Ann Inst Stat Math (2013) 65:721–762
DOI 10.1007/s10463-012-0391-8

Partially linear varying coefficient models with missing at random responses

Francesco Bravo

Received: 27 February 2012 / Revised: 15 October 2012 / Published online: 10 January 2013

Abstract This paper considers partially linear varying coefficient models when the response variable is missing at random. The paper uses imputation techniques to develop an omnibus specification test. The test is based on a simple modification of a Cramer von Mises functional that overcomes the curse of dimensionality often associated with the standard Cramer von Mises functional. The paper also considers estimation of the mean functional under the missing at random assumption. The proposed estimator lies in between a fully nonparametric and a parametric one and can be used, for example, to obtain a novel estimator for the average treatment effect parameter. Monte Carlo simulations show that the proposed estimator and test statistic have good finite sample properties. An empirical application illustrates the applicability of the results of the paper.

Keywords Bootstrap · Imputation · Inverse probability weighting · Missing at random

1 Introduction

Partially linear varying coefficient models are useful extensions of the popular partially linear model considered for example by Engle et al. (1986), Robinson (1988) and Speckman (1988). These models offer additional flexibility compared to partially linear models because they allow interactions between a vector of covariates and a vector of unknown functions depending on another covariate, while avoiding the curse of dimensionality typically associated with partial linear models. Partially linear vary-

F. Bravo (✉)
Department of Economics, University of York, York YO10 5DD, UK
e-mail: francesco.bravo@york.ac.uk
URL: https://sites.google.com/a/york.ac.uk/francescobravo

ing coefficient models are important examples of varying coefficient models (see e.g. Hastie and Tibshirani 1993) which arise in many situations of practical relevance in economics, finance and statistics, and have been used in the context of generalized linear models and quasi-likelihood estimation (Cai et al. 2000a), time series (Cai et al. 2000b), longitudinal data (Fan and Wu 2008), survival analysis (Cai et al. 2008) to name just a few applications-see Fan and Zhang (2008) for a recent review containing further applications and a number of examples. Compared to varying coefficient models, partially linear varying coefficient models contain the additional information that some of the coefficient are in fact constant and this information should be incorporated in the estimation to obtain more efficient estimates. Ahmad et al. (2005) and Fan and Huang (2005) suggest two general estimation techniques based, respectively, on nonparametric series and profile least square estimation. Both procedures result in efficient estimators under the additional assumption of conditional homoskedasticity of the unobservable errors.

In this paper we consider varying coefficient partially linear models when the response variable is not directly observed, but it is assumed to be missing at random (MAR henceforth). MAR is commonly assumed in many statistical models with missing data—see Little and Rubin (2002) for a comprehensive review—and it specifies that the probability of missing depends on variables that are always observed. When the responses are MAR a natural approach to follow is to impute a value for the missing responses and then estimate the parameters of interest as if the imputed responses were the true parameters. Paik (1997) showed that imputation by regression methods can improve the efficiency of estimators for generalized estimating equations models; Chen and Cui (2006) and Wang et al. (2004) considered, respectively, imputation methods with local quasi-likelihood estimators and in the context of partially linear regression. Wang and Rao (2002) and Wang and Chen (2009) combined nonparametric imputation and empirical likelihood to obtain inferences in models with MAR data.

In this paper we consider two problems: first testing for the correct specification of partially linear varying coefficient models; second estimating the mean of a response variable assuming a partially linear varying coefficient specification. The specification test we consider is based on a Cramer von Mises type of functional of a marked empirical process. This type of statistic has been used for checking the correct specification of a number of statistical models including: parametric regressions (Stute 1997; Escanciano 2006), generalized linear models (Stute and Zhu 2002), quantile regressions (He and Zhu 2003; Escanciano and Velasco 2010), partially linear regressions (Zhu and Ng 2003), single index (Stute and Zhu 2005; Xia et al. 2004), conditional moments (Whang 2001). Bravo (2012) proposes some general results for semiparametric conditional moments models, which incorporate as special cases all of above mentioned models. One important characteristic of the test statistic we propose is that it is based on the same dimension reduction approach as that proposed by Escanciano (2006). This approach yields test statistics that do not suffer from the main problem associated with the standard marked empirical process approach, namely the curse of dimensionality related to the potentially high dimension of the conditioning set of covariates. The imputation estimator for the mean functional is similar in spirit to that proposed by Wang et al. (2004) and complement the nonparametric imputation estimator of Cheng (1994), and the fully parametric imputation estimator of Scharfstein et al. (1999). The resulting estimator is rather general and can be used, for example, to construct a novel estimator of the average treatment effect parameter that is central to many causal inference models, see for example Little and Rubin (2002).

In this paper we make the following contributions: first we consider partially linear varying coefficient models with two imputation methods: a standard one based on regression and an alternative one based on inverse probability weighting. Both approaches are based on a complete version of the profile kernel estimator originally suggested by Fan and Huang (2005) when the responses are always observable. It is important however to note that other nonparametric estimators could be used, such as the nonparametric series estimator suggested by Ahmad et al. (2005). We derive the asymptotic distribution of the complete version profile kernel estimator and of the specification tests under the null and both global and local alternative hypotheses. These results extend and/or complement results obtained by Wang et al. (2004), Fan and Huang (2005), Escanciano (2006), Sun et al. (2009) and Koul et al. (2012) among others.

Second we consider two bootstrap procedures that can be used to consistently approximate the critical values of the unknown distributions of the test statistic. The first bootstrap procedure is the wild bootstrap (see e.g. Wu 1986 and Hardle and Mammen 1993), and has been used in the context of the type of specification tests considered in this paper by Stute et al. (1998), Whang (2000), Escanciano (2006) and others. The second one is motivated by the so-called random symmetrization technique (see e.g. Pollard 1984) and by the multiplier central limit theorems (see e.g. Van der Vaart and Wellner 1996), and it has been used by Su and Wei (1991), Delgado et al. (2003), Zhu and Ng (2003) and others.

Third we derive the asymptotic distributions of imputation estimators for a MAR mean functional assuming a partially linear varying coefficient specification, and derive a new semiparametric efficiency bound for this model under the assumption of normality. We also propose a new estimator of the average treatment effect parameter under Rosenbaum and Rubin (1983)'s strong ignorability assumption. These results complement those of Cheng (1994), Hahn (1998), Hirano et al. (2003), Wang et al. (2004) and Muller (2009) among others.

Fourth we use simulations to assess the finite sample properties of the proposed estimators and test statistics. As a by-product of these simulations we are able to compare the relative performances of the two bootstrap procedures. This comparison is, as far as we are aware of, new in the context of specification testing.

Finally we illustrate the practical usefulness of the methods proposed in this paper by investigating the important policy-related question of whether membership to the World Trade Organization (WTO) can have negative effects on the environment.

The remaining part of the paper is structured as follows: next section introduces the model and the estimator. Sect. 3 introduces the Cramer von Mises statistic to test the correct specification of the model and shows the consistency of both bootstrap procedures. Section 4 contains the results on the estimation of the mean functional and the novel semiparametric estimator of the average treatment parameter. Section 5 presents and discusses the results of the Monte Carlo simulations. Section 6 contains the empirical application. Section 7 contains some concluding remarks. An Appendix contains all the proofs.

The following notation is used throughout the paper: "$a.a.$", "$a.s.$" stand for "almost all" and "almost surely" $\Rightarrow$, $\xrightarrow{d}$ denote weak convergence in $l^{\infty}(\cdot)$—the space of all real-valued functions that are uniformly bounded in $\cdot$ (see Pollard 1984 for a definition), and convergence in distribution, respectively. Finally $\|\cdot\|$ is the Euclidean norm, "$'$" denotes transpose and for any vector v $v^{\otimes 2} = vv'$.

2 The model and the estimator

The model we consider is

$$Y = X'\alpha_0(U) + W'\beta_0 + \varepsilon, \tag{1}$$

where $\alpha_0(\cdot)$ is a p-dimensional vector of unknown functions, U is an observable random variable, β_0 is a k-dimensional vector of unknown parameters and the unobservable error ε is such that $E(\varepsilon|U, X, W) = 0$ a.s. and $E\left(\varepsilon^2|U, X, W\right) = \sigma^2(U, X, W)$ a.s.

We assume that some of the Y values in a sample of size n may be MAR whereas the X, U and W values are completely observed, that is the probability of missing, also called propensity score in the causal inference literature, see e.g. Rosenbaum and Rubin (1983), is given by

$$\Pr(\delta|Y, U, X, W) = \Pr(\delta|U, X, W) := \pi(U, X, W) > 0 \quad \text{a.s.}, \tag{2}$$

where $\delta = \{0, 1\}$ is a binary indicator of missingness. Let $(Y_i, U_i, X_i, W_i)_{i=1}^n$ denote an i.i.d. incomplete sample from (Y, U, X, W), and $\delta_i = 0$ indicates that Y_i is missing. To deal with the MAR responses we follow the same approach as that suggested by Wang et al. (2004) and Sun et al. (2009) and propose the following complete case estimators for $\alpha_0(\cdot)$ and β_0

$$\widehat{\beta} = \left[\sum_{i=1}^{n} \delta_i \left(W_i - \widehat{\alpha}_{XW}(U_i)' X_i\right)^{\otimes 2}\right]^{-1} \sum_{i=1}^{n} \delta_i \left(W_i - \widehat{\alpha}_{XW}(U_i)' X_i\right)\left(Y_i - X_i'\widehat{\alpha}_{XY}(U_i)\right) \tag{3}$$

and

$$\widehat{\alpha}(u) = \widehat{\alpha}_{XY}(u) - \widehat{\alpha}_{XW}(u)\widehat{\beta}, \tag{4}$$

where

$$\widehat{\alpha}_{XW}(u) = \left[\sum_{i=1}^{n} \delta_i X_i^{\otimes 2} K_h(U_i - u)\right]^{-1} \sum_{i=1}^{n} \delta_i X_i W_i' K_h(U_i - u),$$
$$\widehat{\alpha}_{XY}(u) = \left[\sum_{i=1}^{n} \delta_i X_i^{\otimes 2} K_h(U_i - u)\right]^{-1} \sum_{i=1}^{n} \delta_i X_i Y_i' K_h(U_i - u)$$

and $K_h(\cdot) := K(\cdot/h)/h$ is a kernel function and $h =: h(n)$ is the bandwidth.

Note that (3) and (4) are like the profile kernel estimators proposed by Fan and Huang (2005) except that they are based on the complete case and conditional heteroskedasticity is allowed. Alternatively we can use nonparametric series estimation and obtain the complete case series based estimator $\widehat{\beta}_S$

$$\widehat{\beta}_S = \left[\sum_{i=1}^{n} \delta_i \left(W_i - W_i P \left(X_i, U_i \right) \right)^{\otimes 2} \right]^{-} \sum_{i=1}^{n} \delta_i \left(W_i - W_i P \left(X_i, U_i \right) \right) \left(Y_i - P \left(X_i, U_i \right) Y_i \right)$$

and

$$\widehat{\alpha} \left(u \right)_S = P \left(X_i, u \right) \left(Y_i - W_i \widehat{\beta}_S \right),$$

where

$$P \left(X_i, u \right) = p \left(X_i, u \right)' \left(\sum_{i=1}^{n} p \left(X_i, u \right) p \left(X_i, u \right)' \right)^{-} p \left(X_i, u \right),$$

$$p \left(X_i, u \right) = \left[X_{1i} p_1^{q_1} \left(u \right)', ..., X_{pi} p_p^{q_p} \left(u \right)' \right]',$$

"−" denotes generalized inverse and $p_j^{q_j} \left(u \right) = \left[p_{j1} \left(x \right), ..., p_{jq_j} \left(u \right) \right]'$ is a $q_j \times 1$ vector of base functions. Under the same regularity conditions of Ahmad et al. (2005) it is possible to show that $\widehat{\beta}_S$ has the same distribution as that of the profile kernel estimator $\widehat{\beta}$ given in the following theorem:

Theorem 1 *Under A1(i), A2–A6(i), A7(i) and A8 listed in the Appendix*

$$n^{1/2} \left(\widehat{\beta} - \beta_0 \right) \xrightarrow{d} N \left(0, \Gamma_0^{-1} \Omega_0 \Gamma_0^{-1} \right)$$

where

$$\Gamma_0 = E \left[\pi \left(U, X, W \right) \left(W - \alpha_{0XW} \left(U \right)' X \right)^{\otimes 2} \right],$$

$$\Omega_0 = E \left[\pi \left(U, X, W \right) \sigma^2 \left(U, X, W \right) \left(W - \alpha_{0XW} \left(U \right)' X \right)^{\otimes 2} \right].$$

Furthermore, under the additional assumption that $\varepsilon \sim N \left(0, \sigma^2 \right)$,

$$n^{1/2} \left(\widehat{\beta} - \beta_0 \right) \xrightarrow{d} N \left(0, \sigma^2 \Gamma_0^{-1} \right), \tag{5}$$

and $\sigma^2 \Gamma_0^{-1}$ *is the semiparametric efficiency bound of Bickel et al. (1993).*

The efficiency result (5) of Theorem 1 is consistent with that of Koul et al. (2012), who derive the efficiency bound for a partially linear model with MAR responses assuming a location density for the error with finite Fisher information. Under the assumption of normality Koul et al. (2012)'s efficiency bound coincides with the one of this paper, and vice-versa under the assumption of a location density (5) can be modified to encompass theirs. It is also interesting to note that (5) is consistent with the results of Kim and Ma (2012) on the semiparametric efficiency of the extended nonlinear least squares estimator of Wang and LeBlanc (2008), under the assumption of homoskedasticity and that the errors have a symmetric density.

Let

$$\widehat{Y}_{1i} = \delta_i Y_i + (1-\delta_i)\left(X_i'\widehat{\alpha}(U_i) + W_i\widehat{\beta}\right),$$
$$\widehat{Y}_{2i} = \frac{\delta_i}{\widehat{\pi}(U_i, X_i, W_i)} Y_i + \left(1 - \frac{\delta_i}{\widehat{\pi}\left(U_i, X_{i,} W_i\right)}\right)\left(X_i'\widehat{\alpha}(U_i) + W_i\widehat{\beta}\right) \tag{6}$$

denote, respectively, the regression and inverse probability weighted imputed values, where $\widehat{\pi}(\cdot)$ is a nonparametric estimator of the propensity score (2) given by

$$\widehat{\pi}(u, x, w) = \frac{\sum_{i=1}^{n} \delta_i L_b(U_i - u, X_i - x, W_i - w)}{\sum_{i=1}^{n} L_b(U_i - u, X_i - x, W_i - w)}, \tag{7}$$

where $L_b(\cdot) := L(\cdot/b)/b$ is a kernel function and $b =: b(n)$ is another bandwidth. Note that $\widehat{Y}_{2i}$ is motivated by the so-called double robustness property, that in the context of estimation with missing data means consistency of the estimator when either the model for the missingness mechanism or the model for the distribution of the complete data is correctly specified (see e.g. Robins et al. 1994; Scharfstein et al. 1999 and Bang and Robins 2005). As noted by Sun et al. (2009), this property is desirable for estimation but not for testing (especially for the correct specification) because it reduces the sensitivity of the test to departures from the null hypothesis. A second potential problem with $\widehat{Y}_{2i}$ is that it might suffer from the curse of dimensionality associated with a possibly high dimensional set of covariates X and W. Of course a natural remedy for this problem would be to consider a fully parametric specification of $\pi(U_i, X_i, W_i)$, such as that of a probit or logit. However the resulting estimator would be characterized by a different asymptotic behaviour and could still be doubly-robust. For these reasons we consider an alternative inverse probability weighting based on $\Pr(\delta = 1|U)$, which we call the partial propensity score. This specification avoids the curse of dimensionality because the covariate U is assumed to be a random variable. Let

$$\widehat{\pi}(u) = \frac{\sum_{i=1}^{n} \delta_i L_b(U_i - u)}{\sum_{i=1}^{n} L_b(U_i - u)} \tag{8}$$

denote the nonparametric estimator, and

$$\widehat{Y}_{3i} = \frac{\delta_i}{\widehat{\pi}(U_i)} Y_i + \left(1 - \frac{\delta_i}{\widehat{\pi}(U_i)}\right)\left(X_i'\widehat{\alpha}(U_i) + W_i\widehat{\beta}\right) \tag{9}$$

denote the resulting inverse probability weighting imputed values. Note that the same type of inverse probability weighting (9) has been considered by Wang et al. (2004) and Sun et al. (2009).

3 Specification analysis

The null hypothesis that (1) is correctly specified is

$$H_0 : E(\varepsilon|U, X, W) = E\left(Y - X'\alpha_0(U) - W'\beta_0|U, X, W\right) = 0 \quad \text{a.s.,} \tag{10}$$

or equivalently

$$H_0 : E[\varepsilon\Psi(U, X, W; u, x, w)] = 0 \quad \text{a.a.}\, u, x, w \in [-\infty, \infty]^{1+k+p} \tag{11}$$

provided the linear span of the function $\Psi(\cdot)$ is dense in the space of bounded measurable functions on the support of U, X, W (Bierens 1982). In this paper we specify $\Psi(\cdot)$ to be the indicator function

$$\begin{aligned} &I\left(\left[U, X', W'\right] \le \left[u, x', w'\right]\right) \\ &\quad = I(U \le u) \prod_{j=1}^{p} I\left(X^{(j)} \le x^{(j)}\right) \prod_{j=1}^{k} I\left(W^{(j)} \le w^{(j)}\right), \end{aligned} \tag{12}$$

but we follow the same projection approach proposed by Escanciano (2006), which avoids the potential practical problem of having a high dimensional set of covariates. To be specific we consider

$$E\left[\varepsilon I\left(U \le u, \theta'\left[X', W'\right]' \le s\right)\right] = 0 \quad \text{a.a.}\, u, s, \theta \in \Pi, \tag{13}$$

where $\Pi = [-\infty, \infty]^2 \times \mathbb{S}^{k+p}$ and by Lemma 1 in Escanciano (2006) $\mathbb{S}^{k+p}$ is the unit sphere in $\mathbb{R}^{k+p}$.

The advantage of specifying $\Psi(U, X, W; u, x, w)$ as in (13) over the standard indicator function (12) is apparent in its dimension reduction character, which implies that potentially high dimensional covariates X, W are not a problem as they would be with (12), i.e. too many zeroes negatively affecting both size and power properties of any test statistic based on it. A more detailed discussion of the merits of (13), including also a comparison with the related approach of Stute and Zhu (2002) in the context of generalized linear models, can be found in Escanciano (2006).

Given (13) a test statistic for the null hypothesis (10) can be constructed by considering a functional of the so-called projected marked empirical process

$$n^{1/2}\widehat{v}_j(u, s, \theta) = \frac{1}{n^{1/2}} \sum_{i=1}^{n} \widehat{\varepsilon}_{ji} I\left(U_i \le u, \theta'\left[X_i', W_i'\right]' \le s\right) \quad j = 1 \text{ or } 3$$

where

$$\widehat{\varepsilon}_{ji} = \widehat{Y}_{ji} - X_i'\widehat{\alpha}(U_i) + W_i'\widehat{\beta} \quad j = 1 \text{ or } 3$$

and $\widehat{\alpha}(\cdot)$ and $\widehat{\beta}$ are the complete case estimators as defined in 4 and 3, respectively.

3.1 Asymptotic null distributions

Theorem 2 *Under A1(i), A2–A6(i), A7(i) and A8 listed in the Appendix*

$$n^{1/2}\widehat{v}_j(u,s,\theta) \Longrightarrow v_j(u,s,\theta) \quad in\ l^{\infty}(\Pi), \quad j=1\ or\ 3,$$

where the $v_j(u,s,\theta)$*s are two centred Gaussian processes with covariance function*

$$E\left[\sigma_j(u_1,s_1,\theta_1)\,\sigma_j(u_2,s_2,\theta_2)\right], \tag{14}$$

$$\begin{aligned}
&\sigma_1(u,s,\theta)\\
&\quad=\delta\varepsilon I\left(U\le u,\theta'\left[X',W'\right]'\le s\right)-D(u,s,\theta)\,\Gamma_0^{-1}\delta\left(W-\alpha_{0WX}(U)'X\right)\varepsilon\\
&\qquad-G(s,\theta,u)\left[E\left(\delta X^{\otimes 2}|U\right)\right]^{-1}\delta X\varepsilon I(U\le u),\\
&\sigma_3(u,s,\theta)\\
&\quad=\frac{\delta\varepsilon}{\pi(U)}I\left(U\le u,\theta'\left[X',W'\right]'\le s\right)-D(u,s,\theta,\pi)\,\Gamma_0^{-1}\delta\left(W-\alpha_{WX}(U)'X\right)\varepsilon\\
&\qquad-G(s,\theta,u)\left[E\left(\delta X^{\otimes 2}|U\right)\right]^{-1}\frac{\delta}{\pi(U)}X\varepsilon I(U\le u),
\end{aligned}$$

and

$$\begin{aligned}
D(u,s,\theta)&=E\left[\delta\left(W'-X'\alpha_{0XW}\right)I\left(U\le u,\theta'\left[X',W'\right]'\le s\right)\right],\\
D(u,s,\theta,\pi)&=E\left[\frac{\delta}{\pi(U)}\left(W'-X'\alpha_{0XW}\right)I\left(U\le u,\theta'\left[X',W'\right]'\le s\right)\right],\\
G(s,\theta,u)&=E\left[\delta X'I\left(\theta'\left[X',W'\right]'\le s\right)|U=u\right].
\end{aligned}$$

Let $F_\theta(u,s)$ and $\widehat{F}_\theta(u,s)$ denote, respectively, the distribution and empirical distribution of U and $\theta'\left[X',W'\right]'$, and let $d\theta$ denote the uniform distribution on the sphere $\mathbb{S}^{k+p}$. Note that the uniform distribution is chosen for computational convenience since the resulting integral admits a simple closed form expression—see (29); other distributions could be used. Given the result of Theorem 2 we can use a Cramer von Mises type of functional to construct a test statistic for the null hypothesis H_0 that is

$$CM_j=n\int_{\Pi}\widehat{v}_j(u,s,\theta)^2\,d\widehat{F}_\theta(u,s)\,d\theta \quad j=1\text{ or }3. \tag{15}$$

A straightforward application of the continuous mapping theorem gives the following:

Corollary 3 *Under A1(i), A2–A6(i), A7(i) and A8-A9 in the Appendix and under the null hypothesis (10)*

$$CM_j \xrightarrow{d} \int_{\Pi} v_j (u, s, \theta)^2 \, dF_\theta (u, s) \, d\theta \quad j = 1 \text{ or } 3. \tag{16}$$

As it is typically the case with this type of approach to specification testing the asymptotic null distributions of the proposed test statistics are not asymptotically distribution free (ADF) since they depend in a complicated way on the influence functions σ_j $(j = 1, 3)$ given in (14). It is also important to note that the CM_j $(j = 1, 3)$ statistics need not to be scaled by a normalizing constant—typically given by a consistent estimator of (14). This is important for two reasons: first because the presence of such normalizing constant can have a negative effect on the power properties of the statistics themselves depending on the type of estimator used. Secondly because the estimation of the normalizing constant is further complicated by the presence of the unknown conditional variance $\sigma^2 (U, X, W)$. In this case one possible estimator that can be used is the same k-nearest neighbourhood-based estimator, say $\widehat{\sigma}^2_{\text{nn}} (U_i, X_i, W_i)$, as that used by Robinson (1987). Note also that we could take the conditional heteroskedasticy directly into account in the estimation of $\widehat{\alpha} (U_i)$ and $\widehat{\beta}$ by scaling the unobservable error ε_i by $\widehat{\sigma}_{\text{nn}} (U_i, X_i, W_i)$, and/or consider a weighted version of the projected marked empirical processes $n^{1/2}\widehat{v}_j (u, s, \theta)$. However in both cases the resulting test would still not be ADF, giving therefore no practical advantage to the scaled test statistic over the original one. In the next subsection we present two resampling techniques that can be used to consistently estimate the null distributions of the CM_j statistics.

3.2 Bootstrap approximation

As mentioned in the Introduction we consider two bootstrap approaches to approximate the distributions of the CM_j statistics. The first one is based on the wild bootstrap (WB henceforth) and the second one is based on the multiplier bootstrap (MB henceforth). Each methods has its own advantage compared to the other: WB is easier to compute as the required multidimensional integration over $\mathbb{S}^{k+p}$ admits a simple closed form expression-see (29) below; at the same time it is rather computationally intensive because of the actual re-estimation involved, especially for the nonparametric component. MB is more complicated to compute because it is directly based on the empirical analog of (14)-see (17) below- but it is less computationally intensive because the model is not actually resampled.

In the context of specification testing with marked empirical processes WB is considered by Stute (1997) and by Escanciano (2006) for linear regressions and parametric regression models, respectively. When the responses are MAR Gonzalez-Manteiga and Perez-Gonzalez (2006) show how to modify the WB for testing the correct specification of regression models using kernel smoothing. In this paper we follow their approach to obtain appropriate bootstrap samples that preserve the MAR assumption. To be specific let $\widehat{\varepsilon}_i$ denote the ith residual when $\delta_i = 1$ and let $\{\eta_i\}_{i=1}^n$ denote a random sample from the distribution of the bounded random variable η with zero mean and unit variance that is independent from U, X and W. Let $\varepsilon_i^* = \widehat{\varepsilon}_i \eta_i$ and note that $E^*\left(\varepsilon_i^*\right) = 0$ and $V^*\left(\varepsilon_i^*\right) = \widehat{\varepsilon}_i^2$, where $E^*(\cdot)$ and $V^*(\cdot)$ denote the expectation and variance operator with respect to the original sample. Let

$$\begin{aligned} Y_i^* &= X_i'\widehat{\alpha}(U_i) + W_i'\widehat{\beta} + \varepsilon_i^* \quad \text{if } \delta_i = 1, \\ Y_i^* &= 0 \quad \text{if} \delta_i = 0 \end{aligned}$$

denote the bootstrap response so that $\left(Y_i^*, U_i, X_i, W_i\right)_{i=1}^n$ represent a bootstrap sample. Let

$$\widehat{\beta}^* = \left[\sum_{i=1}^n \delta_i \left(W_i - \widehat{\alpha}_{WX}(U_i)' X_i\right)^{\otimes 2}\right]^{-1} \sum_{i=1}^n \delta_i \left(W_i - \widehat{\alpha}_{XW}(U_i)' X_i\right)\left(Y_i^* - X_i'\widehat{\alpha}_{XY^*}(U_i)\right),$$

and

$$\widehat{\alpha}(u) = \widehat{\alpha}_{XY^*}(u) - \widehat{\alpha}_{XW}(u)\widehat{\beta}^*$$

denote the bootstrap estimators, and let

$$n^{1/2}\widehat{\upsilon}_j^*(u, s, \theta) = \frac{1}{n^{1/2}} \sum_{i=1}^n \widehat{\varepsilon}_{ji}^* I\left(U_i \le u, \theta'\left[X_i', W_i'\right]' \le s\right) \qquad j = 1 \text{ or } 3$$

denote the bootstrap marked empirical process, where

$$\widehat{\varepsilon}_{ji}^* = \widehat{Y}_{ji}^* - X_i'\widehat{\alpha}^*(U_i) + W_i'\widehat{\beta}^*,$$

and

$$\begin{aligned} \widehat{Y}_{1i}^* &= \delta_i Y_i^* + (1 - \delta_i)\left(X_i'\widehat{\alpha}^*(U_i) + W_i\widehat{\beta}^*\right) \\ \widehat{Y}_{3i}^* &= \frac{\delta_i}{\widehat{\pi}(U_i)} Y_i^* + \left(1 - \frac{\delta_i}{\widehat{\pi}(U_i)}\right)\left(X_i'\widehat{\alpha}^*(U_i) + W_i\widehat{\beta}^*\right). \end{aligned}$$

Let $CM_j^{*(WB)}$ denote the WB version of (15) and let $\Pr^*$ denote the bootstrap probability; the following theorem shows that WB consistently estimates the distribution of (15).

Theorem 4 *Under A1(i), A2–A6(i), A7(i) and A8–A9 listed in the Appendix*

$$\sup_{c \in \mathbb{R}_+} \left|\Pr{}^*\left(CM_j^{*(WB)} \ge c\right) - \Pr\left(CM_j \ge c\right)\right| \xrightarrow{p} 0 \quad j = 1 \textit{ or } 3,$$

where CM_j has the asymptotic distribution as given in (16).

Multiplier bootstrap is similar to WB in that an auxiliary sequence of zero mean and unit variance random variables independent from the original sample is used in the simulation.However as opposed to WB the original model is not resampled, nor re-estimated. MB is considered for example by Su and Wei (1991) in the context of specification tests for generalized linear models and by Zhu and Ng (2003) for partial linear models.

Let

$$n^{1/2}\widehat{\sigma}_j^*(u,s,\theta) = \frac{1}{n^{1/2}}\sum_{i=1}^{n}\widehat{\sigma}_{ji}(u,s,\theta)\,\eta_i \quad j=1 \text{ or } 3 \tag{17}$$

denote the empirical randomized influence function, where

$$\begin{aligned}
&\widehat{\sigma}_{1i}(u,s,\theta)\\
&\quad= \delta_i\widehat{\varepsilon}_i\widehat{\omega}_i I\left(U_i \le u, \theta'\left[X_i', W_i'\right]' \le s\right) - \widehat{D}(u,s,\theta)\,\widehat{\Gamma}^{-1}\delta_i\\
&\quad\left(W_i' - \widehat{\alpha}_{WX}(U_i)' X_i\right)\widehat{\varepsilon}_i\\
&\qquad -\widehat{G}(s,\theta,u)\left[\frac{1}{n}\sum_{j=1}^{n}\delta_j X_j^{\otimes 2} K_h\left(U_j - U_i\right)\right]^{-1}\delta_i X_i\widehat{\varepsilon}_i I(U_i \le u),\\
&\widehat{D}(u,s,\theta) = \frac{1}{n}\sum_{i=1}^{n}\left[\delta_i\left(W_i' - X_i'\widehat{\alpha}_{XW}(U_i)\right) I\left(U_i \le u, \theta'\left[X', W'\right]' \le s\right)\right],\\
&\widehat{G}(s,\theta,u) = \frac{1}{n}\sum_{i=1}^{n}\left[\delta_i X_i' I\left(\theta'\left[X_i', W_i'\right]' \le s\right) K_h(U_i - u)\right],
\end{aligned}$$

where $\widehat{\alpha}_{XW}(U_i)$ and $\widehat{\pi}(U_i)$ are defined in Sect. 2, $\widehat{\sigma}_{3i}(u,s,\theta)$ is defined similarly to $\widehat{\sigma}_{1i}(u,s,\theta)$ and $\{\eta_i\}_{i=1}^n$ is a random sample from the random variable η with the same characteristics as those used in WB. Let

$$CM_j^{*(MB)} = \int_{\Pi}\widehat{\sigma}_j^*(u,s,\theta)^2\, d\widehat{F}_\theta(u,s)\, d\theta$$

denote the MB version of (15).

The following theorem shows that MB consistently estimates the distribution of (15).

Theorem 5 *Under A1(i), A2–A6(i), A7(i) and A8–A9 listed in the Appendix*

$$\sup_{c\in\mathbb{R}_+}\left|\Pr{}^*\left(CM_j^{*(MB)} \ge c\right) - \Pr\left(CM_j \ge c\right)\right| \xrightarrow{p} 0 \quad j=1 \; or \; 3,$$

where CM_j has the asymptotic distribution as given in (16).

3.3 Power properties

We now investigate the power properties of the test statistics CM_j. We first consider global alternatives of the form

$$H_{1g} : \Pr\left(E\left(Y - X'\alpha\left(U\right) - W'\beta|U, X, W\right) \neq 0\right) > 0 \quad \forall \beta, \alpha\left(U\right). \tag{18}$$

Theorem 6 *Under A1(i), A2–A6(i), A7(i), A8–A9 listed in the Appendix and under the alternative global hypothesis (18)*

$$\frac{CM_1}{n} \xrightarrow{d} \int_\Pi E\left\{\delta\left[\left(E\left(Y|U, X, W\right) - X'\alpha^\dagger\left(U\right) - W'\beta^\dagger\right)\right]\right.$$
$$\left.\times I\left(U \le u, \theta'\left[X', W'\right]' \le s\right)\right\}^2 dF_\theta\left(u, s\right) d\theta > 0,$$
$$\frac{CM_3}{n} \xrightarrow{d} \int_\Pi E\left\{\frac{\delta}{\pi\left(U\right)}\left[\left(E\left(Y|U, X, W\right) - X'\alpha^\dagger\left(U\right) - W'\beta^\dagger\right)\right]\right.$$
$$\left.\times I\left(U \le u, \theta'\left[X', W'\right]' \le s\right)\right\}^2 dF_\theta\left(u, s\right) d\theta > 0,$$

where $\left\|\widehat{\beta} - \beta_0\right\| = \beta^\dagger + o_p\left(1\right)$, $\left\|\widehat{\alpha}\left(U\right) - \alpha_0\left(U\right)\right\| = \alpha^\dagger\left(U\right) + o_p\left(1\right)$, *and* $\beta^\dagger \neq 0$, $\alpha^\dagger\left(U\right) \neq 0\ a.s.$

Next we consider local alternatives of the form

$$H_{1l} : E\left(\varepsilon|U, X, W\right) = \frac{\gamma\left(U, X, W\right)}{n^{1/2}} \text{ a.s.} \tag{19}$$

for some known bounded real-valued function $\gamma\left(\cdot\right)$.

Theorem 7 *Under A1(i), A2–A6(i), A7(i), A8–A10 listed in the Appendix and the alternative local hypothesis (19)*

$$CM_j \xrightarrow{d} \int_\Pi \left[v_j\left(u, s, \theta\right) + s_j\left(u, v, \theta\right)\right]^2 dF_\theta\left(u, s\right) d\theta \quad j = 1 \text{ or } 3,$$

where

$$s_1\left(u, v, \theta\right) = E\left[\delta\gamma\left(U, X, W\right) I\left(U \le u, \theta'\left[X', W'\right]' \le s\right)\right]$$
$$-E\left[\delta\left(W' - X'\alpha_{XW}\left(U\right)\right) I\left(U \le u, \theta'\left[X', W'\right]' \le s\right)\right]$$
$$\times\Gamma_0^{-1} E\left\{\pi\left(U, X, W\right)\left(W - X\alpha_{XW}\left(U\right)\right)\right.$$
$$\left.\times\left[\gamma\left(U, X, W\right) - X'\left[E\left(X^{\otimes 2}|U\right)\right]^{-1} E\left(\delta X\gamma\left(U, X, W\right)|U\right)\right]\right\},$$

and $s_3\left(u, v, \theta\right)$ *is as* $s_1\left(u, v, \theta\right)$ *with* δ *replaced by* $\delta/\pi\left(U\right)$.

Theorems 6 and 7 show that the tests (15) are consistent and can detect local alternatives converging to the null hypothesis at the fastest possible rate.

4 Estimation of the mean functional

We now consider the problem of estimating the population mean μ_0 (or any other linear square integrable functional) of the response variable Y using the same i.i.d. incomplete sample $(Y_i, U_i, X_i, W_i)_{i=1}^n$ as that defined in Sect. 2 and the same three different imputed responses $\widehat{Y}_{ji}$ $(j = 1, 2, 3)$ defined in (6) and (9). Let

$$\widehat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n}\widehat{Y}_{ji} \quad j = 1, 2, 3$$

denote the resulting imputation estimators. Recall that $\widehat{\mu}_2$-the inverse probability weighting estimator with estimated full propensity score (7)-enjoys the double robustness property, whereas $\widehat{\mu}_3$ - the inverse probability weighting estimator with estimated partial propensity score (8)- does not (unless $\pi(U, X, W) = \pi(U)$ $a.s.$). It is also important to note that an estimator similar to $\widehat{\mu}_3$ is also considered by Wang et al. (2004), who showed that under the assumption of a partially linear specification, $\widehat{\mu}_3$ is in general more efficient than the corresponding fully nonparametric one of Cheng (1994).

Theorem 8 *Under A1(i), A2–A6(i), A7(i) and A8 listed in the Appendix*

$$n^{1/2}\left(\widehat{\mu}_j - \mu_0\right) \xrightarrow{d} N\left(0, \sigma_j^2 + \operatorname{var}\left(X'\alpha_0(U) + W'\beta_0\right)\right) \quad j = 1 \text{ or } 3,$$

where

$$\sigma_j^2 = E\left[\sigma^2(U, X, W)\,\pi(U, X, W)\left(\Omega_{0j}(U) + \Delta_{0j}\Gamma_0^{-1}\left(W - \alpha_{0XW}(U)'X\right)\right)^2\right],$$

and

$$\begin{aligned}
\Omega_{01}(U) &= [E(X|U) - E(\delta X|U)]'\left[E\left(\delta X^{\otimes 2}|U\right)\right]^{-1}X + 1,\\
\Omega_{03}(U) &= \left[E(X|U) - \frac{E(\delta X|U)}{\pi(U)}\right]'\left[E\left(\delta X^{\otimes 2}|U\right)\right]^{-1}X + \frac{1}{\pi(U)},\\
\Delta_{01} &= E\left[(1-\delta)\left(W' - X'\alpha_{0XW}(U)\right)\right],\\
\Delta_{03} &= E\left[\left(1 - \frac{\delta}{\pi(U)}\right)\left(W' - X'\alpha_{0XW}(U)\right)\right].
\end{aligned} \tag{20}$$

Under A1(ii), A2–A3, A4(ii), A5, A6(ii), A7(ii) and A8 listed in the Appendix

$$n^{1/2}\left(\widehat{\mu}_2 - \mu_0\right) \xrightarrow{d} N\left(0, \frac{\sigma^2(U, X, W)}{\pi(U, X, W)} + \operatorname{var}\left(X'\alpha_0(U) + W'\beta_0\right)\right). \tag{21}$$

The results of Theorem 8 complement those obtained by Cheng (1994) and Wang et al. (2004) and show that the inverse probability weighting estimator with full nonparametric propensity score $\widehat{\mu}_2$ (21) is asymptotically equivalent to the estimator proposed by Cheng (1994), which is known to be semiparametric efficient Hahn (1998), whereas neither of the two other imputation estimators achieve this bound. It is also interesting to investigate the efficiency of the proposed estimators under the assumption that a partially linear varying coefficient specification holds, that is

$$E\left(Y|U,X,W\right)=X'\alpha_0\left(U\right)+W'\beta_0 \quad \text{a.s.} \tag{22}$$

The following theorem derives the semiparametric efficiency bound under (22) and the assumption of normality of the errors.

Theorem 9 *Under A1(i), A2–A6(i), A7(i) and A8 listed in the Appendix, (22) and* $\varepsilon\sim N\left(0,\sigma^2\right)$*, the semiparametric efficiency bound of Bickel et al. (1993) for the mean functional is*

$$\begin{aligned}\sigma_{eff}^2 &=\sigma^2 E\left[E\left(X|U\right)'E\left(\delta X^{\otimes 2}|U\right)^{-1}E\left(X|U\right)\right]\\ &\quad+E\left[\pi\left(U,X,W\right)\left(W-\alpha_{0XW}\left(U\right)'X\right)'\Gamma_0^{-1}\right.\\ &\quad\left.\times E\left(W-\alpha_{0XW}\left(U\right)'X\right)\right]^2+\operatorname{var}\left(X'\alpha_0\left(U\right)+W'\beta_0\right).\end{aligned} \tag{23}$$

The efficiency bound (23) is an important generalization of that derived by Wang et al. (2004). This result complements that of Muller (2009), who considered efficient estimation of the mean functional under MAR assuming a nonlinear regression structure. The following proposition shows that none of the proposed estimators $\widehat{\mu}_j$ ($j=1,2,3$) achieve this bound, and thus they are not semiparametric efficient under (22) and normality. Let $\sigma^2_{j(ho)}$ denote the variances σ_j^2 of Theorem 8 under conditional homoskedasticity.

Proposition 10 *Under the assumptions of Theorem 9*

$$\sigma^2_{j(ho)}-\sigma^2_{eff}=V_{1j}-V_{2j}\geq 0 \quad j=1,2,3,$$

where the $V_{kj}s$ *(*$k=1,2$*) are given in the Appendix.*

Proposition 10 shows that the difference between the variances of the proposed estimators with that of the efficient one consists of two terms: V_{1j} and V_{2j}. The former represents the variance reduction that results from imposing the additional information (22), whereas the latter represents the effect of estimating the additional information. In practice the more precise the estimation is (i.e. the smaller V_{2j}), the larger the efficiency loss of using any of the imputation estimators is. An immediate consequence of Proposition 10 is that the efficiency result of Wang et al. (2004, Theorem 3.4) does not generalize to more complex semiparametric models with a partial linear structure. It is also important to note that Proposition 10 is consistent with the result of Muller (2009), who showed that imputation estimators do not achieve the semiparametric efficiency bound under the additional assumption of a nonlinear regression structure for a MAR mean functional. Indeed to achieve the bound Muller (2009) suggested an alternative weighted estimator with weights based on (profile) empirical likelihood estimation of the additional information. A similar approach could be used for the (22) specification, but its full asymptotic analysis is beyond the scope of this paper.

4.1 Average treatment effect estimation

The results of Theorem 8 can be used in the context of causal inference models to obtain an estimator of the average treatment effect parameter under the so-called strong ignorability condition Rosenbaum and Rubin (1983). To be specific let δ denote a binary indicator of treatment and let $Y^{(\delta)}$ denote the potential outcome. Given a sample $(Y_i, U_i, X_i, W_i)_{i=1}^n$ where

$$Y_i = \delta_i Y_i^{(1)} + (1-\delta_i)\, Y_i^{(0)},$$

is the realised outcome, the average treatment effect parameter is

$$\tau_0 = E\left[Y^{(1)} - Y^{(0)}\right]. \tag{24}$$

The central problem of the treatment literature is that either $Y_i^{(1)}$ or $Y_i^{(0)}$ is observed but never both. Thus without further restrictions, the average treatment effect τ_0 is not identified and hence cannot be consistently estimated. To solve the identification problem, we assume the following strong ignorability condition

$$\Pr\left(\delta | Y^{(\delta)}, U, X, W\right) = \Pr\left(\delta | U, X, W\right) := \pi\left(U, X, W\right) > 0 \text{ a.s.}, \tag{25}$$

which is effectively the MAR assumption in (7). Hahn (1998) and Hirano et al. (2003) consider nonparametric estimation of (24); fully parametric estimation is considered for example by Little and Rubin (2002).

We consider two estimators, one based on imputation and the other based on inverse probability weighting, and assume a varying coefficient partially linear specification, that is

$$\begin{aligned}
\widehat{\tau}_1 &= \frac{1}{n}\sum_{i=1}^{n}\Big[\delta_i Y_i^{(1)} + (1-\delta_i)\left(X_i'\widehat{\alpha}^{(1)}(U_i) + W_i\widehat{\beta}^{(1)}\right) \\
&\qquad - (1-\delta_i)\, Y_i^{(0)} - \delta_i\left(X_i'\widehat{\alpha}^{(0)}(U_i) + W_i\widehat{\beta}^{(0)}\right)\Big], \\
\widehat{\tau}_2 &= \frac{\delta_i}{\widehat{\pi}\left(U_i, X_i, W_i\right)}\left(X_i'\widehat{\alpha}^{(1)}(U_i) + W_i\widehat{\beta}^{(1)}\right) \\
&\quad + \left(1 - \frac{\delta_i}{\widehat{\pi}\left(U_i, X_i, W_i\right)}\right)\left(X_i'\widehat{\alpha}^{(0)}(U_i) + W_i\widehat{\beta}^{(0)}\right).
\end{aligned} \tag{26}$$

The following theorem establishes the asymptotic distribution of the two average treatment parameter estimators $\widehat{\tau}_j$ $j = 1, 2$. Let

$$\sigma_1^{(1)2} = E\left[\sigma^{(1)2}(U, X, W)\pi(U, X, W)\left(\Omega_{01}^{(1)}(U) + \Delta_0^{(1)}\Gamma_0^{(1)-1}\left(W - \alpha_{0XW}^{(1)}(U)' X\right)\right)^2\right],$$
$$\sigma_1^{(0)2} = E\left[\sigma^{(0)2}(U, X, W)(1 - \pi(U, X, W))\left(\Omega_{01}^{(0)}(U) + \Delta_0^{(0)}\Gamma_0^{(1)-1}\left(W - \alpha_{0XW}^{(0)}(U)' X\right)\right)^2\right],$$

where $\Omega_{01}^{(1)}(U)$, $\Delta_0^{(1)}$ are as in (20) and $\Omega_{01}^{(0)}(U)$, $\Delta_0^{(0)}$ are

$$\begin{aligned}\Omega_{01}^{(0)}(u) &= E(\delta X|U)'\left[E\left((1-\delta)X^{\otimes 2}|U\right)\right]^{-1} X + 1,\ \Delta_0^{(0)} \\ &= E\left[\delta\left(W' - X'\alpha_{0XW}(U)\right)\right], \\ \Gamma_0^{(0)} &= E\left[(1 - \pi(U, X, W))\left(W - \alpha_{0XW}^{(0)}(U)' X\right)^{\otimes 2}\right].\end{aligned}$$

Also let

$$\sigma_2^{(1)2} = E\left[\sigma^{(1)2}(U, X, W)\pi(U, X, W)\left(\Omega_{02}^{(1)}(U) + \Delta_{02}^{(1)}\Gamma_0^{(1)-1}\left(W - \alpha_{0XW}^{(1)}(U)' X\right)\right)^2\right],$$
$$\sigma_2^{(0)2} = E\left[\sigma^{(0)2}(U, X, W)(1 - \pi(U, X, W))\left(\Omega_{02}^{(0)}(U) + \Delta_{02}^{(0)}\Gamma_0^{(1)-1}\left(W - \alpha_{0XW}^{(0)}(U)' X\right)\right)^2\right],$$

where

$$\begin{aligned}\Omega_{02}^{(1)}(u) &= \frac{E(X|U)}{\pi(U, X, W)}'\left[E\left(\delta X^{\otimes 2}|U\right)\right]^{-1} X, \\ \Omega_{02}^{(0)}(u) &= \frac{E(X|U)}{\pi(U, X, W)}'\left[E\left((1-\delta)X^{\otimes 2}|U\right)\right]^{-1} X.\end{aligned}$$

Theorem 11 *Under A1(ii), A2, A3, A4(ii), A5, A6(ii), A7(ii) and A8 listed in the Appendix*

$$\begin{aligned}&n^{1/2}(\widehat{\tau}_1 - \tau_0) \\ &\quad\xrightarrow{d} N\left(0, \sigma_1^{(1)2} + \sigma_1^{(0)2} + \operatorname{var}\left(X'\left[\alpha_0^{(1)}(U) - \alpha_0^{(0)}(U)\right] + W'\left[\beta_0^{(1)} - \beta_0^{(0)}\right]\right)\right), \\ &n^{1/2}(\widehat{\tau}_2 - \tau_0) \\ &\quad\xrightarrow{d} N\left(0, \sigma_2^{(1)2} + \sigma_2^{(0)2} + \operatorname{var}\left(X'\left[\alpha_0^{(1)}(U) - \alpha_0^{(0)}(U)\right] + W'\left[\beta_0^{(1)} - \beta_0^{(0)}\right]\right)\right).\end{aligned}$$

5 Monte Carlo evidence

In this section we use simulations to assess the finite sample properties of the proposed estimators and test statistics. We consider two models: a varying coefficient and a partially linear varying coefficient that are given, respectively, by

$$Y_i = U_i + X_{1i} \cos(2\pi U_i) + X_{2i} U_i^2 + \gamma X_{2i}^2 + \varepsilon_i, \tag{27}$$

and

$$Y_i = X_{1i} \sin(2\pi U_i) + X_{2i} \sin(6\pi U_i) + .5 W_{1i} + 2W_{2i} + W_{3i} + \gamma (X_{1i} + W_{2i})^2 + \varepsilon_i, \tag{28}$$

where U is uniformly distributed on [0, 1], X_{ji} are independent standard normal, W_{ji} $(j = 1, 2)$ are jointly normally distributed with mean 0, variance 1 and correlation 2/3, W_{3i} is a Bernoulli random variable taking the value 1 with probability 0.4, ε_i is a normal random variable with mean 0 and variance 0.2 and $\gamma = 0$ under the null hypothesis of correct specification. The two missing probability mechanisms are:

$$\pi_{(1)}(U = u, X = x, W = w) = 0.7 + 0.25(|x_1 - 1| + |w_2 - 1| + |u - 1|)$$
$$\text{if } |x_1 - 1| + |w_2 - 1| + |u - 1| \leq 1/5 \text{ or } 0.9 \text{ otherwise}$$
$$\pi_{(2)}(U = u, X = x, W = w) = 0.6 \text{for all } u, x, w,$$

which imply that the mean probability of missing is approximately 0.1 and 0.4, respectively.

We first compare the finite sample performance of the two test statistics CM_j (15) using bootstrapped critical values; for WB the critical values are based on 500 replications, whereas for MB they are based on 1000 replications. Note that the actual computation of the integral over $\mathbb{S}^{k+p}$ in $CM_j^{*(WB)}$ is based on the following expression

$$\begin{aligned} CM_j &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{k=1}^{n} \sum_{l=1}^{n} \widehat{\varepsilon}_{ij} \widehat{\varepsilon}_{kj} I(U_i \leq U_l) I(U_k \leq U_l) \\ &\quad \times \int_S I\left(\theta' \left[X_i', W_i'\right]' \leq \theta' \left[X_l', W_l'\right]'\right) I\left(\theta' \left[X_k', W_k'\right]' \leq \theta' \left[X_l', W_l'\right]'\right) d\theta \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{k=1}^{n} \sum_{l=1}^{n} \widehat{\varepsilon}_{ij} \widehat{\varepsilon}_{kj} I(U_i \leq U_l) I(U_k \leq U_l) S_{ikl}, \end{aligned} \tag{29}$$

where the integral S_{ikl} is proportional to the volume of a spherical wedge and can be computed as

$$S_{ikl} = \left| \pi - \arccos\left(\frac{\left[(X_i - X_l)', (W_i - W_l)'\right]\left[(X_k - X_l)', (W_k - W_l)'\right]'}{\|X_i - X_l\| \|X_k - X_l\| \|W_i - W_l\| \|W_k - W_l\|} \right) \right| \frac{\pi^{\frac{k+p}{2} - 1}}{\Gamma\left(\frac{k+p}{2} + 1\right)}$$

and $\Gamma(\cdot)$ is the gamma function; see Escanciano (2006) for further details.

One important practical aspect of the results of this paper concerns the choice of the bandwidth, since as pointed out for example by Zhu and Ng (2003), no optimal

Table 1 Finite sample sizes of omnibus tests CM_1 and CM_3 for model 27

	n		b_{cv}		$2b_{cv}$		$4b_{cv}$	
WB								
$\pi_{(1)}$	50	CM1	0.0905	0.0425	0.0899	0.0420	0.0897	0.0420
		CM_3	0.0968	0.0433	0.0965	0.0435	0.0967	0.0436
	100	CM_1	0.0924	0.0454	0.0922	0.0453	0.0921	0.0456
		CM_3	0.0976	0.0461	0.0973	0.0462	0.0975	0.0460
	400	CM_1	0.0953	0.0475	0.0949	0.0470	0.0951	0.0479
		CM_3	0.0990	0.0489	0.0990	0.0486	0.0991	0.0485
$\pi_{(2)}$	50	CM_1	0.0893	0.0398	0.0895	0.0394	0.0895	0.0397
		CM_3	0.0950	0.0425	0.0952	0.0425	0.0957	0.0428
	100	CM_1	0.0925	0.0440	0.0921	0.0442	0.0921	0.0441
		CM_3	0.0965	0.0456	0.0969	0.0456	0.0963	0.0450
	400	CM_1	0.0940	0.0471	0.0931	0.0471	0.0935	0.0470
		CM_3	0.0980	0.0476	0.0982	0.0477	0.0981	0.0478
MB								
$\pi_{(1)}$	50	CM_1	0.0888	0.0403	0.0890	0.0400	0.0887	0.0401
		CM_3	0.0943	0.0426	0.0940	0.0425	0.0938	0.0427
	100	CM_1	0.0910	0.0427	0.0906	0.0426	0.0906	0.0430
		CM_3	0.0949	0.0437	0.0948	0.0438	0.0945	0.0433
	400	CM_1	0.0939	0.0446	0.0938	0.0455	0.0957	0.0447
		CM_3	0.0978	0.0482	0.0973	0.0480	0.0975	0.0479
$\pi_{(2)}$	50	CM_1	0.0883	0.0386	0.0883	0.0389	0.0885	0.0389
		CM_3	0.0925	0.0418	0.0921	0.0420	0.0924	0.0419
	100	CM_1	0.0900	0.0420	0.0902	0.0421	0.0892	0.0419
		CM_1	0.0925	0.0427	0.0926	0.0427	0.0928	0.0422
	400	CM_1	0.0950	0.0441	0.0952	0.0435	0.0954	0.0435
		CM_3	0.0978	0.0469	0.0979	0.0467	0.0978	0.0469

bandwidth selection theory is available in the context of testing. To investigate this problem we consider bandwidths chosen by standard cross-validation (b_{cv}) and then two bandwidths chosen using a fixed grid based on $2b_{\text{cv}}$ and $4b_{\text{cv}}$ (that is twice and fourth times the bandwidth chosen by cross-validation). For the three sample sizes considered the average b_{cv} for model (27) is $[0.084, 0.065, 0.059]$, and for (28) is $[0.058, 0.047, 0.042]$.

Tables 1, 2 report the finite sample size of CM_1 and CM_3 corresponding to a nominal level of 10 and 5 per cent for models (27) and (28) using 1000 replications and a second-order Epanechnikov kernel for both the varying coefficient parameter $\alpha(U)$ and the missing probability mechanisms $\pi_{(j)}(U, X, W)$ $(j = 1, 2)$.

Both tables indicate that the tests are slightly undersized especially for $n = 50$ but they approach the correct nominal size as n increases. Not surprisingly the degree of the size distortion is related to the percentage of missingness, which is higher

Table 2 Finite sample sizes of omnibus tests CM_1 and CM_3 for model 28

π	n		b_{cv}		$2b_{cv}$		$4b_{cv+0.9}$	
WB								
$\pi_{(1)}$	50	CM_1	0.0923	0.0445	0.0920	0.0448	0.0921	0.0449
		CM_3	0.0956	0.0459	0.0958	0.0454	0.0954	0.0457
	100	CM_1	0.0942	0.0465	0.0941	0.0468	0.0943	0.0465
		CM_3	0.0980	0.0470	0.0977	0.0471	0.0980	0.0468
	400	CM_1	0.0983	0.0480	0.0979	0.0479	0.0980	0.0479
		CM_3	0.0991	0.0491	0.0989	0.0491	0.0988	0.0492
	50	CM_1	0.0917	0.0434	0.0918	0.0436	0.0915	0.0435
		CM_3	0.0929	0.0448	0.0928	0.0449	0.0930	0.0445
$\pi_{(2)}$	100	CM_1	0.0938	0.0455	0.0935	0.0455	0.0935	0.0456
		CM_3	0.0950	0.0462	0.0949	0.0460	0.0951	0.0460
	400	CM_1	0.0967	0.0475	0.0966	0.0474	0.0967	0.0475
		CM_3	0.0976	0.0481	0.0975	0.0480	0.0974	0.0480
MB								
$\pi_{(1)}$	50	CM_1	0.0915	0.0435	0.0916	0.0435	0.0914	0.0433
		CM_3	0.0947	0.0448	0.0946	0.0447	0.0947	0.0445
	100	CM_1	0.0935	0.0457	0.0939	0.0455	0.0937	0.0458
		CM_3	0.0954	0.0465	0.0950	0.0463	0.0951	0.0462
	400	CM_1	0.0966	0.0468	0.0964	0.0465	0.0967	0.0465
		CM_3	0.0981	0.0483	0.0978	0.0481	0.0983	0.0482
$\pi_{(2)}$	50	CM_1	0.0915	0.0425	0.0913	0.0429	0.0913	0.0424
		CM_3	0.0938	0.0441	0.0939	0.0440	0.0940	0.0442
	100	CM_1	0.0932	0.0436	0.0938	0.0434	0.0931	0.0438
		CM_3	0.0945	0.0459	0.0943	0.0458	0.0945	0.0456
	400	CM_1	0.0960	0.0463	0.0960	0.0462	0.0961	0.0462
		CM_3	0.0971	0.0475	0.0972	0.0474	0.0970	0.0477

for the second specification $\pi_{(2)}(U, X, W)$. In both tables it appears that WB typically provides a more accurate approximation to the distribution of the test statistics. Finally we note that the bandwidth choice has little effect on the finite sample size.

Figures 1 and 2 illustrate the finite sample power properties of CM_1 and CM_3 for the nominal size 0.05. The power is computed at 12 values of γ in the range $\gamma = [0.7, 3.5]$ for (27), and in the range $\gamma = [0.4, 2.8]$ for (28) using 1000 replications for a sample size of $n = 100$ with the $\pi_{(2)}(U, X, W)$ specification and the same cross-validated bandwidths b_{cv} as those used in Tables 1a and 1b. Results for $n = 50$ and 400 are qualitatively similar to the ones presented here and hence are omitted.

In each figure the left and centre panel show the power of, respectively, the CM_1 and CM_3 statistics with WB (solid line) and MB (dashed line) approximation. In each cases it is evident that the WB approximation yields a test statistic with higher power.

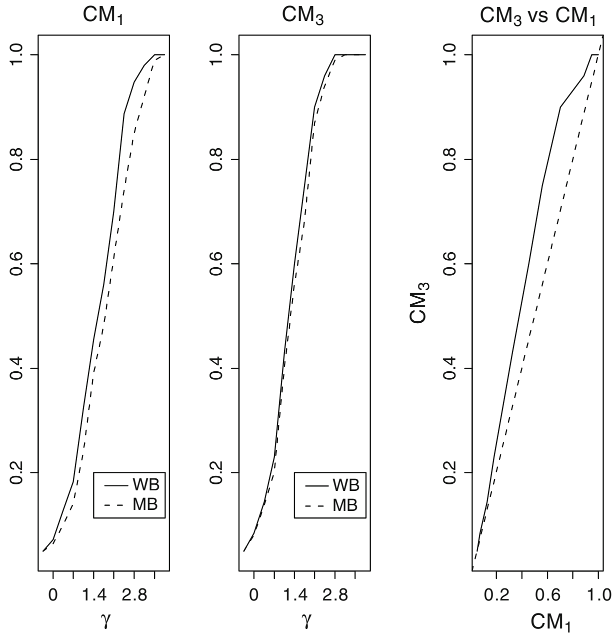

Fig. 1 Finite sample power of omnibus tests CM_1 and CM_3 for model (27) using both WB and MB for $n = 100$

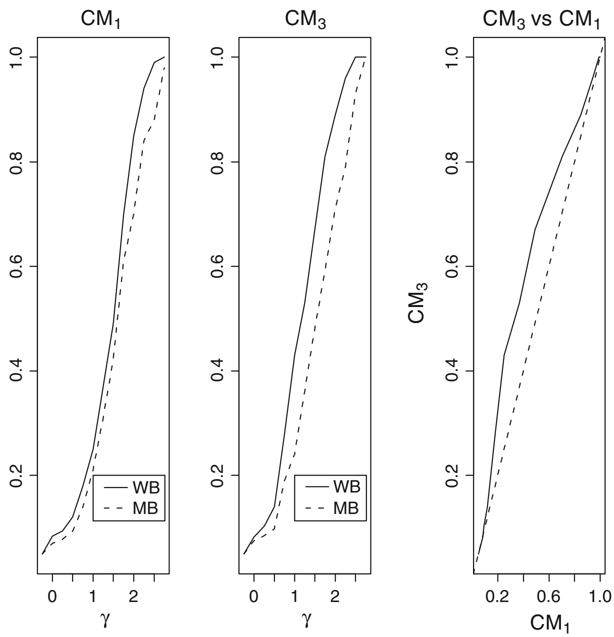

Fig. 2 Finite sample power of omnibus tests CM_1 and CM_3 for model (28) using both WB and MB for $n = 100$

The two right panels compare the relative powers of CM_1 and CM_3 and show that for both models the power of CM_3 is higher than that of CM_1.

Next we consider the three imputed estimators $\widehat{\mu}_j$ for the mean μ_0 and compare them with the same estimator estimator proposed by Horvitz and Thompson (1952) (see also Hirano et al. 2003) that is given by

$$\widehat{\mu}_{\text{HIR}} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i \delta_i.}{\widehat{\pi}\,(U_i, X_i, W_i)},$$

which is asymptotically equivalent to $\widehat{\mu}_2$. We also consider the complete case estimator $\widehat{\mu}_C = \sum_{i=1}^{n} \delta_i Y_i / \sum_{i=1}^{n} \delta_i$.

Tables 3, 4 report the biases and standard deviations for the four estimators based on 5000 replications for sample sizes of $n = 50$, $n = 100$ and $n = 400$ for (27) and (28) respectively, with bandwidth chosen using least squares cross-validation.

Table 3 indicates that $\widehat{\mu}_3$ is characterized by the smallest finite sample bias across both designs and both missing probability specifications; the bias of $\widehat{\mu}_{\text{HIR}}$ is always in between that of $\widehat{\mu}_1$ and those of $\widehat{\mu}_2$ and $\widehat{\mu}_3$. Note also that for $n = 400$ and $\pi_{(1)}(U, X, W)$ the biases of $\widehat{\mu}_2$ and $\widehat{\mu}_3$ are almost identical. Table 3 also indicates that all of the proposed estimators represent an improvement in terms of bias over the complete case estimator $\widehat{\mu}_C$. Table 4 reports the standard deviations of the five estimators. The theoretical values of the standard deviation of the response variable without MAR (i.e. the true unobservable one) are obtained by simulations and are 1.215 and 2.741 for models (27) and (28), respectively. All of the four imputation estimators have a standard deviation that is closer to that of the unobservable response compared to that of the complete case estimator. Thus, bearing in mind that according

Table 3 Bias for imputation estimators for the mean functional

π	n	$\widehat{\mu}_1$	$\widehat{\mu}_2$	$\widehat{\mu}_3$	$\widehat{\mu}_{HIR}$	$\widehat{\mu}_C$
Model M1						
$\pi_{(1)}$	50	−0.00104	−0.00081	−0.00858	−0.00090	−0.00125
	100	−0.00093	−0.00074	−0.00680	−0.00081	−0.00105
	400	−0.00046	−0.00014	−0.00147	−0.00029	−0.00056
$\pi_{(2)}$	50	−0.00283	−0.00159	−0.001508	−0.00187	−0.00299
	100	−0.00143	−0.00113	−0.00105	−0.00126	−0.00164
	400	−0.00078	−0.00059	−0.00058	−0.00068	−0.00088
Model M2						
$\pi_{(1)}$	50	0.00642	0.00328	0.00327	0.00513	0.00751
	100	0.00542	0.00244	0.00223	0.00371	0.00599
	400	0.00322	0.00133	0.00144	0.00222	0.00381
$\pi_{(2)}$	50	0.00758	0.00514	0.00452	0.00557	0.00833
	100	0.00572	0.00384	0.00352	0.00466	0.00633
	400	0.00399	0.00201	0.00192	0.00249	0.00413

Table 4 Standard error for imputation estimators for the mean functional

π	n	$\widehat{\mu}_1$	$\widehat{\mu}_2$	$\widehat{\mu}_3$	$\widehat{\mu}_{\mathrm{HIR}}$	$\widehat{\mu}_C$
Model 27[a]						
$\pi_{(1)}$	50	1.1163	1.1299	1.1324	1.1401	1.0913
	100	1.1203	1.1428	1.1483	1.1479	1.0971
	400	1.1415	1.1541	1.1755	1.1512	1.1164
$\pi_{(2)}$	50	0.99562	1.0041	1.0091	1.0240	0.94937
	100	1.0329	1.0412	1.0954	1.0866	0.96524
	400	1.0645	1.0787	1.0883	1.0893	0.98776
Model 28[b]						
$\pi_{(1)}$	50	2.5870	2.6083	2.6110	2.5935	2.4421
	100	2.6099	2.6240	2.6280	2.6059	2.5559
	400	2.6122	2.6304	2.6475	2.6290	2.6004
$\pi_{(2)}$	50	2.1748	2.2219	2.2344	2.2027	2.0993
	100	2.1959	2.2704	2.2877	2.2657	2.1089
	400	2.2379	2.3073	2.3322	2.3002	2.1413

[a] The simulated standard deviation of unobservable response is 1.215
[b] The simulated standard deviation of unobservable response is 2.741

to Proposition 10 none of the proposed estimators is fully efficient under (27) and (28), we see that the proposed imputation methods provide some improvements on the precision of the resulting estimators. We also note that among the four imputation estimators the one based on inverse probability weighting with partial propensity score (8) has an edge over the others (except for the case of $n = 50$ in model (27) where the Hirano et al. (2003)'s estimator is closer to the standard deviation of the unobservable response).

Taken together the results of Tables 1, 2, 3, 4 and Figures 1 and 2 indicate that the proposed imputation estimators and test statistics are characterized by good finite sample properties which compare favourably with those based on competing estimators and test statistics. The results seem to suggest that WB delivers test statistics that have slightly better finite sample size and power than those based on MB, and that the inverse probability weighted estimators and test statistics based on the partial propensity score (8) are characterized by better finite sample properties than those based on the complete case analogue. These results combined with those of Sect. 4 suggest that, from a practical point of view, partial propensity score inverse probability weighting can be a useful estimation technique with MAR responses, in particular when the dimension of the covariates is high so that the estimation of the full propensity score is subjected to the curse of dimensionality.

6 Empirical application

In this section we illustrate the methods of this paper by considering the question of whether the WTO can have negative effects on the environment. This question has been at the centre of a long standing debate between environmentalists and the trade agreement supporters; see for example Copeland and Taylor (2004) for a review of various theories and some empirical evidence between trade, growth and the environment.

Millimet and Tchernis (2009) and Bravo and Jacho-Chavez (2011) use country-level data from Frankel and Rose (2005) (available at http://faculty.haas.berkeley.edu/arose) to investigate the possible negative causal relationship between trade and environment, by specifying the treatment variable as the GATT/WTO membership and considering fiver different measures of environmental quality: Per capita dioxide (CO_2) emissions, the average annual deforestation rate from 1990–1996, energy depletion, rural access to clean water and urban access to clean water. Both Millimet and Tchernis (2009) and Bravo and Jacho-Chavez (2011) use inverse probability weighed estimators for the average treatment effect with propensity scores estimated, respectively, with parametric and non parametric methods. We use the same data and the same variables as those used by Millimet and Tchernis (2009) and Bravo and Jacho-Chavez (2011), but as opposed to these authors we use a partially linear varying coefficient specification with the same three covariates (log-real per capita GDP ($Log - rgdppc$), a measure of the democratic structure of the government ($Demo - Str$) and land area ($Land$). We consider only four of the five environmental variables, namely the dioxide emissions ($CO_2 pc$), the annual deforestation ($Defor$), rural access ($Rural$) and urban access ($Urban$). This choice is suggested by some preliminary graphical analysis suggesting that for each of these four variables there is evidence of some nonlinear relationship with one of the covariates, as illustrated in Fig. 3.

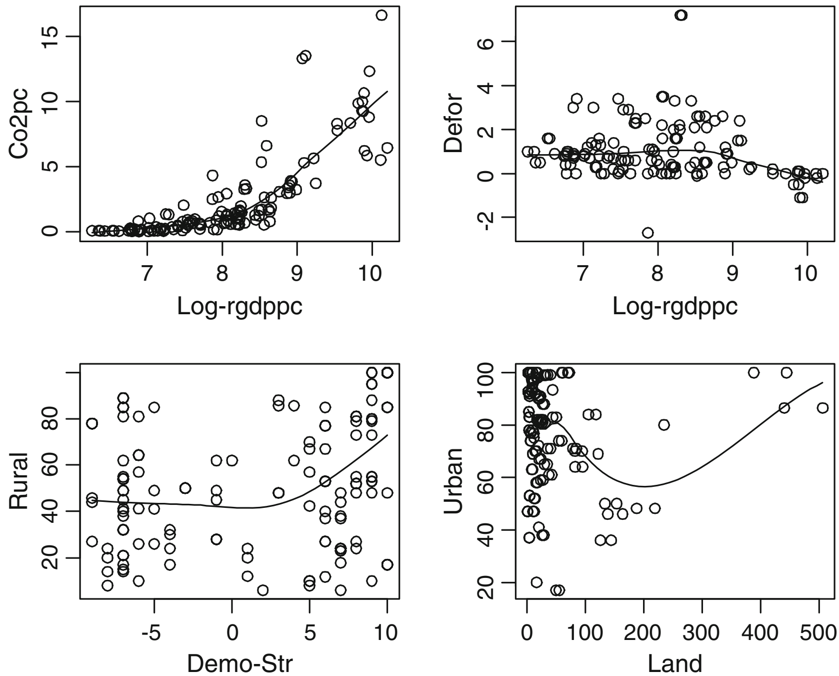

Fig. 3 *Scatter plots* for the four enviromental variables with local linear regression line

We use the results of this paper in two different ways: first to test whether a given partially linear varying coefficient specification is supported by the data; second to estimate the average treatment effect parameter using the corresponding specification. In the estimation we use the same second-order Epanechnikov kernel with cross-validated bandwidth as that used in the previous section. For the specification tests we consider the CM_1 statistic and use bootstrap p-values (p^*) obtained from 500 replications of $CM_1^{*(WB)}$.

The four specifications are

$$
\begin{aligned}
Co2pc_i^{(1)} &= \underset{p=0.145}{0.0302} Land_i{}^{(1)} + Demo - Str_i^{(1)} \widehat{\alpha}^{(1)} (Log - rgdppc_i), \qquad (30)\\
R^2 &= 0.82,\ p^* = 0.562,\\
Co2pc_i^{(0)} &= \underset{p=0.138}{-0.0232} Land_i{}^{(0)} + Demo - Str_i^{(0)} \widehat{\alpha}^{(0)} (Log - rgdppc_i),\\
R^2 &= 0.11,\ p^* = 0.881,\\
Defor_i^{(1)} &= \underset{p=0.021}{-0.003} Land_i^{(1)} + \underset{p=0.001}{0.056} Demo - Str_i^{(0)} + \widehat{\alpha}^{(1)} (Log - rgdppc_i),\\
R^2 &= 0.19,\ p^* = 0.384,\\
Defor_i^{(0)} &= \underset{p=0.321}{-0.0004} Land_i^{(1)} + \underset{p=0.451}{0.003} Demo - Str_i^{(0)} + \widehat{\alpha}^{(0)} (Log - rgdppc_i),\\
R^2 &= 0.069,\ p^* = 0.185,\\
Rural_i^{(1)} &= \underset{p=0.000}{0.0036} Log - rgdppc_i^{(1)} + Land_i^{(1)} \widehat{\alpha}^{(1)} (Demo - Str_i),\\
R^2 &= 0.42,\ p^* = 0.673,\\
Rural_i^{(0)} &= \underset{p=0.122}{-0.0006} Log - rgdppc_i^{(0)} + Land_i^{(0)} \widehat{\alpha}^{(0)} (Demo - Str_i),\\
R^2 &= 0.32,\ p^* = 0.432,\\
Urban_i^{(1)} &= \underset{p=0.000}{0.0026} Log - rgdppc_i^{(1)} + Demo - Str_i{}^{(1)} \widehat{\alpha}^{(1)} (Land_i),\\
R^2 &= 0.323,\ p^* = 0.303,\\
Urban_i^{(0)} &= \underset{p=0.165}{-0.0007} Log - rgdppc_i^{(1)} + Demo - Str_i{}^{(1)} \widehat{\alpha}^{(1)} (Land_i),\\
R^2 &= 0.123,\ p^* = 0.405,
\end{aligned}
$$

that is we have three partially linear varying coefficient models and one partially linear model (for the deforestation variable). The latter is chosen over a partially linear varying coefficient because of a higher R^2 and lower residual standard error.

Table 5 reports the point estimates of the treatment variable GATT/WTO membership, together with the 0.95 confidence interval (C.I.) and the p value (p) of the associated t statistic for the null hypothesis of no treatment effect $H_0 : \tau = 0$, using two estimators: $\widehat{\tau}_1$ as defined in (26) with the specifications of (30), and the inverse probability weighting method of Hirano et al. (2003). Note that as in Bravo and Jacho-

Table 5 Average treatment effect

	τ_1			τ_{HIR}		
	$\widehat{\tau}_1$	$C.I.$	p	$\widehat{\tau}_{\text{HIR}}$	$C.I.$	p
co2pc	0.79	0.34, 1.24	0.00	0.33	−0.61, 1.28	0.484
defor	0.72	0.47, 0.97	0.00	0.38	−0.25, 1.01	0.233
rural	21.44	15.08, 27.39	0.00	8.57	−10.21, 27.37	0.367
urban	28.06	18.89, 37.24	0.00	11.78	−17.26, 40.82	0.422

Chavez (2011) we exclude observations in the averages with an estimated propensity score outside the interval [0.05,0.95] in both sets of estimators.

The results of Table 5 seem to suggest that the WTO might have negative effect on the environment once we take into consideration the three covariates. Interestingly the inverse probability weighted estimator of Hirano et al. (2003) suggests that the environmental effects are statistically insignificant, and the explanation for this seems to stem from the large variability associated with this estimator, as the 0.95 confidence intervals clearly indicate. To assess the sensitivity of the results of Table 5 we have considered a different choice of bandwidth, namely that based on Silverman's rule of thumb, and we have also estimated the average treatment effect using the doubly-robust estimator $\widehat{\tau}_2$. In both cases the point estimates of the treatment parameters varied, but crucially all of them remained statistically significant.

7 Conclusions

In this paper we consider partially linear varying coefficient models with responses missing at random. We consider imputation and inverse probability weighting and propose an omnibus test for the correct specification based on a Cramer von Mises type of statistic. We also consider the problem of estimating the mean of a response variable assumed to be related to a set of covariates via a partially linear varying coefficient specification. As a by-product of this estimator we propose a novel estimator for the average treatment effect parameter that is key for causal inference models. We investigate the finite sample properties of the proposed estimators and test statistics with simulations. The results of the simulations are encouraging and suggest that all of the proposed estimators and test statistics and especially those based on inverse probability weighting with a partial score are characterized by good finite sample properties that compare favourably with those of existing alternatives. We also apply the results of this paper to investigate whether the WTO can have negative effects on the environment, and suggest that this might be the case.

8 Appendix

Throughout this appendix we use the following abbreviations: "CLT", "CMT" and "LNN" denote, respectively, central limit theorem, continuous mapping theorem and (possibly uniform in the sets $\mathcal{U}$ or $\mathcal{Z}$ defined below) law of large numbers. Let $Z = \left[U, X', W'\right]'$.

8.1 Assumptions

- A1 (i) The random variable U has bounded support $\mathcal{U}$ and its density $f_U(\cdot)$ is Lipschitz continuous and bounded away from 0 in $\mathcal{U}$ or (ii) the random vector Z has a compact set support $\mathcal{Z} \subseteq \mathbb{R}^{l+k+1}$ and its density $f_Z(\cdot)$ is Lipschitz continuous and bounded away from 0 in $\mathcal{Z}$,
- A2 The $p \times p$ matrix $E\left(\delta X X'|U\right)$ is nonsingular for each U, and $E\left(\delta W X'|U\right)$, $E\left(\delta X X'|U\right)$ have Lipschitz continuous second derivatives in $U \in \mathcal{U}$, and $E\left(\delta X X'|U\right)^{-1}$ is Lipschitz continuous,
- A3 $E\left(\|X\|^4\right) < \infty$, $E\left(\|W\|^4\right) < \infty$, $E\left(\varepsilon^4\right) < \infty$,
- A4 $E\left[\pi(Z)\left(W - \alpha_{WX}(U)' X\right)^{\otimes 2}\right]$ is positive definite,
- A5 The functions $\alpha_j(U)$ have Lipschitz continuous second derivatives in $U \in \mathcal{U}$,
- A6 (i) $\inf_{u\in\mathcal{U}} \pi(U) > 0$, the function $\pi(U)$ has Lipschitz continuous second derivatives in $U \in \mathcal{U}$ or (ii) $\inf_{z\in\mathcal{Z}} \pi(Z) > 0$, the function $\pi(Z)$ has Lipschitz continuous second derivatives in $Z \in \mathcal{Z}$,
- A7 As $n \to \infty$ (i) $n^{1/2}h^4 \to 0$ and $nh^3/\ln(n) \to \infty$ or (ii) $n^{1/2}b^{2l} \to 0$ and $nh^{k+p+3}/\ln(n) \to \infty$ for $l > k + p + 1$,
- A8 The kernel functions $K(\cdot)$ and $L(\cdot)$ are symmetric densities with compact support,
- A9 The product measure $F_{n,\theta}(u, s)\, d\theta$ is absolutely continuous with respect to the Lebesgue measure in Π,
- A10 The function $E\left[\gamma(Z)|U\right]$ is Lipschitz continuous and $E\left[\gamma(Z)\right] = 0$.

8.2 Proofs of the theorems

Proof of Theorem 1 Let $\alpha_{0XW}(u) = \left[E\left(\delta X^{\otimes 2}|U = u\right)\right]^{-1} E\left[\delta X W'|U = u\right]$ and $\alpha_{0XY}(u) = \left[E\left(\delta X^{\otimes 2}|U = u\right)\right]^{-1} E\left[\delta X Y|U = u\right]$; by results of Fan and Huang (2005) and CMT

$$\|\widehat{\alpha}_{XY}(u) - \alpha_{0XY}(u)\| = o_p(1),$$
$$\|\widehat{\alpha}_{XW}(u) - \alpha_{0XW}(u)\| = o_p(1)$$

uniformly in $\mathcal{U}$, so that by LLN and CMT

$$\left\|\left\{\sum_{i=1}^{n}\left[\delta_i\left(W_i - \alpha_{0XW}(U_i)' X_i\right)\right]^{\otimes 2}/n\right\}^{-1} - \Gamma_0^{-1}\right\| = o_p(1);$$

hence

$$n^{1/2}\left(\widehat{\beta}-\beta_0\right)=\Gamma_0^{-1}\frac{1}{n^{1/2}}\sum_{i=1}^{n}\delta_i\left(W_i-\alpha_{0XW}\left(U_i\right)'X_i\right)\varepsilon_i+o_p\left(1\right), \tag{31}$$

and the result follows by CLT and CMT. Let Q denote the semiparametric model and let Q_χ denote the parametric submodel for the log-density of (Y, δ, U, X, W) that is

$$\ln f_\chi\left(Y,\delta,U,X,W\right)=C-\frac{\delta}{2}\ln\sigma^2-\frac{\delta}{2\sigma^2}\left(Y-X'\alpha_\lambda\left(U\right)-W'\beta\right)^2+\delta\ln\pi\left(Z,\upsilon_1\right)$$
$$+\left(1-\delta\right)\ln\left[1-\pi\left(Z,\upsilon_1\right)\right]+\ln h\left(Z,\upsilon_2\right),$$

where $h(\cdot)$ is the density of the covariates Z and $\chi=\left[\alpha_\lambda', \beta', \sigma^2, \upsilon_1', \upsilon_2'\right]'$ is the finite dimensional parameter. The score functions are $S_\beta=\delta W\varepsilon\sigma^{-2}$, $S_\lambda=\delta\partial\alpha_\lambda/\partial\lambda X\varepsilon\sigma^{-2}$, $S_\sigma=\delta\left(\varepsilon^2-\sigma^2\right)\sigma^{-3}$, $S_{\upsilon_1}=\left(\delta-\pi_{\upsilon_1}\right)\partial\pi_{\upsilon_1}/\partial\upsilon_1\pi_{\upsilon_1}^{-1}\left(1-\pi_{\upsilon_1}\right)^{-1}$ where $\pi_{\upsilon_1}:=\pi_{\upsilon_1}\left(Z,\upsilon_1\right)$, and $S_{\upsilon_2}=h\left(Z,\upsilon_2\right)^{-1}\partial h\left(Z,\upsilon_2\right)/\partial\upsilon_2$. The tangent space (see e.g. Bickel et al. 1993) is

$$T_\chi=\left[\delta W\varepsilon, \delta\partial\alpha_\lambda/\partial\lambda X\varepsilon, \delta\left(\varepsilon^2-\sigma^2\right)\sigma^{-3}, \left(\delta-\pi_{\upsilon_1}\right)a\left(Z\right), b\left(Z\right)\right] \tag{32}$$

where $a(\cdot)$ is such that $E\left\|a\left(Z\right)\right\|^2<\infty$ and $b(Z)$ is such that $E\left[b\left(Z\right)\right]=0$. Then the efficient score function S_β^* for β_0 is the projection -Proj(|)- of S_β into the orthogonal complement of the direct sum $S_\lambda+S_\sigma+S_{\upsilon_1}+S_{\upsilon_2}$ and since S_λ, S_σ, S_{υ_1}, S_{υ_2} are orthogonal to each other and S_β and S_σ are orthogonal to $\left(\delta-\pi_{\upsilon_1}\right)a\left(Z\right)$, $b\left(Z\right)$ and $E\left[S_\beta S_\sigma\right]=0$ by normality, it follows that $S_\beta^*=S_\beta-\text{Proj}\left(S_\beta|S_\lambda\right)$. Simple algebra shows that such projection amounts to finding a vector of the form $\delta\partial\alpha_\lambda/\partial\lambda X$ such that $E\left\|S_\beta-S_\lambda\right\|^2$ is minimised, and that such vector is $E\left(\delta WX'|U\right)E\left(\delta X^{\otimes 2}|U\right)^{-1}\delta X\varepsilon/\sigma^2$. □

Proof of Theorem 2 Note that

$$n^{1/2}\widehat{\upsilon}_1\left(u,s,\theta\right)=\frac{1}{n^{1/2}}\sum_{i=1}^{n}\delta_i\varepsilon_i I\left(U_i\le u,\theta'\left[X_i', W_i'\right]'\le s\right)$$
$$+\frac{1}{n^{1/2}}\sum_{i=1}^{n}\delta_i\left[X_i'\left(\widehat{\alpha}_{XY}\left(U_i\right)-\alpha_{0XY}\left(U_i\right)\right)+X_i'\left(\widehat{\alpha}_{XW}\left(U_i\right)-\alpha_{0XW}\left(U_i\right)\right)\beta_0\right]$$
$$\times I\left(U_i\le u,\theta'\left[X_i', W_i'\right]'\le s\right)$$
$$-\frac{1}{n^{1/2}}\sum_{i=1}^{n}\delta_i\left(W_i'-X_i'\alpha_{XW}\right)\left(\widehat{\beta}-\beta_0\right)I\left(U_i\le u,\theta'\left[X_i', W_i'\right]'\le s\right)$$

$$= \frac{1}{n^{1/2}} \sum_{i=1}^{n} \delta_i X_i' \left(\widehat{\alpha}_{XW}(U_i) - \alpha_{0XW}(U_i)\right)\left(\widehat{\beta} - \beta_0\right)$$
$$\times I\left(U_i \le u, \theta'\left[X_i', W_i'\right]' \le s\right) =: \sum_{j=1}^{4} T_{1j}.$$

By standard kernel calculations

$$T_{12} = \frac{1}{n^{1/2}} \sum_{i=1}^{n} \delta_i X_i' \left\{\left[\sum_{l=1}^{n} \delta_i X_i^{\otimes 2} K_h(U_l - U_i)\right]^{-1}\right.$$
$$\left.\times \sum_{j=1}^{n} \delta_j X_j \left[Y_j - X_j'\alpha_0(U_j) - W_i'\beta_0\right] K_h(U_j - U_i)\right\}$$
$$= \frac{1}{n^{1/2}} \sum_{i=1}^{n} \left[E(X|U_i) - E(\delta X|U_i)\right]' \left[E\left(\delta X^{\otimes 2}|U_i\right)\right]^{-1} \delta_i X_i \varepsilon_i + o_p(1);$$

furthermore by LLN

$$T_{13} = \frac{1}{n^{1/2}} \sum_{i=1}^{n} \delta_i X_i' I\left(U_i \le u, \theta'\left[X_i', W_i'\right]' \le s\right)\left[E\left(\delta X^{\otimes 2}|U_i\right) f(U_i)\right]^{-1}$$
$$\times \frac{1}{n} \sum_{j=1}^{n} \delta_j X_j \varepsilon_j K_h(U_j - U_i) + o_p(1)$$
$$= \frac{1}{n^{1/2}} \sum_{i=1}^{n} G(s, \theta, U_i)\left[E\left(\delta X^{\otimes 2}|U_i\right)\right]^{-1} \delta_i X_i \varepsilon_i + o_p(1),$$

where $G(s, \theta, u) = E\left[\delta X' I\left(\theta'\left[X', W'\right]' \le s\right)|U = u\right]$. By LLN and as in the proof of Theorem 1

$$|T_{14}| \le \sup_{u \in \mathcal{U}} \|\widehat{\alpha}_{XW}(u) - \alpha_{0XW}(u)\| \left\| \frac{1}{n} \sum_{i=1}^{n} X_i I\left(U_i \le u, \theta'\left[X_i', W_i'\right]' \le s\right)\right\|$$
$$\times n^{1/2}\left\|\widehat{\beta} - \beta_0\right\| = o_p(1),$$

and

$$\left\| T_{32} - D(u, s, \theta) n^{1/2}\left(\widehat{\beta} - \beta_0\right) s\right\| = o_p(1),$$

where

$$D(u, s, \theta) = E\left[\delta\left(W' - X'\alpha_{0XY}\right)\omega I\left(U \le u, \theta'\left[X', W'\right]' \le s\right)\right]$$

since the class of function $\{(r, q, t) \to (r - d(q)t)\,\omega(r, q)\, I\left(r \le u, \theta' t \le s\right), u, s, \theta \in \Pi\}$ is Vapnik-Chervonenkis. Thus

$$
\begin{aligned}
& n^{1/2}\widehat{v}_1(u, s, \theta) \\
&= \frac{1}{n^{1/2}} \sum_{i=1}^{n} \delta_i \varepsilon_i I\left(U_i \le u, \theta' \left[X_i', W_i'\right]' \le s\right) \\
&\quad - D(u, s, \theta)\, \Gamma_0^{-1} \frac{1}{n^{1/2}} \sum_{i=1}^{n} \delta_i \left(W_i - \alpha_{0WX}(U_i)' X_i\right) \varepsilon_i \\
&\quad - \frac{1}{n^{1/2}} \sum_{i=1}^{n} G(s, \theta, U_i) \left[E\left(\delta X^{\otimes 2} | U_i\right)\right]^{-1} \delta_i X_i \varepsilon_i I(U_i \le u) + o_p(1). \quad (33)
\end{aligned}
$$

The finite dimensional convergence of (33) follows by CLT, whereas the asymptotic equicontinuity follows by a direct application of Theorem 2.5.2 of Van der Vaart and Wellner (1996), which implies the weak convergence of (33) to $v_1(u, s, \theta)$.

For $n^{1/2}\widehat{v}_3(u, s, \theta)$, (38), CMT and the same arguments as those used to obtain (33) can be used to show

$$
\begin{aligned}
n^{1/2}\widehat{v}_3(u, s, \theta) &= \frac{1}{n^{1/2}} \sum_{i=1}^{n} \frac{\delta_i \varepsilon_i}{\pi(U_i)} I\left(U_i \le u, \theta' \left[X_i', W_i'\right]' \le s\right) \\
&+ \frac{1}{n^{1/2}} \sum_{i=1}^{n} \frac{\delta_i}{\pi(U_i)} \left(W_i' - X_i' \alpha_{XY}\right) \left(\widehat{\beta} - \beta_0\right) I\left(U_i \le u, \theta' \left[X_i', W_i'\right]' \le s\right) \\
&+ \frac{1}{n^{1/2}} \sum_{i=1}^{n} \frac{G(s, \theta, U_i)}{\pi(U_i)} \left[E\left(\delta X^{\otimes 2} | U_i\right)\right]^{-1} \delta_i X_i \varepsilon_i I(U_i \le u) + o_p(1),
\end{aligned}
$$

and the rest of the proof follows by the same arguments as those used to prove the weak convergence of $n^{1/2}\widehat{v}_1(u, s, \theta)$. □

Proof of Theorem 4 Note that as in the proof of Theorem 1 by the bootstrap LLN (see e.g. Bickel and Freedman (1981))

$$
n^{1/2}\left(\widehat{\beta}^* - \widehat{\beta}\right) = \Gamma_0^{-1} \frac{1}{n^{1/2}} \sum_{i=1}^{n} \delta_i \left(W_i - \alpha_{XW}(U_i)' X_i\right) \varepsilon_i^* + o_{p^*}(1),
$$

and

$$
\begin{aligned}
n^{1/2}\widehat{v}_1^*(u, s, \theta) &= \frac{1}{n^{1/2}} \sum_{i=1}^{n} \delta_i \varepsilon_i^* I\left(U_i \le u, \theta' \left[X_i', W_i'\right]' \le s\right) \\
&\quad - \widehat{D}(u, s, \theta)\, \Gamma_0^{-1} \frac{1}{n^{1/2}} \sum_{i=1}^{n} \delta_i \left(W_i - \alpha_{WX}(U_i)' X_i\right) \varepsilon_i^*
\end{aligned}
$$

$$-\frac{1}{n^{1/2}}\sum_{i=1}^{n}\widehat{G}\left(s,\theta,U_i\right)\left[\frac{1}{n}\sum_{j=1}^{n}\delta_j X_j^{\otimes 2}K_h\left(U_j-U_i\right)\right]^{-1}$$
$$\times\delta_i X_i\varepsilon_i^* I\left(U_i\le u\right)+o_{p*}(1), \tag{34}$$

except if the original incomplete sample $(Y_i, U_i, X_i, W_i)_{i=1}^n$ is on a set with probability converging to 0 as $n\to\infty$. As in Stute et al. (1998)

$$\begin{aligned}
&\mathrm{Cov}^*\left(n^{1/2}\widehat{v}_1^*\left(u_1,s_1,\theta\right),n^{1/2}\widehat{v}_1^*\left(u_2,s_2,\theta\right)\right)\\
&=\frac{1}{n}\sum_{i=1}^{n}\delta_i^2\widehat{\varepsilon}_i^2 I\left(U_i\le u_1\wedge u_2,\theta'\left[X_i',W_i'\right]'\le s_1\wedge s_2\right)\\
&\quad-\widehat{D}\left(u_1,s_1,\theta\right)\widehat{\Gamma}^{-1}\frac{1}{n}\sum_{i=1}^{n}\left(W_i-\alpha_{0XW}\left(U_i\right)'X_i\right)^2\delta_i^2\widehat{\varepsilon}_i^2 I\\
&\quad\times\left(U_i\le u_2,\theta'\left[X_i',W_i'\right]'\le s_2\right)\\
&\quad-\widehat{D}\left(u_2,s_2,\theta\right)\widehat{\Gamma}^{-1}\frac{1}{n}\sum_{i=1}^{n}\left(W_i-\alpha_{0XW}\left(U_i\right)'X_i\right)^2\delta_i^2\widehat{\varepsilon}_i^2 I\\
&\quad\times\left(U_i\le u_1,\theta'\left[X_i',W_i'\right]'\le s_1\right)\\
&\quad-\frac{1}{n}\sum_{i=1}^{n}\widehat{G}\left(s_1,\theta,U_i\right)\left[\frac{1}{n}\sum_{j=1}^{n}\delta_j X_j^{\otimes 2}K_h\left(U_j-U_i\right)\right]^{-1}X_i\delta_i^2\widehat{\varepsilon}_i^2 I\\
&\quad\times\left(U_i\le u_2,\theta'\left[X_i',W_i'\right]'\le s_2\right)\\
&\quad-\frac{1}{n}\sum_{i=1}^{n}\widehat{G}\left(s_2,\theta,U_i\right)\left[\frac{1}{n}\sum_{j=1}^{n}\delta_j X_j^{\otimes 2}K_h\left(U_j-U_i\right)\right]^{-1}X_i\delta_i^2\widehat{\varepsilon}_i^2 I\\
&\quad\times\left(U_i\le u_1,\theta'\left[X_i',W_i'\right]'\le s_1\right)\\
&\quad+\widehat{D}\left(u_1,s_1,\theta\right)\Gamma_0^{-1}\frac{1}{n}\sum_{i=1}^{n}\left(W_i-\alpha_{0XW}\left(U_i\right)'X_i\right)^2\delta_i^2\widehat{\varepsilon}_i^2\Gamma_0^{-1}\widehat{D}\left(u_2,s_2,\theta\right)'\\
&\quad-\widehat{D}\left(u_1,s_1,\theta\right)\Gamma_0^{-1}\frac{1}{n}\sum_{i=1}^{n}\left(W_i-\alpha_{0XW}\left(U_i\right)'X_i\right)\delta_i^2\widehat{\varepsilon}_i^2 X_i'\\
&\quad\times\left[\frac{1}{n}\sum_{j=1}^{n}\delta_j X_j^{\otimes 2}K_h\left(U_j-U_i\right)\right]^{-1}\\
&\quad\times\widehat{G}\left(s,\theta,U_i\right)'I\left(U_i\le u_2\right)-\widehat{D}\left(u_2,s_2,\theta\right)\Gamma_0^{-1}
\end{aligned}$$

$$
\begin{aligned}
&\times \frac{1}{n}\sum_{i=1}^{n}\left(W_i - \alpha_{0XW}(U_i)' X_i\right)\delta_i^2 \widehat{\varepsilon}_i^2 X_i' \\
&\times \left[\frac{1}{n}\sum_{j=1}^{n}\delta_j X_j^{\otimes 2} K_h\left(U_j - U_i\right)\right]^{-1} \widehat{G}(s,\theta,U_i)' I(U_i \le u_1) \\
&+\frac{1}{n}\sum_{i=1}^{n}\widehat{G}(s_1,\theta,U_i)\left[\frac{1}{n}\sum_{j=1}^{n}\delta_j X_j^{\otimes 2} K_h\left(U_j - U_i\right)\right]^{-1} \\
&\times X_i X_i' \left[\frac{1}{n}\sum_{j=1}^{n}\delta_j X_j^{\otimes 2} K_h\left(U_j - U_i\right)\right]^{-1} \widehat{G}(s_1,\theta,U_i)' \delta_i^2 \widehat{\varepsilon}_i^2 I(U_i \le u_1 \wedge u_2)
\end{aligned}
$$

and by LLN

$$
\begin{aligned}
&\Big|\mathrm{Cov}^*\left(n^{1/2}\widehat{v}_1^*(u_1,s_1,\theta), n^{1/2}\widehat{v}_1^*(u_2,s_2,\theta)\right) \\
&\quad -\mathrm{Cov}\left(n^{1/2}\widehat{v}_1(u_1,s_1,\theta), n^{1/2}\widehat{v}_1(u_2,s_2,\theta)\right)\Big| = o_p(1).
\end{aligned}
$$

Similarly for $n^{1/2}\widehat{v}_3^*(u,s,\theta)$

$$
\begin{aligned}
&\Big|\mathrm{Cov}^*\left(n^{1/2}\widehat{v}_3^*(u_1,s_1,\theta), n^{1/2}\widehat{v}_3^*(u_2,s_2,\theta)\right) \\
&\quad -\mathrm{Cov}\left(n^{1/2}\widehat{v}_3(u_1,s_1,\theta), n^{1/2}\widehat{v}_3(u_2,s_2,\theta)\right)\Big| = o_p(1).
\end{aligned}
$$

By LLN

$$
n^{1/2}\widehat{v}_j^*(u,s,\theta) = \sum_{i=1}^{n}\xi_i \delta_i \varepsilon_{ji}/n^{1/2} + o_p(1) \quad j = 1 \text{ or } 3,
$$

where

$$
\begin{aligned}
\xi_i &= I\left(U_i \le u, \theta'\left[X_i', W_i'\right]' \le s\right) - D(u,s,\theta)\,\Gamma_0^{-1}\delta_i\left(W_i - \alpha_{WX}(U_i)' X_i\right) \\
&\quad - G(s,\theta,U_i)\, E\left[\delta X^{\otimes 2}|U_i\right] X_i I(U_i \le u),
\end{aligned}
$$

hence, given the conditional independence of ε_i^* from ξ_i, the finite dimensional convergence of $n^{1/2}\widehat{v}_j^*(u,s,\theta)$ follows by Lindeberg's CLT since by A3 and LLN for each $d > 0$

$$\limsup_n \frac{1}{n}\sum_{i=1}^n \int_{\left|\xi_i\varepsilon^*_{ji}\right|\geq dn^{1/2}} \delta_i \left(\xi_i\varepsilon^*_{ji}\right)^2 d\Pr{}^*$$
$$\leq \limsup_n \frac{1}{n}\sum_{i=1}^n \delta_i \left(\xi_i\varepsilon_{ji}\right)^2 I\left(\left|\xi_i\varepsilon_{ji}\geq C\right|\right)\to 0$$

as $C\to\infty$. Finally the asymptotic equicontinuity follows by the same argument as that used in the proof of Theorem 2 hence under H_0

$$n^{1/2}\widehat{\upsilon}^*_j(u,s,\theta)\Longrightarrow \upsilon_j(u,s,\theta)\quad \text{in } l^\infty(\Pi)\quad j=1 \text{ or } 3$$

in probability. The conclusion follows by CMT. □

Proof of Theorem 5 Note that by the definition of $n^{1/2}\widehat{\sigma}^*_j(u,s,\theta)$ it follows that

$$\Big|\mathrm{Cov}^*\left(n^{1/2}\widehat{\sigma}^*_j(u_1,s_1,\theta),n^{1/2}\widehat{\sigma}^*_j(u_2,s_2,\theta)\right)$$
$$-\mathrm{Cov}\left(n^{1/2}\widehat{\sigma}_j(u_1,s_1,\theta),n^{1/2}\widehat{\sigma}_j(u_2,s_2,\theta)\right)\Big|=o_p(1),$$

where $\widehat{\sigma}_{ji}(u,s,\theta)$ are the sample analogues of $\sigma_{ji}(u,s,\theta)$ given in (14), and the same arguments as those used in the proof of Theorem 5 show the finite dimensional convergence and asymptotic equicontinuity of $n^{1/2}\widehat{\sigma}^*_j(u,s,\theta)$. Thus the conclusion follows by CMT. □

Proof of Theorem 6 Let $\beta^\dagger := p\lim\left(\widehat{\beta}\right)$ and $\alpha^\dagger(U) := p\lim(\widehat{\alpha}(U))$. As in the proof of Theorem 2 some calculations show that

$$\widehat{\upsilon}_1(u,s,\theta)=\frac{1}{n}\sum_{i=1}^n\delta_i\varepsilon_i I\left(U_i\leq u,\theta'\left[X_i',W_i'\right]'\leq s\right)$$
$$+\frac{1}{n}\sum_{i=1}^n\delta_i W_i'\beta^\dagger I\left(U_i\leq u,\theta'\left[X_i',W_i'\right]'\leq s\right)$$
$$+\frac{1}{n}\sum_{i=1}^n\delta_i X_i'\alpha^\dagger(U_i) I\left(U_i\leq u,\theta'\left[X_i',W_i'\right]'\leq s\right)$$
$$+\frac{1}{n}\sum_{i=1}^n\delta_i E\left(Y_i|U_i,X_i,W_i\right) I\left(U_i\leq u,\theta'\left[X_i',W_i'\right]'\leq s\right)+o_p(1),$$

and the result follows by LLN and CMT. The result for $n^{1/2}\widehat{\upsilon}_3(u,s,\theta)$ follows similarly hence is omitted. □

Proof of Theorem 7 Note that under (19)

$$Y_i-X_i'\alpha_{0XY}(U_i)-\left(W_i'-X_i'\alpha_{0XW}(U_i)\right)\beta_0=\varepsilon_i+\frac{1}{n^{1/2}}\Big\{\gamma(Z_i)$$
$$-X_i'\left[E\left(\delta_i X_i^{\otimes 2}|U_i\right)\right]^{-1}E\left(\delta_i X_i\gamma(Z_i)|U_i\right)\Big\}.$$

Similarly to (31) by LLN and iterated expectations

$$
\begin{aligned}
n^{1/2}\left(\widehat{\beta}-\beta_0\right) &= \Gamma_0^{-1}\frac{1}{n^{1/2}}\sum_{i=1}^{n}\delta_i\left(W_i-\alpha_{0WX}\left(U_i\right)'X_i\right)\varepsilon_i \\
&\quad +\Gamma_0^{-1}E\left\{\pi\left(Z\right)\left(W'-\alpha_{0XW}\left(U\right)'X\right)\left[\gamma\left(Z\right)\right.\right. \\
&\quad \left.\left. -X'\left[E\left(\delta X^{\otimes 2}|U\right)\right]^{-1}E\left(\delta X\gamma\left(Z\right)|U\right)\right]\right\}+o_p\left(1\right) \\
&= \Gamma_0^{-1}\frac{1}{n^{1/2}}\sum_{i=1}^{n}\delta_i\left(W_i-\alpha_{0WX}\left(U_i\right)'X_i\right)\varepsilon_i+S\left(Z\right)+o_p\left(1\right).
\end{aligned}
$$

By the same arguments as those used in the proof of Theorem 2 expanding $n^{1/2}\widehat{v}_1\left(u,s,\theta\right)$ under the local alternative hypothesis (19), say $n^{1/2}\widehat{v}_1^{(l)}\left(u,s,\theta\right)$, we have

$$
\begin{aligned}
n^{1/2}\widehat{v}_1^{(l)}\left(u,s,\theta\right) &= n^{1/2}\widehat{v}_1\left(u,s,\theta\right)+\frac{1}{n}\sum_{i=1}^{n}\delta_i\gamma\left(Z_i\right)I\left(U\le u,\theta'\left[X',W'\right]'\le s\right) \\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\delta_i\left(W_i'-X_i'\alpha_{0XY}\right)S\left(U,X,W\right)I \\
&\quad \times\left(U_i\le u,\theta'\left[X_i',W_i'\right]'\le s\right)+o_p\left(1\right),
\end{aligned}
$$

and by LLN

$$
n^{1/2}\widehat{v}_1^{(l)}\left(u,s,\theta\right)=n^{1/2}\widehat{v}_{1n}\left(u,s,\theta\right)+s_1\left(u,v,\theta\right)+o_p\left(1\right),
$$

where

$$
\begin{aligned}
s_1\left(u,v,\theta\right) &= E\left[\delta\gamma\left(Z\right)I\left(U\le u,\theta'\left[X',W'\right]'\le s\right)\right] \\
&\quad -E\left[\delta\left(W'-X'\alpha_{XW}\left(U\right)\right)I\left(U\le u,\theta'\left[X',W'\right]'\le s\right)\right] \\
&\quad \times\Gamma_0^{-1}E\left\{\pi\left(Z\right)\left(W-X\alpha_{XW}\left(U\right)\right)\right. \\
&\quad \left.\times\left[\gamma\left(Z\right)-X'\left[E\left(X^{\otimes 2}|U\right)\right]^{-1}E\left(\delta X\gamma\left(Z\right)|U\right)\right]\right\}.
\end{aligned}
$$

The result follows using the same arguments as those used in the proof of Theorem 2 and CMT. The same arguments can be applied to $n^{1/2}\widehat{v}_3^{(l)}\left(u,s,\theta\right)$ to obtain the second conclusion. □

Proof of Theorem 8 Note that

$$n^{1/2}\left(\widehat{\mu}_1-\mu_0\right)=\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left(X_i'\alpha_0\left(U_i\right)+W_i'\beta_0-\mu_0\right)+\frac{1}{n^{1/2}}\sum_{i=1}^{n}\delta_i\varepsilon_i$$
$$+\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left(1-\delta_i\right)\left[X_i'\left(\widehat{\alpha}\left(U_i\right)-\alpha_0\left(U_i\right)\right)+W_i'\left(\widehat{\beta}-\beta_0\right)\right], \quad (35)$$

and that

$$X_i'\left(\widehat{\alpha}\left(U_i\right)-\alpha_0\left(U_i\right)\right)$$
$$=X_i'\widehat{\alpha}_{XX}\left(U_i\right)^{-1}\left[\sum_{j=1}^{n}X_j\delta_j\left(Y_j-X_j'\left(\alpha_{0XY}\left(U_i\right)-\alpha_{0XW}\left(U_i\right)\beta_0\right)-W_j'\beta_0\right)\right]$$
$$\times K_h\left(U_j-U_i\right)+X_i'\alpha_{0XW}\left(\widehat{\beta}-\beta_0\right), \quad (36)$$

where $\widehat{\alpha}_{XX}\left(u\right)=\sum_{i=1}^{n}\delta_iX_i^{\otimes 2}K_h\left(U_i-u\right)/n$, hence by (36), LLN and standard kernel calculations

$$\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left(1-\delta_i\right)\left[X_i'\left(\widehat{\alpha}\left(U_i\right)-\alpha_0\left(U_i\right)\right)+W_i'\left(\widehat{\beta}-\beta_0\right)\right]$$
$$=\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left[E\left(X|U_i\right)-E\left(\delta X|U_i\right)\right]'\left[E\left(\delta X^{\otimes 2}|U_i\right)\right]^{-1}\delta_iX_i\varepsilon_i$$
$$+E\left[\left(1-\delta\right)\left(W'-X'\alpha_{0XW}\left(U\right)\right)\right]n^{1/2}\left(\widehat{\beta}-\beta_0\right)+o_p\left(1\right). \quad (37)$$

Therefore by (31)

$$n^{1/2}\left(\widehat{\mu}_1-\mu_0\right)$$
$$=\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left(X_i'\alpha_0\left(U_i\right)+W_i\beta_0-\mu_0\right)+\frac{1}{n^{1/2}}\sum_{i=1}^{n}\Omega_{01}\left(U_i\right)\delta_i\varepsilon_i$$
$$+\Delta_{01}\Gamma_0^{-1}\frac{1}{n^{1/2}}\sum_{i=1}^{n}\delta_i\left(W_i-\alpha_{0XW}\left(U_i\right)'X_i\right)\varepsilon_i+o_p\left(1\right),$$

and the result follows by CLT and CMT. For the second result note that for v either u or z and $\mathcal{V}$ either $\mathcal{U}$ or $\mathcal{Z}$

$$\sup_{v\in\mathcal{V}}\left|\frac{\pi\left(v\right)}{\widehat{\pi}\left(v\right)}-1\right|=o_p\left(1\right) \quad (38)$$

by Masry (1996), and

$$n^{1/2}\left(\widehat{\mu}_3-\mu_0\right)=\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left(X_i'\alpha_0\left(U_i\right)+W_i\beta_0-\mu_0\right)+\frac{1}{n^{1/2}}\sum_{i=1}^{n}\frac{\delta_i\varepsilon_i}{\pi\left(U_i\right)}$$
$$+\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left(1-\frac{\delta_i}{\pi\left(U_i\right)}\right)\left[X_i'\left(\widehat{\alpha}\left(U_i\right)-\alpha_0\left(U_i\right)\right)+W_i'\left(\widehat{\beta}-\beta_0\right)\right]$$
$$+\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left(\frac{\widehat{\pi}\left(U_i\right)-\pi\left(U_i\right)}{\pi\left(U_i\right)}\right)\delta_i\varepsilon_i+o_p\left(1\right), \tag{39}$$

and

$$\frac{1}{n^{1/2}}\sum_{j=1}^{n}\left(\delta_j-\pi\left(U_j\right)\right)\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_j\varepsilon_i}{\pi\left(U_i\right)f_U\left(U_i\right)}K_h\left(U_j-U_i\right)+o_p\left(1\right)=o_p\left(1\right),$$

hence the result follows by CLT and CMT. The last result follows in a similar manner using (38) and (39) with $\pi\left(U_i\right)$ replaced by $\pi\left(Z_i\right)$, noting that $E\left[E\left(X|U\right)-E\left(\delta X|U\right)/\pi\left(Z\right)\right]=0$. □

Proof of Theorem 9 To obtain the bound we calculate the efficient influence function for an estimator of μ_0. To do so we follow, as in the proof of Theorem 1, the approach of Bickel et al. (1993) and show that the parameter of interest is pathwise differentiable. The efficient influence function is then the projection of this derivative on the tangent space. Let

$$\mu=\int Yf_\varepsilon\left(Y-X'\alpha_\lambda\left(U\right)-W'\beta|Z\right)f\left(Z;\upsilon\right)dYdZ$$

denote parametric (marginal) submodel for the parameter of interest and note that by the implicit function theorem and the normality assumption

$$\partial\mu/\partial\chi$$
$$=\left[-E\left(W\right)',-E\left(\partial\alpha_\lambda\left(U\right)/\partial\lambda\right)',0,E\left[\left(X'\alpha_\lambda\left(U\right)+W'\beta\right)\frac{\partial f\left(Z;\upsilon\right)}{\partial\upsilon}\frac{1}{f\left(Z;\upsilon\right)}\right]'\right].$$

To show the pathwise differentiability of μ we require an $F_\mu\left(Y,U,X,W\right)$ such that $\partial\mu/\partial\chi|_{\chi=\chi_0}=E\left[F_\mu\left(Y,U,X,W\right)s_\chi\right]|_{\chi=\chi_0}$, where s_χ is the score function generating the tangent space T_χ defined in (32). It can be verified that the function $F_\mu\left(Y,U,X,W\right)=\delta\varepsilon/\pi\left(Z\right)-X'\alpha_\lambda\left(U\right)-W'\beta-\mu$ satisfies this condition, and thus μ is pathwise differentiable. Next we show that the function

$$S_\mu^*=E\left(X|U\right)'\left[E\left(\delta X^{\otimes 2}|U\right)\right]^{-1}\delta X\varepsilon$$
$$+E\left(W-\alpha_{0WX}\left(U\right)'X\right)'\Gamma_0^{-1}\left(W-\alpha_{0XW}\left(U\right)'X\right)\delta\varepsilon$$
$$+\left(W+\alpha_0\left(U\right)'X-\mu\right) \tag{40}$$

lies in T_χ, and is the required projection. Note that by the efficiency result of Theorem 1 T_χ is equivalent to $T'_\chi = \left[C\delta\left(W - \alpha_{0XW}(U)' X\right)\varepsilon, \delta g(U) X\varepsilon, ...\right]$ where $g(\cdot)$ is an arbitrary function of U and the other terms are as those given in (32). Then with $C = E\left(W - \alpha_{0XW}(U)' X\right)' \Gamma_0^{-1}$, $g(U) = E(X|U)'\left[E\left(\delta X^{\otimes 2}|U\right)\right]^{-1} - C$ $\alpha_{0XW}(U)$, a zero for S_σ and S_{v_1}, and the last summand in (40) as $b(Z)$, it follows that S_μ^* lies T'_χ. To verify that S_μ^* is the required projection we show that

$$E\left[\left(F_\mu(Y, U, X, W) - S_\mu^*\right) t'_\chi\right] = 0 \quad \forall t'_\chi \in T'_\chi. \tag{41}$$

Note that

$$\begin{aligned} F_\mu(Y, U, X, W) - S_\mu^* = \frac{\delta\varepsilon}{\pi(Z)} - \Bigg[& E(X|U)'\left[E\left(\delta X^{\otimes 2}|U\right)\right]^{-1} X \\ & + E\left(W - \alpha_{0XW}(U)' X\right)' \Gamma_0^{-1} \times \left(W - \alpha_{0XW}(U)' X\right)\Bigg]\delta\varepsilon, \end{aligned}$$

and that

$$\begin{aligned} E\Bigg[& \frac{\delta C\varepsilon^2\left(W - \alpha_{0XW}(U)' X\right)}{\pi(Z)} - E(X|U)'\left[E\left(\delta X^{\otimes 2}|U\right)\right]^{-1} \\ & \times X\left(W - \alpha_{0XW}(U)' X\right)' C'\delta\varepsilon^2 \\ & - E\left(W - \alpha_{0XW}(U)' X\right)' \Gamma_0^{-1}\left(W - \alpha_{0XW}(U)' X\right)^{\otimes 2} C'\delta\varepsilon^2 + \frac{\delta g(U) X\varepsilon^2}{\pi(Z)} \\ & - E(X|U)'\left[E\left(\delta X^{\otimes 2}|U\right)\right]^{-1} XX'g(U)'\delta\varepsilon^2 \\ & - E\left(W - \alpha_{0XW}(U)' X\right)' \Gamma_0^{-1}\left(W - \alpha_{0XW}(U)' X\right) X'g(U)'\delta\varepsilon^2\Bigg] = 0 \end{aligned}$$

since

$$\begin{aligned} & \sigma^2 E\Bigg[\frac{\delta C\left(W - \alpha_{0XW}(U)' X\right)}{\pi(Z)} - E\left(W - \alpha_{0XW}(U)' X\right)' \Gamma_0^{-1} \\ & \qquad \left(W - \alpha_{0XW}(U)' X\right)^{\otimes 2} C'\delta\Bigg] = 0, \\ & \sigma^2 E\left[E(X|U)'\left[E\left(\delta X^{\otimes 2}|U\right)\right]^{-1} X\left(W - \alpha_{0XW}(U)' X\right)' C'\delta\right] = 0, \\ & \sigma^2 E\left[\frac{\delta g(U) X\varepsilon^2}{\pi(Z)} - E(X|U)'\left[E\left(\delta X^{\otimes 2}|U\right)\right]^{-1} Xg(U) X\delta\right] \\ & \quad = \sigma^2 E\left[E(X|U)'\left\{\left[E\left(\delta X^{\otimes 2}|U\right)\right]^{-1} E(X|U) - \alpha_{0XW}(U)' C'\right\}\right. \end{aligned}$$

$$- E(X|U)' \left\{ \left[E\left(\delta X^{\otimes 2}|U\right) \right]^{-1} E(X|U) - \alpha_{0XW}(U)' C' \right\} \right] = 0,$$

$$\sigma^2 E\left[E\left(W - \alpha_{0XW}(U)' X\right)' \Gamma_0^{-1} \left(W - \alpha_{0XW}(U)' X\right) X' g(U)' \delta \right] = 0$$

by the MAR assumption and iterated expectations. It is easy to show the orthogonality of $F_\mu(Y, U, X, W) - S_\mu^*$ with respect to the other components of T'_χ, which implies that (41) is verified so that S_μ^* is the efficient score and the conclusion follows immediately. □

Proof of Theorem 10 We first show the result for $\sigma^2_{2(ho)}$. Note that

$$\sigma^2_{2(ho)} - \sigma^2_{eff} = \sigma^2 E\left(\frac{\delta}{\pi(Z)} - E(X|U)' E\left(\delta X^{\otimes 2}|U\right)^{-1} E(X|U) - \delta\left(W - \alpha_{0XW}(U)' X\right)' \Gamma_0^{-1} E\left(W - \alpha_{0XW}(U)' X\right) \right), \quad (42)$$

and

$$\mathrm{Cov}\left(\frac{\delta\varepsilon}{\pi(Z)} - E(X|U)' E\left(\delta X^{\otimes 2}|U\right)^{-1} X\delta\varepsilon, \left(W - \alpha_{0XW}(U)' X\right)\delta\varepsilon \right) = \sigma^2 E\left(W - \alpha_{0XW}(U)' X\right),$$

so that (42) is equivalent to

$$\mathrm{Var}\left(\frac{\delta}{\pi(Z)} - E(X|U)' E\left(\delta X^{\otimes 2}|U\right)^{-1} E(X|U) \right) - \sigma^2 E\left[\left(W - \alpha_{0XW}(U)' X\right)\right]' \Gamma_0^{-1} \times E\left(W - \alpha_{0XW}(U)' X\right) := V_{12} - V_{22} \geq 0.$$

Similarly for $\sigma^2_{1(ho)}$ we have

$$\sigma^2_{1(ho)} - \sigma^2_{eff} = \sigma^2 E\left(\delta - E(X|U)' E\left(\delta X^{\otimes 2}|U\right)^{-1} E(X|U) - \delta\left(W - \alpha_{0XW}(U)' X\right)' \Gamma_0^{-1} E\left(W - \alpha_{0XW}(U)' X\right) \right), \quad (43)$$

and

$$\mathrm{Cov}\left(\delta\varepsilon - E(X|U)' E\left(\delta X^{\otimes 2}|U\right)^{-1} X\delta\varepsilon, \left(W - \alpha_{0XW}(U)' X\right)\delta\varepsilon \right) = \sigma^2 E\left[\delta\left(W - \alpha_{0XW}(U)' X\right)\right]$$

so that (43) is equivalent to

$$\mathrm{Var}\left(\delta - E(X|U)' E\left(\delta X^{\otimes 2}|U\right)^{-1} E(X|U) \right)$$

$$
-\sigma^2 E\left[\delta\left(W-\alpha_{0XW}(U)'X\right)\right]'\left(\Gamma_{\delta 0}^{-1}\right)\times E\left[\delta\left(W-\alpha_{0XW}(U)'X\right)\right]
$$
$$
:= V_{11}-V_{21}\geq 0,
$$

where $\Gamma_{\delta 0}=E\left[\delta\left(W-\alpha\alpha_{0XW}(U)'X\right)^{\otimes 2}\right]$. Finally for $\sigma^2_{3(ho)}$

$$
\sigma^2_{3(ho)}-\sigma^2_{eff}=\sigma^2 E\left(\frac{\delta}{\pi(U)}-\frac{E(X|U)'}{\pi(U)}E\left(\delta X^{\otimes 2}|U\right)^{-1}E(X|U)\right.
$$
$$
\left.-\frac{\delta}{\pi(U)}\left(W-\alpha_{0XW}(U)'X\right)'\Gamma_0^{-1}E\left(W-\alpha_{0XW}(U)'X\right)\right),
$$

and by the same argument as the one used for (42) and (43) we have that

$$
\sigma^2_{3(ho)}-\sigma^2_{eff}=\mathrm{Var}\left(\frac{\delta}{\pi(U)}-\frac{E(X|U)'}{\pi(U)}E\left(\delta X^{\otimes 2}|U\right)^{-1}E(X|U)\right)
$$
$$
-\sigma^2 E\left[\frac{\delta}{\pi(U)^2}\left(W-\alpha_{0XW}(U)'X\right)\right]'\left(\Gamma^{-1}_{\pi^2 0}\right)
$$
$$
\times E\left[\frac{\delta}{\pi(U)^2}\left(W-\alpha_{0XW}(U)'X\right)\right]:=V_{13}-V_{23}\geq 0,
$$

where $\Gamma_{\pi^2 0}=E\left[\delta\left(W-\alpha_{0XW}(U)'X\right)^{\otimes 2}/\pi^2(U)\right]$. □

Proof of Theorem 11 Similar to the proof of Theorem 8

$$
n^{1/2}\left(\widehat{\tau}_1-\tau_0\right)
$$
$$
=\frac{1}{n^{1/2}}\sum_{i=1}^n\left(X_i'\left[\alpha_0^{(1)}(U_i)-\alpha_0^{(0)}(U_i)\right]+W_i'\left[\beta_0^{(1)}-\beta_0^{(0)}\right]-\tau_0\right)
$$
$$
+\frac{1}{n^{1/2}}\sum_{i=1}^n\left\{\delta_i\varepsilon_i^{(1)}+(1-\delta_i)\left[X_i'\left(\widehat{\alpha}^{(1)}(U_i)-\alpha_0^{(1)}(U_i)\right)+W_i'\left(\widehat{\beta}^{(1)}-\beta_0^{(1)}\right)\right]\right\}
$$
$$
-\frac{1}{n^{1/2}}\sum_{i=1}^n\left\{(1-\delta_i)\varepsilon_i^{(0)}-\delta_i\left[X_i'\left(\widehat{\alpha}^{(0)}(U_i)-\alpha_0^{(0)}(U_i)\right)+W_i'\left(\widehat{\beta}^{(0)}-\beta_0^{(0)}\right)\right]\right\}
$$
$$
:=\sum_{j=1}^3 T_{2j},
$$

and as in (37)

$$
T_{22}=\frac{1}{n^{1/2}}\sum_{i=1}^n\left[\Omega_{01}^{(1)}(U_i)+\Delta_{01}^{(1)}\Gamma_0^{(1)-1}\left(W_i-\alpha_{0XW}^{(1)}(U_i)'X_i\right)\right]\delta_i\varepsilon_i^{(1)}+o_p(1),
$$
$$
T_{23}=\frac{1}{n^{1/2}}\sum_{i=1}^n\left[\Omega_{01}^{(0)}(U_i)+\Delta_{01}^{(0)}\Gamma_0^{(1)-1}\left(W_i-\alpha_{0XW}^{(0)}(U_i)'X_i\right)\right](1-\delta_i)\varepsilon_i^{(0)}+o_p(1),
$$

and the first conclusion follows by CLT noting that $Cov\left(T_{21}, T_{2j}\right) = 0$ for $j = 2, 3$ and $Cov\left(T_{22}, T_{23}\right) = 0$. For the second result note that

$$
\begin{aligned}
n^{1/2}\left(\widehat{\tau}_2 - \tau_0\right) &= \frac{1}{n^{1/2}}\sum_{i=1}^{n}\left[X_i'\left(\alpha_0^{(1)}\left(U_i\right) - \alpha_0^{(0)}\left(U_i\right)\right) + W_i'\left(\beta_0^{(1)} - \beta_0^{(0)}\right) - \tau_0\right] \\
&+\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left\{\frac{\delta_i\left(X_i'\widehat{\alpha}^{(1)}\left(U_i\right) + W_i'\widehat{\beta}^{(1)}\right)}{\widehat{\pi}\left(Z_i\right)} - \left(X_i'\alpha_0^{(1)}\left(U_i\right) + W_i'\beta_0^{(1)}\right)\right\} \\
&+\frac{1}{n^{1/2}}\sum\nolimits_{i=1}^{n}\left\{\frac{\left(1-\delta_i\right)\left(X_i'\widehat{\alpha}^{(0)}\left(U_i\right) + W_i'\widehat{\beta}^{(0)}\right)}{1-\widehat{\pi}\left(Z_i\right)} - \left(X_i'\alpha_0^{(0)}\left(U_i\right) + W_i'\beta_0^{(0)}\right)\right\} \\
&:= \sum_{j=1}^{3} T_{3j}.
\end{aligned}
$$

The second term can be written as

$$
\begin{aligned}
T_{32} &= \frac{1}{n^{1/2}}\sum_{i=1}^{n}\left(\frac{\delta_i}{\pi\left(Z_i\right)} - 1\right)\left(X_i'\alpha_0^{(0)}\left(U_i\right) + W_i'\beta_0^{(0)}\right) \\
&+\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left(\frac{\delta_i}{\widehat{\pi}\left(Z_i\right)} - \frac{\delta_i}{\pi\left(Z_i\right)}\right)x \\
&\times\left[X_i'\left(\widehat{\alpha}^{(1)}\left(U_i\right) - \alpha_0^{(1)}\left(U_i\right)\right) + W_i'\left(\widehat{\beta}^{(1)} - \beta_0^{(1)}\right)\right] \\
&+\frac{1}{n^{1/2}}\sum_{i=1}^{n}\frac{\delta_i\left[X_i'\left(\widehat{\alpha}^{(1)}\left(U_i\right) - \alpha_0^{(1)}\left(U_i\right)\right) + W_i'\left(\widehat{\beta}^{(1)} - \beta_0^{(1)}\right)\right]}{\pi\left(Z_i\right)} \\
&+\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left(\frac{\delta_i}{\widehat{\pi}\left(Z_i\right)} - \frac{\delta_i}{\pi\left(Z_i\right)}\right)\left(X_i'\alpha_0^{(1)}\left(U_i\right) + W_i'\beta_0^{(1)}\right) \\
&:= \sum_{j=1}^{4} T_{32j},
\end{aligned}
$$

and note that by (38), (36) and standard kernel calculations

$$
\begin{aligned}
T_{322} &= \frac{1}{n^{3/2}}\sum_{i=1}^{n}\frac{\delta_i\sum_{j=1}^{n}\left[\pi\left(Z_i\right) - \delta_j\right]K_h\left(Z_j - Z_i\right)}{\pi\left(Z_i\right)^2 f\left(Z_i\right)} \\
&\times\left[E\left(X|U_i\right)'\left[E\left(\delta X^{\otimes 2}|U_i\right)\right]^{-1}\delta_i X_i\varepsilon_i^{(1)} + \left(W_i' + X_i'\alpha_{0XW}\left(U_i\right)\right)\left(\widehat{\beta}^{(1)} - \beta_0^{(1)}\right)\right] \\
&+ o_p(1).
\end{aligned}
$$

By iterated expectation $E\left(T_{322}\right)=0$ and

$$
\begin{aligned}
& E\left|T_{322}\right|^2 \\
& \leq \frac{2}{n^3 h^{2(k+p+1)}}\left(\sum_{i=1}^n \sum_{j=1}^n E\left\{\frac{\delta_i^2\left[E\left(X \mid U_i\right)^{\prime}\left[E\left(\delta X^{\otimes 2} \mid U_i\right)\right]^{-1} \delta_i X_i \varepsilon_i^{(1)}\right]}{\pi\left(Z_i\right)^2 f\left(Z_i\right)}\right\}^2\right. \\
& \left.+\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n E\left\{\frac{\left(W_i^{\prime}+X_i^{\prime} \alpha_{0XW}\left(U_i\right)\right)\left\|n^{1/2}\left(\widehat{\beta}^{(1)}-\beta_0^{(1)}\right)\right\|}{\pi\left(Z_i\right)^2 f\left(Z_i\right)}\right\}\right)^2 K^2\left(Z_j-Z_i\right) \\
& =\frac{C}{n h^{k+p+1}}+o(1),
\end{aligned}
$$

and hence $\left|T_{322}\right|=o_p(1)$. Similarly for T_{324}

$$
\begin{aligned}
T_{324}= & \frac{1}{n^{3/2}} \sum_{i=1}^n \frac{\delta_i\left(X_i^{\prime} \alpha_0^{(1)}\left(U_i\right)+W_i^{\prime} \beta_0^{(1)}\right) \sum_{j=1}^n\left[\pi\left(Z_i\right)-\delta_j\right] K_h\left(Z_j-Z_i\right)}{\pi\left(Z_i\right)^2 f\left(Z_i\right)} \\
& +o_p(1) \\
= & \frac{1}{n} \sum_{i=1}^n \frac{\delta_i\left(X_i^{\prime} \alpha_0^{(1)}\left(U_i\right)+W_i^{\prime} \beta_0^{(1)}\right) K_h\left(Z_j-Z_i\right)}{\pi\left(Z_i\right) f\left(Z_i\right)} \frac{1}{n^{1/2}} \sum_{j=1}^n \frac{\pi\left(Z_i\right)-\delta_i}{\pi\left(Z_i\right)} \\
= & -T_{321}+o_p(1),
\end{aligned}
$$

hence

$$
T_{32}=\frac{1}{n^{1/2}} \sum_{i=1}^n \frac{\delta_i\left[X_i^{\prime}\left(\widehat{\alpha}^{(1)}\left(U_i\right)-\alpha_0^{(1)}\left(U_i\right)\right)+W_i^{\prime}\left(\widehat{\beta}^{(1)}-\beta_0^{(1)}\right)\right]}{\pi\left(Z_i\right)}+o_p(1).
$$

The same arguments show that

$$
\begin{aligned}
T_{33n}= & \frac{1}{n^{1/2}} \sum_{i=1}^n \frac{\left(1-\delta_i\right)\left[X_i^{\prime}\left(\widehat{\alpha}^{(0)}\left(U_i\right)-\alpha_0^{(0)}\left(U_i\right)\right)+W_i^{\prime}\left(\widehat{\beta}^{(0)}-\beta_0^{(0)}\right)\right]}{1-\pi\left(Z_i\right)} \\
& +o_p(1),
\end{aligned}
$$

hence the second conclusion follows using (39) and CLT. □

Acknowledgements I am grateful to an Associate Editor and two referees for useful comments and suggestions that improved the original draft. I am also grateful to Juan Carlos Escanciano and Ursula Müller for some useful conversations. All remaining errors are my own.

References

Ahmad, I., Leelahanon, S., Li, Q. (2005). Efficient estimation of a semiparametric partially linear varying coefficient model. *Annals of Statistics*, *33*, 258–283.
Bang, H., Robins, J. (2005). Doubly robust esimation in missing data and casual inference models. *Biometrics*, *61*, 962–972.
Bickel, P., Freedman, D. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics*, *9*, 1196–1297.
Bickel, P., Klaassen, C., Ritov, Y., Wellner, J. (1993). *Efficient and adaptive estimation for semiparametric models*. Baltimore: John Hopkins University.
Bierens, H. (1982). Consistent model specification tests. *Journal of Econometrics*, *20*, 105–134.
Bravo, F. (2012). Generalized empirical likelihood testing in semiparametric conditional moment restriction models. *Econometrics Journal*, *15*, 1–31.
Bravo, F., Jacho-Chavez, D. (2011). Empirical likelihood for efficient semiparametric average treatment effects. *Econometric Reviews*, *30*, 1–24.
Cai, J., Fan, J., Jiang, J., Zhou, H. (2008). Partially linear hazard regression with varying coefficients for multivariate survival data. *Journal of the Royal Statistical Society B*, pp. 141–158.
Cai, Z., Fan, J., Li, R. (2000a). Efficient estimation and inference for varying-coefficient models. *Journal of the American Statistical Association*, *95*, 888–902.
Cai, Z., Fan, J., Yao, Q. (2000b). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association*, *95*, 941–956.
Chen, S., Cui, H. (2006). On Bartlett correction of empirical likelihood in presence of nuisance parameters. *Biometrika*, *93*, 215–220.
Cheng, P. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association*, *89*, 81–87.
Copeland, B., Taylor, M. (2004). Trade, growth and enviroment. *Journal of Economic Literature*, *42*, 7–71.
Delgado, M., Dominguez, M., Lavergne, P. (2003). Consistent tests of conditional moment restrictions. *Annales de Economie et de Statistique*, *81*, 33–67.
Engle, R., Granger, C., Rice, J., Weiss, A. (1986). Nonparametric estimation of the relation between weather and electricity sales. *Journal of the American Statistical Association*, *81*, 310–320.
Escanciano, J. (2006). On consistent diagnostic tests for regression models using projections. *Econometric Theory*, *22*, 1030–1051.
Escanciano, J., Velasco, C. (2010). Specification tests of parametric dynamic conditional quantiles. *Journal of Econometrics*, *159*, 209–221.
Fan, J., Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, *11*, 1031–1057.
Fan, J., Wu, Y. (2008). Semiparametric estimation of covariance matrices for longitudinal data. *Journal of the American Statistical Association*, *103*, 1520–1533.
Fan, J., Zhang, W. (2008). Statistical methods with varying coefficients models. *Statistics and Its Interface*, *1*, 179–195.
Frankel, J., Rose, A. (2005). Is trade good or bad for the eniviroment? sorting out the causality. *Review of Economics and Statistics*, *87*, 85–91.
Gonzalez-Manteiga, W., Perez-Gonzalez, A. (2006). Goodness-of-fit tests for linear regression models with missing response at random. *Canadian Journal of Statistics*, *34*, 149–170.
Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, *66*(2), 315–331.
Hardle, W., Mammen, E. (1993). Comparing nonparametric versus parametric fits. *Annals of Statistics*, *21*, 1926–1947.
Hastie, T., Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society B*, *55*, 757–796.
He, X., Zhu, L. (2003). A lack-of-fit test for quantile regression. *Journal of the American Statistical Association*, *98*, 1013–1022.
Hirano, K., Imbens, G., Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*, 1161–1189.
Horvitz, D., Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*, 663–685.

Kim, M., Ma, Y. (2012). The efficiency of the second order nonlinear least squares estimator and its extension. *Annals of the Institute of Statistical Mathematics*, *64*, 751–764.
Koul, H., Muller, U., Schick, A. (2012) The transfer principle: A tool for complete case analysis. Unpublished manuscript.
Little, R., Rubin, D. (2002). *Statistical analysis with missing data*. New Jersey, USA: Wiley.
Masry, E. (1996). Multivariate regression estimation: Local polynomial fitting for time series. *Stochastic Processess and their Applications*, *65*, 81–101.
Millimet, D., Tchernis, R. (2009). On the specification of propensity scores, with application to the analysis of trade policies. *Journal of Economic and Business Statistics*, *27*, 397–415.
Muller, U. (2009). Estimating linear functionals in nonlinear regression with responses missing at random. *Annals of Statistics*, *37*, 2245–2277.
Paik, M. (1997). The generalized estimating equation approach when data are not completely missing at random. *Journal of the American Statistical Association*, *92*, 1320–1329.
Pollard, D. (1984). *Convergence of stochastic processes*. New York: Springer and Verlag.
Robins, J., Rotnitzky, A., Zhao, L. (1994). Estimation of regression coefficients when ssome regressors are not always observed. *Journal of the American Statistical Association*, *89*, 846–866.
Robinson, P. (1987). Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica*, *55*, 875–891.
Robinson, P. (1988). Root-n consistent semiparametric regression. *Econometrica*, *56*, 931–954.
Rosenbaum, P., Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.
Scharfstein, D., Rotnizky, A., Robins, J. (1999). Adjusting for nonignorable drop-out using semiparameteric nonresponse models. *Journnal of the American Statistical Association*, *94*, 1096–1120.
Speckman, H. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society B*, *50*, 413–436.
Stute, W. (1997). Nonparametric model checks for regressions. *Annals of Statistics*, *25*, 613–641.
Stute, W., Zhu, L. (2002). Model checks for generalized linear models. *Scandinavian Journal of Statistics*, *29*, 535–545.
Stute, W., Zhu, L. (2005). Nonparametric checks for single index models. *Annals of Statistics*, *33*, 1048–1083.
Stute, W., Gonzalez-Mantiega, W., Quindimil, M. P. (1998). Bootstrap approximations in model checks for regression. *Journal of the American Statistical Association*, *93*, 141–149.
Su, J., Wei, L. (1991). A lack of fit test for the mean function in a generalized linear model. *Journal of the American Statistical Association*, *86*, 420–426.
Sun, Z., Wang, Q., Dai, P. (2009). Model checking for partially linear models with missing responses at random. *Journal of Multivariate Analysis*, *100*, 636–651.
Van der Vaart, A., Wellner, J. (1996). *Weak convergence and empirical processes*. New York: Springer.
Wang, D., Chen, S. (2009). Empirical likelihood for estimating equation with missing values. *Annals of Statistics*, *37*, 490–517.
Wang, L., LeBlanc, A. (2008). Second-order nonlinear least squares estimation. *Annals of the Institute of Statistical Mathematics*, *60*, 883–900.
Wang, Q., Rao, J. (2002). Empirical likelihood-based inference under imputation for missing response data. *Annals of Statistcs*, *30*, 896–924.
Wang, Q., Linton, O., Hardle, W. (2004). Semiparametric regression analysis with missing response at random. *Journal of the American Statistical Association*, *99*, 334–345.
Whang, Y. (2000). Consistent bootstrap tests of parametric regression functions. *Journal of Econometrics*, *98*, 27–46.
Whang, Y. (2001). Consistent specification testing for conditional moment restrictions. *Economics Letters*, *71*, 299–306.
Wu, C. (1986). Jackknife, bootstrap and other resampling methods. *Annals of Statistics*, *14*, 1261–1295.
Xia, Y., Li, W., Tong, H., Zhang, D. (2004). A goodness-of-fit test for single-index models. *Statistica Sinica*, *14*, 1–28.
Zhu, L., Ng, K. (2003). Checking the adequacy of a partial linear model. *Statistica Sinica*, *13*, 763–781.