}{n} - p_n\right) = \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} I(S_i) - p_n\right] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[I(S_i) - p_n\right].$$

Hence the characteristic function of ξ_n is

$$\begin{aligned}
\phi_n(t) &= E\exp(it\xi_n) = E\exp\left(it\frac{1}{\sqrt{n}}\sum_{j=1}^{n}[I(S_j) - p_n]\right) \\
&= \prod_{j=1}^{n} E\exp\left(\frac{it}{\sqrt{n}}[I(S_j) - p_n]\right) \\
&= \prod_{j=1}^{n} E\left\{1 + \frac{it}{\sqrt{n}}[I(S_j) - p_n] - \frac{t^2}{2n}[I(S_j) - p_n]^2 + O\left(\frac{t^3}{n^{3/2}}[I(S_j) - p_n]^3\right)\right\} \\
&= \prod_{j=1}^{n}\left\{1 - \frac{t^2}{2n}p_n(1-p_n) + O\left(\frac{t^3}{n^{3/2}}\right)\right\} \\
&= \prod_{j=1}^{n}\left\{1 - \frac{t^2}{2n}p(1-p) + O\left(\frac{1}{n^{3/2}}\right)\right\} \\
&\rightarrow \exp\left(-\frac{\sigma^2}{2}t^2\right), \quad \text{as} \quad n \rightarrow \infty,
\end{aligned}$$

where $\sigma^2 = p(1-p)$. This implies that $\sqrt{n}\,(d_n/n - p_n) \rightarrow_d N(0, p(1-p))$. □

Now we can derive the asymptotic distributions of $\tilde{p}_{1n}, \tilde{p}_{2n}$ under the above assumption.

Theorem III.3. *Under the assumption that* $p_2 = p_1 + \Delta/\sqrt{n_1}$ *and* $\lim_{n\to\infty} n_2/n_1 = c$, *we have*

$$\sqrt{n_1}(\tilde{p}_{1n} - p_1) \to_d \min\left(W_1, \frac{1}{1+c}W_1 + \frac{\sqrt{c}}{1+c}W_2 + \frac{c}{1+c}\Delta\right) \equiv U_1,$$

and

$$\sqrt{n_2}(\tilde{p}_{2n} - p_2) \to_d \max\left(W_2, \frac{\sqrt{c}}{1+c}W_1 + \frac{c}{1+c}W_2 - \frac{\sqrt{c}}{1+c}\Delta\right) \equiv U_2,$$

as $n \to \infty$.

Proof: The results follow since, by Theorem III.2,

$$\begin{aligned}
\sqrt{n_1}(\tilde{p}_{1n} - p_1) &= \min\left[\sqrt{n_1}\left(\frac{d_1}{n_1} - p_1\right), \sqrt{n_1}\left(\frac{d_1+d_2}{n_1+n_2} - p_1\right)\right] \\
&= \min\left[\sqrt{n_1}\left(\frac{d_1}{n_1} - p_1\right), \frac{n_1}{n_1+n_2}\sqrt{n_1}\left(\frac{d_1}{n_1} - p_1\right)\right. \\
&\qquad \left. + \frac{n_2}{n_1+n_2}\sqrt{n_1}\left(\frac{d_2}{n_2} - p_2\right) + \frac{n_2}{n_1+n_2}\Delta\right] \\
&\to_d \min\left(W_1, \frac{1}{1+c}W_1 + \frac{\sqrt{c}}{1+c}W_2 + \frac{c}{1+c}\Delta\right),
\end{aligned}$$

and

$$\begin{aligned}
\sqrt{n_2}(\tilde{p}_{2n} - p_2) &= \max\left[\sqrt{n_2}\left(\frac{d_2}{n_2} - p_2\right), \sqrt{n_2}\left(\frac{d_1+d_2}{n_1+n_2} - p_2\right)\right] \\
&= \max\left[\sqrt{n_2}\left(\frac{d_2}{n_2} - p_2\right), \frac{n_1}{n_1+n_2}\sqrt{n_2}\left(\frac{d_1}{n_1} - p_1\right)\right. \\
&\qquad \left. + \frac{n_2}{n_1+n_2}\sqrt{n_2}\left(\frac{d_2}{n_2} - p_2\right) - \frac{n_1}{n_1+n_2}\sqrt{\frac{n_2}{n_1}}\Delta\right] \\
&\to_d \max\left(W_2, \frac{\sqrt{c}}{1+c}W_1 + \frac{c}{1+c}W_2 - \frac{\sqrt{c}}{1+c}\Delta\right).
\end{aligned}$$

3.2.2 Construction of Confidence Intervals

We discuss the construction of confidence intervals for p_1 and p_2 under the restriction that $p_1 \leq p_2$. These confidence intervals are based on the asymptotic

distributions of the restricted MLE.

In view of the above theorems, the asymptotic distribution of the restricted MLE of $(p_1, p_2)^T$ is not normal when $p_1 = p_2$ or under the more general assumption $p_2 = p_1 + \Delta/\sqrt{n_1}$. The consequence is that when $p_1 = p_2$ or when p_1 and p_2 are close to each other, the standard normal approximation of the distributions of the estimates of p_1 and p_2 is not appropriate. The appropriate distribution to use is the one given in Theorem III.3. A naive method that approximates the distributions of $\tilde{p}_{1n}$ and $\tilde{p}_{2n}$ by normal distributions and estimate their standard errors by $\sqrt{\tilde{p}_{1n}(1-\tilde{p}_{1n})/n_1}$ and $\sqrt{\tilde{p}_{2n}(1-\tilde{p}_{2n})/n_2}$, respectively, does not make sense in principle.

Based on Theorem III.3, the following procedure of constructing confidence intervals of p_1, p_2 is proposed.

Step 1. After calculating the estimates $\tilde{p}_{1n}$ and $\tilde{p}_{2n}$, estimate Δ by $\hat{\Delta} = \sqrt{n_1}(\tilde{p}_{2n} - \tilde{p}_{1n})$.

Step 2. Let $c = n_2/n_1$. Generate $N = 1000$ i.i.d. samples of $W_1, W_2 \sim N(0, \tilde{p}_{1n})$ and calculate U_1, U_2. Based on the N samples of U_1 and U_2, calculate their 0.025 and 0.975 sample quantiles, respectively. Denote the 0.025 quantiles of U_1 and U_2 as $q_1(0.025)$ and $q_2(0.025)$, and their 0.975 quantiles as $q_1(0.975)$ and $q_2(0.975)$, respectively.

Step 3. The 95% confidence interval for p_1 is set to be $\left[\tilde{p}_{1n} - q_1(0.975)/\sqrt{n_1}, \tilde{p}_{1n} - q_1(0.025)/\sqrt{n_1}\right]$, and the 95% confidence interval for p_2 is $\left[\tilde{p}_{2n} - q_2(0.975)/\sqrt{n_2}, \tilde{p}_{2n} - q_2(0.025)/\sqrt{n_2}\right]$.

Simulation results show that the above confidence intervals give accurate coverage rates when the true values of p_1 and p_2 are equal or almost equal. However, when they are close to each other, but not too close, the coverage rates can be much lower

than the nominal 95% (see Tables 3.2-3.4).

To improve the performance of these confidence intervals, we add the following criterion similar to a test of $H_0 : p_1 = p_2$ to the above procedure. The first two steps remain unchanged. Let

$$T_n = \frac{|\tilde{p}_{1n} - \tilde{p}_{2n}|}{\sqrt{\frac{\tilde{p}_{1n}(1-\tilde{p}_{1n})}{n_1} + \frac{\tilde{p}_{2n}(1-\tilde{p}_{2n})}{n_2}}}.$$

Choose a constant $\delta > 0$, when $T_n \leq \delta$, then the third step is not changed; but when $T_n > \delta$, use the standard confidence intervals $\tilde{p}_{1n} \pm 1.96\sqrt{\tilde{p}_{1n}(1-\tilde{p}_{1n})/n_1}$ and $\tilde{p}_{2n} \pm 1.96\sqrt{\tilde{p}_{2n}(1-\tilde{p}_{2n})/n_2}$, respectively. The optimal choice of δ may depend on the ratio of n_1 and n_2. A relatively good choice can be found by trying a wide range of scenarios in simulation studies, and choose the δ that yields the best coverage rates of resulting confidence intervals. Our choice is $\delta = 0.3$ for sample sizes $n_1 = 50, n_2 = 100$. Simulation studies show that this kind of "optimal" choice of δ only depends on the sample sizes n_1 and n_2.

For one-sample i.i.d. data problems, Andrews (2000) pointed out that the bootstrap is not consistent if the parameter is on a boundary of the parameter space defined by linear or nonlinear constraints. As a remedy, he proposed four alternative methods to construct confidence intervals when this happens. His first method is analogous to the above method with a "test". He first defines $\{\eta_n : n \geq 1\}$ to be a sequence of positive random variables (possibly constant) that satisfies

$$P\left(\lim_{n\to\infty} \eta_n = 0 \quad \text{and} \lim\inf_{n\to\infty} \eta_n(n/(2\log\log n))^{1/2} > 1\right) = 1.$$

According to his first method, if $\tilde{p}_2 - \tilde{p}_1 \leq \eta_n$, then use the asymptotic distribution of $\tilde{p}_1$ and $\tilde{p}_2$ when $p_1 = p_2$ to construct confidence intervals for p_1 and p_2. Otherwise, use the usual normal asymptotic distribution of $\tilde{p}_1$ and $\tilde{p}_2$, i.e., the asymptotic distribution when $p_1 \neq p_2$, to construct confidence intervals for p_1 and p_2. In the

simulation study, we choose $\eta_n = 2\sqrt{(\log\log n)/n}$, which satisfies the requirement for η_n.

The naive method, the method that uses the correct asymptotic distribution without a test, and the method with a test, and Andrews (2000) first method (with $\eta_n = 2\sqrt{(\log\log n)/n}$) are compared in a simulation study (see Tables 3.2-3.4). The procedure with a test criterion, as well as Andrews' first method, improves the performance (empirical coverage rates) of the confidence intervals, although the overall performance is still not ideal. In the following section, we discuss some bootstrap methods that may yield better confidence intervals.

3.3 Bootstrap Confidence Intervals

For simplicity, we still suppose that there are two ordered probabilities. Assume that $d_1 \sim B(n_1, p_1), d_2 \sim B(n_2, p_2)$, and $p_1 \leq p_2$. Bootstrap methods (Efron and Tibshirani, 1993; Andrews, 2000) can be used to construct confidence intervals for p_1 and p_2. Since a Binomial random variable can be treated as the sum of a number of i.i.d. Bernoulli random variables, we can resample from the i.i.d. Bernoulli samples and hence use the nonparametric bootstrap, although only one observation of a Binomial random variable is available. A closer look at the nonparametric bootstrap and parametric bootstrap methods reveals that they are equivalent in our particular problem. Hence, in the following discussion we only consider the nonparametric bootstrap. We consider two types of confidence intervals constructed from the bootstrap sample, which differ in their ways of determining the end points of the intervals. Descriptions of the two types of confidence intervals are given below. Note that this is a two-sample problem and the bootstrap samples are generated in each individual sample respectively, which is different from bootstrapping from a single i.i.d. sample.

3.3.1 Bootstrap percentile confidence intervals

The confidence intervals constructed by the following procedure are called bootstrap percentile confidence intervals.

- Suppose $u_1, \cdots, u_{n_1} \sim Bernoulli(p_1), v_1, \cdots, v_{n_2} \sim Bernoulli(p_2)$. All of them are independent.

- Resample with replacement from $u_1, \cdots, u_{n_1}$ B times. For the kth sample $u_{1k}^*, \cdots, u_{n_1 k}^*$, calculate $d_{1k}^* = \sum_{i=1}^{n_1} u_{ik}^*, 1 \leq k \leq B$.

- Resample with replacement from $v_1, \cdots, v_{n_2}$ B times. For the kth sample $v_{1k}^*, \cdots, v_{n_2 k}^*$, calculate $d_{2k}^* = \sum_{i=1}^{n_2} v_{ik}^*, 1 \leq k \leq B$.

- Calculate

$$p_{1k}^* = \min\left(\frac{d_{1k}^*}{n_1}, \frac{d_{1k}^* + d_{2k}^*}{n_1 + n_2}\right), p_{2k}^* = \max\left(\frac{d_{2k}^*}{n_2}, \frac{d_{1k}^* + d_{2k}^*}{n_1 + n_2}\right), 1 \leq k \leq B.$$

- Calculate the 0.025 and 0.975 quantiles of $p_{11}^*, \cdots, p_{1B}^*$, denoted by $q_1(0.025)$ and $q_1(0.975)$. The 95% confidence interval for p_1 is $[q_1(0.025), q_1(0.975)]$. An analogous method is used to obtain confidence interval for p_2. Note that the closed interval is used.

- Repeat the above procedure $s = 1000$ times, and calculate the coverage rates of the confidence intervals.

3.3.2 Confidence intervals based on bootstrap "tables"

In this bootstrap method, we try to estimate the distribution of $\sqrt{n_1}(\tilde{p}_1 - p_1)$ and $\sqrt{n_2}(\tilde{p}_2 - p_2)$ by the distribution of $\sqrt{n_1}(p_1^* - \tilde{p}_1)$ and $\sqrt{n_2}(p_2^* - \tilde{p}_2)$, given $\tilde{p}_1$ and $\tilde{p}_2$, respectively, where p_1^* and p_2^* are bootstrap estimates of p_1 and p_2.

- Suppose $u_1, \cdots, u_{n_1} \sim Bernoulli(p_1), v_1, \cdots, v_{n_2} \sim Bernoulli(p_2)$. All of them are independent.

- Calculate $d_1 = \sum_{i=1}^{n_1} u_i, d_2 = \sum_{i=1}^{n_2} v_i$, and

$$\tilde{p}_1 = \min\left(\frac{d_1}{n_1}, \frac{d_1+d_2}{n_1+n_2}\right), \tilde{p}_2 = \max\left(\frac{d_2}{n_2}, \frac{d_1+d_2}{n_1+n_2}\right).$$

- Resample with replacement from $u_1, \cdots, u_{n_1}$ B times. For the kth sample $u_{1k}^*, \cdots, u_{n_1 k}^*$, calculate $d_{1k}^* = \sum_{i=1}^{n_1} u_{ik}^*, 1 \le k \le B$.

- Resample with replacement from $v_1, \cdots, v_{n_2}$ B times. For the kth sample $v_{1k}^*, \cdots, v_{n_2 k}^*$, calculate $d_{2k}^* = \sum_{i=1}^{n_2} v_{ik}^*, 1 \le k \le B$.

- Calculate

$$p_{1k}^* = \min\left(\frac{d_{1k}^*}{n_1}, \frac{d_{1k}^*+d_{2k}^*}{n_1+n_2}\right), p_{2k}^* = \max\left(\frac{d_{2k}^*}{n_2}, \frac{d_{1k}^*+d_{2k}^*}{n_1+n_2}\right),$$

  and define

$$z_{1k} = \sqrt{n_1}(p_{1k}^* - \tilde{p}_1), z_{2k} = \sqrt{n_2}(p_{2k}^* - \tilde{p}_2), 1 \le k \le B.$$

- Calculate the 0.025 and 0.975 quantiles of $z_{11}, \cdots, z_{1B}$, denoted by $q_1(0.025)$ and $q_1(0.975)$. The 95% confidence interval for p_1 is set to be $[\tilde{p}_1 - q_1(0.975)/\sqrt{n_1}, \tilde{p}_1 - q_1(0.025)/\sqrt{n_1}]$. Similarly do this for p_2.

- Repeat the above procedure $s = 1000$ times, and calculate the coverage rates of the confidence intervals over all the $s = 1000$ simulations.

3.3.3 A parametric bootstrap with parameter shrunk to the boundary

This method is proposed in Andrews (2000). It is proposed as the second remedy to the usual bootstrap method when parameters can be on the boundary of the parameter space. It is similar to the parametric bootstrap, but the parameter estimator

used to generate the bootstrap samples shrinks to the boundary of the parameter space. Let η_n be defined as above. In our case, we define $\tilde{p}_1 = d_1/n_1$ and $\tilde{p}_2 = d_2/n_2$ if $|d_1/n_1 - d_2/n_2| > \eta_n$ and $d_1/n_1 < d_2/n_2$, and define $\tilde{p}_1 = \tilde{p}_2 = (d_1 + d_2)/(n_1 + n_2)$ if $|d_1/n_1 - d_2/n_2| \leq \eta_n$ or $d_1/n_1 \geq d_2/n_2$. Now generate bootstrap samples using the distributions $B(n_1, \tilde{p}_1)$ and $B(n_2, \tilde{p}_2)$, respectively in a parametric bootstrap procedure.

3.4 Simulation Results

At first, we present simulation results to assess the method in Section 3.2.2. We pick $p_1 = 0.2, 0.5$ and 0.8, and for each fixed p_1, we consider a range of p_2 starting at p_1 and gradually increases to a value close to 1. The sample sizes are $n_1 = 50$ and $n_2 = 100$. For each combination of p_1 and p_2, we calculate the biases for both the unrestricted MLE and the restricted MLE and coverage rates of all types of 95% confidence intervals. In Tables 3.2 to 3.4, the confidence intervals compared include, the standard confidence interval based on the unrestricted MLE of p_1 and p_2 and a normal approximation of its distribution, the naive confidence interval based on the restricted MLE, treating its distribution as normal and using $\sqrt{\tilde{p}_1(1-\tilde{p}_1)/n_1}$ and $\sqrt{\tilde{p}_2(1-\tilde{p}_2)/n_2}$ as the standard errors of $\tilde{p}_1$ and $\tilde{p}_2$, respectively, the confidence intervals based on the asymptotic distribution of the restricted MLE with or without a "test", and Andrews' first method, with η_n chosen as above. In all scenarios, 1000 simulation runs are repeated. From Table 3.1, we see that the restricted MLE, which guarantees the order of the parameters, is more efficient than the unrestricted MLE, although in smaller sample sizes it leads to some negative bias for p_1 and some positive bias for p_2. Moreover, if we use the restricted MLE and approximate its distribution by a normal distribution, the resulting coverage rates of the confidence

intervals can be quite inaccurate when p_1 and p_2 are equal or very close, especially in the case $p_1 = 0.8$ (Tables 3.2 to 3.4). The confidence intervals based on the correct asymptotic distribution when the two probabilities are equal or very close improve the coverage rates where the naive method does not perform well. Nevertheless, they may yield coverage rates that are much lower than the nominal 95% rate when the two true probabilities are close to each other. The reason for this is that, for the method with a test, when p_1 and p_2 are close, neither the asymptotic distribution under the null $p_1 = p_2$, nor the normal distribution approximate the true distribution of the estimator well enough. For the method using the local alternative assumption, we see that when $p_1 = p_2$, then $\Delta = 0$ and the asymptotic distribution derived in Theorem III.3 is an appropriate approximation to the true distribution, according to the result in Theorem III.1; when p_1 and p_2 are well separated, then Δ is a large number, and the asymptotic distributions in Theorem III.3 reduce to normal distributions, which are also the correct (asymptotic) distribution for the estimator. However, when p_1 and p_2 are close to each other, the asymptotic distribution in Theorem III.3 is not a good approximation to the true distribution of the estimator for finite samples, although asymptotically and under the local alternative assumption, it is the correct distribution. Finally, the coverage rates for the confidence intervals based on the restricted MLE under the local alternative assumption $p_2 = p_1 + \Delta/\sqrt{n_1}$ can be quite inaccurate when p_1 and p_2 are far apart. This happens because we treat p_1 as constant, and hence the variances of W_1 and W_2 are $p_1(1 - p_1)$. Due to this formulation, when p_1 and p_2 are far apart, the variance of W_2 is not correctly estimated.

The simulation results of the bootstrap methods are presented in Tables 3.5 to 3.7. We list the coverage rates of the following types of confidence intervals: the

bootstrap percentile confidence interval based on unrestricted MLE, the bootstrap percentile confidence interval based on restricted MLE, the confidence interval based on bootstrap "tables", and Andrews' parametric bootstrap confidence interval. We run 1000 simulations, and in each simulation, we draw 1000 bootstrap samples of the original data. The results show that the percentile bootstrap and the Andrew's parametric bootstrap both work well even for small sample sizes ($n_1 = 20, n_2 = 40$), but the confidence intervals based on bootstrap "tables" does not perform well. This may be due to two reasons. First, according to Andrews (2000), it is impossible to consistently estimate the asymptotic distribution of $\sqrt{n_1}(\tilde{p}_{1n} - p_1)$ and $\sqrt{n_2}(\tilde{p}_{2n} - p_2)$, when $p_2 = p_1 + \Delta/\sqrt{n}$, or in other words, when p_1 and p_2 are close, the bootstrap estimate of the distributions may be not a good approximation to the true distributions. Second, according to Efron and Tibshirani (1993), for the confidence intervals based on bootstrap "tables" to work well, the quantities $\sqrt{n_1}(\tilde{p}_{1n} - p_1)$ and $\sqrt{n_1}(\tilde{p}_{1n} - p_1)$ should be pivotal quantities. However, in our case, their standard deviations still depend on the unknown parameters p_1 and p_2 and hence they are not pivotal quantities.

In Table 3.8, we did a further simulation study to assess the performance of several of the above mentioned methods, including the method based on the asymptotic distribution and a test and all the bootstrap method based on restricted MLE, in larger sample sizes, that is, $n_1 = 500$ and $n_2 = 1000$. The results show that the method based on the asymptotic distribution does not perform well even in a large sample like this (for example, the coverage rate for p_1 is too low when $p_1 = 0.5$ and $p_2 = 0.55$). The confidence interval based on bootstrap "tables" still has some problem, while the two other bootstrap methods both work well. Finally, the simulation results in Table 3.9 show that for sample sizes as small as $n_1 = 10, n_2 = 20$, the boot-

strap percentile confidence interval based on restricted MLE still works well when p_1 is not small. For comparison, we also show the results of the bootstrap percentile confidence interval based on the unrestricted MLE. They alos have lower than ideal coverage rates, indicating that the low coverage rates for the restricted MLE of p_1, when p_1 is small, is due mainly to the small sample size, rather than the restricted nature of the estimator.

Table 3.1: Biases of the restricted MLE and the unrestricted MLE: $n_1 = 50, n_2 = 100$.

	unrestricted MLE		restricted MLE	
	p_1	p_2	p_1	p_2
		bias		
$p_1 = 0.2, p_2 = 0.2$	0.0006	0.0000	-0.018	0.009
$p_1 = 0.2, p_2 = 0.22$	0.0003	-0.0011	-0.013	0.0054
$p_1 = 0.2, p_2 = 0.25$	-0.001	-0.001	-0.003	0.0000
$p_1 = 0.2, p_2 = 0.3$	-0.0001	0.0003	-0.002	0.001
$p_1 = 0.2, p_2 = 0.5$	-0.001	0.0005	-0.001	0.0005
$p_1 = 0.2, p_2 = 0.9$	0.001	0.0000	0.001	0.0000
$p_1 = 0.5, p_2 = 0.5$	-0.0016	-0.0010	-0.024	0.010
$p_1 = 0.5, p_2 = 0.52$	0.0008	-0.0000	-0.017	0.0086
$p_1 = 0.5, p_2 = 0.55$	-0.0012	-0.0005	-0.011	0.0044
$p_1 = 0.5, p_2 = 0.6$	-0.0005	0.0002	-0.004	0.002
$p_1 = 0.5, p_2 = 0.7$	0.0007	0.0006	0.0007	0.0006
$p_1 = 0.5, p_2 = 0.9$	-0.0008	-0.0014	-0.0008	-0.0014
$p_1 = 0.8, p_2 = 0.8$	0.0009	-0.0002	-0.018	0.009
$p_1 = 0.8, p_2 = 0.82$	0.0002	0.0016	-0.012	0.0075
$p_1 = 0.8, p_2 = 0.85$	0.0005	0.0006	-0.0048	0.0032
$p_1 = 0.8, p_2 = 0.9$	0.0003	-0.0003	-0.0007	0.0002
		ratio of empirical variances		
$p_1 = 0.2, p_2 = 0.2$	1	1	0.562	0.784
$p_1 = 0.2, p_2 = 0.22$	1	1	0.620	0.818
$p_1 = 0.2, p_2 = 0.25$	1	1	0.728	0.864
$p_1 = 0.2, p_2 = 0.3$	1	1	0.871	0.934
$p_1 = 0.2, p_2 = 0.5$	1	1	0.993	0.996
$p_1 = 0.2, p_2 = 0.9$	1	1	1	1
$p_1 = 0.5, p_2 = 0.5$	1	1	0.562	0.784
$p_1 = 0.5, p_2 = 0.52$	1	1	0.620	0.818
$p_1 = 0.5, p_2 = 0.55$	1	1	0.728	0.864
$p_1 = 0.5, p_2 = 0.6$	1	1	0.871	0.934
$p_1 = 0.5, p_2 = 0.6$	1	1	0.993	0.996
$p_1 = 0.5, p_2 = 0.9$	1	1	1	1
$p_1 = 0.8, p_2 = 0.8$	1	1	0.589	0.767
$p_1 = 0.8, p_2 = 0.82$	1	1	0.672	0.812
$p_1 = 0.8, p_2 = 0.85$	1	1	0.806	0.884
$p_1 = 0.8, p_2 = 0.9$	1	1	0.945	0.968

Table 3.2: Empirical coverage rates of 95% confidence intervals based on distributions of the estimators: $n_1 = 50, n_2 = 100$, and $p_1 = 0.2$.

		$p_2 = 0.2$	$p_2 = 0.22$	$p_2 = 0.25$	$p_2 = 0.3$	$p_2 = 0.5$	$p_2 = 0.9$
unrestricted MLE with	p_1	0.935	0.941	0.938	0.937	0.938	0.936
normal distribution	p_2	0.931	0.933	0.946	0.951	0.937	0.932
restricted MLE with	p_1	0.952	0.952	0.955	0.948	0.931	0.938
normal distribution	p_2	0.966	0.977	0.962	0.959	0.952	0.945
restricted MLE	p_1	0.937	0.922	0.890	0.881	0.927	0.939
without "test"	p_2	0.932	0.912	0.900	0.887	0.871	0.900
restricted MLE	p_1	0.938	0.943	0.913	0.912	0.939	0.936
with "test"	p_2	0.958	0.962	0.935	0.941	0.937	0.932
Andrews'	p_1	0.922	0.913	0.888	0.855	0.820	0.940
first method	p_2	0.931	0.900	0.868	0.846	0.826	0.939

restricted MLE without "test": CIs based on the asymptotic distribution of the restricted MLE under the local alternative assumption.
restricted MLE with "test": CIs based on the asymptotic distribution of the restricted MLE under the local alternative assumption or normal distribution, the choice of which depends on the result of a "test" of $H_0 : p_1 = p_2$.

Table 3.3: Empirical coverage rates of 95% confidence intervals based on distributions of the estimators: $n_1 = 50, n_2 = 100$, and $p_1 = 0.5$.

		$p_2 = 0.5$	$p_2 = 0.52$	$p_2 = 0.55$	$p_2 = 0.6$	$p_2 = 0.7$	$p_2 = 0.9$
unrestricted MLE with	p_1	0.940	0.930	0.938	0.937	0.940	0.937
normal distribution	p_2	0.945	0.944	0.940	0.944	0.953	0.934
restricted MLE with	p_1	0.975	0.976	0.974	0.963	0.942	0.938
normal distribution	p_2	0.967	0.958	0.961	0.955	0.957	0.931
restricted MLE	p_1	0.937	0.935	0.913	0.871	0.938	0.939
without "test"	p_2	0.960	0.939	0.927	0.926	0.965	0.997
restricted MLE	p_1	0.969	0.948	0.940	0.912	0.936	0.937
with "test"	p_2	0.959	0.955	0.947	0.943	0.950	0.934
Andrews'	p_1	0.947	0.926	0.888	0.854	0.821	0.796
first method	p_2	0.932	0.947	0.937	0.922	0.931	0.964

restricted MLE without "test": CIs based on the asymptotic distribution of the restricted MLE under the local alternative assumption.
restricted MLE with "test": CIs based on the asymptotic distribution of the restricted MLE under the local alternative assumption or normal distribution, the choice of which depends on the result of a "test" of $H_0 : p_1 = p_2$.

Table 3.4: Empirical coverage rates of 95% confidence intervals based on distributions of the estimators: $n_1 = 50, n_2 = 100$, and $p_1 = 0.8$.

		$p_2 = 0.8$	$p_2 = 0.82$	$p_2 = 0.85$	$p_2 = 0.9$
unrestricted MLE with	p_1	0.933	0.934	0.944	0.935
normal distribution	p_2	0.932	0.940	0.936	0.930
restricted MLE with	p_1	0.985	0.982	0.975	0.953
normal distribution	p_2	0.964	0.949	0.939	0.941
restricted MLE	p_1	0.955	0.925	0.876	0.875
without "test"	p_2	0.966	0.951	0.964	0.976
restricted MLE	p_1	0.972	0.960	0.935	0.915
with "test"	p_2	0.941	0.939	0.932	0.928
Andrews'	p_1	0.941	0.913	0.967	0.933
first method	p_2	0.961	0.959	0.965	0.946

restricted MLE without "test": CIs based on the asymptotic distribution of the restricted MLE under the local alternative assumption.
restricted MLE with "test": CIs based on the asymptotic distribution of the restricted MLE under the local alternative assumption or normal distribution, the choice of which depends on the result of a "test" of $H_0 : p_1 = p_2$.

Table 3.5: Comparison of coverage rates of 95% bootstrap confidence intervals: $p_1 = 0.2$.

		$p_1 = 0.2$ $p_2 = 0.2$	$p_1 = 0.2$ $p_2 = 0.22$	$p_1 = 0.2$ $p_2 = 0.25$	$p_1 = 0.2$ $p_2 = 0.3$	$p_1 = 0.2$ $p_2 = 0.5$	$p_1 = 0.2$ $p_2 = 0.9$
				$n_1 = 50, n_2 = 100$			
percentile bootstrap CI	p_1	0.950	0.958	0.953	0.952	0.952	0.952
based on unrestricted MLE	p_2	0.951	0.957	0.952	0.958	0.954	0.956
percentile bootstrap CI	p_1	0.916	0.942	0.954	0.952	0.952	0.952
based on restricted MLE	p_2	0.951	0.963	0.955	0.959	0.954	0.956
CI based on	p_1	0.876	0.880	0.835	0.831	0.882	0.879
bootstrap "tables"	p_2	0.947	0.969	0.987	0.981	0.989	0.832
Andrews' parametric	p_1	0.916	0.942	0.955	0.948	0.957	0.950
bootstrap CI	p_2	0.951	0.961	0.954	0.959	0.954	0.956
				$n_1 = 20, n_2 = 40$			
percentile bootstrap CI	p_1	0.923	0.921	0.925	0.924	0.922	0.920
based on unrestricted MLE	p_2	0.956	0.938	0.957	0.957	0.959	0.916
percentile bootstrap CI	p_1	0.911	0.911	0.925	0.924	0.923	0.920
based on restricted MLE	p_2	0.959	0.942	0.957	0.957	0.959	0.916
CI based on	p_1	0.910	0.912	0.890	0.887	0.888	0.900
bootstrap "tables"	p_2	0.937	0.961	0.987	0.984	0.985	0.797
Andrews' parametric	p_1	0.900	0.911	0.925	0.924	0.922	0.920
bootstrap CI	p_2	0.956	0.940	0.960	0.961	0.959	0.917

Table 3.6: Comparison of coverage rates of 95% bootstrap confidence intervals: $p_1 = 0.5$.

		$p_1 = 0.5$ $p_2 = 0.5$	$p_1 = 0.5$ $p_2 = 0.52$	$p_1 = 0.5$ $p_2 = 0.55$	$p_1 = 0.5$ $p_2 = 0.6$	$p_1 = 0.5$ $p_2 = 0.7$	$p_1 = 0.5$ $p_2 = 0.9$
				$n_1 = 50, n_2 = 100$			
percentile bootstrap CI	p_1	0.964	0.952	0.962	0.960	0.961	0.961
based on unrestricted MLE	p_2	0.958	0.956	0.951	0.953	0.959	0.960
percentile bootstrap CI	p_1	0.938	0.949	0.961	0.961	0.961	0.961
based on restricted MLE	p_2	0.951	0.952	0.949	0.953	0.959	0.960
CI based on	p_1	0.952	0.934	0.915	0.883	0.919	0.950
bootstrap "tables"	p_2	0.967	0.965	0.969	0.960	0.925	0.735
Andrews' parametric	p_1	0.937	0.949	0.962	0.961	0.964	0.960
bootstrap CI	p_2	0.952	0.952	0.952	0.953	0.957	0.958
				$n_1 = 20, n_2 = 40$			
percentile bootstrap CI	p_1	0.958	0.958	0.965	0.960	0.960	0.964
based on unrestricted MLE	p_2	0.954	0.938	0.962	0.950	0.957	0.923
percentile bootstrap CI	p_1	0.934	0.946	0.963	0.962	0.962	0.964
based on restricted MLE	p_2	0.948	0.941	0.958	0.949	0.957	0.934
CI based on	p_1	0.945	0.936	0.927	0.890	0.878	0.933
bootstrap "tables"	p_2	0.961	0.961	0.967	0.956	0.922	0.710
Andrews' parametric	p_1	0.938	0.947	0.963	0.960	0.961	0.964
bootstrap CI	p_2	0.949	0.929	0.956	0.952	0.959	0.920

Table 3.7: Comparison of coverage rates of 95% bootstrap confidence intervals: $p_1 = 0.8$.

$n_1 = 50$ $n_2 = 100$		$p_1 = 0.8$ $p_2 = 0.8$	$p_1 = 0.8$ $p_2 = 0.82$	$p_1 = 0.8$ $p_2 = 0.85$	$p_1 = 0.8$ $p_2 = 0.9$
		$n_1 = 50, n_2 = 100$			
percentile bootstrap CI	p_1	0.948	0.947	0.958	0.948
based on unrestricted MLE	p_2	0.958	0.949	0.960	0.957
percentile bootstrap CI	p_1	0.951	0.962	0.966	0.950
based on restricted MLE	p_2	0.938	0.941	0.951	0.954
CI based on	p_1	0.950	0.927	0.884	0.876
bootstrap "tables"	p_2	0.952	0.940	0.928	0.843
Andrews' parametric	p_1	0.953	0.964	0.968	0.950
bootstrap CI	p_2	0.938	0.938	0.953	0.952
		$n_1 = 20, n_2 = 40$			
percentile bootstrap CI	p_1	0.934	0.924	0.925	0.928
based on unrestricted MLE	p_2	0.960	0.928	0.940	0.928
percentile bootstrap CI	p_1	0.959	0.943	0.961	0.946
based on restricted MLE	p_2	0.942	0.929	0.933	0.925
CI based on	p_1	0.936	0.913	0.887	0.824
bootstrap "tables"	p_2	0.946	0.928	0.906	0.824
Andrews' parametric	p_1	0.956	0.962	0.962	0.945
bootstrap CI	p_2	0.943	0.930	0.933	0.923

Table 3.8: Coverage rates of confidence intervals when sample sizes are $n_1 = 500, n_2 = 1000$.

	$p_1 = 0.2$ $p_2 = 0.2$	$p_1 = 0.2$ $p_1 = 0.22$	$p_1 = 0.2$ $p_2 = 0.25$	$p_1 = 0.5$ $p_2 = 0.5$	$p_1 = 0.5$ $p_2 = 0.52$	$p_1 = 0.5$ $p_2 = 0.55$	$p_1 = 0.8$ $p_2 = 0.8$	$p_1 = 0.8$ $p_2 = 0.82$	$p_1 = 0.8$ $p_1 = 0.85$
CIs based on the asymptotic distribution of the restricted MLE and a test of $p_1 = p_2$									
p_1	0.955	0.923	0.939	0.973	0.932	0.919	0.973	0.940	0.940
p_2	0.958	0.938	0.939	0.968	0.944	0.941	0.957	0.959	0.950
percentile bootstrap CIs based on restricted MLE									
p_1	0.935	0.931	0.936	0.950	0.964	0.957	0.939	0.963	0.958
p_2	0.950	0.941	0.938	0.947	0.941	0.962	0.944	0.969	0.957
CIs based on bootstrap "tables"									
p_1	0.943	0.890	0.923	0.952	0.909	0.882	0.909	0.918	0.958
p_2	0.957	0.958	0.968	0.964	0.962	0.956	0.961	0.921	0.866
Andrews' parametric bootstrap CIs									
p_1	0.922	0.951	0.959	0.924	0.962	0.957	0.936	0.956	0.958
p_2	0.936	0.960	0.952	0.942	0.949	0.963	0.932	0.952	0.956

Table 3.9: Coverage rates of the bootstrap percentile confidence interval based on restricted MLE compared with the bootstrap percentile confidence interval based on unrestricted MLE, when sample sizes are $n_1 = 10, n_2 = 20$.

	$p_1 = 0.2$ $p_2 = 0.2$	$p_1 = 0.2$ $p_1 = 0.22$	$p_1 = 0.2$ $p_2 = 0.25$	$p_1 = 0.5$ $p_2 = 0.5$	$p_1 = 0.5$ $p_2 = 0.52$	$p_1 = 0.5$ $p_2 = 0.55$	$p_1 = 0.8$ $p_2 = 0.8$	$p_1 = 0.8$ $p_2 = 0.82$	$p_1 = 0.8$ $p_1 = 0.85$
				bootstrap percentile CI based on unrestricted MLE					
p_1	0.891	0.867	0.891	0.971	0.964	0.977	0.887	0.881	0.898
p_2	0.916	0.936	0.954	0.949	0.927	0.961	0.930	0.907	0.967
				bootstrap percentile CI based on restricted MLE					
p_1	0.870	0.865	0.885	0.942	0.943	0.960	0.952	0.960	0.953
p_2	0.956	0.939	0.957	0.942	0.920	0.946	0.930	0.901	0.932

3.5 Conclusion

When the true parameters satisfy an order restriction, the restricted MLE is shown to be more efficient than the unrestricted MLE. Moreover, confidence intervals based on the restricted MLE can be constructed using the bootstrap method and have good performances for small sample sizes such as $n_1 = 20$ and $n_2 = 40$. The percentile bootstrap confidence interval and the Andrews' parametric bootstrap confidence intervals have similar performances regarding empirical coverage rates of the intervals.

The bootstrap methods were investigated for the simplest case where there are only two ordered probabilities. However, the approaches can be generalized to problems with any higher dimension of parameters without any difficulty, such as for two-way or three-way contingency tables.

References

Agresti A. and Coull B. (1996). Order-restricted tests for stratified comparisons of binomial proportions. *Biometrics* **52**, 1103-1111.

Agresti A. and Coull B. (1998). Order-restricted inference for monotone trend alternatives in contingency tables. *Computational Statistics and Data Analysis* **28**, 139-155.

Agresti A. and Coull B. (2002). The analysis of contingency tables under inequality constrains. *Journal of Statistical Planning and Inference* **107(1-2)**, 45-73.

Andrews, D. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica* **68,** 399-405.

Barlow, R.E., Bartolomew D.J., Bremner J.M., and Brunk H.D. (2002). *Statistical Inference under Order Restrictions: The Theory and Applications of Isotonic Regression.* Wiley, New York.

Billingsley, P. (1999). *Convergence of Probability measures.* Wiley, New York.

Efron B. and Tibshirani R. (1993). *An Introduction to the Bootstrap.* Chapman and Hall.

Morris, M. (1988). Small-sample confidence limits for parameters under inequality constraints with application to quantal bioassay. *Biometrics* **44**, 1083-1092.

Nair, V., (1987). Chi-squared–type tests for ordered alternatives in contingency tables. *Journal of the American Statistical Association* **82**, 283-291.

Robertson, T., Wright, F. T. and Dykstra R. L., (1988). *Order-Restricted Statistical Inference.* Wiley, New York.

Rojo J. (2004). On the estimation of survival functions under a stochastic order constraint. *The First Eric L. Lehmann Symposium – Optimality, 37-61, Ims Lecture Notes Monogr. Ser. 44, Inst. Math. Statist., Beachwood, OH, 2004.*

Self S.G. and Liang K.Y. (1987). Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. *Journal of the American Statistical Association* **82,** 605-610.

Taylor, J.M.G., Wang L. and Li Z. (2007). Analysis on binary responses with ordered covariates and missing data. *Statistics in Medicine* **26,** 3443-3458.

CHAPTER IV

Future Work

For the missing data problems with grouped survival data or current status data, we assume that the data are independent and identically distributed. However, this assumption may not hold in some important practical situations. A common example is the two-phase stratified sampling with simple random sampling instead of Bernoulli sampling in the second stage. Since the simple random sampling is frequently used in practice, it is important to know the properties of the proposed weighted likelihood estimator in this case. Research in this direction is one of our future research plans. Breslow and Wellner (2007) considered the weighted likelihood estimator in a two-phase stratified sampling with simple random sampling in the second sampling stage, for a general semiparametric problem in which all the parameters are estimable at the $\sqrt{n}$ rate. They show that the weighted likelihood estimator is more efficient if the simple random sampling is used rather than the Bernoulli sampling in the second stage, and the asymptotic variance of the estimator with simple random sampling is equivalent to that of the estimator with estimated weights. Similar properties are expected to hold in our grouped survival data problem and current status data problem. For the grouped survival data problem, the theory in Breslow and Wellner (2007) may apply, since the model is parametric. However, the current status data

problem will be more challenging since it is a semiparametric problem, and part of the parameter is not estimable at the $\sqrt{n}$ rate. Another interesting future work in this setting is exploring the possibility of an analogue of the likelihood ratio test of the parameter β. Banerjee et al. (2007) studied such likelihood ratio tests in the semiparametric binary regression model involving monotonicity constraints, in the full data case, which is connected to the current status data problem with full data. It is worthwhile to explore such possibilities in the missing data case.

For the estimation of ordered probabilities of binomial random variables, we showed via simulation results that the usual percentile bootstrap confidence interval has good properties, and it is the most attractive confidence interval among all those considered. Andrews (2000) claimed that, in the one-sample, i.i.d. data case, the confidence interval based on bootstrap "tables" is not consistent when the parameter is on a boundary of the parameter space defined by linear or nonlinear constraints. However, for our particular problem, simulation results show that the percentile bootstrap confidence intervals have good performances both in smaller sample sizes ($n_1 = 20, n_2 = 40$) and larger sample sizes ($n_1 = 500, n_2 = 1000$), and the empirical coverage rates of the confidence intervals get very close to 95% in the latter case. Thus it seems that the percentile bootstrap confidence interval does work in our problem with parameter on the boundary. A future work to find out theoretical justification for this is desirable. In the setting of hypothesis testing with ordered alternatives, several methods are available, for example, see Morris (1988) and Nair (1987). Confidence intervals (or regions) of ordered parameters can be obtained, possibly by inverting such tests. We plan to investigate such confidence intervals (or regions) and compare them with the ones discussed above. Finally, for sample sizes smaller than $n_1 = 20, n_2 = 40$, such as $n_1 = 5, n_2 = 10$, both the confidence inter-

vals based on asymptotic distribution of the estimator and the bootstrap confidence intervals do not give correct coverage rates in this case. Confidence intervals based on exact distributions of the restricted MLE is possible and may potentially improve the performance of the bootstrap confidence intervals in such small samples. This will also be pursued in our future work.