

# Nonparametric Regression with Selectively Missing Covariates\*

CHRISTOPH BREUNIG<sup>†</sup>

*Emory University*

PETER HAAN<sup>‡</sup>

*DIW Berlin and Freie Universität Berlin*

October 13, 2020

We consider the problem of regression with selectively observed covariates in a nonparametric framework. Our approach relies on instrumental variables that explain variation in the latent covariates but have no direct effect on selection. The regression function of interest is shown to be a weighted version of observed conditional expectation where the weighting function is a fraction of selection probabilities. Nonparametric identification of the fractional probability weight (FPW) function is achieved via a partial completeness assumption. We provide primitive functional form assumptions for partial completeness to hold. The identification result is constructive for the FPW series estimator. We derive the rate of convergence and also the pointwise asymptotic distribution. In both cases, the asymptotic performance of the FPW series estimator does not suffer from the inverse problem which derives from the nonparametric instrumental variable approach. In a Monte Carlo study, we analyze the finite sample properties of our estimator and we compare our approach to inverse probability weighting, which can be used alternatively for unconditional moment estimation. In the empirical application, we focus on two different applications. We estimate the association between income and health using linked data from the SHARE survey and administrative pension information and use pension entitlements as an instrument. In the second application we revisit the question how income affects the demand for housing based on data from the German Socio-Economic Panel Study (SOEP). In this application we use regional income information on the residential block level as an instrument. In both applications we show that income is selectively missing and we demonstrate that standard methods that do not account for the nonrandom selection process lead to significantly biased estimates for individuals with low income.

---

\*The authors gratefully thank the Co-Editor Jianqing Fan and two anonymous referees for their many constructive comments on the previous version of the paper. Support by German Science Foundation through CRC TRR 190 is gratefully acknowledged.

<sup>†</sup>Department of Economics, Emory University, Rich Memorial Building, Atlanta, GA 30322, USA, e-mail: [christoph.breunig@emory.edu](mailto:christoph.breunig@emory.edu)

<sup>‡</sup>DIW Berlin and Freie Universität Berlin, Mohrenstr. 58, 10117 Berlin, Germany, e-mail: [phaan@diw.de](mailto:phaan@diw.de)

*Keywords:* Selection model, instrumental variables, fractional probability weighting, nonparametric identification, partial completeness, incomplete data, series estimation, income distribution, health, housing.

*JEL classification:* C14, C26

## 1. Introduction

Sample selection is a central challenge for empirical evaluation studies. Nonrandom selection can affect the empirical analysis in many ways, for example through nonrandom selection into treatment programs, selective measurement error or through selective nonresponse or missingness of data. In this paper, we propose an instrumental variable approach to address the problem of nonrandom selection. We provide constructive nonparametric identification results and build on those to establish a novel fractional probability weighting estimator.

While the methodology is general and applicable to many situations in which selection might be problematic the leading example in this paper will be selective nonresponse and selective missing data. We are interested in the identification and estimation of the nonparametric regression function  $g(x) = \mathbb{E}[Y|X^* = x]$  where  $Y$  is always observed but  $X^*$  is only selectively observed. In this case, parts of the information of the covariates are missing not at random for some sampling units. Without accounting for selectivity of responses, statements about individual behavior based on such incomplete data might be severely biased.

In this paper, we establish identification of the nonparametric regression function  $g(x) = \mathbb{E}[Y|X^* = x]$  based on instrumental variables that explain variation in the latent covariates but have no direct effect on selection. Such an instrumental variable approach is well suited when selection is driven by the latent variables  $X^*$ . We show that the regression function  $g$  can be written as a weighted version of its observed counterpart. The weighting function is determined by a fraction of selection probabilities, i.e., fractional probability weights (FPW), that depends on latent variables. We propose a novel identification restriction, the so called *partial completeness assumption*, which implies identification of the FPW and thus of the nonparametric regression function  $g$ . In contrast to usual completeness assumptions required for identification of nonparametric instrumental variable models, we are able to provide primitive, functional form conditions for the partial completeness assumption to hold. We emphasize that these functional form conditions do not imply (semi-)parametric restrictions but only impose a nonparametric structure in different forms of separability of  $Y$  and  $X^*$ . Specifically, we show that a nonparametric generalized additive structure of the selection probability is sufficient to obtain identification of the function  $g$ .

Based on the constructive nonparametric identification result we propose a novel nonparametric FPW series estimator that is convenient for implementation. We show that our estimator has a rate of convergence that coincides with usual nonparametric regression estimators, i.e., the asymptotic performance of the estimator is the same as of an estimator with full information of the underlying selection mechanism. We

establish asymptotic normality of the estimator and show that the asymptotic variance is not necessarily enlarged by FPW estimation. We also propose a bootstrap procedure to construct uniform confidence bands. A Monte Carlo simulation study demonstrates the improvements of our approach over missing at random (MAR) estimators. In particular, we highlight our contribution also in a finite sample analysis of linear regression with alternative inverse probability weighting estimators (IPW).

Finally, we use the method in two different empirical applications. Both applications are important for the discussion about income inequality and highly relevant for public policy. First, we use the developed methodology to analyze the association between income and the risk of bad health. The empirical analysis is based on linked data from the Survey of Health, Aging and Retirement in Europe (SHARE). We exploit a specific feature of the data which allows us to link a sub-sample of the survey data to administrative data of the German pension insurance.<sup>1</sup> We find that income in the SHARE data is selectively missing and we demonstrate that standard methods that do not account for the nonrandom selection process are biased, specifically for individuals with low incomes. Assuming linearity of the regression function  $g$  the point estimate of income is significantly negative when imposing MAR however it is not significantly different from zero when accounting for nonrandom nonresponse. In the second example we analyze how housing varies with income. We quantify the relationship between labor earnings and the probability to own a house. For this empirical analysis we use data from the German Socio Economic Panel (SOEP) and exploit information about the regional purchasing power collected on the residential block level (Sub-Zip code level) as an instrument. Again we find that earnings are selectively missing and we demonstrate that the estimates derived in standard methods are biased for individuals with earnings in the lowest decile which is a specifically relevant group for public policy. For individuals with higher earnings the estimates from the different methods do not differ significantly.

Our paper is linked to several strands of the literature. The most common way to deal with missing data is to assume missing at random pioneered by Little and Rubin [2002]. In the context of selectively missing covariates, a sieve semiparametric maximum likelihood estimator was proposed by Chen et al. [2007]. In contrast, an instrumental variable strategy, as proposed in the paper, was used so far only to deal with endogenous missingness of dependent variables, see, for instance, Tang et al. [2003], Ramalho and Smith [2013], and D’Haultfoeuille [2010]. Also Breunig et al. [2018] consider the problem of nonparametric regression with selective nonresponse of the dependent variables; Zhao and Shao [2015] focus on a semiparametric approach. There only has been minor attention to selectively observed covariates. One example is Fang et al. [2018] who consider a semiparametric approach to deal with selectively missing covariates that is crucially different from ours. While Fang et al. [2018] require a parametric specification of the distribution of outcome given potential covariates, we leave these conditional distribution unrestricted. We establish nonparametric identification of the regression function and hence ensure that the identification is not due to specific functional form restrictions that might be violated in practice.

The paper adds as well to the literature on income inequality, more specifically to studies on the income gradient on health outcomes and mortality e.g. Preston [1975],

---

<sup>1</sup>Bingley and Martinello [2017] compare self reported information about income and education from SHARE with matched information from Danish administrative data in order to study the implications of measurement error.

Deaton and Paxson [1998], Cutler et al. [2006], or Cutler et al. [2011], and on the effect of income on housing and on housing demand, see e.g. Quigley and Raphael [2004], Albouy et al. [2016] or Dustmann et al. [2018]. In general these studies are based on survey data in which wealth, income, health and housing information and further demographic variables are self reported. As shown in Breunig [2019] information on income or earnings in surveys is likely to suffer from nonrandom selection which might result in biased estimates of the association between income and health or income and home ownership. In this respect this study extends the previous literature as we account for nonrandom nonresponse of the income information.

The paper is organized as follows. In Section 2 we establish identification of our nonparametric model. In Section 3 we derive the FPW series estimator, establish its rate of convergence and its asymptotic normality, and derive uniform bootstrap confidence bands. Section 4 provides Monte Carlo simulations and discusses implications of FPW estimation to unconditional moments. In Section 5 we apply our methodology to the two different empirical applications. All proofs can be found in Appendix A. Appendix B provides some technical results. Finally, Appendix C provides an extension when also the dependent variable is selectively missing.

## 2. Nonparametric Identification

This section consists of two subsections. In Subsection 2.1, we provide assumptions required for identification. In particular, we introduce a novel restriction, i.e., the partial completeness assumption, and provide primitive conditions for it. Subsection 2.2 establishes identification of the nonparametric regression function.

### 2.1. Setup and Main Assumptions

Given an observable outcome variable  $Y$  and latent covariates  $X^*$  our interest lies in the regression function  $g(x) = \mathbb{E}[Y|X^* = x]$ . Identification relies on instrumental variables  $W$  that explain variations of the latent variable  $X^*$  but are not directly related to the selection mechanism  $D$ . This is formalized in the following. Throughout the paper, we assume that a sample  $(D_1, Y_1, W_1), \dots, (D_n, Y_n, W_n)$  of  $(D, Y, W)$  is observed for each individual. A  $d_x$ -dimensional vector of covariates  $X^*$  is only fully observed depending on a binary indicator variable  $D$ , i.e.,  $X^*$  is observed when  $D = 1$  and missing when  $D = 0$ . We write  $X = DX^*$ .<sup>2</sup> Under the assumptions presented below we see that the selection probability conditional on  $(Y, X^*)$ , i.e.,  $\mathbb{P}(D = 1|Y, X^*)$ , is only partially identified but still point identification of the regression function  $g$  is established.

**Assumption 1** (Exclusion Restriction). *It holds that*

$$\mathbb{P}(D = 1|Y, X^*, W) = \mathbb{P}(D = 1|Y, X^*).$$

Assumption 1 states an exclusion restriction of the random vector  $W$ . It excludes any relation between  $W$  and the selection mechanism  $D$  that is not channeled through

---

<sup>2</sup>The situation can be easily extended to a multivariate version where  $D$  denotes a  $d_x$ -dimensional vector of missing data indicators. In order to keep the notation simple we do not treat this case explicitly.

$(Y, X^*)$ . The setting corresponds to the measurement error set up, where instrumental variables are required to drive the latent, true variable but not the variable that is observed with error. However, identification with nonclassical measurement error requires an additional exclusion restriction which restricts  $W$  to have no information on  $Y$  that is not captured in  $X^*$ , see Assumption 2 (i) in Hu and Schennach [2008]. Interestingly, nonrandom selection as extreme form of nonclassical measurement error simplifies the exclusion restriction imposed on the instruments.

We also emphasize that Assumption 1 allows for dependence of  $D$  and  $Y$ . Thus, our approach captures selection on unobservables that do not only stem from latent characteristics in  $X^*$  but also from unobservables that are unexplained by the regression function  $g(X^*)$ .<sup>3</sup> This is an important feature of our framework, as in many economic environments, selection variables can be driven by unobserved individual characteristics. Related literature on nonrandom nonresponse of covariates does not allow for such a general selection mechanism, see Zhao and Shao [2015]. We introduce the function class  $\mathcal{B} = \{\phi : \mathbb{E}|\phi(Y, X^*)| < \infty \text{ and } \inf_{y,x} \phi(y, x) > 0\}$ .

**Assumption 2** (Partial Completeness). *For all  $\phi \in \mathcal{B}$  it holds:  $\mathbb{E}[\phi(Y, X^*)|Y, W] = 0$  implies that  $\phi$  does not depend on  $Y$ .*

Assumption 2 is less restrictive than the usual completeness assumption which assumes that  $\mathbb{E}[\phi(Y, X^*)|Y, W] = 0$  implies  $\phi(Y, X^*) = 0$ . This assumption is commonly imposed to ensure identification in nonparametric instrumental variable models, see for instance Newey and Powell [2003]. In the context of endogenous selection such completeness assumptions were considered by D'Haultfoeuille [2010] and Breunig et al. [2018]. On the other hand, the partial completeness assumption holds under mild functional form assumptions as shown below.

Assumption 2 is automatically satisfied if  $\phi$  does not depend on  $Y$ . Indeed, if the selection probability  $\mathbb{P}(D = 1|Y, X^*)$  does not depend on  $Y$  the regression function  $g$  is identified as we see in the next subsection and thus, the partial completeness assumption is well suited for our particular selection problem. Moreover, the next result provides functional form restriction under which partial completeness holds. Throughout the paper,  $f_V$  denotes the probability density function of a random variable  $V$ .

**Proposition 2.1.** *Assume that  $f_{X^*|YW} = f_{X^*|W}$ . Assume that for any  $\phi \in \mathcal{B}$  there exist functions  $\phi_1$  and  $\phi_2$  such that either*

$$\phi(Y, X^*) = \phi_1(Y)\phi_2(X^*) - 1 \tag{2.1}$$

or

$$\phi(Y, X^*) = \psi(\phi_1(Y) + \phi_2(X^*)) - 1 \tag{2.2}$$

where  $\psi$  is an analytic function with  $\mathbb{E}[\psi^{(j)}(\phi_1(y) + \phi_2(X^*))|W] \neq 0$  for some  $j \geq 1$  and  $\mathbb{E}[\sum_{j \geq 1} |\psi^{(j)}(\phi_1(y') + \phi_2(X^*))(\phi_1(y) - \phi_1(y'))^j / (j!)|] < \infty$  for all  $y, y'$  in the support of  $Y$ . Then, Assumption 2 is satisfied.

---

<sup>3</sup>In the model  $Y = g(X^*) + U$ , not only  $X^*$  but also unobservables  $U$  are allowed to directly affect the selection mechanism  $D$ .

Proposition 2.1 requires that  $Y$  does not provide information on  $X^*$  that is not contained in the vector  $W$ . Given this mild restriction we see from Proposition 2.1 that functional form restrictions imply the partial completeness assumption to hold. We emphasize that these functional form restrictions do not imply (semi-)parametric specifications but only impose a nonparametric structure in different forms of separability of  $Y$  and  $X^*$ . Note that we subtract by one as the exclusion restriction in Assumption 1 implies the conditional mean restriction  $\mathbb{E}[D/\mathbb{P}(D = 1|Y, X^*) - 1|Y, W] = 0$ . The selection probability  $\mathbb{P}(D = 1|Y, X^*)$  is not necessarily point identified through the former conditional mean restriction given Assumption 2. Equation (2.2) provides a generalized additive type restriction on the functions of interest. The restriction  $\mathbb{E}[\psi^{(j)}(\phi_1(y) + \phi_2(X^*))|W] \neq 0$ , for some integer  $j \geq 1$  and all  $y$ , is satisfied by a broad class of link functions  $\psi$ . This condition holds, in particular, when the first derivative  $\psi'(\phi_1(y) + \phi_2(X^*))$  is strictly positive for all  $y$  in the support of  $Y$ . Consequently, we conclude from Proposition 2.1 that partial completeness holds under a generalized additive restriction when the link function  $\psi$  coincides with any analytic, cumulative distribution function supported on the whole real line. Another case where  $\mathbb{E}[\psi^{(j)}(\phi_1(y) + \phi_2(X^*))|W] \neq 0$ , for some  $j \geq 1$ , holds is when  $\psi$  is a non-constant polynomial function.

While Proposition 2.1 provides a broad class of functions which satisfy the partial completeness assumption, partial completeness can fail for functions which vary in  $y$  with vanishing means and when instruments are independent of  $X^*$ . This is demonstrated in the following illustrative example.

**Example 2.1.** Consider the case where  $\text{Var}(X^*) = 1$ ,  $\mathbb{E}[X^*] = 0$ , and, as in Proposition 2.1,  $f_{X^*|YW} = f_{X^*|W}$ . In addition, we assume that the instrumental variables  $W$  have no information on  $X^*$ , i.e.,  $f_{X^*|W} = f_{X^*}$ . Consider the function  $\phi(y, x) = x^2 - yx - 1$ , then we obtain:

$$\begin{aligned} \mathbb{E}[\phi(Y, X^*)|Y = y, W = w] &= \text{Var}(X^*) - y\mathbb{E}[X^*] - 1 \\ &= \text{Var}(X^*) - 1 \\ &= 0, \end{aligned}$$

but  $\phi$  varies in  $y$  and hence, partial completeness does not hold.

**Assumption 3.** The selection probability  $\mathbb{P}(D = 1|Y, X^*)$  is bounded away from zero uniformly over its support.

Assumption 3 can rule out a selection when it is a deterministic function of  $Y$  and  $X^*$ , such as certain indicator functions. We also emphasize that Assumption 3 can be relaxed if we are only interested in point evaluation at some  $x_0$  in the support of  $X$ . That is, identification of  $\mathbb{E}[Y|X^* = x_0]$  requires only  $\mathbb{P}(D = 1|Y = y, X^* = x_0) > 0$  uniformly over all  $y$  in the support of  $Y$ .

## 2.2. Nonparametric Identification via FPW Weighting

In this subsection, we establish identification of the nonparametric regression function  $g(x) = \mathbb{E}[Y|X^* = x]$ . We show that the function  $g$  can be identified via a fractional probability weight (FPW). In addition, we show that the FPW is identified by making

use of instrumental variables  $W$  which satisfy the previous assumptions. In the next result, we document that the regression function  $g$  can be written as

$$g(x) = \mathbb{E} \left[ Y \omega(Y, x) \middle| D = 1, X^* = x \right] \quad (2.3)$$

where the fractional probability weight (FPW) function  $\omega$  is given by

$$\omega(y, x) = \frac{\mathbb{P}(D = 1 | X^* = x)}{\mathbb{P}(D = 1 | Y = y, X^* = x)}. \quad (2.4)$$

Further, we establish identification of the nonparametric regression function  $g$ .

**Theorem 2.2.** *Let Assumptions 1–3 be satisfied. Then, the FPW function  $\omega$  is identified and thus, identification of the regression function  $g$  follows through (2.3).*

The previous result shows that the FPW function given in (2.4) is point identified although the selection probabilities conditional on latent variables are only partially identified. This is an implication of partial completeness imposed in Assumption 2. Corollary 2.3 presents a useful property of the FPW function  $\omega$ . This result is an immediate consequence of the proof of Theorem 2.2 and hence we omit its proof.

**Corollary 2.3.** *Let Assumption 3 be satisfied. Then, for the FPW function  $\omega$  we obtain*

$$\mathbb{E}[\omega(Y, x) | D = 1, X^* = x] = 1.$$

In empirical applications also the dependent variable might be selectively missing. We discuss this case in Appendix C. The remark below highlights the difference between selectively missing covariates which requires FPW and selectively missing dependent variables which requires inverse probability weighting.

**Remark 2.1** (Relation to selectively missing depend variables). *Consider the case of selectively missing dependent variables, that is,  $Y^*$  is only observed if  $D^Y = 1$  and otherwise missing, while  $X$  is always observed. Further, assume that  $\mathbb{P}(D^Y = 1 | Y^*, X) = \mathbb{P}(D^Y = 1 | Y^*)$ . In this case, (as shown by Breunig et al. [2018]) we obtain*

$$\mathbb{E}[Y^* | X] = \mathbb{E}[Y \psi(Y) | X]$$

where  $Y = Y^* D^Y$  and  $\psi(\cdot) = 1 / \mathbb{P}(D^Y = 1 | Y^* = \cdot)$ . Consequently, the conditional expectation with latent  $Y^*$  corresponds to an inverse probability weighted version of observed counterparts. This correction differs from fractional probability weighting, as considered in this paper, which is required for selectively missing covariates.

While Remark 2.1 highlights the difference between fractional and inverse probability weighting, we note that for unconditional moment estimation both approaches are feasible (see Section 4.2). Still our FPW approach is preferable in this case due to a less restrictive identification requirement and finite sample improvements as shown in Section 4.2.



### 3. The FPW Series Estimator and its Asymptotic Properties

This section consists of four subsections. In Subsection 3.1, we derive the FPW series estimator which stems from our constructive identification result. Subsection 3.2 provides the rate of convergence of the estimator. We establish pointwise asymptotic normality of the FPW series estimator in Subsection 3.3 and provide asymptotic validity of uniform bootstrap confidence bands in Subsection 3.4.

#### 3.1. Estimation

We define the conditional selection probability by

$$\varphi(y, x) := \mathbb{P}(D = 1 | Y = y, X^* = x). \quad (3.1)$$

In particular, the selection probability conditional on the latent regressors  $X^*$  is determined by

$$\mathbb{P}(D = 1 | X^* = x) = \left( \mathbb{E} \left[ \frac{1}{\varphi(Y, x)} \middle| D = 1, X^* = x \right] \right)^{-1},$$

see the proof of Theorem 2.2. We thus obtain the following expression for the FPW function  $\omega$ :

$$\omega(y, x) = \left( \mathbb{E} \left[ \frac{\varphi(y, x)}{\varphi(Y, x)} \middle| D = 1, X^* = x \right] \right)^{-1}.$$

We estimate the regression function  $g(x) = \mathbb{E}[Y | X^* = x] = \mathbb{E}[Y\omega(Y, x) | D = 1, X^* = x]$  using a plug-in series least squares estimator. To do so, we introduce a vector of basis functions  $p^K(\cdot) = (p_1(\cdot), \dots, p_K(\cdot))'$ ;  $K = K(n)$  is an integer which increases with the sample size  $n$ . We further introduce the  $n \times K$ -matrix  $\mathbf{X} = (D_1 p^K(X_1), \dots, D_n p^K(X_n))$ . We estimate the FPW function  $\omega$  via

$$\widehat{\omega}(y, x; \phi) = \left( p^K(x)' (\mathbf{X}' \mathbf{X})^{-1} \sum_{i=1, D_i=1}^n p^K(X_i) \frac{\phi(y, x)}{\phi(Y_i, X_i)} \right)^{-1}.$$

It is common in the context of inverse probability weighting, to normalize the weights to sum up to one. In our context, we normalize  $\omega$  as follows. Employing Corollary 2.3, i.e.,  $\mathbb{E}[\omega(Y, x) | D = 1, X^* = x] = 1$ , we obtain

$$\begin{aligned} \mathbb{E}[D p^K(X) p^K(X)'] &= \mathbb{E}[D p^K(X^*) p^K(X^*)'] \\ &= \mathbb{E}[D p^K(X) \omega(Y, X) p^K(X)']. \end{aligned}$$

Replacing  $\mathbb{E}[D p^K(X) \omega(Y, X) p^K(X)']$  by the empirical matrix  $\mathbf{X}' \boldsymbol{\omega}(\widehat{\varphi}) \mathbf{X}$  we obtain the FPW series estimator of the regression function  $g$  given by

$$\widehat{g}(x) \equiv p^K(x)' (\mathbf{X}' \boldsymbol{\omega}(\widehat{\varphi}) \mathbf{X})^{-1} \sum_{i=1, D_i=1}^n p^K(X_i) Y_i \widehat{\omega}(Y_i, X_i; \widehat{\varphi}) \quad (3.2)$$



where  $\omega(\phi) = \text{diag}(\widehat{\omega}(Y_1, X_1; \phi), \dots, \widehat{\omega}(Y_n, X_n; \phi))$ . Here,  $\widehat{\varphi}$  is a restricted sieve minimum distance estimator of the selection probability  $\varphi$  given as follows. We have the conditional moment restriction induced by the exclusion restriction imposed in Assumption 1, that is,

$$\mathbb{E} \left[ \frac{D}{\varphi(Y, X)} \middle| Y, W \right] = 1. \quad (3.3)$$

(Here, we use that  $\varphi(Y, X^*) = \varphi(Y, X)$  whenever  $D = 1$ .) Consider a vector of tensor product basis functions  $q^L(\cdot, \cdot) = (q_1(\cdot, \cdot), \dots, q_L(\cdot, \cdot))'$  used to approximate the conditional mean in (3.3);  $L = L(n)$  is an integer which increases with the sample size  $n$ . We estimate the conditional mean  $m(y, w; \phi) = \mathbb{E}[D/\phi(Y, X) - 1 | Y = y, W = w]$  by the series least squares estimator

$$\widehat{m}(y, w; \phi) := q^L(y, w) \left( \sum_{i=1}^n q^L(Y_i, W_i) q^L(Y_i, W_i)' \right)^{-1} \sum_{i=1}^n q^L(Y_i, W_i) \left( \frac{D_i}{\phi(Y_i, X_i)} - 1 \right)$$

then we the constrained sieve minimum distance estimator

$$\widehat{\varphi} = \underset{\phi \in \mathcal{B}_L}{\text{argmin}} \sum_{i=1}^n \widehat{m}^2(Y_i, W_i; \phi) \quad (3.4)$$

where  $\mathcal{B}_L = \{\phi = 1/(\beta' q^L) : \min_{y,x} \beta' q^L(y, x) \geq 1\}$  following Chen and Pouzo [2012]. We may assume that  $\mathcal{B}_L$  becomes dense in  $\mathcal{B}$  as  $L$  tends to infinity. Note that in the case of nonparametric estimation, any estimator of  $\varphi$  has a slow rate of convergence since the conditional mean restriction yields in general to a so called ill-posed inverse problem, see Newey and Powell [2003] and Blundell et al. [2007]. In our case,  $\varphi$  is not identified through the conditional mean equation (3.3) but we can always ensure uniqueness of the estimator, for instance, by considering the minimal norm estimator of equation (3.3). Finally, note that FPW series estimation is convenient since control variables which enter the model linearly can be simply included in the empirical matrix  $\mathbf{X}$ . This allows to treat partially linear models as considered in our empirical applications in Section 5.

## 3.2. Rate of Convergence

We now introduce some assumptions. The support of  $X$  is denoted by  $\mathcal{X}$ . We also introduce the  $L_X^2$ -norm  $\|\phi\|_X = \sqrt{\mathbb{E}\phi^2(X)}$  and  $\|\cdot\|$  denotes the Euclidean norm. We make use of the notation  $U = Y - g(X^*)$  and  $h(x, \phi) = \mathbb{E}[1/\phi(V) | D = 1, X^* = x]$ . Recall that  $\omega(V, \varphi) = (\varphi(V)h(X, \varphi))^{-1}$  is identified due to Theorem 2.2. We introduce function class

$$\mathcal{H} = \{\psi : \psi(\cdot, \phi) \in L_X^2 \text{ and } \psi(\cdot, \phi) \geq 1 \text{ for all } \phi \in \mathcal{B}\}.$$

Note that  $h(\cdot, \varphi) \in \mathcal{H}$  due to Assumption 3. In order to achieve the rate of convergence, we have to restrict the complexity of the function classes  $\mathcal{B}$  and  $\mathcal{H}$ , by imposing a finite entropy integral. We first provide primitive conditions by imposing smoothness restrictions on the function classes under consideration.

For any vector  $a = (a_1, \dots, a_d)$  of  $d$  integers, where  $d = \dim(X, Y)$ , define  $D^a = \partial^{|a|} / \partial a_{a_1} \dots \partial a_{a_d}$ , where  $|a| = \sum_{i=1}^d a_i$ . Let  $\mathcal{R}_d$  be a bounded, convex subset of  $\mathbb{R}^d$

with nonempty interior. For a function  $\phi : \mathcal{R}_d \rightarrow \mathbb{R}$  and some  $\alpha > 0$ , let  $\underline{\alpha}$  be the largest integer smaller than  $\alpha$ , and

$$\|\phi\|_{\infty, \alpha} = \max_{|a| \leq \underline{\alpha}} \sup_v |D^a \phi(v)| + \max_{|a| = \underline{\alpha}} \sup_{v \neq v'} \frac{|D^a \phi(v) - D^a \phi(v')|}{\|v - v'\|^{\alpha - \underline{\alpha}}}.$$

We further restrict the function class  $\mathcal{B}$  in the next assumption to contain only functions with bounded  $\|\cdot\|_{\infty, \alpha}$  norm and bounded supremum norm  $\|\cdot\|_{\infty}$ . To do so, we introduce

$$\mathcal{B}_{\alpha} = \{\phi : \mathcal{R}_d \rightarrow \mathbb{R} \text{ continuous with } \inf_{v \in \mathcal{R}_d} \phi(v) > 0 \text{ and } \|\phi\|_{\infty, \alpha} < C\} \quad (3.5)$$

for some  $\alpha > 0$  and constant  $C > 0$ . Similarly for  $\mathcal{H}$  we introduce a smooth function class

$$\mathcal{H}_{\alpha} = \{\psi : \mathcal{R}_{d_x} \rightarrow \mathbb{R} \text{ continuous with } \|\psi(\cdot, \phi)\|_{\infty, \alpha} < C \text{ and } \psi(\cdot, \phi) \geq 1 \text{ for all } \phi \in \mathcal{B}\}$$

for some  $\alpha > 0$  and constant  $C > 0$ , which is a subset of  $\mathcal{H}$  for any constant  $C > 0$ .

**Assumption 4.** (i) We observe a sample  $((D_1, Y_1, X_1, W_1), \dots, (D_n, Y_n, X_n, W_n))$  of independent and identical distributed (i.i.d.) copies of  $(D, Y, X, W)$  where  $X = DX^*$ . (ii) There exist a constant  $C > 0$  and a sequence of positive integers  $K := K(n)$  satisfying  $\sup_{x \in \mathcal{X}} \|p^K(x)\|^2 \leq CK$  such that  $K^2/n = o(1)$ . (iii) The smallest eigenvalue of  $\mathbb{E}[D p^K(X) p^K(X)']$  is bounded away from zero uniformly in  $K$ . (iv) Let  $\mathbb{E}[U^2|X^*] < \infty$ ,  $\mathbb{E}[g^2(X^*)] < \infty$ , and  $\|\gamma' p^K - g\|_X = O(K^{-\alpha/d_x})$  for some  $\gamma \in \mathbb{R}^K$ . (v)  $\mathcal{B} \subset \mathcal{B}_{\alpha}$  with  $\alpha > d/2$  and  $\mathcal{H} \subset \mathcal{H}_{\alpha'}$  with  $\alpha' > d_x/2$ .

Assumption 4 (ii) – (iii) restricts the magnitude of the approximating functions  $\{p_j\}_{j \geq 1}$  and imposes nonsingularity of their second moment matrix. Assumption 4 (ii) holds for instance for polynomial splines, Fourier series and wavelet bases. Assumption 4 (iii) is satisfied if  $p^K$  is a vector of orthonormal basis functions and the probability density function of  $X^*$  given  $D = 1$  is uniformly bounded away from zero on its support. Assumption 4 (iv) determines the sieve approximation error which in turn characterizes the bias of the estimated regression function  $g$ ; see also Chen et al. [2007] for further discussions on sieve bases. Assumption 4 (v) imposes smoothness conditions on the function classes  $\mathcal{B}$  and  $\mathcal{H}$  in order to reduce the complexity of these classes.

The next result establishes the rate of convergence for the FPW series estimator  $\widehat{g}$ .

**Theorem 3.1.** *Let Assumptions 1–4 be satisfied. Then, we have*

$$\|\widehat{g} - g\|_X^2 = O_p\left(\max\left(K^{-2\alpha/d_x}, \frac{K}{n}\right)\right).$$

From Theorem 3.1 we observe that the estimator  $\widehat{g}$  attains the usual bias and variance term in integrated mean square error for nonparametric series regression. If  $K$  is chosen to level variance and bias, i.e.,  $K \sim n^{d_x/(2\alpha+d_x)}$ , then the convergence rate given in Theorem 3.1 coincides with  $n^{-2\alpha/(2\alpha+d_x)}$ . Consequently, we obtain the optimal nonparametric rate of convergence as in the situation where the covariates  $X$  are completely observed, that is, we do not obtain a slower rate of convergence due to the estimation of the FPW function. This result is obtained by the regularity conditions imposed in Assumption 4 (v), which restrict the complexity of the underlying fractional probability function  $\omega$ .

### 3.3. Pointwise Inference

This subsection discusses the inference of the estimator of regression function  $g$  evaluated at some point of the support of  $X$ . In applications, such asymptotic distribution results can be useful to construct approximate confidence intervals. Before stating the result we make the following additional assumptions, in particular, with respect to the error term  $U = Y - g(X^*)$ .

**Assumption 5.** (i)  $\mathbb{E}[|UD\omega(Y, X^*)|^4]$  is bounded from above and  $\mathbb{V}ar(UD\omega(Y, X^*)|X^*)$  is uniformly bounded away from zero. (ii) For some  $x$  in the support of  $X$  it holds  $\sum_{j=1}^K |p_j(x)| = O(\|p^K(x)\|)$ .

The bounds imposed in Assumption 5 (i) are not stronger than the one imposed in Newey [1997]. We also note that it is possible to relax these conditions as noted by Belloni et al. [2015] or Chen and Christensen [2015]. Assumption 5 (ii) is a condition on the basis functions and satisfied for B-splines or wavelets, see Appendix E of Chen and Christensen [2018].

To obtain asymptotic normality of our estimator we require a normalization factor. Therefore, we introduce the sieve variance given by

$$v_K(x) = p^K(x)' \mathbb{E} \left[ p^K(X^*) \mathbb{V}ar(UD\omega(Y, X^*)|X^*) p^K(X^*)' \right] p^K(x).$$

In contrast to the usual series regression in Newey [1997], we see that the sieve variance also contains the FPW function  $\omega$ . As  $\omega$  can take values smaller than one, the sieve variance for our FPW series estimator can be even smaller than the one associated to the usual series estimator. This is in contrast to estimators based on weighting via inverse selection probabilities that always lead to larger sieve variances, see Breunig et al. [2018] for selective outcomes or Das et al. [2003] for propensity score weighting. We replace the sieve variance  $v_K(x)$  by the estimator

$$\widehat{v}_K(x) = p^K(x)' (\mathbf{X}'\mathbf{X})^{-1} n^{-1} \sum_{i=1, D_i=1}^n p^K(X_i) \widehat{U}_i^2 D_i \widehat{\omega}^2(Y_i, X_i; \widehat{\varphi}) p^K(X_i)' (\mathbf{X}'\mathbf{X})^{-1} p^K(x)$$

where  $\widehat{U}_i = Y_i - \widehat{g}(X_i)$ . We now establish the asymptotic distribution of the estimator  $\widehat{g}$  evaluated at some point  $x$  in the support of  $X$ . Similarly, asymptotic distribution results for linear functionals of  $g$  can be obtained. We introduce the supremum norm  $\|\phi\|_\infty = \sup_{x \in \mathcal{X}} |\phi(x)|$ .

**Theorem 3.2.** *Let Assumptions 1–5 be satisfied. If for some  $x$  in the support of  $X$  it holds*

$$n \|\gamma' p^K - g\|_\infty^2 = o(v_K(x)) \tag{3.6}$$

then we have

$$\sqrt{n/\widehat{v}_K(x)} (\widehat{g}(x) - g(x)) \xrightarrow{d} \mathcal{N}(0, 1).$$

Condition (3.6) requires the estimator of  $g$  to be undersmoothed. This ensures that the sieve approximation bias in the second step estimation procedure becomes asymptotically negligible. Theorem 3.2 can be also used to construct pointwise confidence intervals for  $g(x)$  but can also be extended to construct uniform bootstrap confidence bands, as the following remark illustrates.

### 3.4. Uniform Bootstrap Confidence Bands

This section provides a bootstrap procedure to construct uniform confidence bands for  $g$  and establishes asymptotic validity of it. Let  $(\varepsilon_1, \dots, \varepsilon_n)$  be a bootstrap sequence of i.i.d. random variables drawn independently of the data  $\{(D_i, Y_i, X_i, W_i)\}_{1 \leq i \leq n}$ , with  $\mathbb{E}[\varepsilon_i] = 0$ ,  $\mathbb{E}[\varepsilon_i^2] = 1$ , with bounded moments. Common choices of distributions for  $\varepsilon_i$  include the standard Normal, Rademacher, and the two-point distribution of Mammen [1993]. Further,  $\mathbb{P}^*$  denotes the probability distribution of the bootstrap innovations  $(\varepsilon_1, \dots, \varepsilon_n)$  conditional on the data. We introduce the bootstrap process

$$\mathbb{Z}^B(x) = \frac{p^K(x)'(\mathbf{X}'\boldsymbol{\omega}(\hat{\varphi})\mathbf{X}/n)^{-1}}{\sqrt{n\hat{v}_K(x)}} \sum_{i=1, D_i=1}^n p^K(X_i) \left( Y_i \hat{\omega}(Y_i, X_i; \hat{\varphi}) - \hat{g}(X_i) \right) \varepsilon_i.$$

Under regularity conditions it can be shown that the bootstrap process provides a uniform approximation of the influence function of the estimator  $\hat{g}$  and thus, can be used to construct uniform confidence bands (see also Chen and Christensen [2018]).

For the implementation of a  $100(1 - \alpha)\%$  uniform confidence bands, we compute the critical value  $z_{1-\alpha}^B$  as the  $(1 - \alpha)$  quantile of  $\sup_x |\mathbb{Z}^B(x)|$  for a large number of independent bootstrap draws. The resulting  $100(1 - \alpha)\%$  uniform confidence band is then given by

$$x \mapsto \left[ \hat{g}(x) - z_{1-\alpha}^B \sqrt{\frac{\hat{v}_K(x)}{n}}, \hat{g}(x) + z_{1-\alpha}^B \sqrt{\frac{\hat{v}_K(x)}{n}} \right].$$

In the following, we assume that  $\mathcal{C}$  is a closed subset of  $\mathbb{R}^{d_x}$ . Let  $\delta(\cdot, \cdot)$  be the standard deviation semimetric on  $\mathcal{C}$  of the Gaussian Process  $\mathbb{Z}(x) = p^K(x)' \mathcal{Z} / \sqrt{v_K(x)}$  with  $\mathcal{Z} \sim \mathcal{N}(0, \Sigma)$  and  $\Sigma = \mathbb{E}[p^K(X^*) \text{Var}(UD \omega(Y, X^*) | X^*) p^K(X^*)']$  defined as  $\delta(x_1, x_2) = (\mathbb{E}[(\mathbb{Z}(x_1) - \mathbb{Z}(x_2))^2])^{1/2}$ , see *e.g.* van der Vaart and Wellner [2000, Appendix A.2]. Further, we introduce the notation  $N(\varepsilon, \mathcal{F}, \|\cdot\|_{\mathcal{F}})$  for the  $\varepsilon$ -entropy of  $\mathcal{F}$  with respect to a norm  $\|\cdot\|_{\mathcal{F}}$ .

**Assumption 6.** (i)  $\mathcal{C}$  is compact and  $(\mathcal{C}, \delta(\cdot, \cdot))$  is separable for each  $n \geq 1$ . (ii) There exists a sequence of finite positive integers  $c_n$  such that  $1 + \int_0^\infty \sqrt{\log N(\varepsilon, \mathcal{C}, \delta(\cdot, \cdot))} d\varepsilon = O(c_n)$ . (iii) There exists a sequence of positive integers  $r_n$  with  $r_n = o(1)$  such that  $K^{5/2} = o(r_n^3 \sqrt{n})$ ,  $r_n c_n = O(1)$ , and

$$K^2 \sqrt{\frac{\log(n)}{n}} + K n^{-1/4} \log(n) \left( c_n + \sup_{x \in \mathcal{C}} \sqrt{\frac{n}{v_K(x)}} |\gamma' p^K(x) - g(x)| \right) = o(r_n).$$

Assumption 6 is similar to Chen and Christensen [2018, Assumption 6] who establish asymptotic validity of uniform confidence bands in nonparametric instrumental variable estimation. Assumption 6 (ii) is a mild regularity assumption, see also Chen and Christensen [2018, Remark 4.2] for sufficient conditions.

The next theorem establishes asymptotic validity of the bootstrap for constructing uniform confidence bands for the regression function  $g$ . Below,  $\mathbb{P}^*$  denotes a probability measure conditional on the data  $\{(D_i, Y_i, X_i, W_i)\}_{i=1}^n$ .

**Theorem 3.3.** *Let the assumptions of Theorem 3.2 and Assumption 6 hold. Then, we have*

$$\sup_{s \in \mathbb{R}} \left| \mathbb{P} \left( \sup_{x \in \mathcal{C}} \left| \sqrt{\frac{n}{\hat{v}_K(x)}} (\hat{g}(x) - g(x)) \right| \leq s \right) - \mathbb{P}^* \left( \sup_{x \in \mathcal{C}} |\mathbb{Z}^B(x)| \leq s \right) \right| = o_p(1).$$

## 4. Monte Carlo simulation

In this section, we study the finite-sample performance of our estimator by presenting the results of a Monte Carlo simulation. We first focus on the estimation of nonlinear conditional moments, then we turn to linear regression. We perform 1000 Monte Carlo replications in each experiment and the sample size is  $n = 1000$ .

### 4.1. Nonlinear Regression

We consider the estimation of the regression function  $g$  under the following simulation design. The data are generated by  $W = \Phi(\xi)$  and  $X^* = \Phi(\chi)$  where  $\chi = \rho\xi + \sqrt{1 - \rho^2}\nu$  and  $(\xi, \nu)' \sim \mathcal{N}(0, I_2)$ . Here,  $\rho$  characterizes the strength of the instruments and is varied in the experiments below. Further, we draw  $Y$  from the model

$$Y = g(X^*) + U$$

where  $g(x) = \Phi(c_g(x - 0.5))$  with standard normal distribution function  $\Phi$ , a constant  $c_g \in \{5, 20\}$ , and  $U \sim \mathcal{N}(0, \sigma_U^2)$ , where  $\sigma_U^2 \in \{0.5, 1\}$ . We generate realizations of the selection variable  $D$  from the Bernoulli distribution

$$D \sim \text{BERNOULLI}(\Phi(1 + \chi + U/2)). \quad (4.1)$$

Consequently, the selection probability is a function of the latent covariates  $X^*$  and also unobservables  $U$ . The selection probability  $\varphi$  is estimated using the sieve minimum distance procedure in (3.4) with tensor product of quadratic B-splines and zero knots (hence  $L = 9$ ). We estimate the function  $g$  by using the FPW series estimator  $\hat{g}$  given in (3.2). As basis functions we use quadratic B-splines with 2 knots (hence  $K = 5$ ) when  $c_g = 5$  (see Figure 1) and quadratic B-splines with 7 knots (hence  $K = 10$ ) when  $c_g = 20$  (see Figure 2). For the B-splines of  $X$  note that we choose the interior knots located at the quantiles of the observed subset of  $\{X_i\}_{D_i=1, 1 \leq i \leq n}$ .

Figures 1 and 2 depict the median of the FPW series estimator  $\hat{g}$  together with its 95% pointwise confidence bands and a series estimator under the missing at random (MAR) assumption based on listwise deletion under different simulation designs. We vary the parameters  $\rho$  and  $\sigma_U$ ; the first row shows results with  $\sigma_U^2 = 0.5$ , the second row with  $\sigma_U^2 = 1$ , the first column with  $\rho = 0.2$ , and the second column with  $\rho = 0.6$ . In all cases the median of the FPW series estimator is close to the true regression function  $g$  and the MAR series estimator is severely biased. From Figures 1 and 2 we see that the strength of instruments only has a moderate influence on the performance of the FPW series estimator. This is in line with our theoretical results that the asymptotic performance of the estimator is not driven by the correlation of the instruments to the latent covariates. On the other hand, we see that the variance of estimation becomes much larger as  $\sigma_U^2$  increases from 0.5 to 1. For larger values of  $\sigma_U$  the problem of selection on unobservables becomes more severe. In this case, the confidence intervals of the FPW series estimator become larger but also the bias of the MAR series estimator increases.

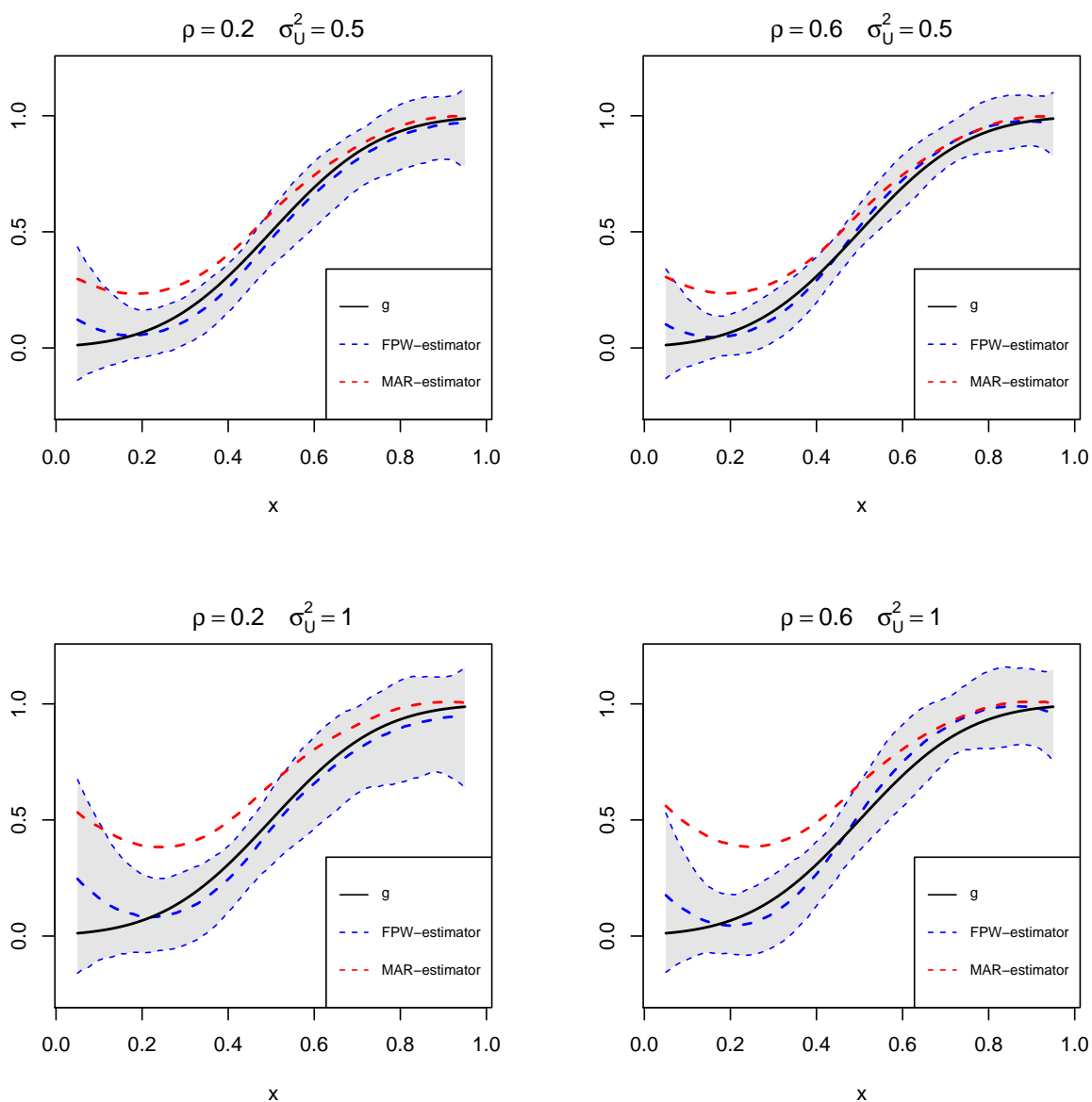


Figure 1: The regression function  $g$  with  $c_g = 5$ , the median of  $\hat{g}$  (blue) with 95% confidence intervals and an estimator under MAR assumption (red).

## 4.2. Linear Regression and Comparison to Inverse Probability Weighting (IPW)

For estimators based on unconditional moments, an alternative approach to FPW is given by IPW. Yet this subsection demonstrates that the FPW approach leads to more accurate estimation results even in linear regression models.

We generate the data as described in the previous subsection. As we are interested in the unconditional mean  $\mathbb{E}[YX^*]$  we could also make use of IPW. Indeed, making use of the notation  $\tilde{\omega}(y, x) = 1/\mathbb{P}(D = 1|Y = y, X^* = x)$  we obtain by the law of

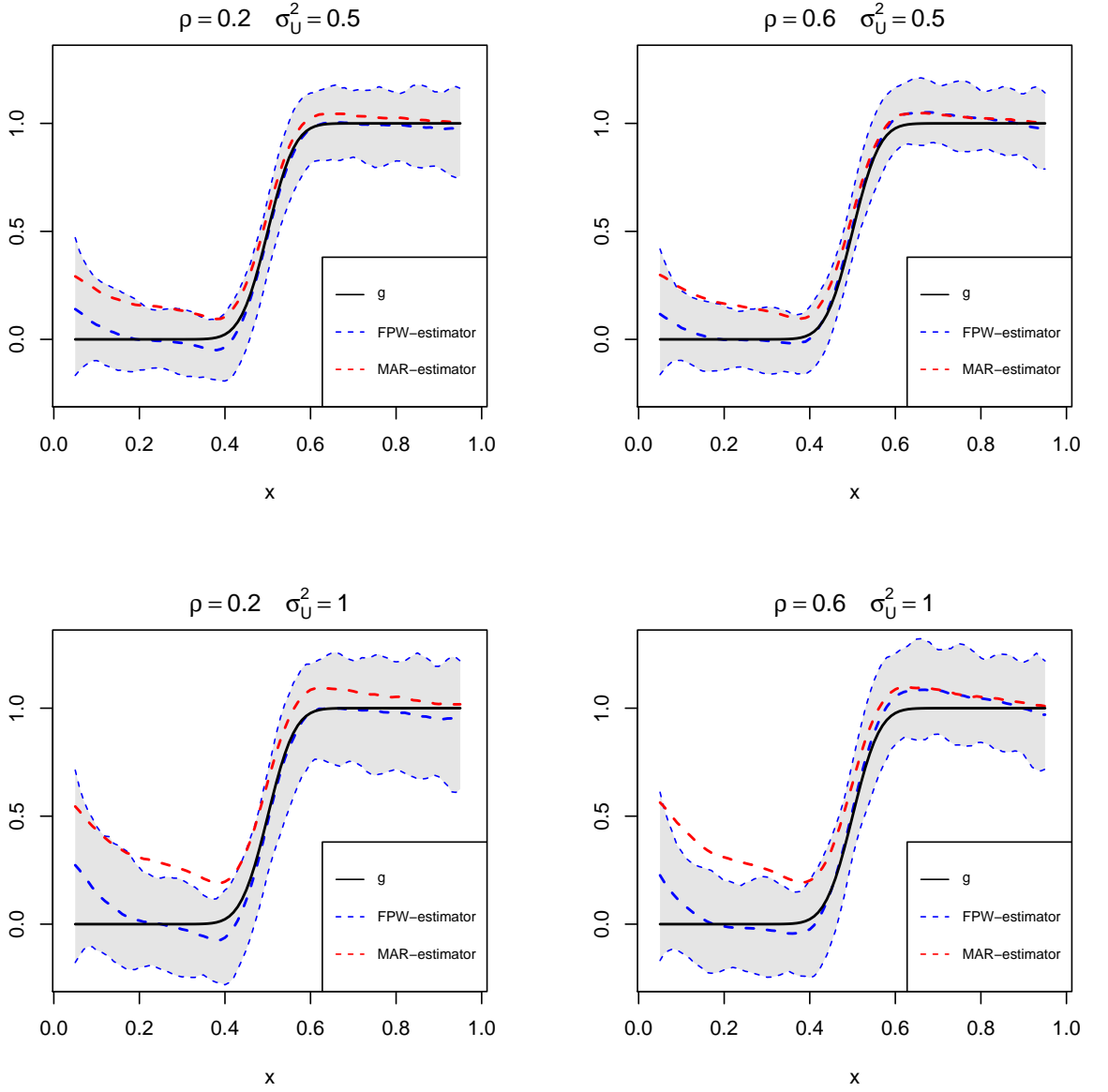


Figure 2: The regression function  $g$  with  $c_g = 20$ , the median of  $\hat{g}$  (blue) with 95% confidence intervals and an estimator under MAR assumption (red).

iterated expectations that

$$\begin{aligned}\mathbb{E}[YX^*] &= \mathbb{E}[YX^*D/\mathbb{P}(D = 1|Y, X^*)] \\ &= \mathbb{E}[YX\tilde{\omega}(Y, X)].\end{aligned}$$

Alternatively, we can apply the FPW function  $\omega(y, x) = \mathbb{P}(D = 1|X^* = x)/\mathbb{P}(D = 1|Y = y, X^* = x)$  and obtain by our nonparametric identification results that

$$\mathbb{E}[YX^*] = \mathbb{E}[YX\omega(Y, X)].$$

Estimating the unconditional mean by FPW has the advantage over IPW that identification of the inverse selection probability is more restrictive than identification of the FPW function  $\omega$ . That is, for identification of the IPW the usual completeness



assumption is required. In addition, we demonstrate in the following the finite sample properties of both approaches in a finite sample analysis. To do so, we consider the linear model

$$Y = \beta_0 + \beta_1 X^* + U \quad (4.2)$$

where  $\beta_0 = 1$  and  $\beta_1 = 3$ . The data is generated as described in the previous subsection with  $\rho = 0.2$  and  $\sigma_U = 1$ . Below, we analyze the absolute median bias and the coverage at the 95% nominal coverage rate for the FPW estimator, the IPW estimator, the MAR estimator based on listwise deletion and the estimator when there is no missing data, that is,  $D \equiv 1$ . We estimate the weights for FPW and IPW nonparametrically as described in the previous subsection. The FPW and IPW estimators coincide then with weighted ordinary least squares (OLS) estimators.

	FPW	IPW	MAR	$D \equiv 1$
Abs. median bias( $\hat{\beta}_0$ )	0.039	0.056	0.403	0.006
Abs. median bias( $\hat{\beta}_1$ )	0.112	0.147	0.447	0.008
Coverage for $\beta_0$	0.873	0.799	0.004	0.943
Coverage for $\beta_1$	0.798	0.728	0.085	0.934

Table 1: Absolute median bias and empirical coverage at 95% nominal coverage rate for the linear model (4.2) for the FPW estimator, the IPW estimator, the MAR estimator, and the OLS estimator without missingness.

In Table 1 we compare the IPW and FPW estimators with the OLS estimator under the MAR hypothesis and the OLS estimator when there is no missing data. The second and third row show the absolute median bias of the estimators of the intercept and the slope parameter. We see that the FPW estimator has smaller median bias than the IPW estimator for both parameters. Not surprisingly, the bias dramatically increases when we ignore selection and consider the MAR estimator. The last two rows depict the coverage of the confidence interval for the intercept and the slope parameter. We see that the FPW estimator has more accurate coverage than the IPW estimator. Yet there is undercoverage of the FPW estimator which is due to the severity of the selectivity of the nonresponse mechanism. Note that the 95% confidence interval of the MAR estimator contains the true intercept only in 4 out of 1000 Monte Carlo Iterations. If we relax the severity of the selectivity then the coverage of the FPW estimator is more accurate. For instance, if in (4.1) the variable  $U/2$  is replaced by  $U/3$  then the empirical coverage of the FPW estimator for  $\beta_1$  increases from 0.798 to 0.887.

## 5. Empirical Applications

In the final section of the paper we apply the developed methodology to study to relevant economic questions. First, we focus on the association between income and health. In the second application we analyze how income affects the demand for housing.

## 5.1. Application I: the association between income and health

As mentioned in the introduction, a large body of literature has documented a positive correlation between income and health, see e.g. Deaton and Paxson [1998].<sup>4</sup> However, in general these studies are based on survey data in which income, health information and further demographic variables are self reported. As shown e.g. in Breunig [2019] information on income in surveys is likely to suffer from nonrandom selection which might result in biased estimates of the association between income and health.

The empirical analysis is based on linked data from the German sample of the Survey of Health, Aging and Retirement in Europe (SHARE, Wave 5, collected in 2013) and the German pension insurance. SHARE is a multinational survey of the elderly population aged 50 and above in Europe, for more information see Börsch-Supan et al. [2013]. The survey includes standard demographic characteristics and self reported information about different income measures and various subjective and objective health outcomes. The key variables for our analysis are individual income and health outcomes. We use a broad definition of income. For non-retired individuals the income includes labor earnings, income from self employment and transfers for unemployed. For retired individuals the income is composed of own pensions, and if applicable widowers pension and additional labor earnings. The health status is described by an objective measurement of the hand grip strength. Previous studies have documented that hand grip strength is a good measure of physical functioning and a predictor of morbidity, disability and mortality, see e.g. Rantanen et al. [1999], Bohannon [2015], or Dodds et al. [2014]. From the grip strength we construct a binary variable which indicates bad health status if the grip strength is below the 25th percentile.

For our analysis we exploit a specific feature of the data which allows us to link a subsample<sup>5</sup> of the survey data to administrative data of the German pension insurance. Thus, in addition to the self-reported income information which might suffer from nonrandom nonresponse the data includes official information about pension entitlements. For pensioners we observe the full pension entitlements, i.e. number of pension points, they have earned during their working life; for non retired individuals we observe the entitlements they have collected so far. Pension entitlements are a deterministic function of the full individual earnings history. The earnings history is a good predictor of current income, however, it contains no direct information about the response behavior for current income. Therefore, this information allows us to construct a suitable instrument to account for potential nonrandom nonresponse of the current self reported income. In fact this instrument is superior to instruments based on self reported lagged employment outcomes which are often used, see e.g. Breunig [2019]. First, the instrument is not affected by transitory shocks since it combines information about the full working life instead of using information of only one period. Second, self reported past information might as well suffer from non-response. This is not the case for the information about pension entitlements in administrative data.

In the empirical analysis we concentrate on 3340 individuals which are younger than

---

<sup>4</sup>In general, it is difficult to identify the causal effect of income on health, therefore most studies focus on the association between income and health. We follow these studies. Notable exceptions are studies that focus on the effect of income or wealth shocks on health, see e.g. Schwandt [2018].

<sup>5</sup>The linkage of the data requires the consent of the individuals, about 2/3 of individuals agreed to the linkage.

80 years and who have agreed to the linkage of the survey data and the information of the pension insurance. Out of this sample, 12.34% do not respond to the income information question.<sup>6</sup> Table 2 provides summary statistics of the relevant variables for the analysis.

	1st Qu.	Median	Mean	3rd Qu.
Grip strength	28.00	35.00	36.53	45.00
Bad health	0	0	0.24	0
log(Income per Year)	8.96	9.48	9.37	9.95
Number of pension points	19.90	33.20	33.59	47.00
Age in years	56.00	62.00	62.39	69.00

Table 2: This table provides summary statistics of the relevant variables for the analysis. The sample includes 3340 individuals which are aged between 50 and 80 years. Grip strength is measured in kilograms. The value of a pension point in 2013 amount to 24.92 (East Germany) and 28.07 (West Germany). Bad health is defined when grip strength is below the 25th percentile.

To quantify the association between income and health we use the following semi-parametric model

$$Badhealth_i = g(\log(Income_i^*)) + \alpha_0 Age_i + \beta_0 Gender_i + U_i, \quad (5.1)$$

where the function  $g$  and the parameters  $\alpha_0$  and  $\beta_0$  are unknown. We assume that  $U_i$  is conditional mean independent of the explanatory variables,  $\log(Income_i^*)$ ,  $Age_i$ , and  $Gender_i$ . We apply the FPW estimator as described in the previous section, i.e., we estimate the nonparametric selection probability using quadratic B-spline basis functions for least square approximations. Specifically, the selection probability  $\varphi$  is estimated using the sieve minimum distance procedure described in (3.4) with tensor product of quadratic B-splines and zero knots, where we additionally control for age and gender. We estimate the function  $g$  using the FPW series estimator  $\hat{g}$  given in (3.2) using quadratic B-splines with 1 knot (placed at the median of observed income) and again controlling for age and gender. The MAR estimator uses the same choice of B-spline basis functions with the same knot placement.

Figure 3 depicts the FPW series estimator with the 95% uniform confidence bands together with the MAR series estimator evaluated at the median age. The uniform confidence bands are computed using the bootstrap procedure as described in Section 3.4 with 1000 bootstrap iterations. For the MAR series estimator, we consider listwise deletion of missing values. The FPW series estimator shows a negative association between income and bad health measured by the grip strength which is moderate. For example we find, that the risk of bad health for individuals at the 25th percentile of income (about 8.9 log income) is about 40% while for individuals at the 75th percentile (log income of 9.9) it is slightly lower (about 37.5%). However, according

<sup>6</sup>In our sample only 4.8% do not provide information about grip strength - we assume that this information is missing at random. Using the test of missing (completely) at random by Breunig [2019] we obtain the value of the test statistic 0.053 with 0.05- level critical value of 0.10 and hence, we fail to reject the MAR hypothesis.

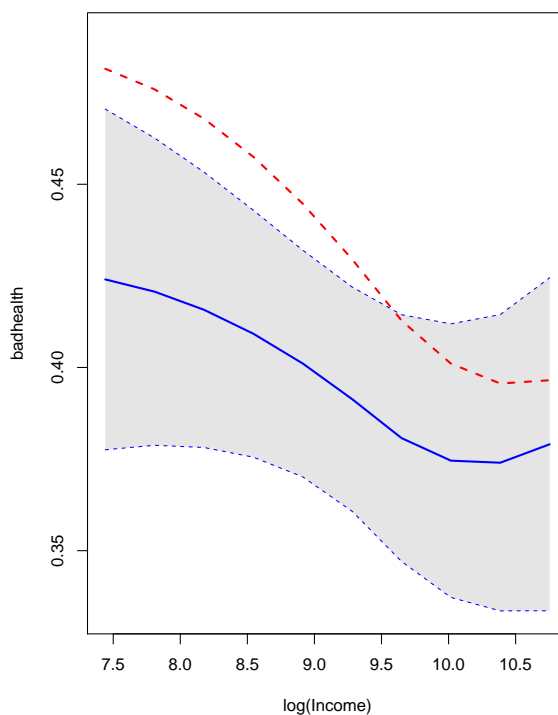


Figure 3: The solid line depicts the FPW series estimator  $\hat{g}$  while the dashed line depicts a MAR series estimator. The range is from 7.46 which is the 5% quantile of observed  $\log Income$  (i.e., the log of 1740 Euros) and 10.78 which is the 95% quantile of observed  $\log Income$  (i.e., the log of 48000 Euros).

to the confidence interval this difference is not significant. For higher incomes the risk only changes moderately and changes are again not statistically different. Importantly, our analysis shows that the MAR assumption leads to biased results and potentially erroneous conclusions about the association between income and health. The negative association obtained in the MAR estimator is far more pronounced than in the estimator which accounts for the nonrandom nonresponse. Specifically, with missing at random we predict a risk of bad health at the 25th percentile of about 44% which is only close to 40% at the 75th percentile of the income distribution. Note, the confidence intervals show that the results of the two different estimators are significantly different for incomes below the median. For higher incomes the differences are not significantly different.

In addition to the non-linear analysis, we assume a linear  $g$  function and present results from a linear model. This linear specification has been used in the literature to test the absolute income hypothesis derived in Preston [1975], see e.g. Adeline and Delattre [2017].

In Table 3 we depict the results using ordinary least squares estimators with and without probability weighting to account for nonrandom nonresponse. For the FPW estimator we leave the functional form of the selection probability completely unrestricted. Overall, this application underlines the importance to account for nonrandom nonresponse in income information when studying the link between income and health. Importantly, while the MAR finds a negative relation between bad health

	PROBABILITY OF BAD HEALTH		
	FPW	MAR	IPW
Constant	-0.104 (0.078)	0.004 (0.085)	-0.164** (0.066)
log(Income)	-0.009 (0.006)	-0.017** (0.007)	-0.006 (0.005)
Gender	-0.382*** (0.013)	-0.413*** (0.014)	-0.395*** (0.013)
Age	0.009*** (0.001)	0.009*** (0.001)	0.010*** (0.001)

Table 3: OLS results for FPW, MAR and IPW. Note: \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$

and income which is significant at the 5% level, the estimators which account for nonrandom nonresponse reject a significant relation between bad health and income in a linear model. Finally, we note that overall the FPW estimator does not lead to larger standard errors relative to MAR, even if the selection probability is estimated via nonparametric instrumental variable method.

## 5.2. Application II: earnings and the demand for housing

In the second application we revisit the question how income affects the demand for housing, for previous studies see e.g. Quigley and Raphael [2004], Albouy et al. [2016] or Dustmann et al. [2018]. For example, Dustmann et al. [2018] show for Germany that about 70% of households in the lowest income quintile are renters whereas in the highest quintile the share is with 30% markedly lower. As in the application of health and income, studies on housing demand are in general based on survey data with self reported income which suffer from nonrandom selection. In the following we estimate the relationship between income and the probability to own a house and quantify the bias when not accounting for nonrandom nonresponse.

The empirical analysis is based on the data of the SOEP. The SOEP is a longitudinal household survey of the German population, for more information see Wagner et al. [2007]. The survey includes self reported standard socio-demographic characteristics including housing and information about different income measures. In contrast to the previous application we focus on a narrow definition of income, labor earnings, which is the most important income component for most individuals. Since a larger fraction of women does not have positive labor earnings, we restrict the analyses to men.

The individual SOEP data can be linked to regional data with information about the average socio-economic situation at the ZIP-code level or even at the residential block. The regional data is provided by a private marketing company which uses administrative information from tax records in combination with credit card infor-

mation and information about local infrastructure, for more details see Goebel et al. [2014]. From the regional data, we use the information about the average purchasing power of households living in a specific residential block to construct an instrument for potentially non-random missings of the self-reported earnings information. This information is well suited to construct an instrument. First there exists a strong positive correlation (0.3025) between the individual labor earning and the average purchasing power of households living in a specific residential block. Second, the regional information is available for all individuals such that the instrument does by definition not suffer from nonresponse.

In the empirical analysis we concentrate on 11735 employed men which are younger than 65 years. Table 4 provides summary statistics of the relevant variables for the analysis.

	1st Qu.	Median	Mean	3rd Qu.
Rate of home owner	0	0	0.45	1
Log Gross Earnings per Month	7.49	7.88	7.76	8.25
Age in years	32.00	41.00	40.29	49.00

Table 4: This table provides summary statistics of the relevant variables for the analysis. The sample includes 11735 individuals which are aged between 16 and 65 years.

To quantify the association between earnings and home ownership we use again a semiparametric model with a binary indicator of home ownership and age as depend variable and log monthly earnings as explanatory variable.

$$Homeowner_i = g(\log(Earning_i^*)) + \alpha_0 Age_i + U_i, \quad (5.2)$$

where the function  $g$  and parameter  $\alpha_0$  are unknown.  $U_i$  is conditional mean independent of the explanatory variables,  $\log(Earning_i^*)$  and  $Age_i$ .

Figure 4 depicts the FPW series estimator with the 95% uniform confidence bands (again described in Section 3.4 using 1000 bootstrap iterations) together with the MAR series estimator evaluated at the median age. The selection probability  $\varphi$  is again estimated using the sieve minimum distance procedure in (3.4) with tensor product of cubic B-splines and one knot. We estimate the function  $g$  using the FPW series estimator  $\hat{g}$  given in (3.2) using cubic B-splines with two knots placed at the quantiles of the observed values of earnings. The MAR estimator is also based on cubic B-splines with the same number and placement of knots. The FPW series estimator shows a strong positive but non-linear relation between earnings and home ownership. For example we find, that the probability to own a house amounts to about 30% at monthly earnings below the median earnings in the sample (about 2400 Euros per months). The probability markedly increases to 60% at monthly earnings at the 75% percentile (8.25 log points or about 3800 Euro per months) and further to close to 70% at earnings above 6500 Euros (8.8 log points). Importantly, our analysis shows that the MAR assumption might lead to biased results about the association between earnings and the home ownership rate for individuals with very low earnings, i.e. individuals in the lowest decile, which is a key group for

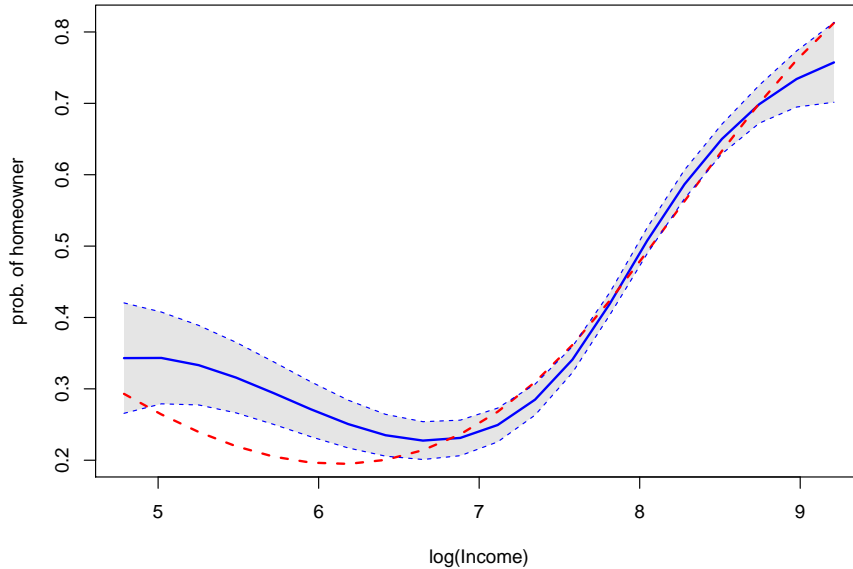


Figure 4: The solid line depicts the FPW series estimator  $\hat{g}$  while the dashed line depicts a MAR series estimator. The uniform confidence bands are based on 1000 bootstrap iterations. For the MAR series estimator, we consider listwise deletion of missing values.

public policy. Specifically, we find significantly lower probabilities of home ownerships when not accounting for nonrandom nonresponse. The difference is with about 10 percentage points economically important. Interestingly, for the rest of the earnings distribution the two estimators do not significantly differ and point at a positive relationship between earnings and home ownership

Finally, we assume again a linear  $g$  function and consider a linear model (see Table 5). Since the missing at random assumption only affected a small part of the earnings distribution in the nonlinear application, it is not surprising that the coefficients in the linear model do not significantly differ. As expected we find a slightly larger point estimator (0.127) when assuming that nonresponse is random.

## 6. Conclusion

In this paper we derive a nonparametric estimator that addresses the problem of nonrandom selection that can be related to nonrandom selection into treatment programs, selective measurement error or through selective nonresponse or missingness of data. Identification of the regression function relies on instrumental variables that are independent of selection conditional on potential covariates. We obtain identification of our nonparametric regression function without restricting the selection probability to belong to a parametric class of functions via a novel partial completeness assumption and provide primitive conditions for it. We achieve optimal rates of nonparametric rates of convergence of our estimator. Moreover, the variance of our estimator is not larger than in the case where the variables are fully observed.

We demonstrate the usefulness and relevance of our method in survey data with



PROBABILITY OF HOME OWNERSHIP			
	FPW	MAR	IPW
Constant	-0.787*** (0.041)	-0.853*** (0.043)	-0.766*** (0.042)
log(Earnings)	0.118*** (0.006)	0.127*** (0.006)	0.117*** (0.006)
Age	0.007*** (0.001)	0.008*** (0.001)	0.008*** (0.001)

Table 5: OLS results for FPW, MAR and IPW. Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

nonrandom missingness in two different applications with different instruments. First, we analyze the association between bad health and income. We show that standard methods that do not account for the nonrandom selection process are strongly upward biased for individuals with below-median income. Moreover, in a linear model the standard estimator finds a negative relation which is significant at the 5% level, however the estimators which account for nonrandom non-response reject a significant relation between bad health and income. In the second application we focus on the relation between housing and earnings. While the different estimators with and without the assumption of random missingness lead to the similar estimates in a linear model and for a large share of the earnings distribution, we document significant and important differences for individuals in the lowest earnings decile, which is a central group for public policy.

## A. Appendix

PROOF OF PROPOSITION 2.1. Consider the functional form restriction (2.1). The conditional mean restriction  $\mathbb{E}[\phi_1(Y)\phi_2(X^*)|Y, W] = 1$  and the conditional independence assumption  $f_{X^*|Y,W} = f_{X^*|W}$  yield

$$\mathbb{E}[\phi_2(X^*)|W = w] - \frac{1}{\phi_1(y)} = 0 \text{ and } \mathbb{E}[\phi_2(X^*)|W = w] - \frac{1}{\phi_1(y')} = 0$$

for all  $y, y'$ . Subtracting both equations gives  $\phi_1(y) = \phi_1(y')$  for all  $y, y'$  in the support of  $Y$ .

Consider the functional form restriction (2.2). From  $\mathbb{E}[\psi(\phi_1(Y) + \phi(X^*))|Y, W] = 1$  and the restriction  $f_{X^*|Y,W} = f_{X^*|W}$  we infer

$$\mathbb{E}[\psi(\phi_1(y) + \phi_2(X^*))|W = w] - \mathbb{E}[\psi(\phi_1(y') + \phi_2(X^*))|W = w] = 0$$

for all  $y, y'$  in the support of  $Y$ . Given  $y'$ , we denote by  $\underline{j} \geq 1$  the smallest integer such that  $\mathbb{E}[\psi^{(\underline{j})}(\phi_1(y') + \phi_2(X^*))|W = w] \neq 0$  for some  $w$ . Taylor series expansion

applied to the analytic function  $\psi$  and Fubini's theorem yield

$$\sum_{j \geq 1} \frac{\mathbb{E}[\psi^{(j)}(\phi_1(y') + \phi_2(X^*)) | W = w]}{j!} (\phi_1(y) - \phi_1(y'))^j = 0$$

Consequently, we have  $\phi_1(y) = \phi_1(y')$  and, since the choice of  $(y, y')$  is arbitrary, the function  $\phi_1$  is constant, which completes the proof.  $\square$

**PROOF OF THEOREM 2.2.** For the proof of the result, we proceed in two steps. **Step 1:** We show that the regression function  $g$  satisfies equation (2.3). Making use of relation  $f_{Y|X^*} = f_{YD|X^*}/f_{D|YX^*}$  and  $f_{YD|X^*} = f_{Y|DX^*}f_{D|X^*}$  we obtain

$$\begin{aligned} g(x) &= \int y f_{Y|X^*}(y|x) dy \\ &= \int y \underbrace{\frac{f_{D|X^*}(1|x)}{f_{D|YX^*}(1|y, x)}}_{=\omega(y, x)} f_{Y|DX^*}(y|1, x) dy \\ &= \mathbb{E} \left[ Y \omega(Y, x) \mid D = 1, X^* = x \right] \end{aligned} \quad (\text{A.1})$$

using the definition of the FPW function  $\omega$  as given in (2.4).

**Step 2:** We show that the FPW function  $\omega$  in (2.4) is identified. Consider the identified set of functions given by

$$\mathcal{I} = \left\{ \phi \in \mathcal{B} : \mathbb{E} [D/\phi(Y, X^*) | Y, W] = 1 \right\}.$$

Clearly, for the true selection probability  $\varphi(y, x) = \mathbb{P}(D = 1 | Y = y, X^* = x)$  we have  $\varphi \in \mathcal{I}$  by Assumption 3 which ensures that  $\varphi$  is uniformly bounded from below. Assumption 1 implies for any function  $\phi \in \mathcal{I}$  that

$$\mathbb{E} \left[ \frac{\varphi(Y, X^*)}{\phi(Y, X^*)} - 1 \mid Y, W \right] = 0.$$

Partial completeness, see Assumption 2, yields

$$\frac{\varphi(Y, X^*)}{\phi(Y, X^*)} - 1 = \psi(X^*)$$

for some unknown function  $\psi$  and thus,

$$\varphi(Y, X^*) = \phi(Y, X^*)(\psi(X^*) + 1). \quad (\text{A.2})$$

Employing relation (A.1) with  $Y\omega(Y, x)$  replaced by  $\omega(Y, x)$ , we obtain

$$1 = \mathbb{E} \left[ \omega(Y, x) \mid D = 1, X^* = x \right]$$

and hence, by the definition of FPW function  $\omega$ , the conditional probability  $\mathbb{P}(D = 1 | X^* = x)$  satisfies

$$\mathbb{P}(D = 1 | X^* = x) = \left( \mathbb{E} \left[ \frac{1}{\varphi(Y, x)} \mid D = 1, X^* = x \right] \right)^{-1}. \quad (\text{A.3})$$

Consequently, we obtain for the FPW function  $\omega$  that

$$\begin{aligned}
\omega(y, x) &= \frac{\mathbb{P}(D = 1 | X^* = x)}{\varphi(y, x)} && \text{(due to Definition in equation (2.4))} \\
&= \left( \mathbb{E} \left[ \frac{\varphi(y, x)}{\varphi(Y, x)} \middle| D = 1, X^* = x \right] \right)^{-1} && \text{(due to equation (A.3))} \\
&= \left( \mathbb{E} \left[ \frac{\phi(y, x)(\psi(x) + 1)}{\phi(Y, x)(\psi(x) + 1)} \middle| D = 1, X^* = x \right] \right)^{-1} && \text{(due to equation (A.2))} \\
&= \left( \mathbb{E} \left[ \frac{\phi(y, x)}{\phi(Y, x)} \middle| D = 1, X^* = x \right] \right)^{-1}
\end{aligned}$$

for all  $\phi \in \mathcal{I}$ . This shows identification of the FPW function  $\omega$  which thus completes the proof.  $\square$

**Additional Notation** For ease of notation, let  $V_i = (Y_i, X_i)$  for  $1 \leq i \leq n$ . Let  $\mathcal{F}$  be a class of measurable functions with a measurable envelope function  $F$ . Then  $N(\varepsilon, \mathcal{F}, \|\cdot\|_V)$  and  $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|_V)$ , respectively, denote the covering and bracketing numbers for the set  $\mathcal{F}$ . The bracketing integral of  $\mathcal{F}$  is denoted by

$$J_{[]}(\mathcal{F}, L_V^2) = \int_0^1 \sqrt{1 + \log N_{[]}(\varepsilon \|F\|_V, \mathcal{F}, \|\cdot\|_V)} d\varepsilon.$$

For ease of notation we write  $\sum_i$  for  $\sum_{i=1}^n$  and  $\sum_{D_i}$  for  $\sum_{i=1, D_i=1}^n$ . We further define  $\widehat{Q} = n^{-1} \sum_{D_i} p^K(X_i) p^K(X_i)'$  and  $\widehat{h}(x, \phi) = p^K(x)' (n\widehat{Q})^{-1} \sum_{D_i} p^K(X_i) / \phi(V_i)$ . We thus have that the estimator of the FPW function coincides with  $\widehat{\omega}(v, \phi) = (\phi(y, x) \widehat{h}(x, \phi))^{-1}$ . Further, let  $\widehat{Q}(\phi) = n^{-1} \sum_{D_i} \widehat{\omega}(V_i, \phi) p^K(X_i) p^K(X_i)'$ . By Assumption 4, the eigenvalues of  $\mathbb{E}[D p^K(X) p^K(X)']$  are bounded away from zero and hence, it may be assumed that  $\mathbb{E}[D p^K(X) p^K(X)'] = I_K$ , where  $I_K$  denotes the  $K \times K$  identity matrix. We also denote  $\gamma = \mathbb{E}[D g(X) p^K(X)]$ . Throughout the proofs, we use the notation  $a_n \lesssim b_n$  to denote  $a_n \leq C b_n$  for some constant  $C > 0$  and for all  $n \geq 1$ .

**PROOF OF THEOREM 3.1.** The proof is based on the upper bound

$$\|\widehat{g} - g\|_X \leq \|\widehat{g} - \gamma' p^K\|_X + \|\gamma' p^K - g\|_X.$$

Since  $\|\gamma' p^K - g\|_X = O(K^{-\alpha/d_x})$  by Assumption 4 (iv) it is sufficient to consider the first term on the right hand side. We observe

$$\begin{aligned}
\|\widehat{g} - \gamma' p^K\|_X^2 &\lesssim \|\widehat{Q}(\widehat{\varphi})^{-1}\|^2 \|I_K - \widehat{Q}(\widehat{\varphi})\|^2 \left\| n^{-1} \sum_{D_i} p^K(X_i) (Y_i - \gamma' p^K(X_i)) \widehat{\omega}(V_i, \widehat{\varphi}) \right\|^2 \\
&\quad + \left\| n^{-1} \sum_{D_i} p^K(X_i) (Y_i - \gamma' p^K(X_i)) \widehat{\omega}(V_i, \widehat{\varphi}) \right\|^2.
\end{aligned}$$

From Lemma B.1 we deduce  $\|\widehat{Q}(\widehat{\varphi}) - I_K\|^2 = K^2/n$  and thus  $\|\widehat{Q}(\widehat{\varphi})^{-1}\|^2 = 1 + o_p(1)$ .

Consequently, it is sufficient to consider

$$\begin{aligned}
& \left\| n^{-1} \sum_{D_i} (Y_i - \gamma' p^K(X_i)) \widehat{\omega}(V_i, \widehat{\varphi}) p^K(X_i) \right\|^2 \\
& \lesssim \underbrace{\left\| n^{-1} \sum_{D_i} Y_i \left( \widehat{\omega}(V_i, \widehat{\varphi}) - \omega(V_i) \right) p^K(X_i) \right\|^2}_I \\
& \quad + \underbrace{\left\| n^{-1} \sum_{D_i} \gamma' p^K(X_i) \left( \widehat{\omega}(V_i, \widehat{\varphi}) - \omega(V_i) \right) p^K(X_i) \right\|^2}_{II} \\
& \quad + \underbrace{\left\| n^{-1} \sum_{D_i} (Y_i - \gamma' p^K(X_i)) \omega(V_i) p^K(X_i) \right\|^2}_{III}.
\end{aligned}$$

Consider  $I$ . We have

$$\begin{aligned}
I &= \left\| n^{-1} \sum_i Y_i D_i \left( \frac{1}{\widehat{\varphi}(V_i) \widehat{h}(X_i, \widehat{\varphi})} - \frac{1}{\varphi(V_i) h(X_i, \varphi)} \right) p^K(X_i) \right\|^2 \\
&\leq \sup_{(\phi, \psi) \in \mathcal{B} \times \mathcal{H}} \left\| n^{-1} \sum_i Y_i D_i \left( \frac{1}{\phi(V_i) \psi(X_i, \phi)} - \frac{1}{\varphi(V_i) h(X_i, \varphi)} \right) p^K(X_i) \right\|^2.
\end{aligned}$$

For  $1 \leq j \leq K$  and  $1 \leq i \leq n$ , we introduce the function

$$h_j(V_i, \phi, \psi) := Y_i D_i \left( \frac{1}{\phi(V_i) \psi(X_i, \phi)} - \frac{1}{\varphi(V_i) h(X_i, \varphi)} \right) p_j(X_i) \quad (\text{A.4})$$

and the class of functions  $\mathcal{F}_j = \{h_j(\cdot, \phi, \psi) : (\phi, \psi) \in \mathcal{B} \times \mathcal{H}\}$ . For all  $(\phi, \psi) \in \mathcal{B} \times \mathcal{H}$  we observe

$$|h_j(V_i, \phi, \psi)| \leq |Y_i D_i p_j(X_i)| \sup_{(\phi, \psi) \in \mathcal{B} \times \mathcal{H}} \left| \frac{1}{\phi(V_i) \psi(X_i, \phi)} - \frac{1}{\varphi(V_i) h(X_i, \varphi)} \right| =: F_j(V_i)$$

and hence,  $F_j$  is an envelope function of the class  $\mathcal{F}_j$ . By the definition of  $\mathcal{B}$ , all functions of  $\mathcal{B}$  and thus also of  $\mathcal{H}$  are uniformly bounded away from zero. Therefore, we obtain the upper bound

$$\|F_j\|_V \lesssim \mathbb{E}[Y^2 D p_j^2(X)] \leq \mathbb{E}[D g^2(X) p_j^2(X)] + \sigma^2 \mathbb{E}[p_j^2(X^*)],$$

for some finite constant  $\sigma > 0$  such that  $\mathbb{E}[D U^2 | X^*] \leq \sigma^2$ . Theorem 2.14.5 of van der Vaart and Wellner [2000] gives

$$\begin{aligned}
& \sum_{j=1}^K \mathbb{E} \left[ \sup_{(\phi, \psi) \in \mathcal{B} \times \mathcal{H}} \left| n^{-1/2} \sum_i h_j(V_i, \phi, \psi) - \mathbb{E} h_j(V, \phi, \psi) \right|^2 \right] \\
& \leq \sum_{j=1}^K \left( \mathbb{E} \sup_{(\phi, \psi) \in \mathcal{B} \times \mathcal{H}} \left| n^{-1/2} \sum_i h_j(V_i, \phi, \psi) - \mathbb{E} h_j(V, \phi, \psi) \right| + \|F_j\|_V \right)^2.
\end{aligned}$$

We further conclude by applying the last display of Theorem 2.14.2 of van der Vaart and Wellner [2000] for  $1 \leq j \leq K$

$$\mathbb{E} \sup_{(\phi, \psi) \in \mathcal{B} \times \mathcal{H}} \left| n^{-1/2} \sum_i h_j(V_i, \phi, \psi) - \mathbb{E} h_j(V, \phi, \psi) \right| \lesssim J_{[]} (1, \mathcal{H}_j, \|\cdot\|_V) \|F_j\|_V. \quad (\text{A.5})$$

Due to Lemma 4.2 (i) of Chen [2007] we have uniformly in  $j$  that

$$\log N_{[]}(\varepsilon, \mathcal{F}_j, \|\cdot\|_V) \leq \log N\left(\frac{\varepsilon}{2C}, \mathcal{B}, \|\cdot\|_V\right) + \log N\left(\frac{\varepsilon}{2C}, \mathcal{H}, \|\cdot\|_X\right).$$

Following Chen et al. [2003, Remark 3],  $\log N(\delta, \mathcal{B}, \|\cdot\|_\infty) \leq C\delta^{-d/\alpha}$  and hence, we obtain  $\int_0^1 \sqrt{\log N(\varepsilon, \mathcal{B}, \|\cdot\|_\infty)} d\varepsilon < \infty$  as long as  $\alpha > d/2$ . Similarly, we conclude  $\int_0^1 \sqrt{\log N(\varepsilon, \mathcal{H}, \|\cdot\|_\infty)} d\varepsilon < \infty$ . Now Assumption 4 (v) together with the inequality  $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$  for  $a, b \geq 0$  implies  $J_{[]}(\varepsilon, \mathcal{F}_j, L_V^2) < \infty$  uniformly in  $j$ . Consequently, we have

$$\begin{aligned} I &\lesssim n^{-1} \sum_{j=1}^K \left( \mathbb{E} \sup_{(\phi, \psi) \in \mathcal{B} \times \mathcal{H}} \left| n^{-1/2} \sum_i h_j(V_i, \phi, \psi) - \mathbb{E} h_j(V, \phi, \psi) \right| + \|F_j\|_V \right)^2 + \|F_j\|_V^2 \\ &\lesssim n^{-1} \sum_{j=1}^K \|F_j\|_V^2 \\ &\lesssim n^{-1} \sup_{x \in \mathcal{X}} \|p^K(x)\|^2 \left( \|g\|_X^2 + \sigma^2 \right) \\ &= O_p(K/n). \end{aligned}$$

Consider  $II$ . It follows  $II = O_p(K/n)$  similarly to the upper bound for the term  $I$  by making use of the following inequality

$$\begin{aligned} \sum_{j=1}^K \mathbb{E} |\gamma' p^K(X) p_j(X)|^2 &\leq \sup_{x \in \mathcal{X}} \|p^K(x)\|^2 \|\gamma' p^K\|_X^2 \\ &= \sup_{x \in \mathcal{X}} \|p^K(x)\|^2 \|\gamma\|^2 \\ &= O(K), \end{aligned}$$

where the last equality is due to Assumption 4 (ii). Consider  $III$ . From Corollary 2.3 we deduce

$$\begin{aligned} \gamma &= \mathbb{E}[Dg(X) p^K(X)] \\ &= \mathbb{E}[Dg(X^*) p^K(X^*) \mathbb{E}[\omega(Y, X^*) | D, X^*]] \\ &= \mathbb{E}[DY p^K(X^*) \omega(Y, X^*)] + \underbrace{\mathbb{E}[D(g(X^*) - Y) p^K(X^*) \omega(Y, X^*)]}_{=0} \\ &= \mathbb{E}[DY \omega(V) p^K(X)] + \underbrace{\mathbb{E}[(g(X^*) - Y) p^K(X^*) \mathbb{P}(D = 1 | X^*)]}_{=0} \\ &= \mathbb{E}[DY \omega(V) p^K(X)] \end{aligned}$$

and consequently

$$\begin{aligned} \mathbb{E}[D(Y - \gamma' p^K(X)) \omega(V) p^K(X)] &= \gamma - \underbrace{\mathbb{E}[Dp^K(X) \omega(V) p^K(X)']}_{=I_K} \gamma \\ &= 0. \end{aligned}$$

Using that the FPW function  $\omega$  is uniformly bounded from above, we thus conclude

$$\begin{aligned}
\mathbb{E}III &= n^{-1} \mathbb{E} \|D(Y - \gamma'p^K(X))\omega(V)p^K(X)\|^2 \\
&\leq 2n^{-1} \sup_{x \in \mathcal{X}} \|p^K(x)\|^2 \sup_v |\omega(v)|^2 \left( \mathbb{E}|D(Y - g(X))|^2 + \mathbb{E}|D(g(X) - \gamma'p^K(X))|^2 \right) \\
&\leq 2n^{-1} \sup_{x \in \mathcal{X}} \|p^K(x)\|^2 \sup_v |\omega(v)|^2 \left( \text{Var}(U) + \|g - \gamma'p^K\|_X^2 \right) \\
&\lesssim n^{-1} K (1 + K^{-2\alpha/d_x}),
\end{aligned}$$

again using Assumption 4 (ii) and the approximation error imposed in Assumption 4 (iv), which completes the proof.  $\square$

**PROOF OF THEOREM 3.2.** We make use of the lower bound of the sieve variance given by

$$\begin{aligned}
v_K(x) &= p^K(x)' \mathbb{E} \left[ p^K(X) \text{Var}(UD\omega(V)|X) p^K(X)' \right] p^K(x) \\
&\gtrsim \|p^K(x)\|^2,
\end{aligned}$$

which is due to the condition that  $\text{Var}(UD\omega(V)|X)$  is uniformly bounded from below, see Assumption 5. The proof is based on the relationship

$$\begin{aligned}
\widehat{g}(x) - \gamma'p^K(x) &= \underbrace{p^K(x)' \widehat{Q}(\widehat{\varphi})^{-1} \frac{1}{n} \sum_{D_i=1} p^K(X_i) U_i \omega(V_i)}_I \\
&\quad + \underbrace{p^K(x)' \widehat{Q}(\widehat{\varphi})^{-1} \frac{1}{n} \sum_{D_i=1} p^K(X_i) U_i (\widehat{\omega}(V_i, \widehat{\varphi}) - \omega(V_i))}_II \\
&\quad + \underbrace{p^K(x)' \widehat{Q}(\widehat{\varphi})^{-1} \frac{1}{n} \sum_{D_i=1} p^K(X_i) (g(X_i) - \gamma'p^K(X_i)) \omega(V_i)}_III,
\end{aligned}$$

where we evaluate each summand on the right hand side separately. Consider *I*. Following the proof of Theorem 3.1 we obtain

$$\begin{aligned}
\sqrt{n/v_K(x)} I &= \sum_{D_i} (n v_K(x))^{-1/2} p^K(x)' p^K(X_i) U_i \omega(V_i) + o_p(1) \\
&= \sum_{D_i} s_{in} + o_p(1).
\end{aligned}$$

In the following, we show that  $s_{in}$ ,  $1 \leq i \leq n$ , satisfy the Lindeberg conditions for the CLT. First, note that  $\mathbb{E}[s_{in}] = 0$  we observe

$$\begin{aligned}
\mathbb{E}[s_{in}] &= \mathbb{E}[DU\omega(V)p^K(X)] \\
&= \mathbb{E}[DU\omega(Y, X^*)p^K(X^*)] \\
&= \mathbb{E}[\mathbb{P}(D = 1|Y, X^*) U\omega(Y, X^*)p^K(X^*)] && \text{(since } U = Y - g(X^*)\text{)} \\
&= \mathbb{E}[U \mathbb{P}(D = 1|X^*)p^K(X^*)] && \text{(due to definition of FPW function } \omega\text{)} \\
&= 0, && \text{(using that } \mathbb{E}[U|X^*] = 0\text{).}
\end{aligned}$$

Thus,  $s_{in}$ ,  $1 \leq i \leq n$ , are centered variables and by the definition of  $v_K(x)$  we have  $n\mathbb{E}[s_{in}^2] = 1$ . Moreover, for all  $\delta > 0$  we observe

$$\begin{aligned} \sum_{D_i} \mathbb{E}[s_{in}^2 \mathbb{1}\{|s_{in}| > \delta\}] &= n\delta^2 \mathbb{E}[|s_{in}/\delta|^2 \mathbb{1}\{|s_{in}/\delta| > 1\}] \\ &\leq n\delta^2 \mathbb{E}|s_{in}/\delta|^4 \\ &\leq Cn^{-1}\delta^{-2}K^2 \mathbb{E}|DU\omega(Y, X^*)|^4 \\ &= o(1) \end{aligned}$$

due to the fourth moments condition imposed in Assumption 5 and the rate condition  $K^2 = o(n)$ . The Lindeberg-Feller CLT thus implies  $\sum_i s_{in} \xrightarrow{d} \mathcal{N}(0, 1)$ . Consider *II*. Recall that  $\omega(V, \varphi) = (\varphi(V)h(X, \varphi))^{-1}$  is identified due to Theorem 2.2. Given consistency of the estimator  $\widehat{\varphi}(\cdot)\widehat{h}(\cdot, \widehat{\varphi})$  it is sufficient to consider the shrinking function class  $\mathcal{A}_n = \{(\phi, \psi) \in \mathcal{B} \times \mathcal{H} : \|\phi(\cdot)\psi(\cdot, \phi) - \varphi(\cdot)h(\cdot, \varphi)\|_\infty \leq r_n\}$  with  $r_n = o(1)$ . Recall the definition of  $h_j$ ,  $1 \leq j \leq K$ , in (A.4) and thus

$$II \leq \sum_{j=1}^K \sup_{(\phi, \psi) \in \mathcal{A}_n} \left| \frac{1}{n} \sum_i p_j(x) h_j(V_i, \phi, \psi) \right| + o_p(\sqrt{v_K(x)}).$$

For all  $(\phi, \psi) \in \mathcal{A}_n$  we observe

$$|h_j(V_i, \phi, \psi)| \leq |Y_i p_j(X_i)| \sup_{(\phi, \psi) \in \mathcal{A}_n} \left| \frac{1}{\phi(V_i)\psi(X_i, \phi)} - \frac{1}{\varphi(V_i)h(X_i, \varphi)} \right| =: F_{jn}(V_i)$$

and hence,  $F_{jn}$  is an envelope function of the class  $\mathcal{F}_{jn} = \{h_j(\cdot, \phi, \psi) : (\phi, \psi) \in \mathcal{A}_n\}$ . In particular, using that  $\varphi(\cdot)h(\cdot, \varphi)$  is uniformly bounded from below (by the definition of the function class  $\mathcal{B}$ ) we obtain by employing inequality (A.5) that

$$\begin{aligned} \sqrt{n}\mathbb{E}|II| &\lesssim \sum_{j=1}^K \sqrt{\mathbb{E}|p_j(x)F_{jn}(V)|^2} \\ &\lesssim \sup_{(\phi, \psi) \in \mathcal{A}_n} \|\phi(\cdot)\psi(\cdot, \phi) - \varphi(\cdot)h(\cdot, \varphi)\|_\infty \sum_{j=1}^K \sqrt{\mathbb{E}|p_j(x)p_j(X)|^2} \\ &\lesssim r_n \sum_{j=1}^K \sqrt{\mathbb{E}|p_j(x)p_j(X)|^2} \\ &= r_n \sum_{j=1}^K |p_j(x)| \\ &= o(\sqrt{v_K(x)}), \end{aligned}$$

where the last bound is due to Assumption 5 (ii). Consider *III*. Using  $\mathbb{E}[\omega(V)|X^*, D = 1] = 1$  we obtain  $\mathbb{E}[p^K(X)D(g(X) - \gamma'p^K(X))\omega(V)] = 0$ . We thus have

$$\begin{aligned} \sqrt{n} III &= n^{-1/2} \|p^K(x)\| \sqrt{\mathbb{E}\|p^K(X)(g(X) - \gamma'p^K(X))\|^2} \times o_p(1) \\ &= O_p\left(n^{-1/2}K\|g - \gamma'p^K\|_\infty\right) \\ &= o_p(1). \end{aligned}$$



Consequently, condition (3.6) implies  $\sqrt{n/v_K(x)}(\hat{g}(x) - g(x)) \xrightarrow{d} \mathcal{N}(0, 1)$ . The result follows by Lemma B.2 which establishes consistency of the sieve variance  $\hat{v}_K(x)$ .  $\square$

PROOF OF THEOREM 3.3. Due to the Chen and Christensen [2018, Proof of Theorem 4.1] it is sufficient to show

$$\left| \sqrt{n/\hat{v}_K(x)}(\hat{g}(x) - g(x)) - \mathbb{Z}(x) \right| = o_p(r_n)$$

since then the result follows by the anti-concentration inequality of Chernozhukov et al. [2014, Theorem 2.1]. We denote  $Z^n = \{(D_1, Y_1, X_1, W_1), \dots, (D_n, Y_n, X_n, W_n)\}$ . **Step 1.** We start by showing that  $\sqrt{n/\hat{v}_K(x)}(\hat{g}(x) - g(x))$  can be uniformly approximated by the process

$$\hat{\mathbb{Z}}(x) = \frac{p^K(x)'}{\sqrt{nv_K(x)}} \sum_{D_i}^n p^K(X_i) U_i \omega(V_i).$$

We observe

$$\begin{aligned} & \left| \sqrt{n/\hat{v}_K(x)}(\hat{g}(x) - g(x)) - \hat{\mathbb{Z}}(x) \right| \\ & \leq \underbrace{\left| \frac{\sqrt{np^K(x)'\hat{Q}(\hat{\varphi})^{-1}}}{\sqrt{v_K(x)}} \sum_{D_i} p^K(X_i) U_i \omega(V_i) - \hat{\mathbb{Z}}(x) \right|}_{I(x)} \\ & + \underbrace{\left| \sqrt{\frac{v_K(x)}{\hat{v}_K(x)}} - 1 \right| \left( \left| \frac{\sqrt{np^K(x)'\mathbf{Q}^{-1}}}{\sqrt{v_K(x)}} \frac{1}{n} \sum_{D_i} p^K(X_i) U_i \omega(V_i) \right| \right)}_{II(x)} \\ & + \underbrace{\left| \frac{\sqrt{np^K(x)'\mathbf{Q}^{-1}}}{\sqrt{v_K(x)}} \frac{1}{n} \sum_{D_i} p^K(X_i) U_i (\hat{\omega}(V_i, \hat{\varphi}) - \omega(V_i)) \right|}_{III(x)} \\ & + \underbrace{\left| \frac{\sqrt{np^K(x)'\mathbf{Q}^{-1}}}{\sqrt{v_K(x)}} \frac{1}{n} \sum_{D_i} p^K(X_i) (g(X_i) - \gamma' p^K(X_i)) \omega(V_i) \right|}_{IV(x)} \\ & + \underbrace{\left| \frac{\sqrt{n}}{\sqrt{v_K(x)}} \left( p^K(x)'\mathbf{Q}^{-1} \frac{1}{n} \sum_{D_i} p^K(X_i) \gamma' p^K(X_i) \omega(V_i) - g(x) \right) \right|}_{V(x)}}. \end{aligned}$$

We have  $\|\hat{Q}(\hat{\varphi}) - \mathbf{I}_K\| = O_p(K\sqrt{\log(n)/n})$ , see Lemma B.1. Further, we obtain

$$\begin{aligned} \sup_{x \in \mathcal{C}} I(x) &= \sup_{x \in \mathcal{C}} \left| \frac{p^K(x)'(\hat{Q}(\hat{\varphi})^{-1} - \mathbf{I}_K)}{\sqrt{nv_K(x)}} \sum_{D_i} p^K(X_i) U_i \omega(V_i) \right| \\ &= O_p(K^2 \sqrt{\log(n)/n}). \end{aligned}$$

Define the process  $\mathbb{Z}(x) = p^K(x)'Q^{-1/2}\mathcal{Z}/\sqrt{v_K(x)}$ . We have

$$\begin{aligned} \sup_{x \in \mathcal{C}} II(x) &\leq \sup_{x \in \mathcal{C}} I(x) + \sup_{x \in \mathcal{C}} |\widehat{\mathbb{Z}}(x)| \\ &= O_p\left(K^2 \sqrt{\log(n)/n}\right) + \sup_{x \in \mathcal{C}} |\widehat{\mathbb{Z}}(x) - \mathbb{Z}(x)| + \sup_{x \in \mathcal{C}} |\mathbb{Z}(x)| \\ &= O_p\left(K^2 \sqrt{\log(n)/n}\right) + o_p(r_n) + \sup_{x \in \mathcal{C}} |\mathbb{Z}(x)| \\ &= o_p(r_n) + O_p(c_n). \end{aligned}$$

where the third bound is due to step 2 below and the last equality is because of the condition  $K^{5/2} = o(r_n^3 \sqrt{n})$  and by Chen and Christensen [2018, Lemma G.5], which is valid under our assumptions and which implies  $\sup_{z \in \mathcal{C}} |\mathbb{Z}(z)| = O_p(c_n)$ . Consider  $III(x)$ . Recall the definition of  $h_j$ ,  $1 \leq j \leq K$  given in (A.4), we obtain

$$\begin{aligned} &\sup_{x \in \mathcal{C}} \frac{\sqrt{n} |III(x)|}{\sqrt{v_K(x)}} \\ &= \sup_{x \in \mathcal{C}} \frac{\|p^K(x)' \widehat{Q}(\widehat{\varphi})^{-1}\|}{\sqrt{v_K(x)}} \left( \sup_{(\phi, \psi) \in \mathcal{B} \times \mathcal{H}} \sum_{j=1}^K \left| \frac{1}{\sqrt{n}} \sum_i \psi_j(V_i, \phi, \psi) \right|^2 \right)^{1/2} \\ &= o_p(K^2 \sqrt{\log(n)/n}) \end{aligned}$$

following the proof of Theorem 3.2. Moreover, we observe

$$\begin{aligned} &\sup_{x \in \mathcal{C}} \frac{\sqrt{n} |IV(x)|}{\sqrt{v_K(x)}} \\ &\leq \sup_{x \in \mathcal{C}} \frac{\|p^K(x)' \widehat{Q}(\widehat{\varphi})^{-1}\|}{\sqrt{v_K(x)}} \left\| \frac{1}{\sqrt{n}} \sum_{D_i} p^K(X_i) (g(X_i) - \gamma' p^K(X_i)) \omega(V_i) \right\| \\ &= O_p\left(n^{-1/2} K \|g - \gamma' p^K\|_\infty\right) \end{aligned}$$

again following the proof of Theorem 3.2. For the last summand we note

$$\sup_{x \in \mathcal{C}} \frac{\sqrt{n} |V(x)|}{\sqrt{v_K(x)}} \leq \sup_{x \in \mathcal{C}} \frac{\sqrt{n}}{\sqrt{v_K(x)}} |\gamma' p^K(x) - g(x)|.$$

Consequently, Lemma B.2, i.e.,  $\sup_{x \in \mathcal{C}} \left| \sqrt{v_K(x)/\widehat{v}_K(x)} - 1 \right| = O_p(\sqrt{n^{-1/2} K^{1/2} \log(n)})$  and the rate requirement in Assumption 6 (iii) imply

$$\left| \sqrt{n/\widehat{v}_K(x)} (\widehat{g}(x) - g(x)) - \widehat{\mathbb{Z}}(x) \right| = o_p(r_n).$$

**Step 2.** We have

$$\begin{aligned} &\sum_i \mathbb{E} \left\| \frac{1}{\sqrt{n}} p^K(X_i) U_i \omega(V_i) \right\|^3 \\ &\lesssim \frac{K^{3/2}}{\sqrt{n}}. \end{aligned}$$

Further, recall that  $r_n$  is a sequence satisfying

$$\frac{K^{5/2}}{r_n^3 \sqrt{n}} = o(1).$$

Hence we may apply Yurinskii's coupling (Pollard [2002, Theorem 10]) and consequently, there exists a sequence of  $\mathcal{N}(0, \Sigma)$  distributed random vectors  $\mathcal{Z}$  such that

$$\left\| \frac{1}{\sqrt{n}} p^K(X_i) U_i \omega(V_i) - \mathcal{Z} \right\| = o_p(r_n). \quad (\text{A.6})$$

Recall the definition  $\mathbb{Z}(x) = p^K(x)' Q^{-1/2} \mathcal{Z} / \sqrt{v_K(x)}$ , which is a centered Gaussian process with covariance function

$$\mathbb{E}[\mathbb{Z}(x_1) \mathbb{Z}(x_2)] = p^K(x_1)' Q^{-1/2} \Sigma Q^{-1/2} p^K(x_2) / \sqrt{v_K(x_1) v_K(x_2)}.$$

Hence, by equation (A.6) we have

$$\sup_{x \in \mathcal{C}} \left| \widehat{\mathbb{Z}}(x) - \mathbb{Z}(x) \right| = o_p(r_n). \quad (\text{A.7})$$

**Step 3.** In this step we approximate the bootstrap process by a Gaussian process. Under the bootstrap distribution  $\mathbb{P}^*$  each term  $Y_i \widehat{\omega}(Y_i, X_i; \widehat{\varphi}) - \widehat{g}(X_i)$  has mean zero for all  $1 \leq i \leq n$ . Moreover, we have

$$\frac{1}{n} \sum_{D_i} \mathbb{E} \left[ \widehat{Q}^{-1} p^K(X_i) \left( Y_i \widehat{\omega}(V_i, \widehat{\varphi}) - \widehat{g}(X_i) \right)^2 \varepsilon_i^2 p^K(X_i)' \widehat{Q}^{-1} \middle| Z^n \right] = \widehat{\Sigma}.$$

Since  $\mathbb{E}[|\varepsilon_i|^3 | Z^n] < \infty$  uniformly in  $i$ , we have

$$\begin{aligned} & \sum_i \mathbb{E} \left[ \left\| \frac{1}{\sqrt{n}} p^K(X_i) D_i \left( Y_i \widehat{\omega}(V_i, \widehat{\varphi}) - \widehat{g}(X_i) \right) \varepsilon_i \right\|^3 \middle| Z^n \right] \\ & \lesssim \frac{1}{\sqrt{n}} \mathbb{E} \|p^K(X)\|^2 \sup_x \|p^K(x)\| \\ & \lesssim \frac{K^{3/2}}{\sqrt{n}}, \end{aligned}$$

with probability approaching one (wpa1). Again using Pollard [2002, Theorem 10], conditional on the data  $Z^n$ , implies existence of a  $\mathcal{N}(0, \widehat{\Sigma})$  distributed random vectors  $\mathcal{Z}^*$  such that

$$\left\| \frac{1}{\sqrt{n}} \sum_{D_i} p^K(X_i) \left( Y_i \widehat{\omega}(V_i, \widehat{\varphi}) - \widehat{g}(X_i) \right) - \mathcal{Z}^* \right\| = o_{p^*}(r_n)$$

wpa1. Therefore,

$$\sup_{x \in \mathcal{C}} \left| \mathbb{Z}^B(x) - \frac{p^K(x)' \mathcal{Z}^*}{\sqrt{\widehat{v}_K(x)}} \right| = o_{p^*}(r_n)$$

wpa1. Define a centered Gaussian process  $\tilde{\mathbb{Z}}(\cdot)$  under  $\mathbb{P}^*$  as

$$\tilde{\mathbb{Z}}(x) = p^K(x)' Q^{-1/2} \Sigma^{1/2} \hat{\Sigma}^{-1/2} \mathcal{Z}^* / \sqrt{v_K(x)}$$

which has the same covariance function as  $\mathbb{Z}(x)$ . By Chen and Christensen [2018, Lemma G.6] below we have:

$$\sup_{x \in \mathcal{C}} \left| \frac{p^K(x)' Q^{-1/2} \mathcal{Z}^*}{\sqrt{\hat{v}_K(x)}} - \tilde{\mathbb{Z}}(x) \right| = o_{p^*}(r_n)$$

wpa1. This and the previous rate of convergence imply that

$$\sup_{x \in \mathcal{C}} \left| \mathbb{Z}^B(x) - \tilde{\mathbb{Z}}(x) \right| = o_{p^*}(r_n)$$

wpa1, which completes the proof.  $\square$

## B. Technical Assertions

**Lemma B.1.** *Under the conditions of Theorem 3.1 it holds*

$$\|\hat{Q}(\hat{\varphi}) - Q(\varphi)\| = O_p(K/\sqrt{n}).$$

*Proof.* From Belloni et al. [2015] we deduce

$$\begin{aligned} \|\hat{Q}(\hat{\varphi}) - Q(\varphi)\| &\leq \|\hat{Q}(\hat{\varphi}) - \hat{Q}(\varphi)\| + \|\hat{Q}(\varphi) - Q(\varphi)\| \\ &= \|\hat{Q}(\hat{\varphi}) - \hat{Q}(\varphi)\| + O_p(\sqrt{(\log n)K/n}). \end{aligned}$$

Further, as the spectral norm is bounded by the Frobenius norm we have

$$\|\hat{Q}(\hat{\varphi}) - \hat{Q}(\varphi)\|^2 \leq \sum_{j,l=1}^K \left| \frac{1}{n} \sum_{D_i} \left( \frac{1}{\hat{\varphi}(V_i) \hat{h}(X_i, \hat{\varphi})} - \frac{1}{\varphi(V_i) h(X_i, \varphi)} \right) p_j(X_i) p_l(X_i) \right|^2,$$

where the term on the right hand side is of the order  $O_p(K^2/n)$  which is due to the analysis preceding inequality (A.5).  $\square$

**Lemma B.2.** *Let Assumptions 1–5 be satisfied. Then,*

$$\left| \sqrt{\frac{\hat{v}_K(x)}{v_K(x)}} - 1 \right| = o_p(1), \tag{B.1}$$

$$\sup_x \left| \sqrt{\frac{\hat{v}_K(x)}{v_K(x)}} - 1 \right| = O_p\left(\sqrt{n^{-1/2} K \log(n)}\right). \tag{B.2}$$

*Proof.* Proof of (B.1). Note that it is sufficient to establish  $\hat{v}_K(x) - v_K(x) = o_p(\|p^K(x)\|^2)$ . We make use of the decomposition

$$\hat{v}_K(x) - v_K(x) = p^K(x)' (\hat{\Sigma} - \tilde{\Sigma}) p^K(x) + p^K(x)' (\tilde{\Sigma} - \Sigma) p^K(x) + o_p(1) \tag{B.3}$$

where

$$\tilde{\Sigma} = \frac{1}{n} \sum_{D_i} p^K(X_j) U_i^2 \omega^2(V_i) p^K(X_j)', \quad \hat{\Sigma} = \frac{1}{n} \sum_{D_i} p^K(X_j) \hat{U}_i^2 \hat{\omega}^2(V_i, \hat{\varphi}) p^K(X_j)',$$

and  $\Sigma = \mathbb{E}\tilde{\Sigma}$ . We further calculate

$$\begin{aligned} & |p^K(x)'(\hat{\Sigma} - \tilde{\Sigma})p^K(x)| \\ & \leq \left| p^K(x)' \frac{1}{n} \sum_{D_i} p^K(X_i) \left( \hat{U}_i^2 \hat{\omega}^2(V_i; \hat{\varphi}) - U_i^2 \omega^2(V_i; \varphi) \right) p^K(X_i)' p^K(x) \right| \\ & \leq n^{-1} \underbrace{\sum_{D_i} \left| p^K(x)' \left( \hat{U}_i \hat{\omega}(V_i; \hat{\varphi}) - U_i \omega(V_i; \varphi) \right) p^K(X_i) \right|^2}_I \\ & \quad + 2 \underbrace{\left| p^K(x)' \frac{1}{n} \sum_{D_i} p^K(X_i) \left( \hat{U}_i \hat{\omega}(V_i; \hat{\varphi}) - U_i \omega(V_i; \varphi) \right) U_i \omega(V_i; \varphi) p^K(X_i)' p^K(x) \right|}_{II}. \end{aligned}$$

Consider  $I$ . Using  $\hat{U}_i \hat{\omega}(V_i, \hat{\varphi}) - U_i \omega(V_i) = (\hat{g}(X_i) - g(X_i))\omega(V_i) + (\hat{\omega}(V_i, \hat{\varphi}) - \omega(V_i))\hat{U}_i$  and that  $\omega$  is uniformly bounded from above, we obtain by following the proof of Theorem 3.2 that

$$\begin{aligned} I & \lesssim \mathbb{E} \|p^K(x)' p^K(X)\|^2 \|\hat{g} - g\|_\infty^2 \\ & \quad + \|p^K(x)\|^2 \left( \sum_{j=1}^K \sup_{(\phi, \psi) \in \mathcal{A}_n} \left| n^{-1} \sum_i h_j(V_i, \phi, \psi) - \mathbb{E} h_j(V, \phi, \psi) \right|^2 \right. \\ & \quad \left. + \sum_{j=1}^K \sup_{(\phi, \psi) \in \mathcal{A}_n} |\mathbb{E} h_j(V, \phi, \psi)|^2 \right) \\ & = O_p \left( \|p^K(x)\|^2 (n^{-1} K + \|\gamma' p^K - g\|_\infty^2) \right) \\ & = o_p(\|p^K(x)\|^2), \end{aligned}$$

using that  $\mathbb{E} \|p^K(x)' p^K(X)\|^2 \leq K$  and  $\|\gamma' p^K - g\|_\infty^2 = o(K/n)$ . Again following the proof of Theorem 3.2 and making use of the Cauchy-Schwarz inequality yields

$$\begin{aligned} I & \leq \sqrt{II} \times \sqrt{n^{-1} \sum_i \left| p^K(x)' U_i \omega(V_i) p^K(X_i) \right|^2} \\ & \leq \sqrt{II} \times O_p(\|p^K(x)\| \sqrt{\mathbb{E}[U^2 \omega^2(V)]}) \\ & = O_p \left( \|p^K(x)\|^2 (n^{-1/2} K^{1/2} + \|\gamma' p^K - g\|_\infty) \right) \\ & = o_p(\|p^K(x)\|^2), \end{aligned}$$

using the upper bound of  $II$ .

Finally, we obtain

$$p^K(x)'(\tilde{\Sigma} - \Sigma)p^K(x) = O_p(\|p^K(x)\|^2 K / \sqrt{n})$$

which is due to the following calculation

$$\begin{aligned}
& \mathbb{E}\|\tilde{\Sigma} - \Sigma\|^2 \\
&= \mathbb{E}\left\|n^{-1} \sum_i p^K(X_i) D_i U_i^2 \omega^2(V_i; \varphi) p^K(X_i)' - \mathbb{E}[p^K(X) D U^2 \omega^2(V) p^K(X)']\right\|^2 \\
&\leq n^{-1} \mathbb{E}\left\|p^K(X) D U \omega(V; \varphi)\right\|^4 \\
&\leq n^{-1} \sup_x \|p^K(x)\|^4 \mathbb{E}|D U \omega(V; \varphi)|^4 \\
&\lesssim n^{-1} K^2,
\end{aligned}$$

based on the fourth moment condition imposed in Assumption 5. This establishes consistency of the sieve variance estimator  $\hat{v}_K(x)$  and hence completes the proof.

The result (B.2) follows analogously.  $\square$

## C. Extension to Selectively Missing Outcomes

In many empirical situations, one may also want to control for selective missingness of the dependent variable. Below,  $D^Y$  and  $D^X$  denote missingness indicators for the variables  $Y^*$  and  $X^*$ , respectively. To account for additional selective missingness of  $Y^*$ , we generalize Assumption 1 to the exclusion restriction

$$\mathbb{P}(D^Y = 1, D^X = 1 | Y^*, X^*, W) = \mathbb{P}(D^Y = 1, D^X = 1 | Y^*, X^*). \quad (\text{C.1})$$

The previous exclusion restriction has the interpretation that  $W$  has no information on the selection indicators  $D^Y$  and  $D^X$  that is not captured in  $(Y^*, X^*)$ . In this section, we make use of the notation  $D = D^Y D^X$ . Table 6 depicts the exclusion conditions required in different sample selection scenarios.

	$Y^*$ obs.	$Y^*$ mis.
$X^*$ obs.		$\mathbb{P}(D^Y = 1   Y^*, X, W) = \mathbb{P}(D^Y = 1   Y^*, X)$
$X^*$ mis.	$\mathbb{P}(D^X = 1   Y, X^*, W) = \mathbb{P}(D^X = 1   Y, X^*)$	$\mathbb{P}(D = 1   Y^*, X^*, W) = \mathbb{P}(D = 1   Y^*, X^*)$

Table 6: Exclusion restrictions for instrument  $W$  depending on missingness.

Under the exclusion restriction imposed in equation (C.1) we can write the regression function  $g$  as

$$\begin{aligned}
g(x) &= \mathbb{E}[Y^* | X^* = x] \\
&= \int y f_{Y^* | X^*}(y | x) \frac{f_{D | Y^* X^*}(1 | y, x)}{f_{D | Y^* X^*}(1 | y, x)} dy \\
&= \int y \frac{f_{D^Y | X^*}(1, y | x)}{f_{D | Y^* X^*}(1 | y, x)} dy \\
&= f_{D^X | X^*}(1 | x) \int y \frac{f_{D^Y | Y^* D^X X^*}(1, y | 1, x)}{f_{D | Y^* X^*}(1 | y, x)} dy \\
&= \mathbb{E}\left[\frac{D^Y Y^* \mathbb{P}(D^X = 1 | X^*)}{\mathbb{P}(D = 1 | Y^*, X^*)} \Big| D^X = 1, X^* = x\right],
\end{aligned}$$

where the right hand side is identified as long as the fractional probability weight  $\mathbb{P}(D^X = 1|X^* = x)/\mathbb{P}(D = 1|Y^*, X^* = x)$  is identified. In the case of potentially missing dependent variable  $Y^*$  and fully observed  $X^*$ , i.e.,  $X^* = X$ , we obtain, in particular,

$$g(x) = \mathbb{E} \left[ \frac{D^Y Y^*}{\mathbb{P}(D^Y = 1|Y^*, X)} \middle| X = x \right]$$

as obtained by Breunig et al. [2018] (see also Remark 2.1 when  $X$  is exogenous to selection  $D^Y$ ). Table 7 provides explicit forms of the regression function  $g$  in different sample selection scenarios.

	$Y^*$ observed ( $Y^* = Y$ )	$Y^*$ selectively missing
$X^*$ obs.		$\mathbb{E} \left[ \frac{D^Y Y^*}{\mathbb{P}(D^Y = 1 Y^*, X)} \middle  X = x \right]$
$X^*$ mis.	$\mathbb{E} \left[ \frac{Y \mathbb{P}(D^X = 1 X^*)}{\mathbb{P}(D^X = 1 Y, X^*)} \middle  D^X = 1, X^* = x \right]$	$\mathbb{E} \left[ \frac{D^Y Y^* \mathbb{P}(D^X = 1 X^*)}{\mathbb{P}(D = 1 Y^*, X^*)} \middle  D^X = 1, X^* = x \right]$

Table 7: Identification results for regression function  $g$  under different sample selection scenarios.

## References

- A. Adeline and E. Delattre. Some microeconomic evidence on the relationship between health and income. *Health economics review*, 7(1):27, 2017.
- D. Albouy, G. Ehrlich, and Y. Liu. Housing demand, cost-of-living inequality, and the affordability crisis. *NBER Working Paper 22816*, 2016.
- A. Belloni, V. Chernozhukov, D. Chetverikov, and K. Kato. Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366, 2015.
- P. Bingley and A. Martinello. Measurement error in income and schooling and the bias of linear estimators. *Journal of Labor Economics*, 35(4):1117–1148, 2017.
- R. Blundell, X. Chen, and D. Kristensen. Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, 75(6):1613–1669, 2007.
- R. W. Bohannon. Muscle strength: clinical and prognostic value of hand-grip dynamometry. *Current Opinion in Clinical Nutrition & Metabolic Care*, 18(5):465–470, 2015.
- A. Börsch-Supan, M. Brandt, C. Hunkler, T. Kneip, J. Korbmacher, F. Malter, B. Schaan, S. Stuck, and S. Zuber. Data resource profile: the survey of health, ageing and retirement in europe (share). *International journal of epidemiology*, 42(4):992–1001, 2013.



- C. Breunig. Testing missing at random using instrumental variables. *Journal of Business & Economic Statistics*, 37(2):223–234, 2019.
- C. Breunig, E. Mammen, and A. Simoni. Nonparametric estimation in case of endogenous selection. *Journal of Econometrics*, 202(2):268–285, 2018.
- Q. Chen, D. Zeng, and J. G. Ibrahim. Sieve maximum likelihood estimation for regression models with covariates missing at random. *Journal of the American Statistical Association*, 102(480):1309–1317, 2007.
- X. Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, 2007.
- X. Chen and T. M. Christensen. Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, 188(2):447–465, 2015.
- X. Chen and T. M. Christensen. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression. *Quantitative Economics*, 9(1):39–84, 2018.
- X. Chen and D. Pouzo. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80(1):277–321, 2012.
- X. Chen, O. Linton, and I. Van Keilegom. Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, 71:1591–1608, 2003.
- V. Chernozhukov, D. Chetverikov, K. Kato, et al. Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics*, 42(5):1787–1818, 2014.
- D. Cutler, A. Deaton, and A. Lleras-Muney. The determinants of mortality. *Journal of economic perspectives*, 20(3):97–120, 2006.
- D. M. Cutler, A. Lleras-Muney, and T. Vogl. Socioeconomic status and health: Dimensions and mechanisms. In *The Oxford Handbook of Health Economics*. 2011.
- M. Das, W. K. Newey, and F. Vella. Nonparametric estimation of sample selection models. *The Review of Economic Studies*, 70(1):33–58, 2003.
- A. Deaton and C. Paxson. Aging and inequality in income and health. *American Economic Review: Papers and Proceedings*, 88(2):248–253, 1998.
- X. D’Haultfoeuille. A new instrumental method for dealing with endogenous selection. *Journal of Econometrics*, 154(1):1–15, 2010.
- R. M. Dodds, H. E. Syddall, R. Cooper, M. Benzeval, I. J. Deary, E. M. Dennison, G. Der, C. R. Gale, H. M. Inskip, C. Jagger, et al. Grip strength across the life course: normative data from twelve british studies. *PloS one*, 9(12):e113637, 2014.
- C. Dustmann, B. Fitzenberger, and M. Zimmermann. Housing expenditures and income inequality. *CReAM DP 16/18*, 2018.

- F. Fang, J. Zhao, and J. Shao. Imputation-based adjusted score equations in generalized linear models with nonignorable missing covariate values. *Statistica Sinica*, 28(4):1677–1701, 2018.
- J. Goebel, C. K. Spiess, N. R. Witte, and S. Gerstenber. Die verknuepfung des soep mit microm-indikatoren: Der microm-soep-datensatz. *SOEP Survey Paper 233*, 2014.
- Y. Hu and S. M. Schennach. Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, 76(1):195–216, 2008.
- R. J. Little and D. B. Rubin. *Statistical analysis with missing data*. 2002.
- E. Mammen. Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, pages 255–285, 1993.
- W. K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147 – 168, 1997.
- W. K. Newey and J. L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71:1565–1578, 2003.
- D. Pollard. *A user’s guide to measure theoretic probability*, volume 8. Cambridge University Press, 2002.
- S. Preston. The changing relation between mortality and level of economic development. *Population studies*, 29(2):231–248, 1975.
- J. M. Quigley and S. Raphael. Is housing unaffordable? why isn’t it more affordable? *Journal of Economic Perspectives*, 18(1):191–214, 2004.
- E. A. Ramalho and R. J. Smith. Discrete choice non-response. *The Review of Economic Studies*, 80(1):343–364, 2013.
- T. Rantanen, J. M. Guralnik, D. Foley, K. Masaki, S. Leveille, J. D. Curb, and L. White. Midlife hand grip strength as a predictor of old age disability. *Jama*, 281(6):558–560, 1999.
- H. Schwandt. Wealth shocks and health outcomes: evidence from stock market fluctuations. Cepr discussion paper no. dp12562, 2018.
- G. Tang, R. J. Little, and T. E. Raghunathan. Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika*, 90(4):747–764, 2003.
- A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics (Springer Series in Statistics)*. Springer, corrected edition, Nov. 2000. ISBN 0387946403.
- G. Wagner, J. Frick, and J. Schupp. The German Socio-Economic Panel Study (SOEP) - Scope, Evolution and Enhancements. *Schmollers Jahrbuch*, 127(1):139–169, 2007.
- J. Zhao and J. Shao. Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association*, 110(512):1577–1590, 2015.