SEMIPARAMETRIC REGRESSION ANALYSIS UNDER IMPUTATION FOR MISSING RESPONSE DATA*

by

Qihua Wang
Chinese Academy of Science, Beijing

Oliver Linton
London School of Economics and Political Science

Wolfgang Härdle
Humboldt-Universität zu Berlin

Contents:

The Suntory Centre
Suntory and Toyota International Centres for Economics and Related Disciplines
London School of Economics and Political Science
Houghton Street
London WC2A 2AE
Tel.: 020 7955 6698

Discussion Paper
No.EM/03/454
May 2003

* The authors thank an associate editor and two referees for their constructive suggestions and comments that led to significant improvements. The research was supported by Humboldt-Universität zu Berlin Sonderforschungsbereich 373, the National Natural Science Foundation of China (No.10231030, No.10241001), and the Economic and Social Research Council of the UK.

Abstract

We develop inference tools in a semiparametric regression model with missing response data. A semiparametric regression imputation estimator, a marginal average estimator and a (marginal) propensity score weighted estimator are defined. All the estimators are proved to be asymptotically normal, with 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, and 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. Furthermore, a real data analysis is given to illustrate our methods.

Keywords: Asymptotic normality; empirical likelihood; semiparametric imputation.

JEL Nos.: C10, C12, C13, C14.

© by the authors. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without special permission provided that full credit, including © notice, is given to the source.

Contact addresses:
Professor Qihua Wang, Academy of Mathematics and System Science, Chinese Academy of Science, Beijing 100080, People's Republic of China.

Professor Oliver Linton, Department of Economics, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK. Email: o.linton@lse.ac.uk

Professor Wolfgang Härdle, Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, 10178 Berlin, Germany.

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 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.

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

$$(X_i, Y_i, \delta_i), i = 1, 2, \ldots, n, \tag{1.1}$$

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 are 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

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

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 $\widehat{\theta}$ and gave a consistent estimator of its asymptotic variance. 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), Silverman (1986)). 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 Tibhirani (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 others being modelled nonparametrically. This motivates us to consider the following semiparametric regression model:

$$Y_i = X_i^\top \beta + g(T_i) + \epsilon_i, \tag{1.2}$$

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. 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. We only treat the case where d^*=1 but our techniques and results apply more generally with slight modification. 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, 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 Gray (1994), and have been studied extensively for complete data settings, see Heckman (1986), Rice (1986), Speckman (1988), Cuzick (1992a, b), Chen (1988) and Severini, Staniswalis (1994) and Härdle, Liang and Gao (2000).

An alternative modelling strategy is to restrict the propensity score function $P(x,t)$ to be semiparametric, say generalized partially linear, and to use the propensity score methods to estimate θ. Propensity score based methods 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. This would be an interesting alternative to our approach, and one that also avoids the curse of dimensionality. One argument in favor of our approach is that modelling of an ancillary quantity like the propensity score does not seem as appealing as modelling the relationship of interest. In addition, reasonable semiparametric models of the propensity score would imply nonlinear semiparametric estimation, which would be less attractive in practice, we believe. Nevertheless, this remains a sensible and interesting alternative to our approach.

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

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

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 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. Under the model specification the estimators are consistent and asymptotically equivalent. We obtain their asymptotic distribution and provide consistent variance estimators based on the jacknife method. We also show that our estimators are semiparametrically efficient in the special case that ϵ_i are homoskedastic and Gaussian. When the model specification (1.2) is incorrect, our estimators are inconsistent; we characterize their biases. One of our 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 advantages 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.

The outline of the paper is as follows. In Section 2, we define the estimators of θ and state their asymptotic properties. In Section 3, we make some comparisons between the proposed estimators and the related estimators and discuss the asymptotic efficiency problem. We then develop methods for inference about θ based on empirical likelihood and bootstrap. 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, and a bootstrap empirical log-likelihood ratio is derived and its distribution is used to approximate that of the imputed empirical log-likelihood ratio. In Section 5, a simulation study is conducted to calculate the biases and the standard errors of the proposed estimators and compare the finite sample properties of the proposed empirical likelihood methods with the normal approximation based method. In Section 6, a real data analysis is given to illustrate the proposed methods. The proofs for the main results are delayed to the Appendix A. In Appendix B, we establish the semiparametric efficiency bound for the case where ϵ is i.i.d.

Gaussian. We use "$\xrightarrow{\mathcal{L}}$" to denote convergence in distribution and "$\xrightarrow{p}$" to denote convergence in probability.

2 Estimation and Asymptotic Normality

We define the three different estimators that we will analyze in this paper. All three are based on only one-dimensional smoothing operations and are closely related.

Premultiplying (1.2) 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 $\widetilde{g}_{2n}(t)$.

The regression imputation estimator of θ is then defined to be

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

We also consider two other estimators. First, 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. Second, the (marginal) propensity score weighted estimator

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

where $\widehat{P}_1(t) = \sum_{j=1}^n \delta_j K\left(\frac{t-T_j}{h_n}\right) / \sum_{j=1}^n K\left(\frac{t-T_j}{h_n}\right)$ is an estimate of $P_1(t) = P(\delta = 1|T = t)$. Estimator $\widehat{\theta}_P$ is different from the usual propensity score weighting method that uses an estimator of the full propensity score.

We next state the properties of $\widehat{\theta}_D$, $D = I, MA, P$, and propose consistent variance estimators. Let $P_1(t) = P(\delta = 1|T = t)$, $P(x,t) = P(\delta = 1|X = x, T = t)$, $m(x,t) = x^\top\beta + g(t)$, $\sigma^2(x,t) = E[(Y - X^\top\beta - g(T))^2|X = x, T = t]$, $u(x,t) = x - g_1(t)$, $\Sigma = E[P(X,T)u(X,T)u(X,T)^\top]$, and $\Omega = E[u(X,T)u(X,T)^\top\sigma^2(X,T)P(X,T)]$.

THEOREM 2.1. *Under all the assumptions listed in the Appendix except for condition (C.K)iii, we have [for $D = I, MA, P$]*

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

where:

$$V = E\left[\frac{P(X,T)}{P_1^2(T)}\sigma^2(X,T)\right] + E[u(X,T)^\top]\Sigma^{-1}\Omega\Sigma^{-1}E[u(X,T)] + \text{var[m(X, T)]}.$$

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 be difficult to 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. To obtain consistent variance estimators we need the influence functions of the three estimators, which are

$$\widehat{\theta}_D - \theta = \frac{1}{n}\sum_{i=1}^{n}\eta(Y_i,\delta_i,X_i,T_i) + o_p(n^{-1/2}), \quad D = I, MA \text{ and } \quad P$$

(see (A.4), (A.7) and (A.15)), where

$$\eta(Y_i,\delta_i,X_i,T_i) = \left(\frac{\delta_i}{P_1(T_i)} + E[(X - g_1(T))^\top]\Sigma^{-1}\delta_i(X_i - g_1(T_i))\right)\epsilon_i + m(X_i,T_i)$$

We replace the unknown quantities $P_1(T_i), \beta, g(T_i), g_1(T_i), \theta$ by estimators and take

$$\widehat{V} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\eta}_i\widehat{\eta}_i^\top.$$

It should be pointed out that this method uses an estimator of the main term of the asymptotic expansion of $\widehat{\theta}_D - \theta$.

An alternative is the jackknife variance estimator. This is more computationally demanding but imposes less conceptual demands on the practitioner. Let $\widehat{\theta}_D^{(-i)}$ be $\widehat{\theta}_D$ based on $\{(Y_j,\delta_j,X_j,T_j)\}_{j=1}^n - \{(Y_i,\delta_i,X_i,T_i)\}$ for $i = 1, 2, \ldots, n$. Let J_{ni} be the jackknife pseudo-values. That is,

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

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

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

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 Discussion

3.1 Comparison of our methods

Note that $\widehat{\theta}_P = \widehat{\theta}_{MA} + n^{-1}\sum_{i=1}^n \delta_i \widehat{\epsilon}_i / \widehat{P}_1(T_i)$, $\widehat{\theta}_I = \widehat{\theta}_{MA} + n^{-1}\sum_{i=1}^n \delta_i \widehat{\epsilon}_i$, where $\widehat{\epsilon}_i = Y_i - X_i^\top \widehat{\beta}_n - \widehat{g}_n(T_i)$, so that both $\widehat{\theta}_I$ and $\widehat{\theta}_P$ can be viewed as different adjustments to the marginal average estimator. The asymptotic equivalence result in Theorem 2.1 is similar to that obtained in Cheng (1994, Theorem 2.1) between the marginal average and the imputation estimator. It is interesting that the propensity score weighting estimator also shares this distribution. The estimators may differ in their higher order properties.

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.

Suppose that the partially linear model assumption (1.2) is incorrect, and let $m_*(x,t)$ be the probability limit of $x^\top \widehat{\beta}_n + \widehat{g}_n(t)$. Then the three estimators are asymptotically biased with

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

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 $\widehat{\theta}_P$ is small. Especially, the asymptotic bias of $\widehat{\theta}_P$ is zero when $m_*(x,t) = m(x,t)$ or $P(x,t) = P_1(t)$ by (3.1). This implies

that $\widehat{\theta}_P$ possesses the 'double robustness' property, namely that even if the mean specification is incorrect, i.e., $m(x,t) \neq \beta^\top x + g(t)$, $\widehat{\theta}_P$ is still consistent provided that $P(X,T) = P_1(T)$. This property has been discussed by Scharfstein, Rotnizky, Robins (1999).

3.2 Comparison with methods for unrestricted regression

We now compare our method with three alternative fully nonparametric procedures: the nonparametric kernel regression imputation estimator $\widetilde{\theta}_c$ due to Cheng (1996), the estimator

$$\widetilde{\theta}_{HIR} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i \delta_i}{\widehat{P}(X_i, T_i)}$$

due to Hirano *et al.* (2000) based on an estimator $\widehat{P}(x,t)$ of the propensity score constructed by kernel smoothing the participation indicator against covariate values, and the weighted estimator

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

where $\widehat{m}_n(X_i, T_i)$ is the nonparametric regression kernel estimator of the regression Y on (X,T). The three nonparametric estimators are all asymptotically equivalent with asymptotic variance

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

This is exactly the so-called 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). As we have pointed out already when X, T are high dimensional a major disadvantage of $\widehat{\theta}_c$, $\widetilde{\theta}_{HIR}$ or $\widetilde{\theta}_W$ is that they require a high-dimensional smoothing operation to compute the regressions 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.

Now consider the cases where $m(x,t)$ is restricted to the partially linear structure. The semiparametric efficiency bound here may be strictly lower than in the unrestricted case. Therefore, the nonparametric estimators may not be asymptotically efficient for the partially linear model. Our estimators all make use of the

partial linear structure in the conditional mean and hence it is possible for them to do better than the above three nonparametric estimators.

Do the estimators $\widehat{\theta}_D$, $D = I, MA, P$ have less asymptotic variance than $\widehat{\theta}_c$, $\widetilde{\theta}_{HIR}$ or $\widetilde{\theta}_W$? We next consider two special cases where the inequality $V \leq V_{UR}$ holds.

First, suppose that $P(X,T) = P_1(T)$. Then, by noting $Eu(X,T) = 0$, we have

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

which is the same as the asymptotic variances of the three nonparametric estimators.

Second, consider the homoskedastic special case where $\sigma^2(x,t) = \sigma^2$, where σ is a constant. In this case,

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

We claim that

$$V \leq V_{UR} \tag{3.2}$$

in this case. First note that

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

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,

$$\frac{\sigma^2 E[u(X,T)^\top](\sigma^2\Sigma)^{-1}\sigma^2 E[u(X,T)]}{\mathrm{var}\left(\left(\frac{\delta}{\mathrm{P(X,T)}} - \frac{\delta}{\mathrm{P_1(T)}}\right)\epsilon\right)} \leq 1, \tag{3.3}$$

because the left hand side can be interpreted as a squared correlation by the above argument. Then note that

$$\mathrm{var}\left[\left(\frac{\delta}{\mathrm{P(X,T)}} - \frac{\delta}{\mathrm{P_1(T)}}\right)\epsilon\right] = \sigma^2 \mathrm{E}\left[\frac{1}{\mathrm{P(X,T)}} - \frac{1}{\mathrm{P_1(T)}}\right]. \tag{3.4}$$

Combining (3.3) and (3.4) we have

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

i.e., $V \leq V_{UR}$ as claimed in (3.2). 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. This shows that our estimator is asymptotically more efficient than the three nonparametric estimators for the special case of homoskedasticity. This also supports the claim that the semiparametric efficiency bound under the partially linear structure may be strictly lower than in the unrestricted case.

We prove in Appendix B that V is the semiparametric efficiency bound for the case that ϵ is i.i.d. Gaussian. This shows that the proposed estimators $\widehat{\theta}_I, \widehat{\theta}_{MA}$ and $\widehat{\theta}_P$ are asymptotically efficient for the special case. Incidentally, $\widehat{\beta}_n$ is also semiparametrically efficient.

We now comment on the general heteroskedastic case. In this case it is clear that none of the estimators $\widehat{\theta}_I$, $\widehat{\theta}_P, \widehat{\theta}_{MA}, \widehat{\theta}_c$, $\widetilde{\theta}_{HIR}$ or $\widetilde{\theta}_W$ are efficient for the partially linear model considered here. One reason for this is 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 our model 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 (1999), 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.

There is another useful comparison with the literature on estimating additive

nonparametric regression. The backfitting methodology of Hastie and Tibshirani (1990) requires only iterative one dimensional smoothing operations and is very popular and has good properties when the error is homoskedastic. When the error is heteroskedastic in some general way, this method can be less efficient than some competitors like the marginal integration estimator (Linton and Nielsen (1995)). Nevertheless, whether it is desirable to pursue efficiency gains by weighting the backfitting iterations is questionable if it requires high dimensional smoothing operations.

3.3 Comparison with methods for Modelled Propensity Score

To consider the partial linear structure and improve the efficiency, one may define an estimator $\widetilde{\theta}_P^*$ to be $\widetilde{\theta}_P$ with $\widehat{m}(X,T)$ replaced by $X^\top\widehat{\beta}_n + \widehat{g}_n(T)$. That is,

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

It can be shown that this estimator is asymptotically normal with the same asymptotic variance as the weighted nonparametric estimator $\widetilde{\theta}_P$ with $m(X,T) = X^\top\beta + g(T)$. This shows that $\widetilde{\theta}_P^*$ cannot be an asymptotic efficient estimator for the model considered here. Also, $\widetilde{\theta}_P^*$ has the same disadvantages as $\widehat{\theta}_{HIR}$, requiring a high dimension smoothing technique to compute the propensity score when the propensity score is unknown completely.

Suppose instead we replaced $\widehat{P}(X_i,T_i)$ by a semiparametric estimator, say one based on fitting the semiparametric model

$$P(X,T) = F(\alpha^\top X + \gamma(T)),$$

where F is a known c.d.f., and the function $\gamma(.)$ and parameters α are unknown. If the model for the propensity score is correct, then we can expect some efficiency gains depending on the model, at least in the homoskedastic case. Also, this method does not require high dimension smoothing operations. However, the estimation procedure to obtain $\alpha, \gamma(.)$ can be quite complicated - it usually involves nonlinear optimization of a criterion function that contains nonparametric estimators. This can be expected to be very time consuming and not perform statistically as well

as is perhaps indicated by the asymptotic theory. This approach has the so-called double robustness property whereby even if one of the two models [for the propensity score or the mean] is incorrect, the estimator is still consistent. Note that our estimator $\widehat{\theta}_P$ also has the double robustness property: when either $P(x,t) = P_1(t)$ or $m(x,t) = \beta^\top x + g(t)$, then $\widehat{\theta}_P$ is consistent. Also, it only requires one dimensional smoothing operations.

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* $H_0 : \theta = \theta_0$,

$$\widehat{l}_n(\theta) \xrightarrow{\mathcal{L}} \frac{V}{\widetilde{V}}\chi_1^2,$$

where χ_1^2 *is a standard chi-square variable with one degree of freedom,* V *is defined in Theorem 2.1 and* $\widetilde{V} = E[P(X,T)\sigma^2(X,T)] + Var(X^\top\beta + g(T))$.

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

$$\gamma\widehat{l}_n(\theta) \xrightarrow{\mathcal{L}} \chi_1^2, \quad (4.4)$$

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

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

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

A consistent estimator of γ_n can be defined as

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

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

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

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 $\widehat{\theta}_P$ and the weighted estimator $\widetilde{\theta}_P^*$ given in Section 3, 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

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

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 $|x-1|+|t-0.5| \le 1$, and 0.95 elsewhere;

Case 2: $P(\delta = 1|X = x, T = t) = 0.9-0.2(|x-1|+|t-0.5|)$ if $|x-1|+|t-0.5| \le 4$, 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. Let $\widetilde{\theta}^*_{P,1}$ be $\widetilde{\theta}^*_P$ with $\widehat{P}(x,t)$ 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 $|u| \le 1$, 0 otherwise; $K_2(v) = \frac{15}{16}(1 - 2t^2 + t^4)$ if $|v| \le 1$, 0 otherwise and $h_{1,n} = h_{2,n} = n^{-\frac{1}{3}}$. Let $\widetilde{\theta}^*_{P,2}$ be $\widetilde{\theta}^*_P$ with

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

for case 1;

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

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 $\bar{T} = n^{-1}\sum_{i=1}^{n} T_i$.

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 5.1. From the 5000 simulated values of $\widehat{\theta}_I, \widehat{\theta}_{MA}, \widehat{\theta}_P$, $\widehat{\theta}^*_{P,1}$ and $\widehat{\theta}^*_{P,2}$, we calculated the biases and standard errors of the five estimators. These simulated results are reported in Tables 5.2 and 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 5.1 here

From Table 5.1, 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.

Insert Tables 5.2 and 5.3 here

From Tables 5.2 and 5.3, 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) $\widehat{\theta}_I, \widehat{\theta}_{MA}$, $\widehat{\theta}_P$ and $\widetilde{\theta}^*_{P,2}$ have smaller SE than $\widetilde{\theta}^*_{P,1}$. Generally, $\widetilde{\theta}^*_{P,1}$ also has slightly bigger bias than other estimators. This suggests that our estimators and $\widetilde{\theta}^*_{P,2}$ outperform $\widetilde{\theta}^*_{P,1}$, a propensity score weighted estimator that uses the nonparametric kernel estimator of the full propensity score. From the simulation results, the weighted estimator $\widetilde{\theta}^*_P$ indeed performs well if the propensity score can be specified correctly.

6 Real Data Analysis

We considered the real data set given in Peixoto (1990). The data gives the normal average January minimum temperature in degrees Fahrenheit (JanTemp) with the latitude (Lat) and longitude (Long) of 56 U.S. cities. (For each year from 1931 to 1960, the daily minimum temperatures in January were added together and divided by 31. Then, the averages for each year were averaged over the 30 years). The data set is also available on http://lib.stat.cmu.edu/DASL/Datafiles/USTemperatures.html.

Peixoto (1990) reports a study in which a linear relationship is assumed between JanTemp and Lat; then, after removing the effects of Lat, a cubic polynomial in Long is used to predict JanTemp. To apply the real data to our problem, we denote the variables for JanTemp, Lat and the natural lagarithm of Long to be Y, X and T respectively. We suppose that Y, X and T satisfy the partial linear model considered here. Figure 6.1 plots the estimated curve $g_n(t)$ of $g(t)$ based on the complete observations (X, T, Y), where $g_n(t)$ is $\widehat{g}_n(t)$, which is defined in Section 2, with δ_i replaced by 1 for $i = 1, 2, \cdots, n$, and the kernel function $K(\cdot)$ taken to be the same as in Section 5 and the bandwidth h_n taken to be $n^{-\frac{1}{3}}$.

Insert Figure 6.1 here

We used this data and deleted 13 Y values given in parentheses in Table 6.1. The deletion mechanism is designed to be MAR with $P(x, t) = 0.75$.

Insert Table 6.1 here

Based on the incomplete data set, we may develop inference on the mean of Y with the methods given before. It is noted that the original data set given by Peixoto (1990) is complete. Inference on the mean of Y with the complete data set doesn't depend on the model assumption and covariables X and T. This just provides us a standard to compare our methods with other methods to handle missing data. For example, we may compare our semiparametric regression imputation estimator $\widehat{\theta}_I$ with the nonparametric kernel regression imputation estimator $\widetilde{\theta}_c$ due to Cheng (1994) by comparing them with the sample mean $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, and compare the empirical likelihood method based on the semiparametric imputation with that based on the nonparametric imputation by comparing them with the standard empirical likelihood based on the complete observations Y.

We calculated $\bar{Y} = 26.5179$ with the complete observations Y, and the estimated values $\widehat{\theta}_I = 26.3131$ and $\widetilde{\theta}_c = 24.4046$, respectively, based on the incomplete data set which are obtained by deleting some Y values with MAR deletion mechanism. When calculating $\widehat{\theta}_I$, the kernel function $K(t)$ and the bandwidth h_n were taken to be the same as in Section 5. For calculation of $\widetilde{\theta}_c$, $\widehat{m}_n(x,t)$ was taken to be

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

where $K_1(\cdot)$, $K_2(\cdot)$, $h_{1,n}$ and $h_{2,n}$ are the same as in Section 5. The value for sample variance estimate of $\bar{Y}$ is 3.1397, and jackknife variances for estimators $\widehat{\theta}_I$ and $\widetilde{\theta}_c$ are 4.3607 and 4.4478 respectively

From the estimated values, $\widehat{\theta}_I$ is closer to the sample mean $\bar{Y}$ than $\widetilde{\theta}_c$. It is also clear $\widehat{\theta}_I$ has smaller jackknife variance estimate than $\widetilde{\theta}_c$.

With the jackknife variance estimators, we calculated the normal aproximation confidence intervals of θ with confidence level 0.95 based on the asymptotic normality of $\bar{Y}, \widehat{\theta}_I$ and $\widetilde{\theta}_c$. The confidence intervals calculated are $(23.0449, 29.9908)$, $(22.2202, 30.4060)$ and $(20.2710, 28.5382)$, respectively. Their lengths are 6.9459, 8.1858 and 8.2672 respectively. Clearly, the confidence intervals based on $\widehat{\theta}_I$ are closer to that based on the complete sample mean $\bar{Y}$ and have shorter lengths.

All in all, the calculation results show that the semiparametric imputation estimator performs better than the nonparametric imputation estimator for the real data example in terms of biases of the two estimators and the lengths of confidence intervals.

Next, we compare the semiparametric imputed empirical likelihood and the nonparametric imputed empirical likelihood with the standard empirical likelihood with complete observations due to Owen (1988). The standard empirical log-likelihood function, $l_{n,S}(\theta)$ is defined by (4.2) and (4.3) with $\widehat{Y}_{in}$ replaced by Y_i and the nonparametric imputed empirical log-likelihood function $\widehat{l}_{n,c}(\theta)$ is defined by (4.2) and (4.3) with $\widehat{Y}_{in}$ replaced by $Y_{in} = \delta_i Y_i + (1-\delta_i)\widehat{m}_n(X_i, T_i)$ for $i = 1, 2, \cdots, n$.

Figure 6.2 plots the curves for the standard empirical log-likelihood function $l_{n,S}(\theta)$, semiparametric imputed empirical log-likelihood function $\widehat{l}_n(\theta)$ and nonparametric empirical log-likelihood function $\widehat{l}_{n,c}(\theta)$. Figure 6.3 plots the curves for $l_{n,S}(\theta)$, $\widehat{l}_{n,ad}(\theta)$ and the adjusted empirical log-likelihood functon of $\widehat{l}_{n,c}(\theta)$ given by

$$\widehat{l}_{nc,ad}(\theta) = \gamma_{n,c}(\theta)\widehat{l}_{n,c}(\theta),$$

where $\gamma_{n,c}(\theta) = V_{n,c}(\theta)/V_{n,cJ}$ with $V_{n,cJ}$ the jackknife variance estimator of $\widetilde{\theta}_c$ and $V_{n,c}(\theta) = n^{-1}\sum_{i=1}^{n}(Y_{in} - \theta)^2$.

Insert Figures 6.2 and 6.3 here

From Figures 6.2 and 6.3, the curves for $\widehat{l}_n(\theta)$ and its adjusted version $\widehat{l}_{n,ad}(\theta)$ are closer to the standard empirical log-likelihood function $l_{n,S}(\theta)$ than $\widehat{l}_{n,c}(\theta)$ and $\widehat{l}_{ac,ad}(\theta)$ respectively. The curves $\widehat{l}_{n,c}(\theta)$ and its adjusted version shift to the left of $\widehat{l}_n(\theta)$. This implies that the nonparametric empirical likelihood method may construct confidence interval with lower coverage than the semiparametric imputed empirical likelihood method. It was calculated that the confidence intervals based on $l_{n,S}(\theta), \widehat{l}_{n,ad}(\theta)$ and $\widehat{l}_{nc,ad}(\theta)$ with confidence level 0.95 are $\{\theta : l_{n,S}(\theta) \leq 3.8415\} = (23.1500, 30.2500)$, $\{\theta : \widehat{l}_{n,ad}(\theta) \leq 3.8415\} = (22.4000, 30.3500)$ and $\{\theta : \widehat{l}_{nc,ad}(\theta) \leq 3.8415\} = (20.6000, 28.6500)$, where 3.8415 is 0.95 quantile of standard chi-square

with one degree of freedom. The lengths for these confidence intervals are 7.1000, 7.9500 and 8.0500 respectively. It is observed that the confidence intervals based on the semiparametric regression imputation are closer to that based on the complete sample mean $\bar{Y}$ and have shorter lengths. This also can be seen from Figure 6.3.

Appendix A: Assumptions and Proofs of Theorems

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

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 2.1. (i) We prove Theorem 2.1 for $\widehat{\theta}_I$. For $\widehat{\theta}_I$, we have

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

Note that

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

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

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

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

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

(ii) We prove Theorem 2.1 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

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

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 2.1 for $\widehat{\theta}_{MA}$.

(iii) We prove Theorem 2.1 for $\widehat{\theta}_P$. For $\widehat{\theta}_P$, we have

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

For T_{n5}, we have

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

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

$$\begin{aligned} T_{n2} &= \tfrac{1}{n}\sum_{i=1}^{n}\tfrac{\delta_i\epsilon_i\{\widehat{P}_1(T_i)-P_1(T_i)\}}{P_1^2(T_i)} \\ &= \tfrac{1}{n}\sum_{i=1}^{n}\tfrac{\delta_i\epsilon_i}{P_1^2(T_i)}\tfrac{1}{nh}\textstyle\sum_{j=1}^{n}K\left(\frac{T_i-T_j}{h_n}\right)\frac{[\delta_j-P_1(T_j)]}{f_T(T_i)}+o_p(n^{-\frac{1}{2}}) \\ &= \tfrac{1}{n}\textstyle\sum_{j=1}^{n}[\delta_j-P_1(T_j)]\tfrac{1}{nh}\sum_{i=1}^{n}\tfrac{\delta_i\epsilon_i}{P_1^2(T_i)}K\left(\frac{T_i-T_j}{h_n}\right)\frac{1}{f_T(T_i)}+o_p(n^{-\frac{1}{2}}) \\ &= o_p(n^{-1/2}). \end{aligned} \tag{A.14}$$

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

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

This together with central limit theorem proves Theorem 2.1 for $\widehat{\theta}_P$.

Proof of Theorem 2.2. Similar to (A.4),(A.8) and (A.15), 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 $\eta(Y,\delta,X,T)$ is defined in Section 2. This proves $\widehat{V}_{nJ} \xrightarrow{p} V(\theta)$.

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 $\lambda_n = O_p(n^{-\frac{1}{2}})$.

By (4.3), we get

$$0 = \sum_{i=1}^{n}\frac{(\widehat{Y}_{in} - \theta)}{1 + \lambda_n(\widehat{Y}_{in} - \theta)} = \sum_{i=1}^{n}[(\widehat{Y}_{in} - \theta)] - \sum_{i=1}^{n}\lambda_n(\widehat{Y}_{in} - \theta)^2 + \sum_{i=1}^{n}\frac{\lambda_n^2(\widehat{Y}_{in} - \theta)^3}{1 + \lambda_n(\widehat{Y}_{in} - \theta)}.$$

This implies

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

and

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

using $\widehat{Y}_{(n)} = o_p(n^{\frac{1}{2}})$ and $\lambda_n = O_p(n^{-\frac{1}{2}})$.

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

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

It can be proved $\widetilde{V}_n \xrightarrow{p} \widetilde{V}$, where $\widetilde{V}_n$ and $\widetilde{V}$ are defined in Section 4. This together with (A.19) and Theorem 2.1 proves Theorem 4.1.

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

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

This together with and Theorem 2.2 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) $0 < \inf_{t\in[0,1]} r_n(t) \leq \sup_{t\in[0,1]} r_n(t) < \infty$; (iii) $\sup_{x,t} E^*[Y^* | X^* = x, T^{**} = t] < \infty$; (iv) $\inf_{x,t} P^*(\delta^* = 1 | X^* = x, T^{**} = t] > 0$; (v) $\Sigma^* = E^*[P(X^*, T^{**})u(X^*, T^*)u(X^*, T^*)^\top]$ 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 infinitey $\widehat{l}^*_m(\widehat{\theta}_n)$ has the same asymptotic scaled chi-square distribution as $\widehat{l}_n(\theta)$. This together with Theorem 4.1 proves Theorem 4.3.

Appendix B: Derivation of Efficiency Bound

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 $\log f_{\beta,g,\sigma^2,P,f_{X,T}}(Y, \delta, X, T; \lambda)$, 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 $\log f_{\beta,g,f_\epsilon,P,f_{X,T}}(Y,\delta,X,T)$ when $\lambda=\lambda_0$. The score functions are:

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

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 the functions

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

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

$$\ell^*_\beta = \delta[X - g_1(T)]\epsilon,$$

as can be immediately verified. The corresponding semiparametric efficiency bound is

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

and no regular estimator can have asymptotic variance less than this. Under these conditions, our estimator $\widehat{\beta}_n$ achieves this bound.

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

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

which has derivatives

$$\begin{aligned}
\frac{\partial \theta}{\partial \beta} &= -\int YX f'_\epsilon(Y - \beta X - g_\gamma(T)|X,T;\sigma^2) f_{X,T}(X,T;\eta_{xt}) dY dX dT \\
&= -\int \epsilon X \frac{f'_\epsilon(\epsilon|X,T)}{f_\epsilon(\epsilon|X,T)} f_{X,T}(X,T) f_\epsilon(\epsilon|X,T) d\epsilon dX dT \\
&= -\frac{1}{\sigma^2} E\left[X\epsilon^2\right] = -E\left[X\right]
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \theta}{\partial \gamma} &= -\int Y \frac{\partial g_\gamma}{\partial \gamma}(T) \frac{f'_\epsilon(\epsilon|X,T)}{f_\epsilon(\epsilon|X,T)} f_{X,T}(X,T) f_\epsilon(\epsilon|X,T) d\epsilon dX dT \\
&= -E\left[\frac{\partial g_\gamma}{\partial \gamma}(T)\right]
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \theta}{\partial \sigma^2} &= \int Y \frac{\partial \log f_Y(Y|X,T)}{\partial \sigma^2} f_{X,T}(X,T) f_Y(Y|X,T) dY dX dT \\
&= -E\left[Y \frac{1}{2\sigma^2}\left(\frac{\epsilon^2}{\sigma^2} - 1\right)\right] = 0
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \theta}{\partial \eta_{xt}} &= \int Y f_Y(Y|X,T) \frac{\partial f_{X,T}(X,T)/\partial \eta_{xt}}{f_{X,T}(X,T)} f_{X,T}(X,T) dY dX dT \\
&= E\left[m(X,T) \frac{\partial f_{X,T}(X,T)/\partial \eta_{xt}}{f_{X,T}(X,T)}\right]
\end{aligned}$$

Define

$$F_\theta = \frac{\delta \epsilon}{P(X,T)} + m(X,T) - \theta.$$

Then it can be seen that

$$E\left[F_\theta s_\lambda\right] = \frac{\partial \theta}{\partial \lambda}$$

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

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

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.

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. (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,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.

Härdle,W., Liang, H. and Gao,J. (2000). Partially Linear Models. Physica-Verlag, Heidelberg.

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
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.

Linton, O.B. and Nielsen, J.P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82**, 93-101.

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.

Mammen, E., Linton, O. B., and Nielsen, J. P. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.* **27**, 1443-1490.

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. & 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.

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-ignroable 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.

Silverman, B. (1986). Density estimation for statistics and data analysis. London, Chapman and Hall.

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. (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.

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

Stone, C.J. (1985). Additive regression and other nonparametric models. *Ann. Statist.*, 13, 685-705.

Stone, C.J. (1986). The dimensionality reduction principle for Generalized additive models. *Ann. Statist.* 14, 592-606.

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

Table 5.1. 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

Table 5.2. Biases of $\widehat{\theta}_I$, $\widehat{\theta}_{MA}$, $\widehat{\theta}_P$, $\widetilde{\theta}^*_{P,1}$ and $\widetilde{\theta}^*_{P,2}$ under different missing functions $P(x)$ and different sample sizes n

$P(x)$	n	$\widehat{\theta}_I$	$\widehat{\theta}_{MA}$	$\widehat{\theta}_P$	$\widetilde{\theta}^*_{P,1}$	$\widetilde{\theta}^*_{P,2}$
	30	-0.0089	-0.0098	-0.0089	-0.0088	-0.0091
$P_1(x)$	60	0.0008	0.0003	0.0007	0.0027	0.0008
	100	0.0003	0.0001	0.0004	-0.0031	0.0003
	30	-0.0038	-0.0039	-0.0037	-0.0047	-0.0033
$P_2(x)$	60	-0.0017	-0.0022	-0.0013	-0.0034	-0.0011
	100	0.0013	0.0008	0.0016	0.0007	0.0015
	30	-0.0056	-0.0059	-0.0057	-0.0055	-0.0054
$P_3(x)$	60	0.0049	0.0049	0.0050	0.0066	0.0049
	100	0.0045	0.0043	0.0044	0.0060	0.0047

Table 5.3. Standard errors (SE) of $\widehat{\theta}_I$, $\widehat{\theta}_{MA}$, $\widehat{\theta}_P$, $\widetilde{\theta}^*_{P,1}$ and $\widetilde{\theta}^*_{P,2}$ under different missing functions $P(x)$ and different sample sizes n

$P(x)$	n	$\widehat{\theta}_I$	$\widehat{\theta}_{MA}$	$\widehat{\theta}_P$	$\widetilde{\theta}^*_{P,1}$	$\widetilde{\theta}^*_{P,2}$
	30	0.3144	0.3146	0.3145	0.3361	0.3145
$P_1(x)$	60	0.2233	0.2232	0.2236	0.2456	0.2234
	100	0.1745	0.1748	0.1747	0.2189	0.1744
	30	0.3459	0.3458	0.3480	0.3604	0.3476
$P_2(x)$	60	0.2402	0.2401	0.2415	0.2780	0.2410
	100	0.1887	0.1886	0.1899	0.2544	0.1902
	30	0.3610	0.3608	0.3632	0.3787	0.3613
$P_3(x)$	60	0.2526	0.2549	0.2522	0.2910	0.2530
	100	0.1985	0.1983	0.2000	0.2386	0.1988

Table 6.1.The normal average January minimum temperature in degrees Fahrenheit with the latitude and longitude of 56 U.S. cities.

X	31.2	32.9	33.6	35.4	34.3	38.4	40.7	41.7	40.5	39.7	31.0	25.0	26.3	33.9
T	4.48	4.46	4.72	4.53	4.78	4.81	4.66	4.30	4.33	4.35	4.41	4.41	4.39	4.44
Y	44	38	(35)	31	47	42	15	22	26	30	45	65	58	37
δ	1	1	0	1	1	1	1	1	1	1	1	1	1	1
X	43.7	42.3	39.8	41.8	38.1	39.0	30.8	44.2	39.7	42.7	43.1	45.9	39.3	47.1
T	4.76	4.48	4.47	4.54	4.58	4.46	4.50	4.26	4.35	4.27	4.43	4.54	4.51	4.72
Y	22	(19)	21	11	(22)	27	45	12	25	23	(21)	2	24	8
δ	1	0	1	1	0	1	1	1	1	1	0	1	1	1
X	41.9	43.5	39.8	35.1	42.6	40.8	35.9	36.4	47.1	39.2	42.3	35.9	45.6	40.9
T	4.57	4.28	4.32	4.67	4.30	4.31	4.40	4.37	4.62	4.44	4.41	4.58	4.81	4.35
Y	13	(11)	27	24	14	27	34	(31)	(0)	(26)	(21)	(28)	33	24
δ	1	0	1	1	1	1	1	0	0	0	0	0	1	1
X	40.9	33.3	36.7	35.6	29.4	30.1	41.1	45.0	37.0	48.1	48.1	43.4	43.3	41.2
T	4.32	4.39	4.47	4.62	4.56	4.56	4.72	4.30	4.34	4.81	4.77	4.50	4.48	4.65
Y	24	38	31	24	(49)	44	18	7	32	33	(19)	9	13	(14)
δ	1	1	1	1	0	1	1	1	1	1	0	1	1	0

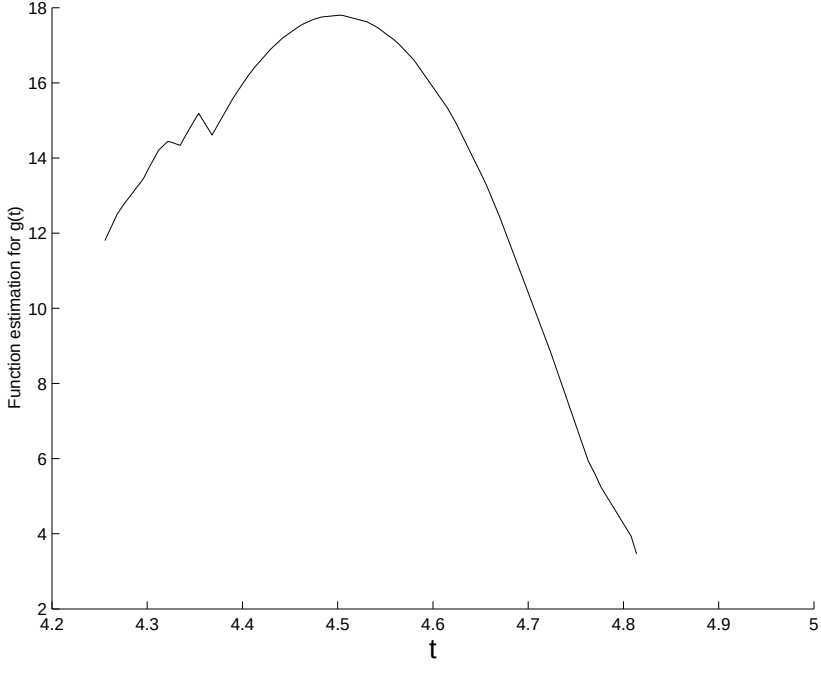

Figure 6.1. Curve for $g_n(t)$

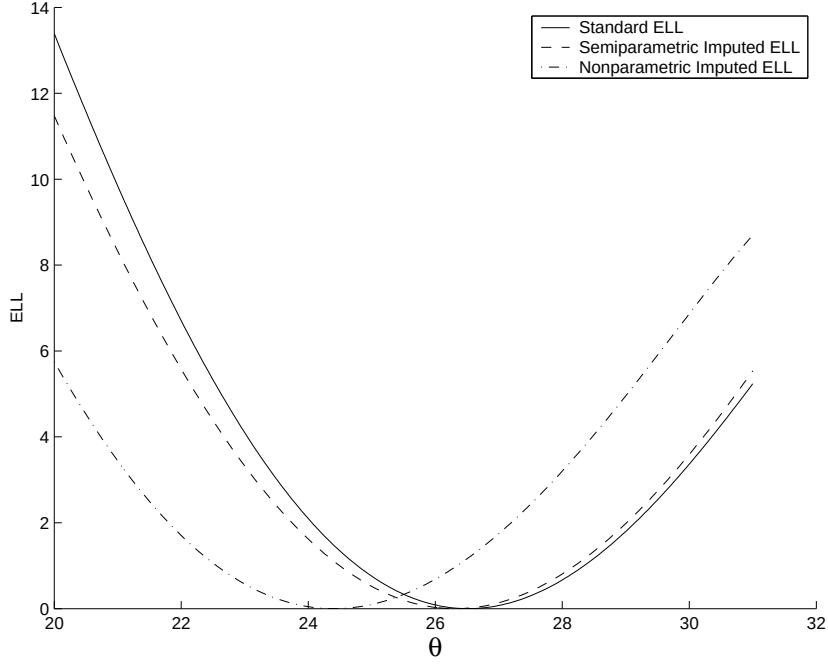

Figure 6.2. Curves for standard empirical log-likelihood (ELL) function with complete observations Y, semiparametric imputed ELL and nonparametric imputed ELL functions

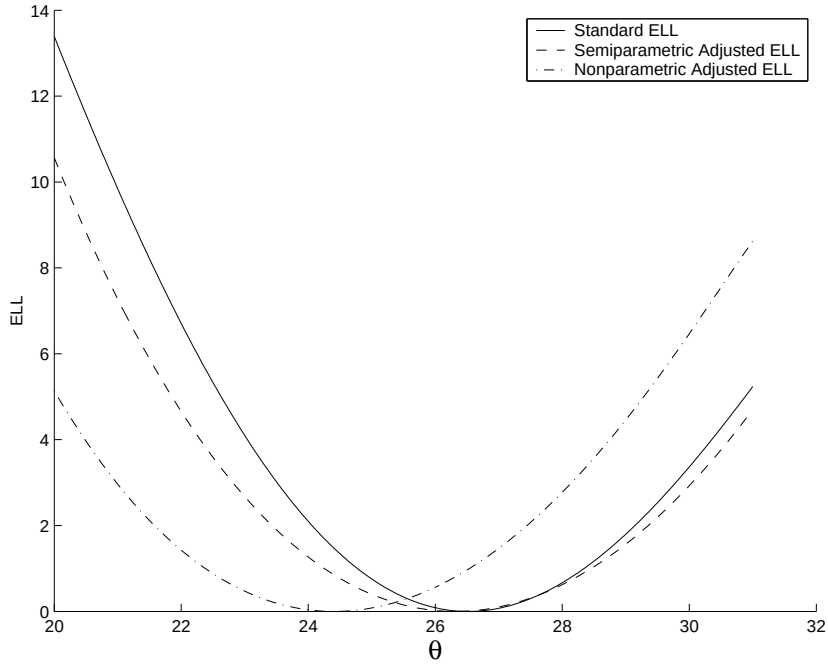

Figure 6.3. Curves for standard empirical log-likelihood (ELL) function with complete observations Y, semiparametric adjusted ELL and nonparametric adjusted ELL functions