

Semiparametric Regression Analysis under Imputation for Missing Response Data

Qihua Wang

*Institute of Applied Mathematics, Academy of Mathematics and System Science
Chinese Academy of Science, Beijing 100080, P. R. China*

Wolfgang Härdle

*Institut für Statistik und Ökonometrie, Humboldt-Universität
10178 Berlin, Germany*

Oliver Linton

*Department of Economics, London School of Economics
London WC2A 2AE, United Kingdom*

Abstract

We develop inference tools in a semiparametric regression model with missing response data. A semiparametric regression imputation estimator and an empirical likelihood based one for the mean of the response variable are defined. Both the estimators are proved to be asymptotically normal, with asymptotic variances estimated with Jackknife method. The empirical likelihood method is developed. It is shown that when missing responses are imputed using the semiparametric regression method the empirical log-likelihood is asymptotically a scaled chi-square variable or a weighted sum of chi-square variables with unknown weights in the absence of auxiliary information or in the presence of auxiliary information. An adjusted empirical log-likelihood ratio, which is asymptotically standard chi-square, is obtained. Also, a bootstrap empirical log-likelihood ratio is also derived and its distribution is used to approximate that of the imputed empirical log-likelihood ratio. A simulation study is conducted to compare the imputed, adjusted and bootstrap empirical likelihood with the normal approximation based methods in terms of coverage accuracies and average lengths of confidence intervals. Based on biases and standard errors, a comparison is also made by simulation between the proposed two estimators. The simulation indicates that the empirical likelihood methods developed perform competitively and the use of auxiliary information provides improved inference.

Key words and phrases: Asymptotic normality; Empirical likelihood; Semiparametric imputation.

Short Title. Semiparametric Imputation Regression Analysis

AMS 2000 subject classifications. Primary 62J99, Secondary 62E20.

1 Introduction

In many scientific areas, a basic task is to assess the simultaneous influence of several factors (covariates) on a quantity of interest (response variable). Regression models provide a powerful framework, and associated parametric, semiparametric and nonparametric inference theories are well established. However, in practice, often not all responses may be available for various reasons such as unwillingness of some sampled units to supply the desired information, loss of information caused by uncontrollable factors, failure on the part of investigator to gather correct information, and so forth. In this case, the usual inference procedures cannot be applied directly. A common method for handling missing data in a large dataset is to impute (i.e., fill in) a plausible value for each missing datum, and then analyze the result as if they were complete. Commonly used imputation methods for missing response include linear regression imputation (Yates (1993); Healy and Westmacott (1996)), kernel regression imputation (Cheng (1994)) and ratio imputation (Rao (1996)) and among others.

Let X be a d -dimensional vector of factors and Y be a response variable influenced by X . In practice, one often obtains a random sample of incomplete data

$$(X_i, Y_i, \delta_i), i = 1, 2, \dots, n, \quad (1.1)$$

where all the X_i 's are observed and $\delta_i = 0$ if Y_i is missing, otherwise $\delta_i = 1$. It is desired to estimate the mean of Y , say θ . This kind of sampling scheme can arise due to double or two-stage sampling, where first a complete sample of response and covariate variables is obtained and then some additional covariate values obtained, perhaps because it is expensive to acquire more Y 's.

Cheng (1994) applied kernel regression imputation to estimate the mean of Y , say θ . Cheng (1994) imputed every missing Y_i by kernel regression imputation and estimated θ by

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n (\delta_i Y_i + (1 - \delta_i) \widehat{m}_n(X_i)),$$

where $\widehat{m}_n(\cdot)$ is the Nadaraya-Watson kernel estimator based on (X_i, Y_i) for $i \in \{i : \delta_i = 1\}$. Under the assumption that the Y values are missing at random (MAR),

Cheng (1994) established the asymptotic normality of a trimmed version $\hat{\theta}$ and gave a consistent estimator of its asymptotic variance. In practice, however, it may be difficult to estimate θ well by the kernel regression imputation because the dimension of X may be high and hence the curse of dimensionality may occur. Although this does not affect the first order asymptotic theory, it does affect the practical performance of estimators, and the reliability of the asymptotic approximations; indeed, this effect shows up dramatically in the higher order asymptotics, see Linton (1995) for example. Wang and Rao (1999) considered the linear regression model and developed empirical likelihood method by filling in all the missing response values with linear regression imputation. In many practical situations, however, the linear model is not complex enough to capture the underlying relation between the response variables and its associated covariates. A natural generalization of the linear model is to allow only some of the predictors to be modelled linearly, with others being modelled nonlinearly. This motivates us to consider the following semiparametric regression model

$$Y_i = X_i^T \beta + g(T_i) + \epsilon_i, \quad (1.2)$$

where Y_i 's are i.i.d. scalar response variables, X_i 's are i.i.d. d -variable random covariate vectors, T_i 's are i.i.d. scalar covariates, the function $g(\cdot)$ is unknown and the model errors ϵ_i are independent with conditional mean zero given the covariates. The semiparametric regression model was introduced by Engle, Granger, Rice and Weiss (1986) to study the effect of weather on electricity demand. The implicit asymmetry between the effects of X and T may be attractive when X consists of dummy or categorical variables, as in Stock (1989, 1991). This specification arises in various sample selection models that are popular in econometrics, see Ahn and Powell (1993), and Newey, Powell, and Walker (1990). In fact, the partially linear model has also been applied in many other fields such as biometrics (see, e.g., Gray (1994)) and have been studied extensively for complete data settings (see, e.g., Heckman (1986), Rice (1986), Speckman (1988), Cuzick (1992a, b), Chen (1988) and Severini and Staniswalis (1994)).

In this paper, we are interested in inference on the mean of Y , say θ , under regression imputation of missing responses based on the semiparametric regression

model (1.2). For this model, we consider the case where some Y -values in a sample of size n may be missing, but X and T are observed completely. That is, we obtain the following incomplete observations

$$(Y_i, \delta_i, X_i, T_i), \quad i = 1, 2, \dots, n$$

from model (1.2), where all the X_i 's and T_i 's are observed and $\delta_i = 0$ if Y_i is missing, otherwise $\delta_i = 1$. Throughout this paper, we assume that Y is missing at random (MAR). The MAR assumption implies that δ and Y are conditionally independent given X and T . That is, $P(\delta = 1|Y, X, T) = P(\delta = 1|X, T)$. MAR is a common assumption for statistical analysis with missing data and is reasonable in many practical situations (see Little and Rubin (1987), Chapter 1). We propose an estimator of θ in the partially linear model that does not rely on high dimensional smoothing and thereby avoids the curse of dimensionality. We also develop empirical likelihood and bootstrap empirical likelihood methods that deliver better inference than standard asymptotic approximations. The empirical likelihood method, introduced by Owen (1988), has many advantages over normal approximation methods and the usual bootstrap approximation approaches for constructing confidence intervals when data are observed completely. How does empirical likelihood method work in the presence of missing responses for the semiparametric regression model? This is just one of the problems we need to investigate.

The outline of the paper is as follows. In Section 2, we define the estimator of θ and states the main results. Section 3 defines an improved estimator of θ and states the corresponding results when auxiliary information is available. In Section 4, an adjusted empirical log-likelihood ratio is derived and its asymptotic distribution is shown to be a standard chi-square with one degree of freedom. In Section 5, we define an adjusted empirical log-likelihood ratio, which is shown to be asymptotically distributed as a standard chi-square, when auxiliary information on X is available. In Section 6, a simulation study is conducted to calculate the bias and the standard errors of the proposed estimators and compare the finite sample properties of the proposed empirical likelihood methods with the normal approximation methods based on the different estimators. The proofs for the main results are delayed to the Appendix. We use " $\xrightarrow{\mathcal{L}}$ " to denote convergence in

distribution and " \xrightarrow{p} " to denote convergence in probability.

2 Semiparametric Imputation Estimator and Asymptotic Normality

Let $K(\cdot)$ be a kernel function and h_n be a bandwidth sequence tending to zero as $n \rightarrow \infty$. Let

$$W_{nj}(t) = \frac{K\left(\frac{t-T_j}{h_n}\right)}{\sum_{j=1}^n \delta_j K\left(\frac{t-T_j}{h_n}\right)}.$$

Let $\tilde{g}_{1n}(t) = \sum_{j=1}^n \delta_j W_{nj}(t) X_j$, $\tilde{g}_{2n}(t) = \sum_{j=1}^n \delta_j W_{nj}(t) Y_j$. Then, for every fixed β , the fact that $g(t) = E[Y - X^\tau \beta | T = t]$ suggests an estimator of $g(t)$ can be defined to be

$$\tilde{g}_{n0}(t, \beta) = \tilde{g}_{2n}(t) - \tilde{g}_{1n}^\tau(t) \beta, \quad (2.1)$$

based on the observed triples (X_i, T_i, Y_i) for $i \in \{i : \delta_i = 1\}$. The estimator of β is then defined as the one satisfying

$$\min_{\beta} \sum_{i=1}^n \delta_i (Y_i - X_i^\tau \beta - \tilde{g}_{n0}(T_i, \beta))^2 \quad (2.2)$$

From (2.2), it is easy to obtain that the estimator of β is given by

$$\hat{\beta}_n = \left[\sum_{i=1}^n \delta_i (X_i - \tilde{g}_{1,n}(T_i))(X_i - \tilde{g}_{1,n}(T_i))^\tau \right]^{-1} \sum_{i=1}^n \delta_i (X_i - \tilde{g}_{1,n}(T_i))(Y_i - \tilde{g}_{2,n}(T_i))$$

based on the observed triples (X_i, T_i, Y_i) for $i \in \{i : \delta_i = 1\}$. This is the Robinson (1988) estimator based on the complete subsample. The final estimator of $g(\cdot)$ is then given by

$$\hat{g}_n(t) = \tilde{g}_{2n}(t) - \tilde{g}_{1n}^\tau(t) \hat{\beta}_n$$

by replacing β in (2.1) by $\hat{\beta}_n$. By regression imputation, the estimator of θ is then defined to be

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n [\delta_i Y_i + (1 - \delta_i)(X_i^\tau \hat{\beta}_n + \hat{g}_n(T_i))] \quad (2.3)$$

Let $P_1(t) = P(\delta = 1|T = t)$, $P(x, t) = P(\delta = 1|X = x, T = t)$, $m(x, t) = x^\tau \beta + g(t)$, $\sigma^2(x, t) = E[(Y - X^\tau \beta - g(T))^2|X = x, T = t]$, $u(x, t) = P(x, t)(x - E[X|T = t])$, and $\Sigma = E[P(X, T)(X - E[X|T])(X - E[X|T])^\tau]$.

THEOREM 2.1. *Under all the assumptions listed in the Appendix except for condition (C.K)iii, we have*

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} N(0, V(\theta)),$$

where

$$\begin{aligned} V(\theta) = & E \left[\left(\frac{P(X, T)}{P_1(T)} \right)^2 \frac{\sigma^2(X, T)}{P(X, T)} \right] + \text{var}[m(X, T)] \\ & + E[u(X, T)^\tau] \Sigma^{-1} E \left[u(X, T) u(X, T)^\tau \frac{\sigma^2(X, T)}{P(X, T)} \right] \Sigma^{-1} E[u(X, T)] \\ & - 2E[u(X, T)^\tau] \Sigma^{-1} E \left[u(X, T) \frac{P(X, T)}{P_1(T)} \frac{\sigma^2(X, T)}{P(X, T)} \right]. \end{aligned}$$

There are a number of other estimators here that compete with ours in addition to the Cheng estimator that is also consistent here. First, $\tilde{\theta}_r = n^{-1} \sum_{i=1}^n [X_i^\tau \hat{\beta}_n + \hat{g}_n(T_i)]$, the average of the semiparametric regression function. It can be shown that $\tilde{\theta}_r$ has the same asymptotic distribution as $\hat{\theta}$. Second, the estimator $\tilde{\theta}_{HIR} = n^{-1} \sum_{i=1}^n Y_i \cdot \delta_i / \hat{P}(X_i, T_i)$ based on an estimator of the propensity score $\hat{P}(x, t)$ constructed by kernel smoothing the participation indicator against covariate values. This estimator is considered in Hirano, Imbens, and Ridder (2000); it is a version of propensity score matching, which is very popular in applied work. They show that $\tilde{\theta}_{HIR}$ achieves the semiparametric efficiency bound of Hahn (1998) [for the case where $m(x, t)$ is unrestricted], which is

$$V_{HIR} = E \left[\frac{\sigma^2(X, T)}{P(X, T)} \right] + \text{var}[m(X, T)].$$

This is exactly the same variance as obtains for the Cheng estimator (1994, Theorem 2.1). We rewrite the first line of $V(\theta)$ as

$$E \left[\left(\frac{P(X, T)}{P_1(T)} \right)^2 \frac{\sigma^2(X, T)}{P(X, T)} \right]$$

$$= E \left[\frac{\sigma^2(X, T)}{P(X, T)} \right] \left\{ 1 + E \left[\frac{\text{var}[P(X, T)|T]}{P_1^2(T)} \right] \right\} + \text{cov} \left[\left(\frac{P(X, T)}{P_1(T)} \right)^2, \frac{\sigma^2(X, T)}{P(X, T)} \right],$$

where the first two terms are positive but the last term can be negative. Also, the other terms in $V(\theta)$ could collectively be positive or negative, so there is no uniform ranking of the variances of the two estimators. In the special case that $\sigma^2(X, T) = \sigma^2(T)$ and $P(X, T) = P_1(T)$ we have

$$V(\theta) = E \left[\frac{\sigma^2(T)}{P_1(T)} \right] + \text{var}[m(X, T)],$$

which is the same as V_{HIR} . The disadvantage of $\tilde{\theta}_{HIR}$ here is that it requires a high-dimensional smoothing operation to compute the propensity score, and so its actual distribution may be very different from that predicted by the asymptotic theory due to the curse of dimensionality.

To define a consistent estimator of $V(\theta)$, we may first define estimators of $P(X, T)$, $P_1(T)$, $\sigma^2(X, T)$ and $E[X|T = t]$ by kernel regression method and then define a consistent estimator of $V(\theta)$ by “plug in” method. However, this method may be difficult to estimate $V(\theta)$ well when the dimension of X is high. Instead, take

$$\hat{V} = \frac{1}{n} \sum_{i=1}^n \hat{\eta}_i \hat{\eta}_i^\tau,$$

where, with $\hat{\epsilon}_i = Y_i - X_i^\tau \hat{\beta}_n - \hat{g}_n(T_i)$:

$$\hat{\eta}_i = \left[\frac{\delta_i}{\hat{P}_1(T_i)} + \hat{\Gamma}^\tau \hat{\Sigma}^{-1} \delta_i (X_i - \tilde{g}_{1n}(T_i)) \right] \hat{\epsilon}_i + (X_i^\tau \hat{\beta} + \hat{g}_n(T_i) - \hat{\theta})$$

$$\hat{\Gamma} = \frac{1}{n} \sum_{i=1}^n [(1 - \delta_i)(X_i - \tilde{g}_{1n}(T_i)); \quad \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n \delta_i (X_i - \tilde{g}_{1n}(T_i))(X_i - \tilde{g}_{1n}(T_i))^\tau].$$

It should be pointed out that this method uses an estimator of the main term of the asymptotic expansion of $\hat{\theta}_n - \theta$ (see (A.1) to construct asymptotic variance. Hence, it is not a natural method.

Another alternative is the jackknife variance estimator. Let $\hat{\theta}_n^{(-i)}$ is $\hat{\theta}$ based on $\{(Y_j, \delta_j, X_j, T_j)\}_{j=1}^n - \{(Y_i, \delta_i, X_i, T_i)\}$ for $i = 1, 2, \dots, n$. Let J_{ni} be the jackknife pseudo-values. That is,

$$J_{ni} = n\hat{\theta} - (n-1)\hat{\theta}_n^{(-i)}, \quad i = 1, 2, \dots, n$$

Then, the jackknife variance estimator can be defined as

$$\widehat{V}_{nJ} = \frac{1}{n} \sum_{i=1}^n (J_{ni} - \bar{J}_n)^2,$$

where $\bar{J}_n = \frac{1}{n} \sum_{i=1}^n J_{ni}$.

THEOREM 2.2. *Under assumptions of Theorem 2.1, we have*

$$\widehat{V}_{nJ} \xrightarrow{p} V(\theta).$$

By Theorem 2.1 and 2.2, the normal approximation based confidence interval with confidence level $1 - \alpha$ is $\widehat{\theta} \pm \sqrt{\frac{\widehat{V}_{nJ}}{n}} u_{1-\frac{\alpha}{2}}$, where $u_{1-\frac{\alpha}{2}}$ is the $1 - \frac{\alpha}{2}$ quantile of standard normal distribution.

3 Semiparametric Empirical Likelihood Based Estimator and Asymptotic Normality

In this section, we will construct an empirical likelihood based estimator to improve $\widehat{\theta}$ when auxiliary information on X is available. We assume that auxiliary information on X of the form

$$EA(X) = 0$$

is available, where $A(\cdot) = (A_1(\cdot), \dots, A_r(\cdot))^T$, $r \geq 1$ is a known vector (or scalar) function. For example, when the mean or median of X is known in the scalar X case.

To use the auxiliary information, we first maximize $\prod_{i=1}^n p_i$ subject to $\sum_{i=1}^n p_i = 1$ and $\sum_{i=1}^n p_i A(X_i) = 0$. Provided that the origin is inside the convex hull of $A(X_1), \dots, A(X_n)$, by the method of Lagrange multipliers, we get

$$p_i = \frac{1}{n} \frac{1}{1 + \zeta_n^T A(X_i)}$$

where ζ_n is the solution of the following equation

$$\frac{1}{n} \sum_{i=1}^n \frac{A(X_i)}{1 + \zeta_n^T A(X_i)} = 0. \quad (3.1)$$

An empirical likelihood-based semiparametric estimator (BLSE) of θ is then defined by

$$\hat{\theta}_{n,AU} = \sum_{i=1}^n p_i [\delta_i Y_i + (1 - \delta_i)(X_i^\tau \hat{\beta}_n + \hat{g}_n(T_i))]. \quad (3.2)$$

THEOREM 3.1. *Under the assumption of Theorem 2.1, if $EA(X)A^\tau(X)$ is a positive definite matrix, then we have*

$$\sqrt{n}(\hat{\theta}_{n,AU} - \theta) \xrightarrow{\mathcal{L}} N(0, V_{AU}(\theta))$$

where $V_{AU}(\theta) = V(\theta) - V_0(\theta)$ with

$$V_0(\theta) = E[(X^\tau \beta + g(T) - \theta)A(X)]^\tau (EA(X)A^\tau(X))^{-1} E[(X^\tau \beta + g(T) - \theta)A(X)]$$

and $V(\theta)$ defined in Theorem 2.1.

Clearly, $\hat{\theta}_{n,AU}$ is asymptotically more efficient than $\hat{\theta}$ due to the use of auxiliary information. Similar to the definition of $\hat{V}_{n,J}$, we can define a jackknife consistent variance estimator, say $\hat{V}_{n,J,AU}$, for $V_{AU}(\theta)$. Based on Theorem 3.1, the normal approximation based confidence interval is then defined to be $\hat{\theta}_{n,AU} \pm \sqrt{\frac{\hat{V}_{n,AU,J}}{n}} u_{1-\frac{\alpha}{2}}$.

4 Estimated, Adjusted and Bootstrap Empirical Likelihood

4.1 Estimated and adjusted empirical likelihood

In this section, we derive an adjusted empirical likelihood (ADEL) method to make global inference for θ . Let $\tilde{Y}_i = \delta_i Y_i + (1 - \delta_i)(X_i^\tau \beta + g(T_i))$. We have $E\tilde{Y}_i = \theta_0$ under the MAR assumption if θ_0 is the true value of θ . This implies that the problem of testing $H_0 : \theta = \theta_0$ is equivalent to testing $E\tilde{Y}_i = \theta_0$. If β and $g(\cdot)$ were known, then one could test $E\tilde{Y}_i = 0$ using the empirical likelihood of Owen (1990):

$$l_n(\theta) = -2 \sup \left\{ \sum_{i=1}^n \log(np_i) \mid \sum_{i=1}^n p_i \tilde{Y}_i = \theta, \sum_{i=1}^n p_i = 1, p_i > 0, i = 1, 2, \dots, n \right\}.$$

It follows from Owen (1990) that, under $H_0 : \theta = \theta_0$, $l_n(\theta)$ has an asymptotic central chi-square distribution with one degree of freedom. An essential condition for this result to hold is that the \tilde{Y}_i 's in the linear constraint are i.i.d. random variables.

Unfortunately, β and $g(\cdot)$ are unknown, and hence $l_n(\theta)$ cannot be used directly to make inference on θ . To solve this problem, it is natural to consider an estimated empirical log-likelihood by replacing β and $g(\cdot)$ with their estimators. Specifically, let $\hat{Y}_{in} = \delta_i Y_i + (1 - \delta_i)(X_i^\tau \hat{\beta}_n + \hat{g}_n(T_i))$. An estimated empirical log-likelihood evaluated at θ is then defined by

$$\hat{l}_n(\theta) = -2 \sup \left\{ \sum_{i=1}^n \log(np_i) \mid \sum_{i=1}^n p_i \hat{Y}_{in} = \theta, \sum_{i=1}^n p_i = 1, p_i > 0, i = 1, 2, \dots, n \right\}. \quad (4.1)$$

By using the Lagrange multiplier method, when $\min_{1 \leq i \leq n} \hat{Y}_{in} < \theta < \max_{1 \leq i \leq n} \hat{Y}_{in}$ with probability tending to one, $\hat{l}_n(\theta)$ can be shown to be

$$\hat{l}_n(\theta) = 2 \sum_{i=1}^n \log(1 + \lambda(\hat{Y}_{in} - \theta)), \quad (4.2)$$

where λ is the solution of the equation

$$\frac{1}{n} \sum_{i=1}^n \frac{(\hat{Y}_{in} - \theta)}{1 + \lambda(\hat{Y}_{in} - \theta)} = 0. \quad (4.3)$$

Unlike the standard empirical log-likelihood $l_n(\theta)$, $\hat{l}_n(\theta)$ is based on \hat{Y}_{in} 's that are not independent. Consequently, $\hat{l}_n(\theta)$ does not have an asymptotic standard chi-square distribution. Actually, $\hat{l}_n(\theta)$ is asymptotically distributed as a scaled chi-squared variable with one degree of freedom. Theorem 4.1 states the result.

THEOREM 4.1. *Assuming conditions of Theorem 2.1. Then, under $H_0 : \theta = \theta_0$,*

$$\hat{l}_n(\theta) \stackrel{\mathcal{L}}{\longrightarrow} \frac{V(\theta)}{\tilde{V}(\theta)} \chi_1^2,$$

where χ_1^2 is a standard chi-square variable with one degree of freedom, $V(\theta)$ is defined in Theorem 2.1 and $\tilde{V}(\theta)$ is defined in Lemma A.1.

By Theorem 4.1, we have under $H_0 : \theta = \theta_0$

$$\gamma(\theta) \hat{l}_n(\theta) \stackrel{\mathcal{L}}{\longrightarrow} \chi_1^2, \quad (4.4)$$

where $\gamma(\theta) = \tilde{V}(\theta)/V(\theta)$. If one can define a consistent estimator, say $\gamma_n(\theta)$, for $\gamma(\theta)$, an adjusted empirical log-likelihood ratio is then defined as

$$\hat{l}_{n,ad}(\theta) = \gamma_n(\theta) \hat{l}_n(\theta) \quad (4.5)$$

with adjustment factor $\gamma_n(\theta)$. It readily follows from (4.4) and (4.5), $\widehat{l}_{n,ad}(\theta_0) \stackrel{\mathcal{L}}{\longrightarrow} \chi_1^2$ under $H_0 : \theta = \theta_0$.

We now provide a consistent estimator $\gamma_n(\theta)$ of $\gamma(\theta)$. By Theorem 2.2 and Lemma A.1, a consistent estimator of $\gamma_n(\theta)$ can be defined as

$$\gamma_n(\theta) = \frac{\widetilde{V}_n(\theta)}{\widehat{V}_{nJ}}$$

where \widehat{V}_{nJ} is defined in Section 2 and

$$\widetilde{V}_n(\theta) = \frac{1}{n} \sum_{i=1}^n (\widehat{Y}_{in} - \theta)^2. \quad (4.6)$$

It should be pointed out that it may increase efficiency that we leave θ in $\gamma_n(\theta)$ not to be estimated.

THEOREM 4.2. *Assume the conditions in Theorem 2.1. Then, under $H_0 : \theta = \theta_0$*

$$\widehat{l}_{n,ad}(\theta_0) \stackrel{\mathcal{L}}{\longrightarrow} \chi_1^2.$$

From Theorem 4.2, it follows immediately that an approximation $1-\alpha$ confidence region for θ is given by

$$\{\theta : \widehat{l}_{n,ad}(\theta) \leq \chi_{1,\alpha}^2\}$$

where $\chi_{1,\alpha}^2$ is the upper α percentile of the χ_1^2 distribution. Theorem 4.2 can also be used to test the hypothesis $H_0 : \theta = \theta_0$. One could reject H_0 at level α if

$$\widehat{l}_{n,ad}(\theta_0) > \chi_{1,\alpha}^2.$$

4.2 Partially Smoothed Bootstrap Empirical Likelihood

Next, we develop a bootstrap empirical likelihood method. Let $\{(X_i^*, T_i^*, \delta_i^*, Y_i^*), 1 \leq i \leq m\}$ be the bootstrap sample from $\{(X_j, T_j, \delta_j, Y_j), 1 \leq j \leq n\}$. Let \widehat{Y}_{im}^* be the bootstrap analogy of $\{\widehat{Y}_{in}\}$. Then, the bootstrap analogy of $\widehat{l}_n(\theta)$ can be defined to be

$$\widehat{l}_m^*(\widehat{\theta}_n) = 2 \sum_{i=1}^m \log\{1 + \lambda_m^* (\widehat{Y}_{im}^* - \widehat{\theta}_n)\},$$

where λ^* satisfies

$$\frac{1}{m} \sum_{i=1}^m \frac{\widehat{Y}_{im}^* - \widehat{\theta}_n}{1 + \lambda^*(\widehat{Y}_{im}^* - \widehat{\theta}_n)} = 0.$$

To prove that the asymptotic distribution of $\widehat{l}_m^*(\widehat{\theta})$ approximates to that of $\widehat{l}_n(\theta)$ with probability one, we need that T_1^*, \dots, T_m^* have a probability density. This motivates us to use smooth bootstrap. Let $T_i^{**} = T_i^* + h_n \zeta_i$ for $i = 1, 2, \dots, m$, where h_n is the bandwidth sequence used in Section 2 and $\zeta_i, i = 1, 2, \dots, n$ are independent and identically distributed random variables with common probability density $K(\cdot)$, the kernel function in Section 2. We define $\widehat{l}_m^{**}(\widehat{\theta})$ to be $\widehat{l}_m^*(\widehat{\theta})$ with T_i^* replaced by T_i^{**} for $1 \leq i \leq m$. This method is termed as partially smoothed bootstrap since it used smoothed bootstrap sample only partially.

THEOREM 4.3. *Assuming conditions of Theorem 2.1 and condition (C.K)iii. Then, under $H_0 : \theta = \theta_0$, we have with probability one*

$$\sup_x |P(\widehat{l}_n(\theta) \leq x) - P^*(\widehat{l}_m^{**}(\widehat{\theta}_n) \leq x)| \rightarrow 0$$

as $n \rightarrow \infty$ and $m \rightarrow \infty$, where P^* denotes the bootstrap probability.

The bootstrap distribution of $\widehat{l}_m^{**}(\widehat{\theta}_n)$ can be calculated by simulation. The result of Theorem 4.3 can then used to construct a bootstrap empirical likelihood confidence interval for θ . Let c_α^* be the $1 - \alpha$ quantile of the distribution of $\widehat{l}_m^{**}(\widehat{\theta}_n)$. We can define a bootstrap empirical log-likelihood confidence region to be

$$\{\theta : \widehat{l}_n(\theta) \leq c_\alpha^*\}.$$

By Theorem 4.3, the bootstrap empirical likelihood confidence interval has asymptotically correct coverage probability $1 - \alpha$.

Compared to the estimated empirical likelihood and the adjusted empirical likelihood, an advantage of the bootstrap empirical likelihood is that it avoids estimating the unknown adjusting factor. This is especially attractive in some cases when the adjustment factor are difficult to estimate efficiently.

5 Estimated, Adjusted and Bootstrap Empirical likelihood with Auxiliary Information

5.1 Estimated and adjusted empirical likelihood

In this section, we develop an adjusted empirical likelihood method to construct confidence interval for θ when auxiliary information on X of the form $EA(X) = 0$ is available, where $A(X)$ is as defined in Section 3. This problem is to maximize $\prod_{i=1}^n np_i$ subject to $\sum_{i=1}^n p_i = 1$, $\sum_{i=1}^n p_i A(X_i) = 0$ and $\sum_{i=1}^n p_i (\hat{Y}_{in} - \theta) = 0$, where \hat{Y}_{in} is as defined in Section 4. An empirical log-likelihood evaluated at θ is then defined by

$$\hat{l}_{n,AU}(\theta) = -2 \sup \left\{ \sum_{i=1}^n \log(np_i) \mid \sum_{i=1}^n p_i h_{ni}(\theta) = 0, \sum_{i=1}^n p_i = 1, i = 1, 2, \dots, n \right\},$$

where $h_{ni}(\theta) = (A^\tau(X_i), \hat{Y}_{in} - \theta)^\tau$. Provided that the origin is inside the convex hull of points $h_{n1}(\theta), \dots, h_{nn}(\theta)$ with probability tending to one, the method of Lagrange multipliers may be used to show

$$\hat{l}_{n,AU}(\theta) = 2 \sum_{i=1}^n \log(1 + \eta_n^\tau h_{ni}(\theta)), \quad (5.1)$$

where η_n satisfies the following equation

$$\frac{1}{n} \sum_{i=1}^n \frac{h_{ni}(\theta)}{1 + \eta_n^\tau h_{ni}(\theta)} = 0. \quad (5.2)$$

Let $V_1(\theta) = E(A(X)A^\tau(X))$, $V_2(\theta) = E[(A(X)(X^\tau\beta + g(T) - \theta)]$, $V_3(\theta) = V_2(\theta)$, and let

$$V_{1,AU}(\theta) = \begin{pmatrix} V_1(\theta), & V_2(\theta) \\ V_2^\tau(\theta), & \tilde{V}(\theta) \end{pmatrix} \quad \text{and} \quad V_{2,AU}(\theta) = \begin{pmatrix} V_1(\theta), & V_3(\theta) \\ V_3^\tau(\theta), & V(\theta) \end{pmatrix},$$

where $V(\theta)$ and $\tilde{V}(\theta)$ are as defined in Theorem 2.1 and Lemma A.1 respectively.

THEOREM 5.1. *Assume conditions of Theorem 2.1. If $EA(X)A^\tau(X)$ is a positive definite matrix, then, under $H_0 : \theta = \theta_0$*

$$\hat{l}_{n,AU}(\theta) \stackrel{\mathcal{L}}{\longrightarrow} w_1 \chi_{1,1}^2 + \dots + w_{r+1} \chi_{1,r+1}^2,$$

where the weights w_i for $1 \leq i \leq d+1$ are the eigenvalues of $V_{0,AU}(\theta) = V_{1,AU}^{-1}(\theta)V_{2,AU}(\theta)$, and $\chi_{1,i}^2$ for $1 \leq i \leq d+1$ are independent χ_1^2 variables.

To apply Theorem 5.1 to construct confidence intervals for θ , we must estimate the unknown weights w_i consistently. Let $V_{n1}(\theta) = \frac{1}{n} \sum_{i=1}^n A(X_i)A^\tau(X_i)$, $V_{n2}(\theta) = \frac{1}{n} \sum_{i=1}^n A(X_i)(\hat{Y}_{in} - \theta)$, $V_{n3}(\theta) = \frac{1}{n} \sum_{i=1}^n A(X_i)(X_i^\tau \hat{\beta}_n + \hat{g}_n(T_i) - \theta)$ and denote $V_{n1,AU}(\theta)$ and $V_{n2,AU}(\theta)$ to be $V_{1,AU}(\theta)$ and $V_{2,AU}(\theta)$ with $V_1(\theta)$, $V_2(\theta)$, $V_3(\theta)$, $\tilde{V}(\theta)$ and $V(\theta)$ in $V_{1,AU}(\theta)$ and $V_{2,AU}(\theta)$ replaced by $V_{n1}(\theta)$, $V_{n2}(\theta)$, $V_{n3}(\theta)$, $\tilde{V}_n(\theta)$ and $\hat{V}_n(\theta)$ respectively.

By the ‘‘plug in’’ method, $V_{1,AU}(\theta)$ and $V_{2,AU}(\theta)$ can be estimated consistently by $V_{n1,AU}(\theta)$ and $V_{n2,AU}(\theta)$ respectively. This implies that the eigenvalues of $V_{n0,AU}(\theta) = V_{n1,AU}^{-1}(\theta)V_{n2,AU}(\theta)$, say \hat{w}_i , estimate w_i consistently for $i = 1, 2, \dots, r+1$. Let \hat{c}_α be the $1 - \alpha$ quantile of the conditional distribution of the weighted sum $\hat{S}_n = \hat{w}_1 \chi_{1,1}^2 + \dots + \hat{w}_{1+r} \chi_{1,r+1}^2$ given the data. Then, the confidence interval for θ with asymptotically correct coverage probability $1 - \alpha$ can be defined to be

$$I_{\alpha,AU}(\theta) = \{\theta : \hat{l}_{n,AU}(\theta) \leq \hat{c}_\alpha\}.$$

In practice, the conditional distribution of the weighted sum \hat{S}_n given data $\{(X_i, T_i, Y_i, \delta_i)_{i=1}^n\}$ can be obtained using Monte Carlo simulation by repeatedly generating independent samples $\chi_{1,1}^2, \dots, \chi_{1,r+1}^2$ from χ_1^2 distribution. Following Rao & Scott (1981), the distribution of $\tilde{r}(\beta) \left(\sum_{i=1}^{r+1} w_i \chi_{1,i}^2 \right)$ can be approximated by $\chi_{\tilde{r}+1}^2$, where $\tilde{r}(\beta) = (r+1)/tr\{V_{0,AU}(\theta)\}$ and $tr(A)$ denotes the trace of a certain matrix A . This implies that the asymptotic distribution of $\tilde{l}_{n,AU}(\theta) = \tilde{r}_n(\theta)\hat{l}_n(\theta)$ can be approximated by $\chi_{\tilde{d}}^2$ by Theorem 5.1 and the consistency of $V_{n1,AU}(\theta)$ and $V_{n2,AU}(\theta)$, where $\tilde{r}_n(\theta) = (r+1)/tr\{V_{n0,AU}(\theta)\}$. However, this provides only approximation distribution of the asymptotic distribution and this accuracy of this approximation depends on the values of w'_i s. Next, we give an adjusted empirical log-likelihood whose asymptotic distribution is exactly a standard chi-squares. Note that

$$\tilde{r}_n(\theta) = \frac{tr\{V_{n2,AU}^{-1}(\theta)V_{n2,AU}(\theta)\}}{tr\{V_{n1,AU}^{-1}(\theta)V_{n2,AU}(\theta)\}}. \quad (5.3)$$

By examining the asymptotic expansion of $\hat{l}_{n,AU}(\theta)$, we replace $V_{n2,AU}(\theta)$ in (5.3) by

$H_n(\theta) = \left(\frac{1}{n} \sum_{i=1}^n h_{ni}(\theta)\right) \left(\frac{1}{n} \sum_{i=1}^n h_{ni}(\theta)\right)^\tau$ and get a different adjustment factor

$$\hat{r}_n(\theta) = \frac{\text{tr} \left\{ V_{n2,AU}^{-1}(\theta) H_n(\theta) \right\}}{\text{tr} \left\{ V_{n1,AU}^{-1}(\theta) H_n(\theta) \right\}}.$$

By replacing $\tilde{r}_n(\theta)$ in $\tilde{l}_{n,AU}(\theta)$ by $\hat{r}_n(\theta)$, we can define an adjusted empirical log-likelihood by

$$\hat{l}_{ad,AU}(\theta) = \hat{r}_n(\theta) \hat{l}_{n,AU}(\theta).$$

The following theorem proves that $\hat{l}_{ad,AU}(\theta)$ is asymptotically standard χ^2 .

THEOREM 5.2. *Assume the conditions of Theorem 5.1. Then, under $H_0 : \theta = \theta_0$,*

$$\hat{l}_{ad,AU}(\theta) \xrightarrow{\mathcal{L}} \chi_d^2.$$

Based on Theorem 5.2, $l_{ad,AU}(\theta)$ can be used to construct a confidence interval for θ , $\{\theta : \hat{l}_{ad,AU}(\theta) \leq \chi_{p,\alpha}^2\}$, where $\chi_{p,\alpha}^2$ is the upper α percentile of the χ_p^2 distribution.

5.2 Partially smoothed bootstrap empirical likelihood

Let $\{(X_i^*, T_i^*, \delta_i^*, Y_i^*), 1 \leq i \leq m\}$ be the bootstrap sample from $\{(X_j, T_j, \delta_j, Y_j), 1 \leq j \leq n\}$. Similar to Subsection 4.2, the partially smoothed bootstrap analogy of $\hat{l}_{n,AU}(\theta)$ can be defined to be

$$\hat{l}_{m,AU}^{**}(\hat{\theta}_n) = 2 \sum_{i=1}^m \log \{1 + \eta_m^{**\tau} (h_{mi}^{**}(\hat{\theta}_n))\},$$

where $h_{mi}^{**}(\hat{\theta}_n) = (A^\tau(X_i^*), \hat{Y}_{im}^{**} - \hat{\theta}_n)^\tau$, \hat{Y}_{im}^{**} is the \hat{Y}_{im}^* with T_i^* in it replaced by T_i^{**} , where \hat{Y}_{im}^* and T_i^{**} are as defined in Subsection 4.2 for $i = 1, 2, \dots, m$. the partially smoothed bootstrap bootstrap analogy of \hat{Y}_{in}^* as defined in Subsection 4.2 and η_m^* satisfies

$$\frac{1}{m} \sum_{i=1}^m \frac{h_{mi}^{**}(\hat{\theta}_n)}{1 + \eta_m^{**\tau} h_{mi}^{**}(\hat{\theta}_n)} = 0.$$

Theorem 5.3. *Assuming conditions of Theorem 4.3. If $EA(X)A^\tau(X)$ is a positive definite matrix, then, under $H_0 : \theta = \theta_0$, we have with probability one*

$$\sup_x |P(\hat{l}_{n,AU}(\theta) \leq x) - P^*(\hat{l}_{m,AU}^{**}(\hat{\theta}_n) \leq x)| \rightarrow 0$$

as $n \rightarrow \infty$ and $m \rightarrow \infty$, where P^* denotes the bootstrap probability.

Similar to Theorem 4.3, Theorem 5.3 can be used to define the confidence interval for θ . Let $c_{\alpha,AU}^*$ be the $1 - \alpha$ quantile of the distribution of $\hat{l}_m^{**}(\hat{\theta}_n)$. We can define a bootstrap empirical likelihood confidence interval to be $I_{\alpha,AU}^*$ with $\hat{l}_{n,AU}(\theta)$ and \hat{c}_α in $I_{\alpha,AU}$ replaced by $\hat{l}_m^{**}(\hat{\theta}_n)$ and $c_{\alpha,AU}^*$ respectively. Then, by Theorem 5.3, the bootstrap empirical likelihood confidence interval, $I_{AU,\alpha}^*$, has asymptotically correct coverage probability $1 - \alpha$.

6 Simulation Results

In this section, we conducted simulation to understand the finite-sample performance of the proposed estimators and estimated, adjusted and bootstrap empirical likelihood methods. We compare the three empirical likelihood methods with the normal approximation-based methods in terms of coverage accuracies of confidence intervals in the two cases where auxiliary information is available or not.

The simulation used the partial linear model $Y = X\beta + g(T) + \epsilon$ with X and T simulated from the normal distribution with mean 1 and variance 1 and the uniform distribution $U[0, 1]$ respectively, and ϵ generated from the standard normal distribution, where $\beta = 1.5$, $g(t) = 3.2t^2 - 1$ if $t \in [0, 1]$, $g(t) = 0$ otherwise. The kernel function was taken to be

$$K(t) = \begin{cases} \frac{15}{16}(1 - 2t^2 + t^4), & -1 \leq t \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

and the bandwidth h_n was taken to be $n^{-2/3}$.

We generated 1000 Monte Carlo random samples of size $n=30, 60$ and 100 based on the following three cases respectively:

Case 1: $P(\delta = 1|X = x, T = t) = 0.8 + 0.2(|x - 1| + 1 - T)$ if $|x - 1| + 1 - T \leq 1$, and 0.95 elsewhere;

Case 2: $P(\delta = 1|X = x, T = t) = 0.9 - 0.2|x - 1| + 1 - T$ if $|x - 1| + 1 - T \leq 4$, and 0.1 elsewhere;

Case 3: $P(\delta = 1|X = x, T = t) = 0.6$ for all x and t .

The average missing rates corresponding to the above three cases are approximately $0.10, 0.25$ and 0.40 respectively. For nominal confidence level $1 - \alpha = 0.95$,

using the simulated samples, we calculated the coverage probabilities and the average lengths of the confidence intervals, which are reported in Tables 1 and 2. From the 5000 simulated values of $\hat{\theta}_n$ and $\hat{\theta}_{n,AU}$, we calculated the biases and standard errors of the two estimators. These simulated results are reported in Table 3.

For convenience, in what follows AEL and AAUEL represent the adjusted empirical likelihood confidence interval given in Subsection 4.1 and Subsection 5.1 respectively. BEL and BAUEL denote the smoothed bootstrap empirical likelihood confidence intervals given in Subsection 4.2 and 5.2 respectively. AUEL denotes the estimated empirical likelihood confidence interval given in Subsection 5.1. NA and NAAU denote the normal approximation based confidence intervals given in Section 2 and 3 respectively. The auxiliary information $EX = 1$ was used when we calculated the empirical coverages and average lengths of AUEL, BAUEL, AAUEL and NAAU.

Insert Tables 1 and 2 here

From Tables 1 and 2, we observe the following:

(1) BAUEL, NAAU, AAUEL and AUEL achieve higher coverage accuracies but similar or shorter average lengths than AEL, BEL and NA. This suggests the use of auxiliary improves inference.

(2) BAUEL do perform competitively in comparison to AUEL, AAUEL and NAAU since BAUEL have generally higher coverage accuracies but only slightly bigger average lengths. NAAU has higher slightly coverage accuracy than AUEL and AAUEL. But. it does this using much longer intervals. This implies that AUEL and AAUEL might be preferred over NAAU. This also applies to the comparison between NA and AEL.

(3) BEL has generally higher coverage accuracy, but bigger slightly average length than AEL and NA as $n = 60$ and 100 . This suggests, for $n = 60$ and 100 , BEL perform relatively better. For $n = 30$, AEL might be preferred since it has much smaller average length and the coverage accuracy is also not so low.

(4) All the coverage accuracies increase and the average lengths decreases as n increase for every fixed missing rate. Clearly, the missing rate also affects the

coverage accuracy and average length. Generally, the coverage accuracy decreases and average length increases as the missing rate increases for every fixed sample size.

Insert Table 3 here

From Table 3, we observe:

(a) Biases and SE decrease as n increases for every fixed censoring rate. Also, SE increases with missing rate for every fix sample size n .

(b) $\hat{\theta}_{n,AU}$ has not only smaller SE but also smaller bias than $\hat{\theta}_n$. This further suggests that the use of auxiliary information improve inference.

7 Concluding Remarks

We have proposed a new method for estimating the average effect parameter in a semiparametric model with missing response data. Our estimator is not generally efficient but has the considerable practical advantage of not requiring high dimensional smoothing operations. Our simulation results confirm the enhanced performance of the various empirical likelihood and bootstrap procedures that were used to obtain inference.

8 Appendix: Assumptions and Proofs of Theorems

Appendix: Assumptions and Proofs of Theorems

Let $g_1(t) = E[X|T = t]$, $g_2(t) = E[Y|T = t]$. Denote by $g_{1r}(\cdot)$ the r th component of $g_1(\cdot)$. Let $\|\cdot\|$ be the Euclid norm. The following assumptions are needed for the asymptotic normality of $\hat{\theta}_n$.

(C.X): $\sup_t E[\|X\|^2|T = t] < \infty$,

(C.T): The density of T , say $r(t)$, exists and satisfies

$$0 < \inf_{t \in [0,1]} r(t) \leq \sup_{t \in [0,1]} r(t) < \infty.$$

(C.Y): $\sup_{x,t} E[Y^2|X = x, T = t] < \infty$.

(C.g): $g(\cdot), g_{1r}(\cdot)$ and $g_2(\cdot)$ satisfy Lipschitz condition of order 1.

(C.P₁): i: $P_1(t)$ has bounded partial derivatives up to order 2 almost surely.

ii: $\inf_{x,t} P(x, t) > 0$.

(C.Σ) $\Sigma = E[P(X, T)(X - E[X|T])(X - E[X|T])^\tau]$ is a positive definite matrix.

(C.K)i: There exist constant $M_1 > 0, M_2 > 0$ and $\rho > 0$ such that

$$M_1 I[|u| \leq \rho] \leq K(u) \leq M_2 I[|u| \leq \rho].$$

ii: $K(\cdot)$ is a kernel function of order 2.

iii: $K(\cdot)$ has bounded partial derivatives up to order 2 almost surely.

(C.h_n): $nh_n \rightarrow \infty$ and $nh_n^2 \rightarrow 0$.

REMARK: Condition (C.T) implies that T is a bounded random variable on $[0, 1]$. (C.K)i implies that $K(\cdot)$ is a bounded kernel function with bounded support.

Sketch of Proof of Theorem 2.1 Standard arguments can be used to prove

$$\hat{\theta}_n - \theta = \frac{1}{n} \sum_{i=1}^n \eta(Y_i, \delta_i, X_i, T_i) + o_p(n^{-\frac{1}{2}}), \quad (\text{A.1})$$

where

$$\begin{aligned} \eta(Y_i, \delta_i, X_i, T_i) &= \left\{ \frac{\delta_i}{P_1(T_i)} + E[(1 - \delta)(X - E[X|T])^\tau] \Sigma^{-1} \delta_i (X_i - E[X_i|T_i]) \right\} \epsilon_i \\ &\quad + (X_i^\tau \beta + g(T_i) - \theta). \end{aligned}$$

By central limit theorem and some direct calculation, Theorem 2.1 is then proved.

Sketch of Proof of Theorem 2.2. Similar to (A.1), we can get

$$\hat{V}_{nJ} = \frac{1}{n} \sum_{i=1}^n (\eta(Y_i, \delta_i, X_i, T_i) - \frac{1}{n} \sum_{i=1}^n \eta(Y_i, \delta_i, X_i, T_i))^2 + o_p(1).$$

This proves $\hat{V}_{nJ} \xrightarrow{p} V(\theta)$.

Sketch of Proof of Theorem 3.1 $\hat{\theta}_{n,AU}$ can be represented as

$$\hat{\theta}_{n,AU} = \frac{1}{n} \sum_{i=1}^n \hat{Y}_{in} - E[(X^\tau \beta + g(T) - \theta) A^\tau(X)] (E A(X) A^\tau(X))^{-1} \left(\frac{1}{n} \sum_{i=1}^n A(X_i) \right) + o_p(n^{-\frac{1}{2}}). \quad (\text{A.3})$$

It is easy to get

$$Cov\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n(\widehat{Y}_{in} - \theta), \frac{1}{\sqrt{n}}\sum_{i=1}^n A(X_i)\right) \longrightarrow E[A^\tau(X)(X^\tau\beta + g(T) - \theta)]. \quad (\text{A.4})$$

(A.3) and (A.4) together prove Theorem 3.1

Sketch of Proofs of Theorem 4.1 and 4.2. Standard arguments can be used to prove

$$\widehat{l}_n(\theta) = \widetilde{V}_n^{-1}(\theta) \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n (\widehat{Y}_{in} - \theta) \right]^2 + o_p(1). \quad (\text{A.5})$$

and $\widetilde{V}_n(\theta) \xrightarrow{p} \widetilde{V}(\theta)$, where $\widetilde{V}(\theta)$ is defined in Theorem 4.1. This together with Theorem 2.1 and 2.2 proves Theorem 4.1.

Recalling the definition of $\widehat{l}_{n,ad}(\theta)$, by (A.5) we get

$$\widehat{l}_{n,ad}(\theta) = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\widehat{Y}_{in} - \theta}{\sqrt{\widehat{V}_{nJ}}} \right)^2 + o_p(1). \quad (\text{A.6})$$

Hence, (A.6), Theorem 2.1 and Theorem 2.2 together prove Theorem 4.2.

Sketch of Proof of Theorem 4.3 Under assumptions (C.X), (C.T), (C.Y), (C.P₁), (C.Σ) and (C.K)iii, standard arguments can be used to prove with probability 1: (i) $\sup_t E^*[\|X^*\|^2 | T^{**} = t] < \infty$; (ii) $0 < \inf_{t \in [0,1]} r_n(t) \leq \sup_{t \in [0,1]} r_n(t) < \infty$; (iii) $\sup_{x,t} E^*[Y^* | X^* = x, T^{**} = t] < \infty$; (iv) $\inf_{x,t} P^*(\delta^* = 1 | X^* = x, T^{**} = t) > 0$; (v) $\Sigma^* = E^*\{P((X^*, T^{**})(X^* - E^*[X^* | T^{**}])(X^* - E^*[X^* | T^{**}]])\}^\tau$ is a positive definite matrix; (vi): $P_1^*(t) = P^*(\delta^* = 1 | T^{**} = t)$ has bounded partial derivatives up to order 2 almost surely. By (i)–(vi), conditions (C.g), (C.K) and (C.h_n) and similar arguments to those used in the proof of Theorem 4.1, we can prove that along almost all sample sequences, given $(X_i, T_i, Y_i, \delta_i)$ for $1 \leq i \leq n$, as m and n go to infinity $\widehat{l}_m^*(\widehat{\theta}_n)$ has the same asymptotic scaled chi-square distribution as $\widehat{l}_n(\theta)$. This together with Theorem 4.1 proves Theorem 4.3.

Sketch of Proofs of Theorem 5.1 and 5.2 By Lemma A.4(b) and Lagrange multiplier method, (5.1) and (5.2) follows from the definition of $\widehat{l}_{n,AU}(\theta)$. Applying Taylor's expansions to (5.1) and (5.2) and using the results $\max_{1 \leq i \leq n} h_{ni}(\theta) = o_p(n^{\frac{1}{2}})$, $\frac{1}{\sqrt{n}} \sum_{i=1}^n h_{ni}(\theta) = O_p(1)$ and $\eta_n = O_p(n^{\frac{1}{2}})$, it follows that

$$\widehat{l}_{n,AU}(\theta) = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n V_{2,AU}^{-\frac{1}{2}}(\theta) h_{ni}(\theta) \right)^\tau V_{0,AU}(\theta) \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n V_{2,AU}^{-\frac{1}{2}}(\theta) h_{ni}(\theta) \right) + o_p(1) \quad (\text{A.7})$$

Let $D = \text{diag}(w_1, \dots, w_{r+1})$, where $w_i, 1 \leq i \leq r+1$ are defined in Theorem 5.1. Then, an orthonormal matrix Q exists such that $Q^T D Q = V_{0,AU}$. This together with (A.7) yields

$$\hat{l}_{n,AU}(\theta) = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n Q V_{2,AU}^{-\frac{1}{2}}(\theta) h_{ni}(\theta) \right)^\tau D \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n Q V_{2,AU}^{-\frac{1}{2}}(\theta) h_{ni}(\theta) \right) + o_p(1). \quad (\text{A.8})$$

Standard arguments can be used to prove

$$\frac{1}{\sqrt{n}} Q V^{-\frac{1}{2}}(\theta) \sum_{i=1}^n h_{ni}(\theta) \xrightarrow{\mathcal{L}} N(0, I_{r+1}), \quad (\text{A.9})$$

where I_p is the $p \times p$ identity matrix. (A.8) and (A.9) together prove Theorem 5.1

Next, we prove Theorem 5.2. Recalling the definition of $\hat{l}_{ad,AU}(\theta)$, we have

$$\hat{l}_{ad,AU}(\theta) = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n h_{ni}(\theta) \right)^\tau \hat{V}_{n2,AU}^{-\frac{1}{2}}(\theta) \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n h_{ni}(\theta) \right) + o_p(1). \quad (\text{A.10})$$

where $\hat{V}_{n2,AU}(\theta)$ is defined in Section 5. It can be proved that $\hat{V}_{n2,AU}(\theta) \xrightarrow{p} V_{2,AU}(\theta)$, This together with (A.9) and (A.10) proves Theorem 5.2.

Sketch of Proof of Theorem 5.3 Similar to Theorem 4.3, we can prove Theorem 5.3.

Acknowledgements. The research was supported by Humboldt-Universität Berlin– Sonderforschungsbereich 373 and the National Natural Science Foundation of China.

REFERENCES

- Ahn, H., and J.L. Powell (1993). Estimation of Censored Selection Models with a Nonparametric Selection Mechanism. *Journal of Econometrics*, 58, 3-30.
- Chen, H (1988). Convergence rates for parametric components in a partly linear model. *Ann. Statist.* 16 136-146.
- Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *J. Amer. Statist. Assoc.*, **89**, 81-87.
- Cuzik, J. (1992a). Semiparametric additive regression. *Journal of the Royal Statistical Society, Series B*, **54**, 831-843.

- Cuzick, J. (1992b). Efficient estimates in semiparametric additive regression models with unknown error distribution. *Annals of Statistics*, **20**, 1129-1136.
- Engle, Granger, Rice and Weiss (1986). Nonparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.* **81** 310-3
- Gray, R. (1994). Spline-based tests in survival analysis. *Biometrics*, **50**, 640-652.
- Hahn, J (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66**, 315-331.
- Healy, M.J.R. and Westmacott, M. (1956). Missing values in experiments analyzed on automatic computers. *Appl. Statist.*
- Heckman, N. (1986). Spline smoothing in partly linear models. *J. Roy. Statist. Soc. Ser B*, **48**, 244-248.
- Hirano, K., G. Imbens, G. Ridder, (2000). Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score. NBER Technical Working Paper 251.
- Kitamura, Y., and M. Stutzer (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica* **65**, 861-874.
- Kong, A., Liu, J.S. and Wong, W.H. (1994). Sequential imputation and Bayesian missing data problems. *J. Amer. Statist. Assoc.*, **89**, 278-288.
- Linton, O.B. (1995). Second Order Approximation in the Partially Linear Regression Model. *Econometrica* **63**, 1079-1112.
- Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. John Wiley & Sons, New York.
- Newey, W.K., J.L. Powell, and J.R. Walker, (1990): "Semiparametric Estimation of Selection Models: Some Empirical Results." *American Economic Review*, **80**, 324-328.
- Owen, A. (1988). Empirical likelihood ratio confidence intervals for single functional. *Biometrika* **75**, 237-249.
- Owen, A. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18**, 90-120.
- Owen, A. (1991). Empirical likelihood for linear models. *Ann. Statist.* **19**, 1725-1747.
- RAO, J.N.K. & SCOTT, A.J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fits and independence in two-way tables. *J. Amer. Statist. Assoc.* **76**, 221-230.
- Rice, J. (1986). Convergence rates for partially splined models. *Statistics & Probability Letters*, **4**, 203-208.

Table 1. Empirical coverages of the confidence intervals on θ under different missing functions $P(x)$ and sample sizes n when nominal level is 0.95

$P(x)$	n	AEL	BEL	AUEL	BAUEL	AAUEL	NA	NAAU
$P_1(x)$	30	.9200	.9750	.9320	.9640	.9330	.9220	.9690
	60	.9240	.9620	.9390	.9520	.9390	.9280	.9420
	100	.9450	.9580	.9460	.9540	.9460	.9440	.9480
$P_2(x)$	30	.9160	.9770	.9220	.9670	.9240	.9190	.9700
	60	.9220	.9640	.9370	.9580	.9380	.9250	.9400
	100	.9430	.9590	.9490	.9540	.9490	.9450	.9530
$P_3(x)$	30	.9140	.9820	.9190	.9720	.9190	.9170	.9770
	60	.9210	.9690	.9330	.9600	.9350	.9230	.9660
	100	.9390	.9580	.9420	.9560	.9440	.9390	.9580

- Robins, J., and A. Rotnitzky (1995). Semiparametric Efficiency in Multivariate Regression Models with missing data. *Journal of the American Statistical Association* 90, 122-129.
- Robinson, P. M. (1988). Root-N-Consistent Semiparametric Regression. *Econometrica*, 56, 931-954.
- Severini, T. A. and Staniswalis, J. G. (1994). Quasilikelihood estimation in semiparametric models. *Journal of the American Statistical Association*, 89, 501-511.
- Speckman, J. H. (1988). Kernel smoothing in partial linear models. *J. Roy. Statist. Soc. Ser B*, 50, 413-436.
- Stock, J. H. (1989): "Nonparametric Policy Analysis," *Journal of the American Statistical Association*, 84, 567-576.
- Stock, J. H. (1991): "Nonparametric Policy Analysis: An Application to Estimating Hazardous Waste Cleanup Benefits," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Eds Barnett, Powell, and Tauchen.
- Wang, Q. H. and Rao, J.N.K. (2002). Empirical Likelihood-based Inference in Linear Models with Missing Data. *Scand. J. Statist.* To appear.
- Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Emp. J. Exp. Agric.* 1, 129-142.

Table 2. Average lengths of the confidence intervals on θ under different missing functions $P(x)$ and sample sizes n when nominal level is 0.95

$P(x)$	n	AEL	BEL	AUEL	BAUEL	AAUEL	NA	NAAU
$P_1(x)$	30	0.8700	1.1400	0.8400	1.1000	0.8100	1.1734	1.1002
	60	0.6900	0.7900	0.6200	0.7400	0.6500	0.8539	0.7342
	100	0.5400	0.6000	0.4600	0.5600	0.5200	0.6691	0.5635
$P_2(x)$	30	0.9900	1.4500	0.9100	1.4300	0.9100	1.3599	1.3194
	60	0.7700	0.9500	0.6100	0.8700	0.7200	0.9460	0.8829
	100	0.6000	0.7300	0.4900	0.6900	0.5900	0.7290	0.6293
$P_3(x)$	30	1.1200	1.5100	0.9800	1.5300	0.9200	1.4587	1.3985
	60	0.7800	1.0500	0.6400	0.9600	0.7000	0.9983	0.9084
	100	0.6200	0.7600	0.4700	0.7000	0.6100	0.7664	0.7901

Table 3. Biases and standard errors (SE) of $\hat{\theta}_n$ and $\hat{\theta}_{n,AU}$ under different missing functions $P(x)$ and different sample sizes n

$P(x)$	n	Bias		SE	
		$\hat{\theta}_n$	$\hat{\theta}_{n,AU}$	$\hat{\theta}_n$	$\hat{\theta}_{n,AU}$
$P_1(x)$	30	-0.0040	-0.0034	0.3172	0.3020
	60	0.0038	0.0027	0.2208	0.1980
	100	-0.0021	-0.0011	0.1707	0.1403
$P_2(x)$	30	-0.0088	0.0052	0.3441	0.3337
	60	-0.0073	-0.0025	0.2438	0.2115
	100	-0.0040	-0.0021	0.1860	0.1599
$P_3(x)$	30	0.0055	0.0041	0.3606	0.3589
	60	-0.0043	-0.0036	0.2520	0.2224
	100	0.0023	-0.0011	0.1939	0.1700