Wang, Qihua; Linton, Oliver Bruce; Härdle, Wolfgang

Working Paper

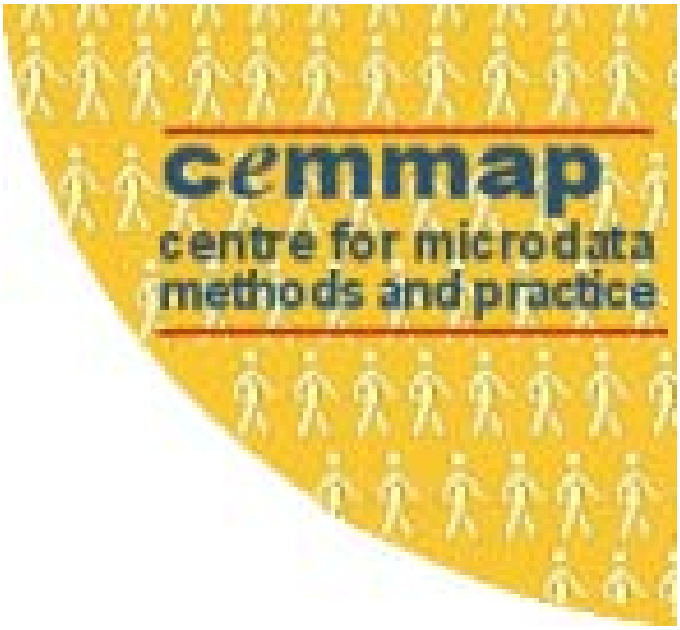
Semiparametric regression analysis with missing response at random

cemmap working paper, No. CWP11/03

Provided in Cooperation with:
The Institute for Fiscal Studies (IFS), London

Suggested Citation: Wang, Qihua; Linton, Oliver Bruce; Härdle, Wolfgang (2003) : Semiparametric regression analysis with missing response at random, cemmap working paper, No. CWP11/03, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2003.1103

This Version is available at:
https://hdl.handle.net/10419/79344



SEMIPARAMETRIC REGRESSION ANALYSIS WITH MISSING RESPONSE AT RANDOM

Qihua Wang
Oliver Linton
Wolfgang Härdle

THE INSTITUTE FOR FISCAL STUDIES
DEPARTMENT OF ECONOMICS, UCL
cemmap working paper CWP11/03

Semiparametric Regression Analysis with Missing Response at Random

Qihua Wang, Oliver Linton and Wolfgang Härdle *

Abstract

We develop inference tools in a semiparametric partially linear regression model with missing response data. A class of estimators is defined that includes as special cases: a semiparametric regression imputation estimator, a marginal average estimator and a (marginal) propensity score weighted estimator. We show that any of our class of estimators is asymptotically normal. The three special estimators have the same asymptotic variance. They achieve the semiparametric efficiency bound in the homoskedastic Gaussian case. We show that the Jackknife method can be used to consistently estimate the asymptotic variance. Our model and estimators are defined with a view to avoid the curse of dimensionality, that severely limits the applicability of existing methods. 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. An adjusted empirical log-likelihood ratio, which is asymptotically standard chi-square, is obtained. Also, a bootstrap empirical log-likelihood ratio is derived and its distribution is used to approximate that of the imputed empirical log-likelihood ratio. A simulation study is conducted to compare the adjusted and bootstrap empirical likelihood with the normal approximation based method 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 estimators and the related estimators.

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

Short Title. Semiparametric Imputation Regression Analysis

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

*Qihua Wang is Professor, Academy of Mathematics and System Science, Chinese Academy of Science, Beijing 100080, P. R. China. Oliver Linton is Professor, Department of Economics, London School of Economics, London WC2A 2AE, UK. Wolfgang Härdle is Professor,Center for Applied Statistics and Economics, Humboldt-Universität, 10178 Berlin,Germany. The authors thank Frank Samaniego, an Associate Editor and two referees for their constructive suggestions and comments that led to significant improvements. The research was supported by Humboldt-Universität Berlin– Sonderforschungsbereich 373, the National Natural Science Foundation of China (Key grant: 10231030) and the Economic and Social Science Research Council of the UK.

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.

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

$$(X_i, Y_i, \delta_i), i = 1, 2, \ldots, n,$$

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 are obtained, perhaps because it is expensive to acquire more $Y's$.

A common method for handling missing data in a large data set 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)), ratio imputation (Rao (1996)) and among others. 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

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

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 of $\widehat{\theta}_c$ and

gave a consistent estimator of its asymptotic variance. An alternative to imputation is the propensity score based methods that are very popular in applied studies, especially in measuring 'treatment effects', following the influential paper by Rosenbaum and Rubin (1983). See Heckman, Ichimura, and Todd (1998) for a recent discussion from an economists point of view and a semiparametric application to the evaluation of social programs. Hahn (1998) has established the semiparametric efficiency bound for estimation of θ, and he constructs an estimator based on the propensity score $P(x)$ that achieves the bound. Actually, Cheng's estimator is also asymptotically efficient. With the nonparametric kernel regression imputation scheme, Wang and Rao (2002a) develop imputed empirical likelihood approaches for constructing confidence intervals of θ.

In practice, however, the nonparametric kernel regression imputation estimator of Cheng and the imputed empirical likelihood may not work well because the dimension of X may be high and hence the curse of dimensionality may occur, Stone (1980). Although this does not affect the first order asymptotic theory, it does show up dramatically in the higher order asymptotics, see Linton (1995) for example. More importantly, dimensionality substantially affects the practical performance of estimators, and the reliability of the asymptotic approximations. Similar comments apply to the propensity score weighting methods when the propensity score itself depends on many covariates. Without further restrictions nonparametric regression methods only work well in low dimensional situations. Indeed, much recent work in statistics has been devoted to intermediate structures like additivity, index models, or semiparametric functional form, in which the curse of dimensionality is mitigated. See for example Hastie and Tibshirani (1990) for a discussion.

Wang and Rao (2001, 2002b) considered the linear regression models and developed the empirical likelihood inference 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 compromise between the linear model and the fully nonparametric model, is to allow only some of the predictors to be modelled linearly, with oth-

ers being modelled nonparametrically. This motivates us to consider the following semiparametric regression model:

$$Y_i = X_i^\top \beta + g(T_i) + \epsilon_i, \tag{1.1}$$

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. d^*-variable random covariate vectors, the function $g(\cdot)$ is unknown and the model errors ϵ_i are independent with conditional mean zero given the covariates. Clearly, the partially linear models contain at least the linear models as a special case. Suppose that the model is linear, but we specify it as partially linear models. The resulting estimator based on the partially linear model is still consistent. Hence, the partially linear model is a flexible one and allows one to focus on particular variables that are thought to have very nonlinear effects. The partially linear 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). 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 Gray (1994), and have been studied extensively for complete data settings, see Heckman (1986), Rice (1986), Speckman (1988), Cuzick (1992), Chen (1988) and Severini, Staniswalis (1994).

In this paper, we are interested in inference on the mean of Y, say θ, when there missing responses in the semiparametric regression model (1.1). Specifically, 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

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

from model (1.1), 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 several estimators of θ in the partially linear model that are simple to compute and do not rely on high dimensional smoothing, thereby avoiding the curse of dimensionality. Our class of estimators includes an imputation estimator and a number of propensity score weighting estimators. Under the model specification the estimators are consistent and asymptotically normal. We obtain their asymptotic distribution and provide consistent variance estimators based on the jacknife method. We also show that a special subclass of our estimators are semiparametrically efficient in the special case that ϵ_i are homoskedastic and Gaussian. When the model specification (1.1) is incorrect, our estimators are inconsistent; we characterize their biases. One of the efficient estimators has a version of the double robustness property of Scharfstein, Rotnizky, Robins (1999).

We also develop empirical likelihood and bootstrap empirical likelihood methods that deliver better inference than standard asymptotic approximations. Though empirical likelihood approaches are also developed with the nonparametric imputation scheme of Cheng in Wang and Rao (2002a) and linear regression imputation scheme in Wang and Rao (2001, 2002b), this paper uses semiparametric regression imputation scheme and use semiparametric techniques to develop an adjusted empirical likelihood and a partially smoothed bootstrap empirical likelihood. The developed partially smoothed bootstrap empirical likelihood method has an advantage over the adjusted empirical likelihood. That is, it avoids estimating the unknown adjusting factor. This is especially attractive in some cases when the adjustment factor is difficult to estimate efficiently. This method is also very useful for the problem considered by Wang and Rao (2002a) since the adjusted factors are difficult to estimate well for nonparametric regression imputation scheme because of "curse of dimensionality". Unfortunately, Wang and Rao (2002a,b) do not develop such a method. Wang and Rao (2001) considers a different inference problem from this paper. They do not consider inference on the response mean, but develops empirical likelihood inference for regression coefficient only in linear regression models with fixed design.

The empirical likelihood method, introduced by Owen (1988), has many advan-

tages over normal approximation methods and the usual bootstrap approximation approaches for constructing confidence intervals. For example, the empirical likelihood confidence intervals do not have a predetermined shape, whereas confidence intervals based on the asymptotic normality of an estimator have a symmetry implied by asymptotic normality. Also, empirical likelihood confidence intervals respect the range of the parameter: if the parameter is positive, then the confidence intervals contains no negative values. Another preferred characteristic is that the empirical likelihood confidence interval is transformation respecting; that is, an empirical likelihood confidence interval for $\phi(\theta)$ is given by ϕ applied to each value in the confidence interval for θ.

2 Estimation

In this section we define the estimators that we will analyze in this paper. We first describe how to estimate the regression function.

Premultiplying (1.1) by the observation indicator we have

$$\delta_i Y_i = \delta_i X_i^\top \beta + \delta_i g(T_i) + \delta_i \epsilon_i,$$

and taking conditional expectations given T we have

$$E\left[\delta_i Y_i | T_i = t\right] = E\left[\delta_i X_i^\top | T_i = t\right] \beta + E\left[\delta_i | T_i = t\right] g(t),$$

from which it follows that

$$g(t) = g_2(t) - g_1(t)^\top \beta, \tag{2.1}$$

where

$$g_1(t) = \frac{E\left[\delta X | T = t\right]}{E\left[\delta | T = t\right]} \text{ and } g_2(t) = \frac{E\left[\delta Y | T = t\right]}{E\left[\delta | T = t\right]}.$$

It follows that

$$\delta_i \left[Y_i - g_2(T_i)\right] = \delta_i \left[X_i - g_1(T_i)\right]^\top \beta + \delta_i \epsilon_i, \tag{2.2}$$

which suggests that an estimator of β can be based on a least squares regression using $\delta_i = 1$ observations and estimated $g_j(\cdot)$, $j = 1, 2$.

Let $K(\cdot)$ be a kernel function and h_n be a bandwidth sequence tending to zero as $n \to \infty$, and define the weights

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

Then $\widetilde{g}_{1n}(t) = \sum_{j=1}^{n} \delta_j W_{nj}(t) X_j$ and $\widetilde{g}_{2n}(t) = \sum_{j=1}^{n} \delta_j W_{nj}(t) Y_j$ are consistent estimates of $g_1(t)$ and $g_2(t)$ respectively. From (2.2), the estimator of β is then defined as the one satisfying:

$$\min_{\beta} \sum_{i=1}^{n} \delta_i \{(Y_i - \widetilde{g}_{2n}(T_i)) - (X_i - \widetilde{g}_{1n}(T_i))\beta\}^2. \tag{2.3}$$

From (2.3), it is easy to obtain that the estimator of β is given by

$$\widehat{\beta}_n = \left[\sum_{i=1}^{n} \delta_i \{(X_i - \widetilde{g}_{1n}(T_i))(X_i - \widetilde{g}_{1n}(T_i))^\top\}\right]^{-1} \sum_{i=1}^{n} \delta_i \{(X_i - \widetilde{g}_{1n}(T_i))(Y_i - \widetilde{g}_{2n}(T_i))\}$$

based on the observed triples (X_i, T_i, Y_i) for $i \in \{i : \delta_i = 1\}$. This is like the Robinson (1988) estimator of β except that it is based on the complete subsample [note also that g_j are not simple conditional expectations as in his case]. (2.1) suggests that an estimator of $g(t)$ can be defined to be

$$\widehat{g}_n(t) = \widetilde{g}_{2n}(t) - \widetilde{g}_{1n}^\top(t)\widehat{\beta}_n$$

by replacing β, $g_1(t)$ and $g_2(t)$ in (2.1) by $\widehat{\beta}_n$, $\widetilde{g}_{1n}(t)$ and $\widetilde{g}_{2n}(t)$.

We now turn to the estimation of θ. Consider now the general class of estimators

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i Y_i}{P_n^*(X_i, T_i)} + \frac{1}{n}\sum_{i=1}^{n} \left(1 - \frac{\delta_i}{P_n^*(X_i, T_i)}\right)(X_i^\top \widehat{\beta}_n + \widehat{g}_n(T_i)),$$

where $P_n^*(x,t)$ is some sequence of quantities with probability limits $P^*(x,t)$. We are particularly interested in some special cases. First, when $P_n^*(x,t) = 1$, we have the regression imputation estimator of θ:

$$\widehat{\theta}_I = \frac{1}{n}\sum_{i=1}^{n} \{\delta_i Y_i + (1-\delta_i)(X_i^\top \widehat{\beta}_n + \widehat{g}_n(T_i))\}.$$

When $P_n^*(x,t) = \infty$, we have the marginal average estimator

$$\widehat{\theta}_{MA} = \frac{1}{n}\sum_{i=1}^{n} (X_i^\top \widehat{\beta}_n + \widehat{g}_n(T_i)),$$

which just averages over the estimated regression function. Define the marginal propensity score $P_1(t) = P(\delta = 1|T = t)$.When $P_n^*(x,t) = \widehat{P}_1(t) = \sum_{j=1}^n \delta_j K\left(\frac{t-T_j}{h_n}\right) / \sum_{j=1}^n K\left(\frac{t-T_j}{h_n}\right)$, we have the (marginal) propensity score weighted estimator

$$\widehat{\theta}_{P_1} = \frac{1}{n}\sum_{i=1}^n \left[\frac{\delta_i Y_i}{\widehat{P}_1(T_i)} + \left(1 - \frac{\delta_i}{\widehat{P}_1(T_i)}\right)(X_i^\top \widehat{\beta}_n + \widehat{g}_n(T_i))\right].$$

Estimator $\widehat{\theta}_{P_1}$ is different from the usual propensity score weighting method that uses an estimator of the full propensity score. Let $\widehat{\theta}^*$ denote either $\widehat{\theta}_I, \widehat{\theta}_{MA}$, or $\widehat{\theta}_{P_1}$. These estimators only rely on one-dimensional smoothing operations and are explicitly defined. These two properties are desirable from a computational and statistical point of view.

The marginal average and imputation estimators do not depend on any 'estimate' of the propensity score, and so are intellectually less demanding. One computational advantage of the imputation estimator is that in case the data are augmented with additional single Y observations, the extra values can be directly included in the average of the observed Y's.

The class of estimators $\{\widehat{\theta}\}$ also includes the unrestricted estimate of the propensity score

$$\widehat{\theta}_P = \frac{1}{n}\sum_{i=1}^n \frac{Y_i\delta_i}{\widehat{P}(X_i,T_i)} + \frac{1}{n}\sum_{i=1}^n \left(1 - \frac{\delta_i}{\widehat{P}(X_i,T_i)}\right)\{X_i\widehat{\beta}_n + \widehat{g}_n(T_i)\}.$$

when $P_n^*(x,t) = \widehat{P}(x,t)$, where $\widehat{P}(x,t)$ a high-dimensional kernel estimator of the propensity score defined by

$$\widehat{P}(x,t) = \frac{\sum_{j=1}^n \delta_j W\left(\frac{x-X_j}{b_n}, \frac{t-T_j}{b_n}\right)}{\sum_{j=1}^n W\left(\frac{x-X_j}{b_n}, \frac{t-T_j}{b_n}\right)}$$

with $W(\cdot,\cdot)$ the weighting function and b_n the bandwidth sequence. However, this estimator depends on high dimensional smoothing. The well known "curse of dimensionality" may restrict the use of this estimator.

Suppose we have an auxiliary semiparametric or parametric model for $P(x,t)$ denoted $P_\tau(x,t)$, where τ can contain finite dimensional and infinite dimensional parameters, Bickel, Klaassen, Ritov, and Wellner (1993), and let $\widehat{P}_{\widehat{\tau}}(x,t)$ be an estimate of $P_\tau(x,t)$. Define

$$\widehat{\theta}_{P_{\hat{\tau}}} = \frac{1}{n}\sum_{i=1}^n \frac{Y_i\delta_i}{\widehat{P}_{\widehat{\tau}}(X_i,T_i)} + \frac{1}{n}\sum_{i=1}^n \left(1 - \frac{\delta_i}{\widehat{P}_{\widehat{\tau}}(X_i,T_i)}\right)\{X_i\widehat{\beta}_n + \widehat{g}_n(T_i)\}.$$

A leading case would be the parametric probit model. In this case, $\widehat{P}_{\widehat{\tau}}(x,t)$ is easy to compute and likely to have a distribution well approximated by its limit. Semiparametric cases of interest include where the index inside the probit link function is allowed to be partially linear or the semiparametric index model. In either of these cases the method need not require high dimension smoothing operations. However, the estimation procedure to obtain $\widehat{\tau}$ can be quite complicated - it usually involves nonlinear optimization of a criterion function, and if it also contains nonparametric estimators then the properties may be poor [reference to the average derivative case]. In this case, the probability limits $P^*(x,t)$ depend on the estimation method; when a likelihood method is used, $P^*(x,t)$ is the Kullback-Liebler minimizing distance from $P(x,t)$ - it can be a quite complicated function different from any of the special cases listed above.

3 Asymptotic Normality

We next state the properties of $\widehat{\theta}$ with $P_n^*(x,t) \in \{1, \infty, \widehat{P}_1(t), \widehat{P}_n(x,t)\}$ and propose consistent variance estimators. Let $P_1(t) = P(\delta = 1|T = t)$, $P(x,t) = P(\delta = 1|X = x, T = t)$, $m(x,t) = x^\top\beta + g(t)$, and $\sigma^2(x,t) = E[(Y - X^\top\beta - g(T))^2|X = x, T = t]$. Then define $u(x,t) = x - g_1(t)$, $\Sigma = E[P(X,T)u(X,T)u(X,T)^\top]$,

THEOREM 3.1. *Under all the assumptions listed in the Appendix except for condition (C.K)iii, we have*

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

where

$$V = E\left[(\pi_0(X,T) + \pi_1(X,T))^2 P(X,T)\sigma^2(X,T)\right] + \text{var[m(X, T)]}.$$

with $\pi_0(x,t) = \frac{1}{P_1(t)}$ *and* $\pi_1(x,t) = E\left[u(X,T)^\top\right]\Sigma^{-1}u(x,t)$ *when* $P_n^*(x,t) \in \{1, \infty, \widehat{P}_1(t)\}$, *and* $\pi_0(x,t) = \frac{1}{P(x,t)}$ *and* $\pi_1(x,t) = 0$ *when* $P_n^*(x,t)$ *is taken to be* $\widehat{P}(x,t)$.

Our estimators $\widehat{\theta}^*$, for which $P_n^*(x,t) \in \{1, \infty, \widehat{P}_1(t)\}$, have a common asymptotic variance $V^* = V$ with $\pi_0(x,t) = \frac{1}{P_1(t)}$ and $\pi_1(x,t) = E\left[u(X,T)^\top\right]\Sigma^{-1}u(x,t)$. The asymptotic equivalence result is similar to that obtained in Cheng (1994, Theorem 2.1) between the marginal average and the imputation estimator. It is interesting

that the marginal propensity score weighting estimator also shares this distribution. The estimators may differ in their higher order properties. The full propensity score weighting estimator with $P_n^*(x,t) = \widehat{P}_n(x,t)$ has a different asymptotic variance from our estimators $\widehat{\theta}^*$.

If $P(x,t)$ is specified to be a parametric or semiparametric model $P_\tau(x,t)$, $\widehat{\theta}_{P_{\widehat{\tau}}}$ can be proved to be asymptotically normal with zero mean and the same asymptotic variance as $\widehat{\theta}_P$ if τ is estimated consistently with an appropriate rate. However, the conditions to prove the asymptotic normality depend on the specified model for $P(x,t)$. We don't investigate the asymptotic property of the estimator further here.

To define a consistent estimator of V, we may first define estimators of $P(x,t)$, $P_1(t)$, $\sigma^2(x,t)$ and $g_1(t)$ by kernel regression method and then define a consistent estimator of V by "plug in" method. However, this method may not estimate V well when the dimension of X is high. This can be avoided because both $P(x,t)$ and $\sigma^2(x,t)$ only enter in the numerator and can be replaced by squared residuals or the indicator function where appropriate.

An alternative is the jackknife variance estimator. Let $\widehat{\theta}^{(-i)}$ be $\widehat{\theta}$ based on $\{(Y_j, \delta_j, X_j, T_j)\}_{j\neq i}$ for $i = 1, 2, \ldots, n$. Let J_{ni} be the jackknife pseudo-values. That is,

$$J_{ni} = n\widehat{\theta} - (n-1)\widehat{\theta}^{(-i)}, \quad i = 1, 2, \cdots, n$$

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

THEOREM 3.2. *Under assumptions of Theorem 3.1, we have*

$$\widehat{V}_{nJ} \xrightarrow{p} V.$$

By Theorem 3.1 and 3.2, 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.

3.1 Efficiency

We assume throughout this section that the partial linear structure is true. We compare the efficiency of our estimators $\widehat{\theta}^*$ with other members of $\widehat{\theta}$ and with estimators that do not consider the partially linear structure. Specifically, consider the class of estimators

$$\widetilde{\theta} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i \delta_i}{P_n^*(X_i, T_i)} + \frac{1}{n}\sum_{i=1}^{n} \left(1 - \frac{\delta_i}{P_n^*(X_i, T_i)}\right) \widehat{m}_n(X_i, T_i),$$

where $\widehat{m}_n(X_i, T_i)$ is the nonparametric regression kernel estimator of the regression of Y on (X, T). This class includes Cheng's (1996) estimator when $P_n^*(X_i, T_i) = \infty$, the imputation estimator when $P_n^*(X_i, T_i) = 1$, and a full propensity score weighting estimator when $P_n^*(X_i, T_i) = \widehat{P}(X_i, T_i)$. Let $\widetilde{\theta}^*$ denote either of these three special cases. These three nonparametric estimators are all asymptotically equivalent and equivalent to an estimator $\widetilde{\theta}_{HIR} = n^{-1}\sum_{i=1}^{n} Y_i \delta_i / \widehat{P}(X_i, T_i)$ due to Hirano *et al.* (2000). The common asymptotic variance denoted V_{UR}^* is

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

This is exactly the semiparametric efficiency bound of Hahn (1998) for the case where $m(x, t)$ is unrestricted. Hence, all three nonparametric estimators are asymptotically efficient in the sense of Hahn (1998) in this more general model. However, the restrictions implied by the partially linear structure reduce the semiparametric efficiency bound. Therefore, the nonparametric estimators $\widetilde{\theta}$ are not asymptotically efficient for the partially linear model. Another disadvantage of the three nonparametric estimators is that they require a high-dimensional smoothing operation to compute the regression of Y or δ on X, T. Therefore, their actual distributions may be very different from that predicted by the asymptotic theory due to the curse of dimensionality. Our estimators $\widehat{\theta}^*$ all make use of the partial linear structure in the conditional mean and hence it is possible for them to be more efficient. We show two efficiency results for our estimators $\widehat{\theta}^*$.

THEOREM 3.3 *Suppose that ϵ is conditionally homoskedastic with $\sigma^2(x, t) = \sigma^2$, where σ is a constant. Then*

$$V^* \leq V_{UR}^*. \tag{3.2}$$

The equality holds only when $(\delta/P(X,T) - \delta/P_1(T))\epsilon = a\delta(X - g_1(T))\epsilon + b$.

This shows that our estimator is asymptotically more efficient than the three nonparametric estimators for the special case of homoskedasticity. It also holds in this case that V^* is the smallest V in our class $\widehat{\theta}$.

THEOREM 3.4. *When* ϵ *is i.i.d. Gaussian,* V^* *is the semiparametric efficiency bound, and* $V^* \leq V^*_{UR}$.

This shows that the proposed estimators $\widehat{\theta}^*$ are asymptotically efficient for this special case. They have lower variance than any other member of $\widehat{\theta}$ or $\widetilde{\theta}$.

We now discuss the efficiency bound in the general heteroskedastic case. It is possible that V^* is the semiparametric efficiency bound in the general case with $\epsilon|X,T$ unrestricted other than its mean being zero. However, note that in the presence of heteroskedasticity, the Robinson type least squares estimator of β is inefficient; the efficient estimator is a weighted least squares version of this where the weights are some consistent estimate of $\sigma^{-2}(x,t)$, a high dimensional problem. We speculate that the semiparametric efficiency bound for θ in that case is very complicated and that, significantly, the efficient score function (Bickel, Klaassen, Ritov, and Wellner (1986)) would require estimation of the high dimensional regression functions $P(x,t)$ and $\sigma^2(x,t)$ as well as perhaps solving an integral equation. See *inter alia:* Nan, Emond, and Wellner (2000), Rotnizky and Robins (1997), Scharfstein, Rotnizky, and Robins (1997), Robins, Hsieh, and Newey (1995), Robins, Rotnizky, and Zhao (1994). Thus, we are left with the trade-off between the promise of large sample efficiency and the practical reality imposed by the curse of dimensionality, which says that an enormous sample may be needed in order to achieve those gains. In practical situations, it may be preferable to have an estimator that only depends on one dimensional smoothing operations. This is certainly a view commonly expressed in applied statistics, see for example Hastie and Tibshirani (1990) and Robins and Ritov (1997). In addition, our estimators are very simple to compute and are explicitly defined.

Finally, consider the estimator $\widehat{\theta}_{P\widehat{\tau}}$. As we have shown, $\widehat{\theta}_{P\widehat{\tau}}$ may have the same asymptotic variance as $\widehat{\theta}_P$. Hence, this estimator is generally inefficient and less efficient than our estimators $\widehat{\theta}^*$ at least for the main case we consider.

3.2 Robustness

Suppose that the partially linear model assumption (1.1) may be incorrect. Let $m_*(x,t)$ be the probability limit of $x^\top \widehat{\beta}_n + \widehat{g}_n(t)$, and recall that $P^*(x,t)$ is the probability limit of $P_n^*(x,t)$. Then

$$p\lim_{n\to\infty} \widehat{\theta} = \theta + E\left[\left(\frac{P^*(X,T) - P(X,T)}{P^*(X,T)}\right)(m_*(X,T) - m(X,T))\right].$$

This shows that the bias of any member of the class $\widehat{\theta}$ depends on both $m_*(X,T) - m(X,T)$ and $P^*(X,T) - P(X,T)$. Specifically, the three estimators in $\widehat{\theta}^*$ are asymptotically biased and have different biases

$$\begin{array}{rcl} p\lim_{n\to\infty} \widehat{\theta}_{P_1} & = & \theta + E\left[\left(1 - \frac{P(X,T)}{P_1(T)}\right)(m_*(X,T) - m(X,T))\right] \\ p\lim_{n\to\infty} \widehat{\theta}_I & = & \theta + E\left[(1 - P(X,T))(m_*(X,T) - m(X,T))\right] \\ p\lim_{n\to\infty} \widehat{\theta}_{MA} & = & \theta + E\left[(m_*(X,T) - m(X,T))\right]. \end{array} \tag{3.1}$$

Likewise

$$p\lim_{n\to\infty} \widehat{\theta}_{P_\tau} = \theta + E\left[\left(\frac{P_\tau(X,T) - P(X,T)}{P_\tau(X,T)}\right)(m_*(X,T) - m(X,T))\right].$$

There is no necessary ranking among the magnitudes of the biases, nor specific predictions about their directions. However, when $P(x,t)$ is close to 1 the bias of $\widehat{\theta}_I$ is likely to be smaller than the bias of $\widehat{\theta}_{MA}$, while when $P(X,T)$ does not vary much about its conditional mean $P_1(T)$, the bias of $\widehat{\theta}_{P_1}$ is small. When $P_\tau(X,T)$ is a good approximation to $P(X,T)$, the bias of $\widehat{\theta}_{P_\tau}$ is likely to be small.

The two estimators $\widehat{\theta}_{P_1}$ and $\widehat{\theta}_{P_\tau}$ have a credible 'double robustness' property, namely that even if the mean specification is incorrect, i.e., $m(x,t) \neq \beta^\top x + g(t)$, $\widehat{\theta}_{P_1}$ is still consistent provided that $P(X,T) = P_1(T)$ a.s., while $\widehat{\theta}_{P_\tau}$ is consistent whenever $P(X,T) = P_\tau(X,T)$ a.s. This property has been discussed by Scharfstein, Rotnizky, Robins (1999). The other estimators $\widehat{\theta}_I$ and $\widehat{\theta}_{MA}$ do not share this property.

4 Estimated, Adjusted and Bootstrap Empirical Likelihood

In this section and the next we provide methods to conduct global inference on θ using empirical likelihood and bootstrap empirical likelihood. Specifically, we

consider the problem of testing $H_0 : \theta = \theta_0$, where θ_0 is a specific value. This sort of application arises a lot in the program evaluation literature, see Hahn (1998). The methods we develop are preferable to the naive confidence intervals developed in section 2 as is well known from other contexts. We also show the advantages of these refined methods in simulations below.

4.1 Estimated and adjusted empirical likelihood

Here, we derive an adjusted empirical likelihood (ADEL) method to develop global inference for θ. Let $\widetilde{Y}_i = \delta_i Y_i + (1-\delta_i)\{X_i^\top \beta + g(T_i)\}$. We have $E\widetilde{Y}_i = \theta_0$ 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 $E\widetilde{Y}_i = \theta_0$. If β and $g(\cdot)$ were known, then one could test $E\widetilde{Y}_i = 0$ using the empirical likelihood of Owen (1990):

$$l_n(\theta) = -2\sup\{\sum_{i=1}^n \log(np_i) | \sum_{i=1}^n p_i \widetilde{Y}_i = \theta, \sum_{i=1}^n p_i = 1, p_i > 0, i = 1, 2, \ldots, n\}.$$

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 $\widetilde{Y}_i's$ 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 $\widehat{Y}_{in} = \delta_i Y_i + (1-\delta_i)\{X_i^\top \widehat{\beta}_n + \widehat{g}_n(T_i)\}$. An estimated empirical log-likelihood evaluated at θ is then defined by

$$\widehat{l}_n(\theta) = -2\sup\{\sum_{i=1}^n \log(np_i) | \sum_{i=1}^n p_i \widehat{Y}_{in} = \theta, \sum_{i=1}^n p_i = 1, p_i > 0, i = 1, 2, \ldots, n\}. \quad (4.1)$$

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

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

where λ is the solution of the equation

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

Unlike the standard empirical log-likelihood $l_n(\theta)$, $\widehat{l}_n(\theta)$ is based on $\widehat{Y}'_{in}s$ that are not independent. Consequently, $\widehat{l}_n(\theta)$ does not have an asymptotic standard chi-square distribution. Actually, $\widehat{l}_n(\theta)$ 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* $H_0 : \theta = \theta_0$,

$$\widehat{l}_n(\theta) \xrightarrow{\mathcal{L}} \frac{V(\theta)}{\widetilde{V}(\theta)}\chi^2_1,$$

where χ^2_1 *is a standard chi-square variable with one degree of freedom,* $V(\theta)$ *is defined in Theorem 3.1 and* $\widetilde{V}(\theta) = E[P(X,T)\sigma^2(X,T)] + Var(X^\top\beta + g(T))$.

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

$$\gamma(\theta)\widehat{l}_n(\theta) \xrightarrow{\mathcal{L}} \chi^2_1, \tag{4.4}$$

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

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

with adjustment factor $\gamma_n(\theta)$. It readily follows from (4.4) and (4.5), $\widehat{l}_{n,ad}(\theta_0) \xrightarrow{\mathcal{L}} \chi^2_1$ under $H_0 : \theta = \theta_0$.

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

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

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* $H_0 : \theta = \theta_0$

$$\widehat{l}_{n,ad}(\theta_0) \xrightarrow{\mathcal{L}} \chi^2_1.$$

From Theorem 4.2, it follows immediately that an approximation $1-\alpha$ confidence region for θ is given by $\{\theta : \widehat{l}_{n,ad}(\theta) \leq \chi^2_{1,\alpha}\}$ where $\chi^2_{1,\alpha}$ is the upper α percentile of the χ^2_1 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 $\widehat{l}_{n,ad}(\theta_0) > \chi^2_{1,\alpha}$.

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 $\widehat{l}_m^*(\widehat{\theta}_n)$ approximates to that of $\widehat{l}_n(\theta)$ with probability one, we need that $T_1^*, \ldots, T_m^*$ have a probability density. This motivates us to use smooth bootstrap. Let $T_i^{**} = T_i^* + h_n\zeta_i$ for $i = 1, 2, \ldots, m$, where h_n is the bandwidth sequence used in Section 2 and $\zeta_i, i = 1, 2, \ldots, m$ 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*

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

as $n \to \infty$ and $m \to \infty$, 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 be 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 is difficult to estimate efficiently.

5 Simulation Results

We conducted a simulation to analyze the finite-sample performances of the proposed estimators $\widehat{\theta}_I, \widehat{\theta}_{MA}$ and $\widehat{\theta}_{P_1}$ and the weighted estimator $\widehat{\theta}_P$ and $\widehat{\theta}_{P_\tau}$ given in Section 2, and compare the two empirical likelihood methods, namely the adjusted empirical likelihood and the partly smoothed bootstrap empirical likelihood, with the normal approximation-based method in terms of coverage accuracies of confidence intervals.

The simulation used the partial linear model $Y = X^\top\beta + g(T) + \epsilon$ 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 $K(t) = \frac{15}{16}(1-2t^2+t^4)$ if $|t| \leq 1, 0$ otherwise, and the bandwidth h_n was taken to be $n^{-2/3}$.

We generated 5000 Monte Carlo random samples of size $n = 30, 60$ and 100 based on the following three cases respectively:

Case 1: $P(\delta = 1|X = x, T = t) = 0.8+0.2(|x-1|+|t-0.5|$ if $|x-1|+|t-0.5| \leq 1$, and 0.95 elsewhere;

Case 2: $P(\delta = 1|X = x, T = t) = 0.9-0.2(|x-1|+|t-0.5|)$ if $|x-1|+|t-0.5| \leq 4$, 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 approxi-

mately 0.10, 0.25 and 0.40 respectively.

For calculating $\widehat{\theta}_P$, $\widehat{P}(x,t)$ was taken to be the nonparametric kernel estimator given by

$$\widehat{P}(x,t) = \frac{\sum_{i=1}^{n} \delta_i K_1\left(\frac{x-X_i}{h_{1,n}}\right) K_2\left(\frac{t-T_i}{h_{2,n}}\right)}{\sum_{i=1}^{n} K_1\left(\frac{x-X_i}{h_{1,n}}\right) K_2\left(\frac{t-T_i}{h_{2,n}}\right)}$$

where $K_1(u) = -\frac{15}{8}u^2 + \frac{9}{8}$ if $|u| \leq 1$, 0 otherwise; $K_2(v) = \frac{15}{16}(1 - 2t^2 + t^4)$ if $|v| \leq 1$, 0 otherwise and $h_{1,n} = h_{2,n} = n^{-\frac{1}{3}}$.

Let $\widehat{\theta}_{P_{\widehat{\tau}},1}$ be $\widehat{\theta}_{P_{\widehat{\tau}}}$ with

$\widehat{P}(x,t) = 0.8 + 0.2(|x - \bar{X}| + |t - \bar{T}|)$ if $|x - \bar{X}| + |t - \bar{T}| \leq 1$, and 0.95 elsewhere for case 1;

$\widehat{P}(x,t) = 0.9 - 0.2(|x - \bar{X}| + |t - \bar{T}|)$ if $|x - \bar{X}| + |t - \bar{T}| \leq 4$, and 0.1 elsewhere for case 2 and

$\widehat{P}(x,t) = 0.6$ for case 3, respectively, where $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$ and $\bar{T} = n^{-1}\sum_{i=1}^{n} T_i$.

Let $\widehat{\theta}_{P_{\widehat{\tau}},2}$ be $\widehat{\theta}_{P_{\widehat{\tau}}}$ with $\widehat{P}(x,t)$ taken to be

$$\widehat{P}(x,t) = exp\{-(|x - \bar{X}| + |t - \bar{T}|)\} \text{ if } |x - \bar{X}| + |t - \bar{T}| \leq 2, \text{ and } 0.70 \text{ elsewhere}$$

for all the three cases considered here. Clearly, $\widehat{\theta}_{P_{\widehat{\tau}},2}$ is defined based on an incorrectly specified propensity score model.

From the 5000 simulated values of $\widehat{\theta}_I, \widehat{\theta}_{MA}, \widehat{\theta}_{P_1}, \widehat{\theta}_P$, $\widehat{\theta}_{P_{\widehat{\tau}},1}$ and $\widehat{\theta}_{P_{\widehat{\tau}},2}$, we calculated the biases and standard errors of the six estimators. These simulated results are reported in Tables 5.1 and 5.2.

Insert Tables 5.1 and 5.2 here

From Tables 5.1 and 5.2, we observe:

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

(b) $\widehat{\theta}_I, \widehat{\theta}_{MA}$, $\widehat{\theta}_{P_1}$ have smaller SE than $\widehat{\theta}_P$, $\widehat{\theta}_{P_{\widehat{\tau}},1}$ and $\widehat{\theta}_{P_{\widehat{\tau}},2}$. $\widehat{\theta}_P$ and $\widehat{\theta}_{P_{\widehat{\tau}},1}$ have less SE than $\widehat{\theta}_{P_{\widehat{\tau}},2}$. Generally, $\widehat{\theta}_P$ and $\widehat{\theta}_{P_{\widehat{\tau}},2}$ also have bigger bias than other estimators, and $\widehat{\theta}_{P_{\widehat{\tau}},2}$ has bigger bias than $\widehat{\theta}_P$. This suggests that our estimators and $\widehat{\theta}_{P_{\widehat{\tau}},1}$ outperform $\widehat{\theta}_P$ and $\widehat{\theta}_{P_{\widehat{\tau}},2}$, and our estimators perform better than $\widehat{\theta}_{P_{\widehat{\tau}},1}$ in terms of

SE. From the simulation results, the weighted estimator $\widehat{\theta}_{P_{\widehat{\tau}},2}$ doesn't perform well if the propensity score is incorrectly specified.

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 Table 5.3. For convenience, in what follows AEL represents the adjusted empirical likelihood confidence interval given in subsection 4.1. BEL denotes the smoothed bootstrap empirical likelihood confidence intervals given in subsections 4.2. NA denotes the normal approximation based confidence intervals given in Section 2 based on $\widehat{\theta}_I$.

Insert Table 5.3 here

From Table 5.3, we observe the following:

(1) BEL does perform competitively in comparison to AEL and NA since BEL has generally higher coverage accuracies but only slightly bigger average lengths. NA has higher slightly coverage accuracy than AEL. But. it does this using much longer intervals. This implies that AEL might be preferred over NA.

(2) 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 performs 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.

(3) All the coverage accuracies increase and the average lengths decrease as n increases 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.

Appendix A: Assumptions and Proofs of Theorems 3.1, 3.2, 3.3 and 3.4

Denote by $g_{1r}(\cdot)$ the rth component of $g_1(\cdot)$. Let $\|\cdot\|$ be the Euclid norm. The following assumptions are needed for the asymptotic normality of $\widehat{\theta}_n$.

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

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

(C.Σ) $\Sigma = E[P(X,T)u(X,T)u(X,T)^\top]$ 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.W)(i): The kernel function $W(\cdot)$ is a bounded kernel function with bounded support and bounded variation.

(ii): $W(\cdot)$ is a kernel of order $k(> d+1)$.

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

(C.b_n): $nb_n^{2(d+1)}/\log n \to \infty$ and $nb_n^{2k} \to 0$.

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.

Proof of Theorem 3.1. (i) We prove Theorem 3.1 for $\widehat{\theta}_I$. For $\widehat{\theta}_I$, we have

$$\begin{array}{rcl} \widehat{\theta}_I & = & \frac{1}{n}\sum_{i=1}^n \{\delta_i Y_i + (1-\delta_i)(X_i^\top\beta + g(T_i))\} \\ & & +\frac{1}{n}\sum_{i=1}^n (1-\delta_i)X_i^\top(\widehat{\beta}_n - \beta) + \frac{1}{n}\sum_{i=1}^n (1-\delta_i)(\widehat{g}_n(T_i) - g(T_i)). \end{array} \tag{A.1}$$

Note that

$$\widehat{\beta}_n - \beta = \Sigma^{-1}\frac{1}{n}\sum_{i=1}^n \delta_i \left[X_i - g_1(T_i)\right]\epsilon_i + o_p(n^{-1/2}). \tag{A.2}$$

$$\frac{1}{n}\sum_{i=1}^n (1-\delta_i)(\widehat{g}_n(T_i) - g(T_i)) = \frac{1}{n}\sum_{j=1}^n \delta_j\epsilon_j\frac{(1-P_1(T_j))}{P_1(T_j)} - \frac{1}{n}\sum_{j=1}^n (1-\delta_j)g_1(T_j)(\widehat{\beta}_n - \beta) + o_p(n^{-1/2}) \tag{A.3}$$

By (A.1), (A.2) and (A.3), we get

$$\begin{array}{rcl} \widehat{\theta}_I - \theta & = & \frac{1}{n}\sum_{i=1}^n \left\{\frac{\delta_i}{P_1(T_i)} + E[u(X,T)^\top]\Sigma^{-1}\delta_i(X_i - g_1(T_i))\right\}\epsilon_i \\ & & +\frac{1}{n}\sum_{i=1}^n (X_i^\top\beta + g(T_i) - \theta) + o_p(n^{-1/2}), \end{array} \tag{A.4}$$

By (A.4) and the central limit theorem, $\widehat{\theta}_I$ has the stated asymptotic normality.

(ii) We prove Theorem 3.1 for $\widehat{\theta}_{MA}$. For $\widehat{\theta}_{MA}$, we have

$$\widehat{\theta}_{MA}-\theta=\frac{1}{n}\sum_{i=1}^{n}(X_i^\top\beta+g(T_i))-\theta+E(X)^\top(\widehat{\beta}_n-\beta)+\frac{1}{n}\sum_{i=1}^{n}(\widehat{g}_n(T_i)-g(T_i))+o_p(n^{-1/2}), \tag{A.5}$$

where

$$\begin{aligned}\tfrac{1}{n}\sum_{i=1}^{n}(\widehat{g}_n(T_i)-g(T_i)) &= \tfrac{1}{n}\sum_{i=1}^{n}\textstyle\sum_{j=1}^{n}\delta_j W_{nj}(T_i)\epsilon_j-\tfrac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_j W_{nj}(T_i)X_j^\top(\widehat{\beta}_n-\beta)+o_p(n^{-1/2})\\ &= \tfrac{1}{n}\sum_{i=1}^{n}\epsilon_i\tfrac{\delta_i}{P_1(T_i)}-E[g_1(T_i)^\top](\widehat{\beta}_n-\beta)+o_p(n^{-1/2}).\end{aligned} \tag{A.6}$$

Therefore, (A.2), (A.5) and (A.6) together prove

$$\begin{aligned}\widehat{\theta}_{MA}-\theta &= \tfrac{1}{n}\sum_{i=1}^{n}\epsilon_i\tfrac{\delta_i}{P_1(T_i)}+E(u(X,T))^\top\Sigma^{-1}\tfrac{1}{n}\sum_{i=1}^{n}\delta_i\left[X_i-g_1(T_i)\right]\epsilon_i\\ &\quad+\tfrac{1}{n}\sum_{i=1}^{n}(X_i^\top\beta+g(T_i)-\theta)+o_p(n^{-1/2}).\end{aligned} \tag{A.7}$$

This together with central limit theorem proves Theorem 3.1 for $\widehat{\theta}_{MA}$.

(iii) We prove Theorem 3.1 for $\widehat{\theta}_{P_1}$. For $\widehat{\theta}_{P_1}$, we have

$$\begin{aligned}\widehat{\theta}_{P_1} &= \theta+\tfrac{1}{n}\sum_{i=1}^{n}\tfrac{\delta_i\epsilon_i}{P_1(T_i)}+\tfrac{1}{n}\sum_{i=1}^{n}\tfrac{\delta_i\epsilon_i\{\widehat{P}_1(T_i)-P_1(T_i)\}}{P_1^2(T_i)}\\ &\quad+\tfrac{1}{n}\sum_{i=1}^{n}(X_i^\top\beta+g(T_i)-\theta)\\ &\quad+\tfrac{1}{n}\sum_{i=1}^{n}\left(1-\tfrac{\delta_i}{P_1(T_i)}\right)X_i^\top(\widehat{\beta}_n-\beta)\\ &\quad+\tfrac{1}{n}\sum_{i=1}^{n}\left(1-\tfrac{\delta_i}{P_1(T_i)}\right)(\widehat{g}_n(T_i)-g(T_i))+o_p(n^{-1/2})\\ &= \theta+T_{n1}+T_{n2}+T_{n3}+T_{n4}+T_{n5}+o_p(n^{-1/2}).\end{aligned} \tag{A.8}$$

For T_{n5}, we have

$$\begin{aligned}T_{n5} &= \tfrac{1}{n}\sum_{i=1}^{n}\left(1-\tfrac{\delta_i}{P_1(T_i)}\right)\sum_{j=1}^{n}\delta_j\tfrac{1}{nh}\tfrac{K\left(\frac{T_i-T_j}{h_n}\right)}{P_1(T_i)f_T(T_i)}\epsilon_j\\ &\quad-\tfrac{1}{n}\sum_{i=1}^{n}\left(1-\tfrac{\delta_i}{P_1(T_i)}\right)\sum_{j=1}^{n}\delta_j\tfrac{1}{nh}\tfrac{K\left(\frac{T_i-T_j}{h_n}\right)}{P_1(T_i)f_T(T_i)}g_1(T_j)^\top(\widehat{\beta}_n-\beta)+o_p(n^{-\frac{1}{2}})\end{aligned} \tag{A.9}$$

Note that$E\left[1-\frac{\delta_i}{P_1(T_i)}|T_i\right]=0$. We have

$$\begin{aligned}&\tfrac{1}{n}\sum_{i=1}^{n}\left(1-\tfrac{\delta_i}{P_1(T_i)}\right)\sum_{j=1}^{n}\delta_j\tfrac{1}{nh}\tfrac{K\left(\frac{T_i-T_j}{h_n}\right)}{P_1(T_i)f_T(T_i)}\epsilon_j\\ =\ &\tfrac{1}{n}\sum_{j=1}^{n}\delta_j\epsilon_j\tfrac{1}{nh}\sum_{i=1}^{n}\left(1-\tfrac{\delta_i}{P_1(T_i)}\right)\tfrac{K\left(\frac{T_i-T_j}{h_n}\right)}{P_1(T_i)f_T(T_i)}=o_p(n^{-1/2})\end{aligned} \tag{A.10}$$

and

$$\begin{aligned} & \tfrac{1}{n}\sum_{i=1}^{n}\left(1-\tfrac{\delta_i}{P_1(T_i)}\right)\sum_{j=1}^{n}\delta_j\tfrac{1}{nh}\tfrac{K\left(\frac{T_i-T_j}{h_n}\right)}{P_1(T_i)f_T(T_i)}g_1(T_j)^\top \\ = & \tfrac{1}{n}\sum_{j=1}^{n}\delta_j g_1(T_j)^\top\tfrac{1}{nh}\sum_{i=1}^{n}\left(1-\tfrac{\delta_i}{P_1(T_i)}\right)\tfrac{K\left(\frac{T_i-T_j}{h_n}\right)}{P_1(T_i)f_T(T_i)}=o_p(1) \end{aligned} \tag{A.11}$$

(A.9), (A.10) and (A.11) together with the fact that $\widehat{\beta}_n-\beta=O_p(n^{-\frac{1}{2}})$ prove

$$T_{n5}=o_p(n^{-\frac{1}{2}}). \tag{A.12}$$

Furthermore, $E\left[\left(1-\frac{\delta_i}{P_1(T_i)}\right)X_i\right]=E\left[(X-g_1(T))\right]$ so that the term

$$T_{n4}=\frac{1}{n}\sum_{i=1}^{n}\left(1-\frac{\delta_i}{P_1(T_i)}\right)X_i^\top(\widehat{\beta}_n-\beta)=E\left[(X-g_1(T))^\top\right](\widehat{\beta}_n-\beta)+o_p(n^{-1/2}). \tag{A.13}$$

For T_{n2}, we have

$$T_{n2}=\frac{1}{n}\sum_{j=1}^{n}\left[\delta_j-P_1(T_j)\right]\frac{1}{nh}\sum_{i=1}^{n}\frac{\delta_i\epsilon_i}{P_1^2(T_i)}K\left(\frac{T_i-T_j}{h_n}\right)\frac{1}{f_T(T_i)}+o_p(n^{-\frac{1}{2}})=o_p(n^{-1/2}). \tag{A.14}$$

(A.8), (A.12),(A.13) and (A.14) together prove

$$\begin{aligned} \widehat{\theta}_{P_1}-\theta = & \tfrac{1}{n}\sum_{i=1}^{n}\tfrac{\delta_i\epsilon_i}{P_1(T_i)}+\tfrac{1}{n}\sum_{i=1}^{n}X_i^\top\beta+g(T_i)-\theta \\ & +E\left[(X-g_1(T))\right]^\top\Sigma^{-1}\tfrac{1}{n}\sum_{i=1}^{n}\delta_i\left[X_i-g_1(T_i)\right]\epsilon_i+o_p(n^{-1/2}). \end{aligned} \tag{A.15}$$

This together central limit theorem proves Theorem 3.1 for $\widehat{\theta}_{P_1}$.

(iv) We prove Theorem 3.1 for $\widehat{\theta}_P$. For simplicity, let $Z_i=(X_i,T_i)$. Observe that

$$\widehat{\theta}_P=\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_iY_i}{\widehat{P}(Z_i)}+\frac{1}{n}\sum_{i=1}^{n}(1-\frac{\delta_i}{\widehat{P}(Z_i)})(X_i^\top\widehat{\beta}_n+\widehat{g}_n(T_i)). \tag{A.16}$$

Next, we prove

$$\frac{1}{n}\sum_{i=1}^{n}(1-\frac{\delta_i}{\widehat{P}(Z_i)})(X_i^\top\widehat{\beta}_n+\widehat{g}_n(T_i))\longrightarrow 0. \tag{A.17}$$

By assumptions (C.W), (C.b_n), (C.K), (C.h_n), (C.T) and (C.r), it can be proved that

$$\begin{aligned} & n^{-1}\sum_{i=1}^{n}\left(1-\tfrac{\delta_i}{\widehat{P}(Z_i)}\right)(X_i^\top\widehat{\beta}_n+\widehat{g}_n(T_i)) \\ = & n^{-1}\sum_{i=1}^{n}\left(1-\tfrac{\delta_i}{P(Z_i)}\right)(X_i^\top\beta+g(T_i))+n^{-1}\sum_{i=1}^{n}\tfrac{\delta_i(\widehat{P}(Z_i)-P(Z_i))}{P^2(Z_i)}(X_i^\top\beta+g(T_i))+o_p(n^{-\frac{1}{2}}) \end{aligned} \tag{A.18}$$

Let L_{n2} be the second term of the right hand side of the equality in (A.18). Then

$$\begin{aligned} L_{n2} &= n^{-1}\sum_{j=1}^{n}(\delta_j - P(Z_j))(nb_n^{d+1})^{-1}\sum_{i=1}^{n}\frac{\delta_i W\left(\frac{Z_i-Z_j}{b_n}\right)(X_i^\top\beta+g(T_i))}{P^2(Z_i)f(Z_i)} \\ &= n^{-1}\sum_{j=1}^{n}\frac{\delta_j-P(Z_j)}{P(Z_j)}(X_j^\top\beta+g(T_j))+o_p(n^{-\frac{1}{2}}). \end{aligned} \tag{A.19}$$

where $f(z)$ is probability density of Z and $\widehat{f}_n(z)=\frac{1}{nb_n^{d+1}}\sum_{i=1}^{n}W\left(\frac{z-Z_i}{b_n}\right)$.

(A.18) and (A.19) together prove (A.17). The first term in (A.16) is just $\widetilde{\theta}_{HIR}$, which is proved by Hirano *et al* (2000) to be asymptotically normal with mean zero and variance $V_P=E\left[\frac{\sigma^2(X,T)}{P(X,T)}\right]+Var[m(X,T)]$. This proves Theorem 3.1.

Proof of Theorem 3.2. Similar to (A.4),(A.8),(A.15) and Hirano *et al* (2000), we can get

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

where $\eta(Y,\delta,X,T)=(\pi_0(X,T)+\pi_1(X,T))\delta\epsilon+m(X,T)-\theta$ with $\pi_0(x,t)$ and $\pi(x,t)$ defined in Section 3. This proves $\widehat{V}_{nJ}\xrightarrow{p}V(\theta)$.

Proof of Theorem 3.3. Under conditions of Theorem 3.3, we have

$$\begin{aligned} V^* &= \sigma^2E\left[\tfrac{1}{P_1(T)}\right]+\sigma^2E[u(X,T)^\top]\Sigma^{-1}E[u(X,T)]+\mathrm{var[m(X,T)]} \\ V^*_{UR} &= \sigma^2E\left[\tfrac{1}{P(X,T)}\right]+\mathrm{var[m(X,T)]}. \end{aligned}$$

Note that

$$\begin{aligned} \sigma^2E\left[u(X,T)\right] &= \sigma^2E\left[\left(\tfrac{\delta}{P(X,T)}-\tfrac{\delta}{P_1(T)}\right)\delta\left(X-g_1(T)\right)\right] \\ &= \mathrm{cov}\left(\left(\tfrac{\delta}{\mathrm{P(X,T)}}-\tfrac{\delta}{\mathrm{P_1(T)}}\right)\epsilon,\delta\left(\mathrm{X}-\mathrm{g_1(T)}\right)\epsilon\right) \end{aligned}$$

because $E[\delta\left(X-g_1(T)\right)/P_1(T)]=0$ and $E[\delta\left(X-g_1(T)\right)/P(X,T)]=E[X-g_1(T)]$. Furthermore,

$$\frac{\sigma^2E[u(X,T)^\top](\sigma^2\Sigma)^{-1}\sigma^2E[u(X,T)]}{\mathrm{var}\left(\left(\frac{\delta}{\mathrm{P(X,T)}}-\frac{\delta}{\mathrm{P_1(T)}}\right)\epsilon\right)}\leq 1, \tag{A.20}$$

because the left hand side is a squared correlation. Then note that

$$\mathrm{var}\left[\left(\frac{\delta}{\mathrm{P(X,T)}}-\frac{\delta}{\mathrm{P_1(T)}}\right)\epsilon\right]=\sigma^2\mathrm{E}\left[\frac{1}{\mathrm{P(X,T)}}-\frac{1}{\mathrm{P_1(T)}}\right]. \tag{A.21}$$

Combining (A.20) and (A.21) we have

$$\sigma^2E[u(X,T)^\top]\Sigma^{-1}E[u(X,T)]\leq\sigma^2E\left[\frac{1}{P(X,T)}-\frac{1}{P_1(T)}\right],$$

i.e., $V^* \leq V^*_{UR}$ as claimed in Theorem 3.3. Clearly, the equality holds only when $(\delta/P(X,T) - \delta/P_1(T))\epsilon = a\delta(X - g_1(T))\epsilon + b$, where both a and b are constants.

Proof of Theorem 3.4. We follow the approach of Bickel, Klaassen, Ritov, and Wellner (1993, section 3.3), as applied by Hahn (1998). The log density of (Y, δ, X, T) is

$$\begin{aligned}\log f_{\beta,g,f_\epsilon,P,f_{X,T}}(Y,\delta,X,T) &= \delta \log f_\epsilon(Y - \beta X - g(T)|X,T) + \delta \log P(X,T) \\ &\quad +(1-\delta)\log(1 - P(X,T)) + \log f_{X,T}(X,T),\end{aligned}$$

where $f_\epsilon(e|X,T)$ denotes the conditional density of ϵ given X, T, and $f_{X,T}$ is the covariate density. Let $\mathbf{Q}$ denote the semiparametric model. Now consider any regular parametric submodel $\mathbf{Q}_\lambda$ with $\epsilon \sim N(0, \sigma^2)$ and parameters $\lambda = (\beta, \gamma, \sigma^2, \eta_p, \eta_{xt})$, such that the log density $\log f_{\beta,g,\sigma^2,P,f_{X,T}}(\delta Y, X, T, \delta; \lambda)$, which we denote by ℓ_{sub} is

$$\begin{aligned}&\delta\frac{-1}{2\sigma^2}(Y - \beta X - g_\gamma(T))^2 + \delta\frac{-1}{2}\log\sigma^2 + \delta\log P(X,T;\eta_p) \\ &+(1-\delta)\log(1 - P(X,T;\eta_p)) + \log f_{X,T}(X,T;\eta_{xt}),\end{aligned}$$

which equals $\log f_{\beta,g,f_\epsilon,P,f_{X,T}}(\delta Y, X, T, \delta)$ when $\lambda = \lambda_0$. The score functions are:

$$\begin{aligned}\tfrac{\partial \ell_{sub}}{\partial \beta} &= -\delta\frac{1}{\sigma^2}X\epsilon, \quad \frac{\partial \ell_{sub}}{\partial \gamma} = -\delta\frac{1}{\sigma^2}\frac{\partial g_\gamma}{\partial \gamma}(T)\epsilon, \quad \frac{\partial \ell_{sub}}{\partial \sigma^2} = -\delta\frac{1}{2\sigma^2}\left(\frac{\epsilon^2}{\sigma^2} - 1\right), \\ \tfrac{\partial \ell_{sub}}{\partial \eta_p} &= \frac{\delta - P(X,T)}{P(X,T)(1-P(X,T))}\frac{\partial P}{\partial \eta_p}(X,T), \quad \frac{\partial \ell_{sub}}{\partial \eta_{xt}} = \frac{\partial f_{f_{X,T}}(X,T)/\partial \eta_{xt}}{f_{f_{X,T}}(X,T)},\end{aligned}$$

where $\epsilon = Y - \beta X - g_\gamma(T)$. The semiparametric model is the union of all such parametric models, and so the tangent space of $\mathbf{Q}$, denoted $\mathcal{T}$, is generated by

$$\left\{\delta X\epsilon, \delta\gamma(T)\epsilon, \delta\left(\frac{\epsilon^2}{\sigma^2} - 1\right), a(X,T)(\delta - P(X,T)), b(X,T)\right\},$$

where: $E\epsilon = 0, E\epsilon^2 = \sigma^2$, and $Eb(X,T) = 0$, while $a(X,T)$ is any square integrable measurable function of X, T.

We first consider what is the efficiency bound for estimation of β in the semiparametric model. We follow Bickel et al. (1993, section 2.4) and find the efficient score function for β in the presence of the nuisance functions $P, f_{X,T}, g$, and parameter σ^2. The efficient score function for estimation of β has to be orthogonal to all of the other score functions and in particular orthogonal to any function of the form

$\delta\gamma(T)\epsilon$ [which is a candidate score function for the parameters of g]. The efficient score function for β in the semiparametric model is $\ell^*_\beta = \delta[X - g_1(T)]\epsilon$, and the semiparametric efficiency bound is

$$I^{*-1}_{\beta\beta} = \sigma^2 \left(E\left[\delta[X - g_1(T)][X - g_1(T)]^\top\right]\right)^{-1},$$

and no regular estimator can have asymptotic variance less than this. Since our estimator $\widehat{\beta}_n$ achieves this bound and is hence efficient.

We now turn to the efficiency bound for the parameter λ. We first show pathwise differentiability of the parameter θ. For the parametric submodel

$$\theta = \int Y f_\epsilon(Y - \beta X - g_\gamma(T)|X, T; \sigma^2) f_{X,T}(X, T; \eta_{xt}) dY dX dT,$$

which has derivatives $\frac{\partial\theta}{\partial\beta} = -E[X]$, $\frac{\partial\theta}{\partial\gamma} = -E\left[\frac{\partial g_\gamma}{\partial\gamma}(T)\right]$, $\frac{\partial\theta}{\partial\sigma^2} = 0$ and $\frac{\partial\theta}{\partial\eta_{xt}} = E\left[m(X,T)\frac{\partial f_{X,T}(X,T)/\partial\eta_{xt}}{f_{X,T}(X,T)}\right]$

Define $F_\theta = \frac{\delta\epsilon}{P(X,T)} + m(X,T) - \theta$. Then it can be seen that $E[F_\theta s_\lambda] = \frac{\partial\theta}{\partial\lambda}$ for parameters λ, where s_λ is the corresponding element of $\mathcal{T}$. Therefore, θ is a differentiable parameter.

To find the variance bound we must find the mean square projection of F_θ onto the tangent space $\mathcal{T}$. In view of the above arguments, $\mathcal{T}$ is equivalently generated from the functions $\delta[X - g_1(T)]\epsilon, \delta\gamma(T)\epsilon, \ldots$. Furthermore, we can effectively ignore the second term $m(X,T) - \theta$ in F_θ, since this is already in $\mathcal{T}$. Without loss of generality we find κ to minimize the variance of

$$\left\{\frac{\delta}{P(X,T)} - \frac{\delta}{P_1(T)} - \kappa\delta(X - g_1(T))\right\}\epsilon.$$

The solution is $\kappa = \frac{E[X - g_1(T)]}{E\left[\delta(X - g_1(T))^2\right]}$ because

$$\left\{\frac{\delta}{P(X,T)} - \frac{\delta}{P_1(T)} - \kappa\delta(X - g_1(T))\right\}\epsilon$$

is then orthogonal to any function in $\mathcal{T}$ as can easily be verified. Therefore, the efficient influence function is

$$\left\{\frac{\delta}{P_1(T)} + \kappa\delta(X - g_1(T))\right\}\epsilon + m(X,T) - \theta,$$

which is the influence function of our estimators $\widehat{\theta}_I, \widehat{\theta}_{MA}$ and $\widehat{\theta}_P$. This shows that our estimators are asymptotically efficient for the special case where ϵ is i.i.d. Gaussian.

Appendix B: Proofs of Theorem 4.1 and 4.2

Proofs of Theorem 4.1 and 4.2. It can be proved that $\min_{1\le i\le n}\widehat{Y}_{in} < \theta < \max_{1\le i\le n}\widehat{Y}_{in}$ with probability tending to 1 when $n \to \infty$. Hence, by Lagrange multiplier method, (4.2) and (4.3) are then obtained from (4.1). Applying Taylor's expansion to (4,2), we get

$$\widehat{l}_n(\theta) = 2\sum_{i=1}^{n}\{\lambda_n(\widehat{Y}_{in}-\theta) - \frac{1}{2}[\lambda_n(\widehat{Y}_{in}-\theta)]^2\} + o_p(1) \tag{A.16}$$

by the facts that $\widehat{Y}_{(n)} = o_p(n^{\frac{1}{2}})$ and $\lambda_n = O_p(n^{-\frac{1}{2}})$. Applying Taylor's expansion to (4.3), we get

$$\sum_{i=1}^{n}\lambda_n(\widehat{Y}_{in}-\theta) = \sum_{i=1}^{n}[\lambda_n(\widehat{Y}_{in}-\theta)]^2 + o_p(1) \tag{A.17}$$

and

$$\lambda_n = \left(\sum_{i=1}^{n}(\widehat{Y}_{in}-\theta)^2\right)^{-1}\sum_{i=1}^{n}(\widehat{Y}_{in}-\theta) + o_p(n^{-\frac{1}{2}}). \tag{A.18}$$

(A.16), (A.17) and (A.18) together yield

$$\widehat{l}_n(\theta) = \widetilde{V}_n^{-1}(\theta)\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\widehat{Y}_{in}-\theta)\right]^2 + o_p(1). \tag{A.19}$$

This together with Theorem 3.1 proves Theorem 4.1.

Recalling the definition of $\widehat{l}_{n,ad}(\theta)$, by (A.19) we get

$$\widehat{l}_{n,ad}(\theta) = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\widehat{Y}_{in}-\theta}{\sqrt{\widehat{V}_{nJ}}}\right)^2 + o_p(1). \tag{A.20}$$

It can be proved that $\widetilde{V}_n \xrightarrow{p} \widetilde{V}(\theta)$. This together with (A.20) and Theorem 3.2 proves Theorem 4.2.

Proof of Theorem 4.3 Under assumptions (C.X), (C.T), (C.Y), (C.P_1), (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) $0 < \inf_{t\in[0,1]} r_n(t) \le \sup_{t\in[0,1]} r_n(t) < \infty$;

(iii) $\sup_{x,t} E^*[Y^*|X^* = x, T^{**} = t] < \infty$; (iv) $\inf_{x,t} P^*(\delta^* = 1|X^* = x, T^{**} = t] > 0$; (v) $\Sigma^* = E^*[P((X^*, T^{**})u(X^*, T^*)u(X^*, T^*)^\top]$ 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)i,ii 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.

REFERENCES

Ahn, H., and J.L. Powell (1993). Estimation of Censored Selection Models with a Nonparametric Selection Mechanism. *Journal of Econometrics*, 58, 3-30.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and J. A. Wellner (1993). *Efficient and adaptive estimation for semiparametric models*. The John Hopkins University Press, Baltimore and London.

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.

Cuzick, J. (1992). Semiparametric additive regression. *Journal of the Royal Statistical Society, Series B* , **54**, 831-843.

Engle,R.F., C.W.J. Granger, J. Rice and A. Weiss (1986). Nonparametric estimates of the relation between weather ang 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.

Hastie, T.J. and Tibshirani, R. J.(1990). Generalized Additive Models. Chapman and Hall.

Healy, M.J.R. and Westmacott,M.(1956). Missing values in experiments analyzed on automatic computers. *Appl. Statist.*

Heckman, J., H. Ichimura, and P. Todd (1998). Matching as an Econometric Estimator. Review of Economic Studies, 65, 261–294.

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 Working Paper 251. Forthcoming in *Econometrica*.

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.

Nan, B., M. Emond, and J.A. Wellner (2000). Information bounds for regression models with missing data.

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.

Peixoto, J. L. (1990). A property of well-formulated polynormial regression models. *American Statistician* **44**, 26-30.

Rao,J. N. K. (1996). On variance estimation with imputed survey data (with discussion). *J. Amer. Statist. Assoc.* **91** 499-520.

Rice, J. (1986). Convergence rates for partially splined models. *Statistics & Probability Letters*, **4**, 203-208.

Robins, J., and A. Rotnizky (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.

Rosenbaum, P. and Rubin, D.B. (1983). The central role of the propensity score in

observational studies for causal effects. *Biometrika,* 70, pp. 41-55.

Rotnizky, A., and J. Robins (1997). Analysis of semi-parametric regression models with non-ignorable non-response. *Statistics in Medicine* 16, 81-102.

Robins, J., and Y. Ritov (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semiparametric models. *Statistics in Medicine* 16, 285-319.

Robins, J., F. Hsieh, and W. Newey (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *Journal of the Royal Statistical Society,* Ser B 57, 409-424.

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

Scharfstein, D.O., Rotnizky, A., and J. Robins (1999). Adjusting for nonignroable drop out in semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association* 94, 1096-1146.

Severini, T. A. and Staniswalis, J. G. (1994). Quasilikelihood estimation in semiparametric models. *Journal of the American Statistical Asssociation*, **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.

Stone, C.J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* 8, 1348-1360.

Wang, Q. H. and Rao, J.N.K. (2001). Empirical Likelihood for linear regression models under imputation for missing responses. *The Canadian Journal of Statistics*, 29, 597-608.

Wang, Q. H. and Rao, J.N.K. (2002a). Empirical Likelihood-based Inference under imputation with Missing Response. *Ann. Statist.*, 30, 896-924.

Wang, Q. H. and Rao, J.N.K. (2002b). Empirical likelihood-based inference in linear models with missing data. *Scandinavian Journal of Statistics*, 29, 563-576.

Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Emp. J. Exp. Agric.*. **1**, 129-142.

Table 5.1. Biases of $\widehat{\theta}_I$, $\widehat{\theta}_{MA}$, $\widehat{\theta}_{P_1}$, $\widehat{\theta}_P$, $\widehat{\theta}_{P_{\widehat{\tau}},1}$ and $\widehat{\theta}_{P_{\widehat{\tau}},2}$ under different missing functions $P(x)$ and different sample sizes n

$P(x)$	n	$\widehat{\theta}_I$	$\widehat{\theta}_{MA}$	$\widehat{\theta}_{P_1}$	$\widehat{\theta}_P$	$\widehat{\theta}_{P_{\widehat{\tau}},1}$	$\widehat{\theta}_{P_{\widehat{\tau}},2}$
	30	-0.0085	-0.0078	-0.0079	-0.0066	-0.0089	-0.0102
$P_1(x)$	60	0.0007	-0.0008	0.0005	-0.0022	0.0006	-0.0040
	100	0.0004	0.0003	0.0004	-0.0017	0.0004	-0.0029
	30	-0.0021	-0.0018	-0.0016	0.0027	0.0020	0.0086
$P_2(x)$	60	0.0016	0.0011	0.0011	0.0014	0.0013	0.0045
	100	0.0009	0.0007	0.0007	0.0011	0.0009	-0.0024
	30	0.0039	0.0037	0.0037	0.0094	0.0040	-0.0074
$P_3(x)$	60	-0.0031	-0.0029	-0.0028	-0.0072	-0.00026	-0.0068
	100	0.0025	0.0025	0.0022	0.0057	0.0023	-0.0052

Table 5.2. Standard errors (SE) of $\widehat{\theta}_I$, $\widehat{\theta}_{MA}$, $\widehat{\theta}_{P_1}$, $\widehat{\theta}_P$, $\widehat{\theta}_{P_{\widehat{\tau}},1}$, $\widehat{\theta}_{P_{\widehat{\tau}},2}$ under different missing functions $P(x)$ and different sample sizes n

$P(x)$	n	$\widehat{\theta}_I$	$\widehat{\theta}_{MA}$	$\widehat{\theta}_{P_1}$	$\widehat{\theta}_P$	$\widehat{\theta}_{P_{\widehat{\tau}},1}$	$\widehat{\theta}_{P_{\widehat{\tau}},2}$
	30	0.3227	0.3231	0.3230	0.3268	0.3246	0.3828
$P_1(x)$	60	0.2206	0.2202	0.2202	0.2501	0.2355	0.2741
	100	0.1776	0.1718	0.1718	0.1915	0.1806	0.2098
	30	0.3355	0.3350	0.3351	0.3392	0.3376	0.3850
$P_2(x)$	60	0.2340	0.2343	0.2343	0.2614	0.2403	0.2870
	100	0.1866	0.1866	0.1867	0.2087	0.1946	0.2112
	30	0.3502	0.3496	0.3499	0.3621	0.3508	0.3933
$P_3(x)$	60	0.2534	0.2528	0.2528	0.2672	0.2539	0.2891
	100	0.1911	0.1910	0.1911	0.2119	0.1983	0.2215

Table 5.3. Empirical coverages and average lengths of the confidence intervals on θ under different missing functions $P(x)$ and sample sizes n when nominal level is 0.95

		Empirical Coverages			Average Lengths		
$P(x)$	n	AEL	BEL	NA	AEL	BEL	NA
	30	.9200	.9750	.9220	0.8700	1.1400	1.1734
$P_1(x)$	60	.9240	.9620	.9280	0.6900	0.7900	0.8539
	100	.9450	.9580	.9440	0.5400	0.6000	0.6691
	30	.9160	.9770	.9190	0.9900	1.4500	1.3599
$P_2(x)$	60	.9220	.9640	.9250	0.7700	0.9500	0.9460
	100	.9430	.9590	.9450	0.6000	0.7300	0.7290
	30	.9140	.9820	.9170	1.1200	1.5100	1.4587
$P_3(x)$	60	.9210	.9690	.9230	0.7800	1.0500	0.9983
	100	.9390	.9580	.9390	0.6200	0.7600	0.7664