Econometric Reviews

Editor: Esfandiar Maasoumi

Included in this print edition:
Number 1

Volume 42 2023

Econometric Reviews

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lecr20

Yet another look at the omitted variable bias

Masayuki Hirukawa, Irina Murtazashvili & Artem Prokhorov

To cite this article: Masayuki Hirukawa, Irina Murtazashvili & Artem Prokhorov (2023) Yet another look at the omitted variable bias, Econometric Reviews, 42:1, 1-27, DOI: 10.1080/07474938.2022.2157965

To link to this article: https://doi.org/10.1080/07474938.2022.2157965

ECONOMETRIC REVIEWS
2023, VOL. 42, NO. 1, 1–27
https://doi.org/10.1080/07474938.2022.2157965

Yet another look at the omitted variable bias

Masayuki Hirukawa[a], Irina Murtazashvili[b], and Artem Prokhorov[c]

[a]Ryukoku University, Kyoto, Japan; [b]Drexel University, Philadelphia, PA, USA; [c]University of Sydney, Sydney, NSW, Australia; CEBA, St. Petersburg State University, St. Petersburg, Russia; CIREQ, Université de Montréal, Montreal, QC, Canada

ABSTRACT

When conducting regression analysis, econometricians often face the situation where some relevant regressors are unavailable in the data set at hand. This article shows how to construct a new class of nonparametric proxies by combining the original data set with one containing the missing regressors. Imputation of the missing values is done using a nonstandard kernel adapted to mixed data. We derive the asymptotic distribution of the resulting semiparametric two-sample estimator of the parameters of interest and show, using Monte Carlo simulations, that it dominates the solutions involving instrumental variables and other parametric alternatives. An application to the PSID and NLS data illustrates the importance of our estimation approach for empirical research.

KEYWORDS
Data combination; endogeneity; instrumental variable estimation; omitted variable bias; two-sample estimation

JEL CLASSIFICATION CODES
C13; C14; C31

1. Introduction

Omission of relevant variables leads to challenging problems in applied work. If the correlation between the omitted variable and included regressors is strong, least squares estimates are biased and inconsistent. This issue has a long history in economics.

A classic example of the missing regressor is an ability measure in Mincer's (1974) wage regression, where estimates suffer from the so called "ability bias" unless the regressors include a variable representing ability (see, e.g., Card, 1995, for details). Micro-level data sets such as the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID) do not generally contain individual test scores that can be used as (a proxy for) ability.

Another example can be found in the study of gender wage gap (e.g., Black et al., 1999; Zabalza and Arrufat, 1985). Work experience is an important regressor in the wage regression. Although the General Household Survey (GHS) and CPS contain wages and other predictors, they do not include a variable for actual work experience.

This article argues that availability of a suitable auxiliary sample offers an underexploited opportunity. We use the sample to construct a new semiparametric estimator and we study its properties. In a nutshell, our two-sample estimation procedure is to run the ordinary least squares (OLS) after replacing the missing regressor with a nonparametric estimate of its conditional mean, obtained using a nonstandard kernel function tailored for both discrete and continuous data.

Our approach is different from but not unrelated to the work on matched-sample estimators (see, e.g., Abadie and Imbens, 2006, 2011; Hirukawa and Prokhorov, 2018). One key advantage of

CONTACT Masayuki Hirukawa ✉ hirukawa@econ.ryukoku.ac.jp Faculty of Economics, Ryukoku University, 67 Tsukamoto-cho, Fukakusa, Fushimi-ku, Kyoto 612-8577, Japan.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/07474938.2022.2157965.

our estimator is that the imputed proxy does not rely on a combined sample constructed via the nearest-neighbor matching (NNM) or another ad hoc matching scheme. Matching-free estimation is also the purpose of some well-known moment-based two-sample methods (e.g., Angrist and Krueger, 1992, 1995; Arellano and Meghir, 1992; Inoue and Solon, 2010; Klevmarken, 1982; Murtazashvili et al., 2015; Pacini and Windmeijer, 2016). Unfortunately, these approaches are not applicable in the setting of a linear regression where some regressors are missing in the primary sample, and two-sample moment-based estimation is infeasible. Our setup is closer to what is considered by Pacini (2019), Chen et al. (2008) and Graham et al. (2016), who also use an auxiliary sample to impute missing regressors. However, Pacini's (2019) main focus is on relaxing identification conditions for two-sample estimation, while Chen et al. (2008) and Graham et al. (2016) study parametric and semi-parametric efficiency bounds for GMM.

We establish large-sample properties of the proposed semiparametric two-sample estimator. Its finite-sample properties are examined through extensive Monte Carlo simulations (largely deferred to an online supplement to save space), where we show that our estimator dominates the competitors, in particular, instrumental variable (IV) estimators using linear and nonlinear instruments and a parametric two-sample alternative. In an empirical application, we look for a setting where we are able to obtain estimates based on both the IV estimator and proposed estimator at the same time. We do so for the return to schooling and offer new insights arising from the use of IVs based on one sample as well as from the use of proxies based on two samples.

The nonparametric imputation we propose is important in several practical dimensions. First, having access to a second sample, empirical economists commonly impute the variables that are necessary for estimation but missing in the first sample. For example, Fang et al. (2008) impute medical expenditure of Health and Retirement Study (HRS) respondents using the information from the Medicare Current Beneficiary Survey (MCBS), and Flavin and Nakagawa (2008) impute the square footage of the homes of PSID respondents using data on housing from the American Housing Survey (AHS). Instead of estimating the conditional mean of the missing variable nonparametrically, these papers adopt a linear projection-based imputation. As a matter of course, if the functional form turns out to be incorrectly specified, then consistency of the estimators can no longer be guaranteed. Furthermore, estimation errors associated with imputation need to be explicitly accounted for, as imputed values cannot be treated as error-free and adjustments need to be made to the standard errors. A remarkable observation is that our nonparametric imputation approach performs no worse even under correct parametric specification. The result on inconsistency of the linear projection approach and the circumstances under which it still works may be of independent interest. We explore this in the Monte Carlo simulations of this article and in the online supplement.

Second, our nonparametric imputation method is important in the context of the literature on two-step semiparametric regression estimation with a plug-in first-step nonparametrically generated regressor (see, e.g., Hsu et al., 2022; Newey, 2009; Rilstone, 1996; Stengos and Yan, 2001). The literature explores statistical properties of regression estimators when the regressors that cannot be observed directly are estimated by nonparametric methods, similarly to Pagan's (1984) classical fully parametric case. However, these approaches are not directly applicable when imputation needs to be based on two samples or when there are no common observations in them.

Finally, we employ a new uniform consistency result for estimators based on a nonstandard kernel and mixed data, which is of independent interest. A key aspect of the result is compactness of the support of the kernel function, a feature that often arises in economics and finance either by construction or as a theoretical requirement. For example, economic and financial variables such as expenditure and budget shares, unemployment rates and default and recovery rates are typically bounded from above and below by construction. Theoretical compactness requirements are used in the context of partial linear regression models (see, e.g., Yatchew, 1997), NNM (see, e.g., Abadie and Imbens, 2006), and estimation of first-price auctions (see, e.g., Guerre et al., 2000), to name just a few examples. Interested readers may consult Hirukawa et al. (2022) for further details.

The remainder of this article is organized as follows. Section 2 describes the model, defines the proposed estimator, and derives and discusses its convergence properties. Section 3 presents results of selected Monte Carlo simulations to compare finite-sample properties of the proposed estimator with the IV and other alternatives. As an empirical example, in Section 4, we apply the two-sample estimator to a version of Mincer's (1974) wage regression. Both in simulations and the application, we compare our estimator (a) with the IV approach, (b) with an approach where the missing variable is imputed using a fully parametric method, and (c) with the approach of Hirukawa and Prokhorov (2018) based on NNM. Section 5 concludes with a few questions for future research. Technical proofs are provided in the Appendix. The online supplement contains details of the alternative estimators (a)–(c), comprehensive simulation results for a variety of designs, and additional discussions on the empirical example.

The article adopts the following notational conventions: $\|A\| = \left\{\mathrm{tr}(A'A)\right\}^{1/2}$ is the Euclidean norm of matrix A; $\mathbf{1}\{\cdot\}$ denotes an indicator function; $0_{p\times q}$ signifies the $p \times q$ zero matrix, where the subscript may be suppressed if $q=1$; $B(p,q) = \int_0^1 y^{p-1}(1-y)^{q-1}dy$ for $p,q > 0$ is the beta function; and the symbol $>$ applied to matrices means positive definiteness.

2. Two-sample two-step estimation

2.1. Model

We consider the following linear regression model

$$Y = \beta_0 + X_1'\beta_1 + X_2'\beta_2 + X_{3I}'\beta_3 + u, \tag{1}$$

where $X_1 \in \mathbb{R}^{d_1}$, $X_2 \in \mathbb{R}^{d_2}$, $X_{3I} \in \mathbb{R}^{d_{3I}}$ (the distinction between these regressors will be made clear shortly). Throughout, we assume that either β_1 or β_3 is the parameter of primary interest. Let $d := d_1 + d_2 + d_{3I}$. When $(1, X_1', X_2', X_{3I}')' \in \mathbb{R}^{d+1}$ is exogenous and a single random sample of (Y, X_1, X_2, X_{3I}) is available, the OLS estimator of $\beta = (\beta_0, \beta_1', \beta_2', \beta_3')'$ is consistent under the usual assumptions.

Let $\mathcal{S}_1$ denote the data set at hand and let $\mathcal{S}_2$ denote a second data set which will be required for a two-sample estimation. We assume that $\mathcal{S}_1 = (Y, X_1, X_3)$ and $\mathcal{S}_2 = (X_2, X_3)$, where $X_3 := (X_{3I}, X_{3E})$ is the vector of common variables across the two samples that are partitioned into those included (X_{3I}) and those excluded (X_{3E}) from regression (1). We note that X_2 is missing in $\mathcal{S}_1$ although it is assumed to be relevant, i.e., $\beta_2 \neq 0$. This setup was recently considered by Pacini (2019) but his focus was on relaxing the identification assumptions involving the joint distribution of (Y, X_1, X_2, X_{3I}).

The distinction between X_1 and X_3 will become important later. For now, we can think of X_3 as variables that are common across $\mathcal{S}_1$ and $\mathcal{S}_2$, enter as regressors and are used, due to their presence in both samples, for imputation of the missing regressors in X_2. On the other hand, when a common variable serves merely as a regressor but is not used for imputation, it is classified as X_1. The vector X_1 can be empty; the vector X_3 cannot.

Even though $\mathcal{S}_1$ and $\mathcal{S}_2$ contain common variables X_3, this does not mean that they need to have common observations, i.e., the two samples do not have to overlap. Finally, denote $d_3 := \dim(X_3) = \dim(X_{3I}) + \dim(X_{3E}) := d_{3I} + d_{3E}$, where $d_3 > 0$ must be the case, and either d_{3I} or d_{3E} is allowed to be zero.

An example of this sampling arrangement is when a micro-level data set (e.g., CPS or PSID) can be complemented by an auxiliary data set which is focused on some important variables that are not available elsewhere, e.g., a psychometric data set with ability measures.

In such cases it is common to instrument for the unobserved variable (see Section 1.1 of the online supplement for details of IV estimation). However, the IV estimation is unavailable if there are no valid instruments in the same sample and it may be desirable to avoid the IV estimation for a host of other reasons discussed in Section 2.5.

Denote $\mathcal{S}_1 = \mathcal{S}_{1n} = \left\{(Y_i, X_{1i}, X_{3i})\right\}_{i=1}^{n}$ and $\mathcal{S}_2 = \mathcal{S}_{2m} = \left\{(X_{2j}, X_{3j})\right\}_{j=1}^{m}$. We now discuss how we construct a proxy for X_2 using $\mathcal{S}_{2m}$.

2.2. Additional notation

We start by introducing some additional notation that will help define the semiparametric two-sample estimator and derive its asymptotic properties. First, we allow for the vector of common variables $X_3 = (X_{3I}, X_{3E})$ to consist of both continuous (C) and discrete (D) variables so that $d_3 = \dim(X_3) = \dim(X_{3C}) + \dim(X_{3D}) =: d_{3C} + d_{3D}$. While $d_{3D} = 0$ is allowed, $d_{3C} > 0$ must be the case for the subsequent asymptotic analysis. This distinction will be needed for the application of kernel smoothing to mixed continuous and discrete data. Second, let $\mathbb{X}_3 := \mathbb{X}_{3C} \times \mathbb{X}_{3D}$, where $\mathbb{X}_{3C} := \text{supp}(X_{3C})$ and $\mathbb{X}_{3D} := \text{supp}(X_{3D})$. Third, denote $g_2(X_3) := E(X_2|X_3)$ and $\eta_2 := X_2 - g_2(X_3)$ with $E(\eta_2|X_3) = 0$. The conditional expectation function $g_2(\cdot)$ and the error term η_2 will determine the asymptotic adjustments required to account for the use of nonparametric proxies. Finally, for a pair of vectors (Z, W), let $\Phi_{Z,W} := E(Z_i W_i')$, $S_{n,Z,W} := (1/n)\sum_{i=1}^{n} Z_i W_i'$ and $S_{m,Z,W} := (1/m)\sum_{j=1}^{m} Z_j W_j'$, where summations for $S_{n,Z,W}$ and $S_{m,Z,W}$ are assumed to be taken within $\mathcal{S}_{1n}$ and $\mathcal{S}_{2m}$, respectively, and where the subscripts n and m emphasize the relevant sample sizes.

2.3. Estimation of β and $g_2(\cdot)$

We reformulate the regression (1) as

$$Y = X'\beta + \epsilon,$$

where $X := (1, X_1', g_2(X_3)', X_{3I}')'$ and $\epsilon := u + \eta_2'\beta_2$.

In essence, we wish to obtain an estimator of the unknown function $g_2(X_3)$ using $\mathcal{S}_2$ and construct a proxy for X_2 using some distance measure between the common variables X_3 in $\mathcal{S}_1$ and $\mathcal{S}_2$. We will use a kernel function (as opposed to NNM or another matching scheme) to achieve that.

Let $\hat{g}_2(\cdot)$ be some consistent nonparametric estimator of $g_2(\cdot)$. Then, the estimation procedure we propose consists of the following two steps:

Step 1: Regard $\left\{(X_{2j}, X_{3j})\right\}_{j=1}^{m}$ in $\mathcal{S}_2$ and $\{X_{3i}\}_{i=1}^{n}$ in $\mathcal{S}_1$ as m data points and n design points, respectively, and obtain n nonparametric estimates $\left\{\hat{g}_2(X_{3i})\right\}_{i=1}^{n}$.

Step 2: Run OLS for the regression of Y_i on $\hat{X}_i := (1, X_{1i}', \hat{g}_2(X_{3i})', X_{3Ii}')'$.

The estimator of β is then

$$\hat{\beta}_{PILS} := S_{n,\hat{X},\hat{X}}^{-1} S_{n,\hat{X},Y} = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{X}_i \hat{X}_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} \hat{X}_i Y_i.$$

Because this estimator is the OLS with $\hat{g}_2(X_{3i})$ plugged in place of the missing regressor X_{2i}, we call it the *plug-in least squares* (PILS) estimator hereinafter.

The part of X_3 that is used for imputation but excluded from the original model can be used to form exclusion restrictions $E(X_{3E}'u) = 0$. If $g_2(X_3)$ is not linear in X_{3E}, these restrictions can, in principle, be used to improve precision of the estimation of β (or to construct specification tests). However, since our main goal is to use X_{3E} in the construction of a proxy for X_2, we do not pursue the issues of efficiency or specification testing here.

At a first glance, the PILS approach—just like its fully parametric counterpart (see Section 1.2 of the online supplement for details)—may be viewed as a variant of generated regressors, which have been extensively studied by many researchers. However, two samples and nonparametric imputation raise new and nontrivial issues, which require careful consideration. We study the issues in the next section.

A remaining task is to deliver a consistent estimator of $g_2(\cdot)$. Taking into consideration that $g_2(\cdot)$ may depend on both continuous and discrete covariates, we choose the kernel regression smoother for mixed continuous and categorical data proposed by Racine and Li (2004). In this sense, PILS may be viewed as a partial mean estimator of Newey (1994), although Newey (1994) does not consider the situation where design and data points for kernel smoothing come from different data sources.

A key part of the estimator is the construction of a multivariate kernel. Let $\mathcal{K}(t_{3C}; x_{3C}, \mathbf{h})$ and $\mathcal{L}(t_{3D}; x_{3D}, \boldsymbol{\lambda})$ denote product kernels for the continuous and discrete components of X_3, respectively, where $t_{3\cdot}$ and $x_{3\cdot}$ are data points and design points, and $\mathbf{h}$ and $\boldsymbol{\lambda}$ are vectors of bandwidths. We provide details of how these kernels are constructed in Appendix A.1. Then, the product kernel for X_3 is

$$\mathcal{W}(t_3; x_3, \mathbf{h}, \boldsymbol{\lambda}) := \mathcal{K}(t_{3C}; x_{3C}, \mathbf{h})\mathcal{L}(t_{3D}; x_{3D}, \boldsymbol{\lambda}).$$

It follows that a nonparametric estimator of $g_2(\cdot)$ can be defined as

$$\hat{g}_2(X_{3i}) := \frac{\sum_{j=1}^{m} X_{2j}\mathcal{W}\left(X_{3j}; X_{3i}, \mathbf{h}, \boldsymbol{\lambda}\right)}{\sum_{j=1}^{m} \mathcal{W}\left(X_{3j}; X_{3i}, \mathbf{h}, \boldsymbol{\lambda}\right)}, \quad i = 1, \dots, n.$$

There are many conventional options for the specific univariate kernel to use for continuous variables due to the requirement for compactness of $\mathbb{X}_{3C}$ (see Assumption 2 below). Assuming without loss of generality that the compact set is a d_{3C}-dimensional unit hypercube $[0, 1]^{d_{3C}}$, we propose to employ Chen's (1999) univariate beta kernel with support on $[0, 1]$ as an attractive alternative to standard symmetric kernels. Our application of this kernel is motivated by the uniform convergence result on a compact support which we prove in a companion paper (Hirukawa et al., 2022) and which is, as far as we know, new. This asymmetric kernel is free of the boundary bias by construction and its shape varies across design points even under a fixed value of the smoothing parameter b. The latter property implies that the amount of smoothing by this kernel changes in an adaptive manner.

2.4. Convergence properties of PILS

2.4.1. Regularity conditions

Now we explore asymptotic properties of the PILS estimator as both n and m diverge. For this purpose, the following regularity conditions are imposed.

Assumption 1. The two random samples $(\mathcal{S}_1, \mathcal{S}_2) = (\mathcal{S}_{1n}, \mathcal{S}_{2m})$ are drawn independently from the joint distribution of (Y, X_1, X_2, X_3) with finite fourth-order moments.

Assumption 2. X_{3C} is continuously distributed with a convex and compact support $\mathbb{X}_{3C}$, and its density is bounded and bounded away from zero on $\mathbb{X}_{3C}$.

Assumption 3.

(i) $E(u|X_1,X_3)=0$ and $\sigma_u^2(X_1,X_3):=E(u^2|X_1,X_3)\in(0,\infty)$.

(ii) $(\eta_2 \perp\!\!\!\perp X_1)|X_3$.

(iii) $g_2(\cdot)$ is non-constant on $\mathbb{X}_{3C}$ if X_{3E} contains at least one continuous variable, and $g_2(\cdot)$ is strictly nonlinear on $\mathbb{X}_{3C}$ otherwise.

Assumption 4.

(i) Let $f(\cdot)$ be the marginal pdf of X_{3C}. Then, for some integer $\nu\geq 2$, the νth-order derivatives of $f(\cdot)$ and $f(\cdot)g_2(\cdot)$ with respect to X_{3C} are continuous and bounded uniformly on $\mathbb{X}_{3C}$.

(ii) $\Phi_{X,X}:=E(XX')>0$.

(iii) There exist some constants $\gamma\in(0,\infty)$ and $C\in[1,\infty)$ so that

$$\sup_{x_3\in\mathbb{X}_3} E\left(||X_2||^{2+\gamma}|X_3=x_3\right)\leq C.$$

Assumption 5. The univariate continuous kernel K is either: (a) a symmetric and bounded kernel function of order ν that satisfies (i) $\int K(t)dt=1$, $\int t^l K(t)dt=0$ for $l=1,\ldots,\nu-1$, $\int t^\nu K(t)dt\neq 0$, (ii) the first-order Lipschitz condition, and (iii) $\int|t|^k|K(t)|dt<\infty$ for some $k>2\nu$; or (b) the beta kernel of Chen (1999).

Assumption 6. Sequences of smoothing parameters $h_p(=h_p(m)>0)$, $b_p(=b_p(m)>0)$ and $\lambda_q(=\lambda_q(m)\in(0,1))$ and boundary parameters $\theta_p(=\theta_p(m)>0)$ for $p=1,\ldots,d_{3C}$ and $q=1,\ldots,d_{3D}$ satisfy one of the following conditions as $m\to\infty$: (a) for a symmetric νth-order kernel, $h_p,\lambda_q\to 0$ and $\log m/(m\prod_{p=1}^{d_{3C}}h_p)\to 0$; and (b) for the beta kernel, $b_p,\lambda_q,\theta_p\to 0$, $b_p/\theta_p\to 0$ and $\log m/\left(m\sqrt{\prod_{p=1}^{d_{3C}}b_p\theta_p}\right)\to 0$.

The first three assumptions are often encountered in matching-based estimation (see, e.g., Hirukawa and Prokhorov, 2018) and are natural in our settings. In particular, it follows from Assumption 3(i) that even when some part of X_3 does not enter regression (1) due to an exclusion restriction, X_1 and X_3 are exogenous in the full model. Nonlinearity of $g_2(\cdot)$ in Assumption 3(iii) is required only when all continuous common variables are included as regressors in (1). Otherwise, excluded continuous common variables introduce additional randomness in $g_2(\cdot)$, which helps identify β.

Assumption 3(ii) is also standard for the proxy variable literature. It implies that $\eta_2|(X_1,X_3)\overset{d}{=}\eta_2|X_3$, where "$\overset{d}{=}$" denotes equality in distribution. It is sometimes possible to use a weaker assumption, e.g., conditional mean independence $E(\eta_2|X_1,X_3)=E(\eta_2|X_3)$ (see, e.g., Wooldridge, 2010, Section 4.3.2). However, we focus on conditional independence for two reasons. First, it implies orthogonality of η_2 and X_1, which is a key requirement for consistency of PILS in Theorem 1. Second, it simplifies the covariance estimation in Section 2.4.3. A similar assumption can be found in Inoue and Solon (2010, Assumption c) and Hirukawa and Prokhorov (2018, Assumption 4).

Moreover, this assumption involves quantities that are either unobserved (η_2) or unobserved in the same sample (η_2 and X_1). Shah and Peters (2020) have recently demonstrated that (even if they were observed) a uniformly valid conditional independence test does not exist. This may look like a disadvantage of our method. However, as discussed by Wooldridge (2010, p.68), a conditional independence assumption is routinely made or implied in all proxy-based estimators, and our approach is no different. As we argue in Section 4.3, the assumption is easier to claim given enough observables in X_3.

Along with compactness of $\mathbb{X}_{3C}$ in Assumption 2, Assumptions 4–6 are required to establish weak uniform consistency of the nonparametric regression estimator $\hat{g}_2(\cdot)$ on $\mathbb{X}_3$. Similar conditions can be found, for instance, in Li and Ouyang (2005), Hansen (2008) and Su et al. (2013) for a symmetric kernel, and Hirukawa et al. (2022) for the beta kernel. Notice that the beta kernel is nonnegative so that the order $\nu = 2$ is the case in Assumption 4(i).

Strictly speaking, uniform consistency holds on $\mathbb{X}_3$ for a symmetric kernel. In contrast, for the beta kernel, the boundary parameters $\theta_1, \dots, \theta_{d_{3C}}$ play a key role in uniform consistency of $\hat{g}_2(\cdot)$. Specifically, uniform consistency in this case is established on $\mathbb{S}_{X_{3C}} \times \mathbb{X}_{3D}$, where $\mathbb{S}_{X_{3C}} := \prod_{p=1}^{d_{3C}} [\theta_p, 1-\theta_p]$ is a compact set expanding to $\mathbb{X}_{3C} = [0,1]^{d_{3C}}$ (more slowly than the smoothing parameters $b_1, \dots, b_{d_{3C}}$) so that $\mathbb{S}_{X_{3C}} \times \mathbb{X}_{3D}$ is expanding to $\mathbb{X}_3$ as $m \to \infty$. On the other hand, because the regression estimator $\hat{g}_2(\cdot)$ is a Nadaraya-Watson-type estimator, it suffers from the so-called boundary bias when a symmetric kernel is employed. When the beta kernel is used, $\hat{g}_2(\cdot)$ is free of this issue by construction.

We note that Assumption 5(a) allows for higher-order kernels (i.e., $\nu > 2$). In theory, such kernels could be a remedy for the curse of dimensionality implied by Corollary 1 below. In practice, however, it is well known that higher-order kernels are unlikely to dominate nonnegative ones for moderate sample sizes, and that the advantage, if any, is typically marginal (see, e.g., Marron, 1994; Marron and Wand, 1992). In fact, our simulation results indicate poor finite-sample performance of PILS using higher-order kernels.

2.4.2. Consistency and asymptotic normality of PILS

The next two theorems establish consistency and asymptotic normality of $\hat{\beta}_{PILS}$. In particular, asymptotic normality of $\hat{\beta}_{PILS}$ is obtained after correcting for B_{g_2}, the bias term due to kernel smoothing. Each theorem holds regardless of the number of common variables and regardless of the divergence patterns in (n, m).

Theorem 1. *If Assumptions 1–6 hold, then $\hat{\beta}_{PILS} \xrightarrow{p} \beta$ as $n, m \to \infty$.*

Theorem 2. *If Assumptions 1–6 hold, then*

$$\sqrt{n}\left(\hat{\beta}_{PILS} - \beta - B_{g_2}\right) \xrightarrow{d} N\left(0_{(d+1)\times 1}, V\right) := N\left(0_{(d+1)\times 1}, \Phi_{X,X}^{-1}\Omega\Phi_{X,X}^{-1}\right)$$

as $n, m \to \infty$, where

$$\begin{aligned} B_{g_2} &:= S_{n,\hat{X},\hat{X}}^{-1} S_{n,\hat{X},[g_2(X_3) - E\{\hat{g}_2(X_3)|X_3\}]'\beta_2} \\ &= \left(\frac{1}{n}\sum_{i=1}^{n}\hat{X}_i\hat{X}_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n}\hat{X}_i\left[g_2(X_{3i}) - E\{\hat{g}_2(X_{3i})|X_{3i}\}\right]'\beta_2 \\ &= \begin{cases} O_p\left(\sum_{p=1}^{d_{3C}} h_p^{\nu} + \sum_{q=1}^{d_{3D}} \lambda_q\right) & \textit{for a symmetric } \nu\textit{th-order kernel} \\ O_p\left(\sum_{p=1}^{d_{3C}} b_p + \sum_{q=1}^{d_{3D}} \lambda_q\right) & \textit{for the beta kernel} \end{cases}, \quad \textit{and} \\ \Omega &:= \Omega_1 + \Omega_2 := E\left(XX'\epsilon^2\right) + E\left(XX'\beta_2'\eta_2\eta_2'\beta_2\right). \end{aligned}$$

As derived in the proof of Theorem 1, there is the asymptotically negligible bias term B_{g_2}, due to kernel smoothing. There are also two $\sqrt{n}$-asymptotically normal terms. One comes from the sampling error in regression (1), and it has the asymptotic variance $\Phi_{X,X}^{-1}\Omega_1\Phi_{X,X}^{-1}$. The other is due to the approximation error of g_2, and its asymptotic variance is $\Phi_{X,X}^{-1}\Omega_2\Phi_{X,X}^{-1}$. As a consequence, the asymptotic variance Ω in Theorem 2 has similar structure to that of the double kernel estimator by Stengos and Yan (2001, Theorem 1) and the two-step series estimator by Hsu et al. (2022, Theorem 2.1). A difference from Stengos and Yan (2001) and Hsu et al. (2022) is that the two terms are uncorrelated. Each of them can be obtained within two independent samples $\mathcal{S}_1$ and $\mathcal{S}_2$, and thus Ω is free of covariance terms.

2.4.3. Covariance estimation

Covariance estimation is essential for inference. The problem of estimating V consistently boils down to proposing consistent estimators of Ω_1 and Ω_2. For the PILS residual $\hat{\epsilon}_i := Y_i - \hat{X}_i'\hat{\beta}_{PILS}$, an estimator of Ω_1 is

$$\hat{\Omega}_1 := \frac{1}{n}\sum_{i=1}^{n} \hat{X}_i\hat{X}_i'\hat{\epsilon}_i^2.$$

Consistency of $\hat{\Omega}_1$ can be established in the same manner as in the proofs of Theorems 1 and 2.

On the other hand, $\ddot{\Omega}_2 := (1/n)\sum_{i=1}^{n} \hat{X}_i\hat{X}_i'\hat{\beta}_{2,PILS}'\hat{\eta}_{2i}\hat{\eta}_{2i}'\hat{\beta}_{2,PILS}$ might work as a consistent estimator for Ω_2, where $\hat{\beta}_{2,PILS}$ is the PILS estimator of β_2 and $\hat{\eta}_2 := X_2 - \hat{g}_2(X_3)$. In reality, it turns out to be difficult to obtain $\{\hat{\eta}_{2i}\}_{i=1}^{n}$ because of the absence of X_2 in $\mathcal{S}_1$. Instead, using $\eta_2|(X_1, X_3) \overset{d}{=} \eta_2|X_3$ yields

$$\Omega_2 = E\{XX'\beta_2'E(\eta_2\eta_2'|X_3)\beta_2\} =: E\{XX'\beta_2'\Sigma_2(X_3)\beta_2\}.$$

Then, we can alternatively consider the estimator

$$\hat{\Omega}_2 := \frac{1}{n}\sum_{i=1}^{n} \hat{X}_i\hat{X}_i'\hat{\beta}_{2,PILS}'\hat{\Sigma}_2(X_{3i})\hat{\beta}_{2,PILS},$$

where

$$\hat{\Sigma}_2(X_{3i}) := \frac{\sum_{j=1}^{m}\hat{\eta}_{2j}\hat{\eta}_{2j}'\mathcal{W}(X_{3j}; X_{3i}, \mathbf{h}, \boldsymbol{\lambda})}{\sum_{j=1}^{m}\mathcal{W}(X_{3j}; X_{3i}, \mathbf{h}, \boldsymbol{\lambda})}, \quad i = 1, \ldots, n,$$

and the nonparametric regression residuals $\{\hat{\eta}_{2j}\}_{j=1}^{m} = \{X_{2j} - \hat{g}_2(X_{3j})\}_{j=1}^{m}$ can be obtained within $\mathcal{S}_2$. The estimation procedure for $\hat{\Sigma}_2(\cdot)$ is inspired by the Algorithm in Section 2.4 of Fan and Yao (1998). To show consistency of $\hat{\Omega}_2$, we can see that a similar argument to the proof of Lemma A1 in Appendix A.2 establishes uniform consistency of $\hat{\eta}_{2j}$ for η_{2j}, which in turn leads to uniform consistency of $\hat{\Sigma}_2(X_{3i})$ for $\Sigma_2(X_{3i})$. Then, $\hat{\Omega}_2 = (1/n)\sum_{i=1}^{n} X_iX_i'\beta_2'\Sigma_2(X_{3i})\beta_2 + o_p(1) \xrightarrow{p} \Omega_2$.

In conclusion, V can be consistently estimated by

$$\hat{V} := S_{n,\hat{X},\hat{X}}^{-1}\hat{\Omega}S_{n,\hat{X},\hat{X}}^{-1} := S_{n,\hat{X},\hat{X}}^{-1}(\hat{\Omega}_1 + \hat{\Omega}_2)S_{n,\hat{X},\hat{X}}^{-1}.$$

2.4.4. $\sqrt{n}$-consistency of PILS

Theorems 1 and 2 jointly imply that although PILS is consistent, its convergence is affected by the bias term B_{g_2} generated by kernel smoothing. For a large d_{3C}, the order of magnitude in the bias term dominates and the convergence rate of $\hat{\beta}_{PILS}$ becomes inferior. It appears that this problem is unavoidable in a regression that uses a nonparametric component. A similar phenomenon arises in the context of imputation bias correction (see, e.g., Hirukawa and Prokhorov, 2018).

However, $\sqrt{n}$-consistency of $\hat{\beta}_{PILS}$ automatically holds for small values of d_{3C}. To illustrate such cases, we modify Assumption 6 in order to control the bias and variance convergence of $\hat{g}_2(\cdot)$ more easily.

Assumption 6′. Sequences of the smoothing and boundary parameters satisfy one of the following conditions as $m \to \infty$: (a) for a νth-order symmetric kernel, there is a bandwidth $h (= h(m) > 0)$ so that $h_1, ..., h_{d_{3C}} \propto h$, $\lambda_1, ..., \lambda_{d_{3D}} \propto h^{\nu}$, $h \to 0$, and $\log m / (m h^{d_{3C}}) \to 0$; and (b) for the beta kernel, there are a smoothing parameter $b (= b(m) > 0)$ and a boundary parameter $\theta (= \theta(m) > 0)$ so that $b_1, ..., b_{d_{3C}} \propto b$, $\lambda_1, ..., \lambda_{d_{3D}} \propto b$, $\theta_1, ..., \theta_{d_{3C}} \propto \theta$, $b, \theta \to 0$, $b/\theta \to 0$, and $\log m / \left\{ m (b\theta)^{d_{3C}/2} \right\} \to 0$.

The corollary below describes the cases in which $\hat{\beta}_{PILS}$ becomes $\sqrt{n}$-asymptotically normal.

Corollary 1. *Let $h \propto (\log m/m)^{\alpha}$ and $b \propto (\log m/m)^{2\alpha}$ for some constant $\alpha > 0$. Also suppose that one of the following divergence patterns in (n, m) is true: (i) $n/m \to \kappa \in (0, \infty)$; (ii) $n/m \to 0$; or (iii) $n/m \to \infty$ and $n/m^{4\alpha} \to 0$. Then, under Assumptions 1–5 and 6′, as $n, m \to \infty$, $\sqrt{n}(\hat{\beta}_{PILS} - \beta) \xrightarrow{d} N(0_{(d+1)\times 1}, V)$ holds either (a) if a νth-order symmetric kernel is employed and $d_{3C} \leq 2\nu - 1$ or (b) if the beta kernel is employed and $d_{3C} \leq 3$.*

Corollary 1 presents an upper bound of the number of continuous common variables for PILS to be $\sqrt{n}$-consistent. This is recognized as the curse of dimensionality in continuous common variables, and it can be relaxed, in theory, by using higher-order kernels. In particular, when employing a nonnegative kernel (which is either a symmetric pdf or the beta kernel), we must limit the number of continuous common variables to 3 or less to make PILS $\sqrt{n}$-consistent. To see this in more detail, note that for the most realistic case $n/m \to \kappa$, shrinkage rates of the smoothing parameters are $h \propto (\log m/m)^{\alpha}$ and $b \propto (\log m/m)^{2\alpha}$ for some $\alpha \in (1/4, 1/d_{3C})$ (see Appendix A.4). These rates are faster than what yields Stone's (1982) optimal global rate of convergence, i.e., $h^* \propto (\log m/m)^{1/(4+d_{3C})}$ and $b^* \propto (\log m/m)^{2/(4+d_{3C})}$ that can be obtained through balancing the squared bias and variance of $\hat{g}_2(\cdot)$. This is because to attain $\sqrt{n}$-consistency we should keep the convergence rate of the dominant bias term in $\hat{g}_2(\cdot)$ sufficiently fast by undersmoothing, or more specifically, by setting the exponent α within the aforementioned range. We give some examples of h and b that fulfill the rate requirement in Sections 3 and 4.

Finally, it may be the case that the missing regressor X_2 can be observed with a measurement error. This issue occurs, for example, when an ability measure is unreliable in the wage regression. Suppose that we can observe $\breve{X}_2 = X_2 + \upsilon$ at best, where υ is the measurement error. As long as $E(\upsilon | X_3) = 0_{d_2 \times 1}$, the effect of υ is filtered out via kernel smoothing and

$$\breve{g}_2(X_{3i}) := \frac{\sum_{j=1}^{m} \breve{X}_{2j} \mathcal{W}(X_{3j}; X_{3i}, \mathbf{h}, \boldsymbol{\lambda})}{\sum_{j=1}^{m} \mathcal{W}(X_{3j}; X_{3i}, \mathbf{h}, \boldsymbol{\lambda})} = \hat{g}_2(X_{3i}) + o_p(1), \quad i = 1, ..., n$$

holds. In the end, consistency of PILS is maintained.

2.5. Comparison of PILS with competing estimators

We conclude this section by comparing PILS with competing estimators. The competing estimators we consider are IV estimators using linear and nonlinear instruments, a fully parametric alternative to PILS (called PARA hereinafter), and the matched-sample indirect inference (MSII) and fully-modified MSII (MSII-FM) estimators. PARA is often a method of choice (see, e.g., Fang et al. 2008). MSII(-FM) was proposed by Hirukawa and Prokhorov (2018) as an alternative to the inconsistent OLS estimator using the matched sample (called MSOLS hereinafter). Definitions and statistical properties of these estimators are reviewed in the online supplement. Finite-sample properties of all the estimators will be assessed in comparison with PILS in the next section.

2.5.1. Comparison with IV

The use of IVs has been associated with at least two problems. First, instruments that are not strongly correlated with the endogenous variables—weak instruments—lead to large inconsistencies in the IV estimators even if a weak correlation exists between the instruments and the structural equation error, i.e., even if the instruments only slightly violate the validity assumption (see, e.g., Bound et al., 1995). When we are ready to assume that the IVs are valid, the weak instrument problem results in large asymptotic and finite-sample biases and, when many such instruments are available, the distribution of IV estimators is known to deviate substantially from the large-sample approximation (see, e.g., Bekker, 1994; Chao and Swanson, 2005).

Second, finite-sample properties of IV estimators are known to be generally poor, especially when the instruments are weak (see, e.g., Flores-Lagunes, 2007). Various bias reduction techniques have had limited success and even very inventive uses of instruments have, for these reasons, been criticized in terms of their validity, strength and finite-sample performance.

Furthermore, valid and strong instruments may be unavailable in the same sample. A number of ingenious methods have been proposed to combine data from more than one source in such cases (see, e.g., Angrist and Krueger, 1992, 1995; Arellano and Meghir, 1992; Inoue and Solon, 2010; Klevmarken, 1982; Murtazashvili et al., 2015). However, the two-sample IV estimators inherit the same problems as their single-sample counterparts (see, e.g., Choi et al., 2018).

The identification strategy for our estimator is different. The central identification conditions are that $g_2(\cdot)$ either depends on additional exogenous variables X_{3E} or is nonlinear. If there are no additional exogenous variables, then identification relies on nonlinearity in the first step. The weak identification issue, similar to the weak IV, can appear in our setting if $g_2(\cdot)$ is linear and depends very weakly on X_{3E} so that $g_2(\cdot)$ is close to being a linear function of X_{3I} only. We consider these cases in the extensive simulation comparisons.

There is a rich literature proposing ingenious methods such as integrated regression functions, basis functions, sieves and local smoothing, to incorporate nonlinearities in an IV estimation based on conditional moment restrictions (see, e.g., Dominguez and Lobato, 2004; Kitamura et al., 2004; Lavergne and Patilea, 2013; Mammen et al., 2016). For example, Mammen et al. (2016) provide a general "three-step" estimation framework that uses a nonparametric estimator in each step. A full account of nonlinearity would involve one of these methods. However, our use of several functions of instruments is quite close to how these nonlinearities are commonly handled in practice.

Let Z denote the instruments. The validity of Z as instruments can be checked. Note that X_2 admits the reduced form $X_2 = g_2(X_3) + \eta_2 = g_2(X_{3I}, X_{3E}) + \eta_2$. If X_{3I} is non-empty (i.e., some common variables are included as regressors) and some elements of Z are relevant, then those elements are either a part of X_{3I} or correlated with X_{3I}. As a result, Z and X_2 are also correlated, and the IV estimator of the omitted variable regression becomes inconsistent. Then, we are tempted to test the null of consistency of IV estimation indirectly by testing the null hypothesis $H_0 : g_2(X_{3I}, X_{3E}) = g_2(X_{3E})$ *a.e.* against the alternative $H_1 : g_2(X_{3I}, X_{3E}) \neq g_2(X_{3E})$ for a non-empty set on $\text{supp}(X_{3I})$. Several versions of the test of significance in nonparametric regressions have been proposed in the literature. Examples include Fan and Li (1996) and Racine (1997) for continuous regressors and Lavergne (2001) and Racine et al. (2006) for discrete or categorical regressors, to name a few.

2.5.2. Comparison with PARA

PARA imputes a parametric, linear predictor of X_2. It seems attractive, as it is not subject to the curse of dimensionality in continuous common variables unlike PILS. However, it requires

$\dim(X_{3E}) > 0$ for identification because otherwise, the imputed predictor of X_2 is a linear function of X_{3I}. In contrast, as in Assumption 3(iii), PILS achieves identification by nonlinearity even if X_{3E} is empty.

PARA shares similar properties with the two-sample (TS) two-stage least squares (2SLS) estimator of Inoue and Solon (2010). An interpretation of PARA in the context of TS2SLS is that the missing regressor X_2 in the former plays the role of an endogenous variable in the latter. In particular, in the special case when X_1 is empty, PARA is numerically equivalent to TS2SLS.[1]

As demonstrated in Proposition S1 of the online supplement, consistency of PARA depends on whether X_1 is present in (1). It necessarily becomes consistent when X_1 is empty. When X_1 is nonempty, which is the most common setting in practice, consistency of PARA fails if $E(X_1X_2') \neq E(X_1X_3')A'$, where A is the coefficient matrix in the linear projection of X_2 on X_3.[2] On the other hand, consistency of PILS fails if $E(X_1X_2'|X_3) \neq E(X_1|X_3)E(X_2'|X_3)$, which is a violation of an implication of Assumption 3(ii). Since $E(X_1X_2') \neq E(X_1X_3')A'$ does not imply $E(X_1X_2'|X_3) \neq E(X_1|X_3)E(X_2'|X_3)$, PILS can be consistent when $E(X_1X_2') \neq E(X_1X_3')A'$ (i.e., PARA is inconsistent) but still $E(X_1X_2'|X_3) = E(X_1|X_3)E(X_2'|X_3)$ (i.e., Assumption 3(ii) is valid). This could be an explanation for the significant differences observed between the PILS and PARA estimates of the return to schooling in Section 4.

2.5.3. Comparison with MSII(-FM)

While PILS and MSII(-FM) are both two-sample semiparametric estimators, the former has three novel features relative to the latter. First, while both MSII(-FM) and PILS are $\sqrt{n}$-consistent and asymptotically normal two-sample estimators, their approaches to restoring consistency differ. Consistency of MSII(-FM) is established by imputing X_2 from $\mathcal{S}_2$ and then eliminating the nonvanishing bias caused by the imputation. The bias correction requires a consistent, nonparametric estimate of Σ_2, and that the estimation error in $\hat{\Sigma}_2$ is $O_p(n^{-1/2})$. As a consequence, the asymptotic variance of MSII(-FM) tends to be large and highly complicated because of multiple asymptotically normal terms with the same $O_p(n^{-1/2})$ rate.

In contrast, consistency of PILS comes from directly imputing a consistent estimate of $g_2(X_3) = E(X_2|X_3)$ in place of the missing X_2. By construction, it does not require bias correction that is a key ingredient in MSII(-FM).[3] The asymptotic variance of PILS also comes from two $O_p(n^{-1/2})$ terms, namely, the sampling error as in OLS and the approximation error from replacing the unobservable g_2 with its kernel estimate $\hat{g}_2$. Although it is hard to compare asymptotic variances of PILS and MSII(-FM) analytically, Monte Carlo simulations below indicate that the former tends to be smaller than the latter.

Second, the asymptotic analysis we develop for PILS explicitly incorporates discrete matching variables. This is in contrast to the asymptotic analysis of Hirukawa and Prokhorov (2018), who do not accommodate discrete variables explicitly but argue that, similarly to the treatment effect literature (see, e.g., Abadie and Imbens, 2006), the inclusion of discrete matching variables with a finite number of support points does not affect convergence rates of MSII(-FM).

Third, we clarify the role of excluded continuous matching variables for identification. Hirukawa and Prokhorov (2018) maintain the assumption that all common variables enter the

[1] We thank an anonymous referee for pointing this out to us.

[2] The coefficient matrix A, as opposed to the one given in Proposition S1 of the online Supplement, is based on the assumption that the linear projection has no intercepts. Our argument below is not affected with or without intercepts in the projection.

[3] From the viewpoint of imputation, the relationship between MSII(-FM) and PILS may be compared to the one between the NNM-based estimators studied by Abadie and Imbens (2006, 2011) and kernel-based matching estimators studied by Heckman et al. (1998).

regression and are used for both estimation and matching. The price to pay for this assumption is that we need to impose nonlinearity of the conditional mean of the missing regressor given the matching variables in order to achieve identification. This article relaxes the assumption so that some common variables may be employed only for imputation. The existence of such common variables allows for linearity in the conditional mean.

3. Finite-sample performance

3.1. Monte Carlo setup

In this section we present selected results of an extensive Monte Carlo study comparing finite-sample properties of PILS with the alternative estimators. The design of the Monte Carlo study is meant to reflect the two scenarios depending on whether or not X_1 is empty. Due to the limitation of space, we focus on the more realistic case in which X_1 is non-empty. Monte Carlo designs and results when X_1 is empty are available in the online supplement.

The Monte Carlo study is designed to mimic important aspects of the return to schooling application in Section 4. These are: (a) allowing for both included and excluded continuous matching variables (*educ* and *age* in the application); (b) maintaining $n/m = 2$ and considering the sample sizes of $(n, m) = (2000, 1000)$; (c) permitting the included continuous matching variable X_{3IC} (*educ*) to become endogenous when X_2 (*abil*) is omitted; and (d) choosing as an instrument Z a variable that differs from the matching variable (*fatheduc*) and setting $corr(X_{3IC}, Z)$ and $corr(X_{3IC}, Z^2)$ (roughly) equal to 0.4.[4] Of primary interest is estimation of β_3, the coefficient on X_{3IC} (*educ*).

Our design is largely inspired by the study on nonlinear instruments of Dieterle and Snell (2016). Consider the two forms of our regression

$$
\begin{aligned}
&(L) \quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_{3I_C} + u, \\
&(S) \quad Y = \beta_0 + \beta_1 X_1 + \beta_3 X_{3I_C} + v, \ \ v = u + \beta_2 X_2,
\end{aligned}
$$

where true parameter values are $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 1$, and β_3 is the parameter of interest. Let $u, \eta_1 \overset{iid}{\sim} N(0, 1)$, $X_{3E_C}, \xi \overset{iid}{\sim} U[-2, 2]$, $Z \overset{iid}{\sim} U[0, 1]$, and $X_{3E_D} \overset{iid}{\sim} Bernoulli(1/2) - 1/2$, where all these random variables are mutually independent. Next, X_1 and X_{3I_C} are generated as

$$X_1 = 2X_{3I_C} X_{3E_C}^2 X_{3E_D} + \eta_1, \ \text{and}$$

$$X_{3I_C} = 4\rho_1 \left(Z - \frac{1}{2} \right) + \sqrt{1 - \rho_1^2}\,\xi,$$

where $\rho_1 = corr(X_{3I_C}, Z)$ and $\rho_1 \in \{0.1, 0.4\}$. The cases with $\rho_1 = 0.1$ and 0.4 correspond to weak and strong instruments, respectively. The specification of X_{3I_C} ensures compactness of $\operatorname{supp}(X_{3I_C})$, $E(X_{3I_C}) = E(X_{3E_C}) = E(X_{3E_D}) = 0$ and $Var(X_{3I_C}) = Var(X_{3E_C}) = 4/3$ regardless of ρ_1. A straightforward calculation also yields

$$\rho_2 = corr\left(X_{3I_C}, Z^2\right) = \frac{\sqrt{15}}{4} \rho_1 = \begin{cases} 0.0968 & \text{for } \rho_1 = 0.1 \\ 0.3873 & \text{for } \rho_1 = 0.4 \end{cases}.$$

The missing regressor X_2 is generated as one of the following two models:

[4] Sample correlations between *educ* and *fatheduc* and between *educ* and *fatheduc*^{2} are 0.434 and 0.420, respectively.

Model A:

$$X_2 = \frac{X_{3I_C}}{5\rho_1} + X_{3E_C} + X_{3E_D} + \eta_2, \text{ and}$$
$$\eta_2 = -E\left(\frac{X_{3I_C}}{5\rho_1}\middle| Z\right) = -\frac{2}{5}(2Z-1).$$

Model B:

$$X_2 = \frac{1}{12\sqrt{C}}\left(\frac{X_{3I_C}}{2\rho_1}\right)^3 + \frac{1}{4}\sin\left(\frac{\pi}{2}X_{3E_C}\right)X_{3E_D} + \eta_2,$$
$$\eta_2 = -\frac{1}{12\sqrt{C}}E\left\{\left(\frac{X_{3I_C}}{2\rho_1}\right)^3\middle| Z\right\} = -\frac{1}{12\sqrt{C}}\left\{(2Z-1)^3 + \left(\frac{1}{\rho_1^2}-1\right)(2Z-1)\right\}, \text{ and}$$
$$C = \frac{8}{105} - \frac{4}{15\rho_1^2} + \frac{1}{3\rho_1^4}.$$

The constant C for Model B is the variance of $E\{(X_{3I_C}/(2\rho_1))^3|Z\}$. Hence, the term $12\sqrt{C}$ in the denominator of η_2 is intended to make $Var(\eta_2)$ invariant to ρ_1. Also observe that X_2 is linear in $X_3 = (X_{3I_C}, X_{3E_C}, X_{3E_D})$ for Model A. This guarantees consistency of PARA. On the other hand, X_2 is nonlinear and even additively non-separable in X_3 for Model B. As a result, PARA becomes inconsistent in this case.

The above procedure provides us with two observable samples

$$\mathcal{S}_1 = \left\{(Y_i, X_{1i}, X_{3I_Ci}, X_{3E_Ci}, X_{3E_Di}, Z_i)\right\}_{i=1}^n \text{ and } \mathcal{S}_2 = \left\{(X_{2j}, X_{3I_Cj}, X_{3E_Cj}, X_{3E_Dj})\right\}_{j=1}^m.$$

The complete sample

$$\mathcal{S}^* = \left\{(Y_i, X_{1i}, X_{2i}, X_{3I_Ci}, X_{3E_Ci}, X_{3E_Di}, Z_i)\right\}_{i=1}^n$$

is the sample that would not be observed in practice. Finally, the matched sample

$$\mathcal{S} = \left\{\left(Y_i, X_{1i}, X_{2j_1(i)}, \ldots, X_{2j_K(i)}, X_{3I_Ci}, X_{3E_Ci}, X_{3E_Di}, Z_i\right)\right\}_{i=1}^n$$

is constructed via the NNM with respect to $X_3 = (X_{3I_C}, X_{3E_C}, X_{3E_D})$ as in Hirukawa and Prokhorov (2018). The NNM is based on the Mahalanobis distance, and the number of matches is set equal to $K=1$ (single match) that is most commonly chosen. Sample sizes of $\mathcal{S}_1$ and $\mathcal{S}_2$ are $n \in \{1000, 2000\}$ and $m = n/2$. The number of replications is 1000.

There are three options to estimate β_3 consistently. The first option is to estimate regression (S) using $\mathcal{S}_1$ only. While $E(v) = E(X_1 v) = 0$ holds, $E(X_{3I_C}v) \neq 0$ (i.e., X_{3I_C} is endogenous in (S)) is the case and OLS for (S) is inconsistent. However, $E(v|Z) = 0$ and $\rho \neq 0$. Hence, we can estimate β_3 consistently by using Z and its functions as instruments for X_{3I_C}. The remaining two options rely on both $\mathcal{S}_1$ and $\mathcal{S}_2$ to estimate regression (L). Option two is to construct the matched sample $\mathcal{S}$ from $\mathcal{S}_1$ and $\mathcal{S}_2$ and run MSII or MSII-FM using $\mathcal{S}$. Option three is to employ PILS using both $\mathcal{S}_1$ and $\mathcal{S}_2$.

Based on the above estimation strategies, we compare the following estimators of β_3: (i) the infeasible OLS estimator for regression (L) using $\mathcal{S}^*$ [OLS*];[5] (ii) the inconsistent OLS estimator for regression (S) using $\mathcal{S}_1$ only [OLS-S]; (iii) the IV estimator for (S) using Z as an instrument for X_{3I_C} [IV1-S]; (iv) the IV estimator for (S) using Z^2 as an instrument for X_{3I_C} [IV2-S]; (v) the two-step GMM estimator for (S) using (Z, Z^2) as instruments for X_{3I_C} [GMM-S], along with the

[5] OLS* is consistent, because $E(u) = E(X_1 u) = E(X_2 u) = E(X_{3I_C}u) = 0$.

initial 2SLS estimator [2SLS-S]; (vi) the inconsistent MSOLS estimator for (L) using $\mathcal{S}$ with $K=1$ [MSOLS]; (vii) the MSII-FM estimator for (L) using $\mathcal{S}$ with $K=1$ and second-order polynomial [MSII-FM], along with the first-step MSII estimator [MSII]; (viii) the PARA estimator for (L) using $\mathcal{S}_1$ and $\mathcal{S}_2$ [PARA]; (ix) the PILS estimator for (L) using $\mathcal{S}_1$, $\mathcal{S}_2$ and the Epanechnikov kernel $K_E(t) = (3/4)(1-t^2)\mathbf{1}\{|t| \le 1\}$ [PILS-E];[6] and (x) the PILS estimator for (L) using $\mathcal{S}_1$, $\mathcal{S}_2$ and the beta kernel [PILS-B].

Implementing the two PILS estimators requires a choice of the smoothing parameters. We use $\hat{h} = \hat{\sigma}_X(\log m/m)^{0.3}$ for the Epanechnikov kernel, $\hat{b} = \hat{\sigma}_U(\log m/m)^{0.6}$ for the beta kernel, and $\hat{\lambda} = (\log m/m)^{0.6}$ for the discrete kernel, where $\hat{\sigma}_X$ and $\hat{\sigma}_U$ are sample standard deviations of the continuous common variable in the original scale $X(= X_{3I_C}, X_{3E_C})$ and in the transformed scale $U := (X+2)/4 \in [0,1]$, respectively. These shrinkage rates fulfill the requirements in Section 2.4.4.

For each estimator of β_3, the following performance measures are computed: (i) *Mean* (simulation average of the parameter estimate); (ii) *SD* (simulation standard deviation of the parameter estimate); (iii) *RMSE* (root mean-squared error of the parameter estimate); (iv) $\overline{SE}$ (simulation average of the standard error); and (v) *CR* (coverage rate for the nominal 95% confidence interval). Formulae for standard errors are as follows: (a) heteroskedasticity-robust standard errors by Eicker (1963) and White (1980) are calculated for OLS* and IV-S; (b) the formula for MSII-FM follows Theorem 4 of Hirukawa and Prokhorov (2018), (c) the formula for PILS appears in Theorem 2; and (d) the formula for PARA is given in Proposition S3 of the online supplement. Because of inconsistency we do not compute $\overline{SE}$ or *CR* of OLS-S and MSOLS. Those of the first-step MSII are also omitted due to its nonparametric convergence rate. Likewise, we present $\overline{SE}$ and *CR* of PARA only when it is consistent, i.e., for Model A.

We also consider a variation to the above design which examines how the estimators behave near identification failure. Identification of two-sample estimators heavily relies on the functional form of $g_2(\cdot)$ and existence of excluded matching variable X_{3E}. In particular, identification failure occurs if $g_2(\cdot)$ reduces to a linear function of the included matching variable X_{3I} only.

Therefore, we allow for weaker identification by changing how we generate X_2:

Model A′:

$$X_2 = \frac{X_{3I_C}}{5\rho_1} + w(X_{3E_C} + X_{3E_D}) + \eta_2,$$

Model B′:

$$X_2 = \frac{1}{12\sqrt{C}}\left(\frac{X_{3I_C}}{2\rho_1}\right)^3 + \frac{1}{4}w\sin\left(\frac{\pi}{2}X_{3E_C}\right)X_{3E_D} + \eta_2,$$

where the weight $w \in \{1.0, 0.1\}$ controls the strength of identification in two-sample estimators. We fix $\rho_1 = 0.4$, $(n, m) = (2000, 1000)$ and the number of replications is 1000. All other aspects of the simulation setup remain unchanged. Therefore, $w = 1.0$ corresponds to the results for $\rho_1 =$

[6] In addition to the second-order Epanechnikov kernel, we consider the fourth and sixth-order Epanechnikov kernels by Hansen (2005), where their functional forms are

$$K_{E,4}(t) = \frac{15}{8}\left(1 - \frac{7}{3}t^2\right)K_E(t) \text{ and } K_{E,6}(t) = \frac{175}{64}\left(1 - 6t^2 + \frac{33}{5}t^4\right)K_E(t),$$

respectively. However, our early simulation results indicate huge finite-sample biases from these estimators, and thus are not reported.

Table 1. Monte Carlo results.

	$(n, m) = (1000, 500)$					$(n, m) = (2000, 1000)$				
Estimator	*Mean*	*SD*	*RMSE*	$\bar{SE}$	*CR*	*Mean*	*SD*	*RMSE*	$\bar{SE}$	*CR*
				Model A ($\rho_1 = 0.1$)						
OLS*	0.9986	0.0559	0.0559	0.0560	95%	1.0020	0.0397	0.0398	0.0397	95%
OLS-S	2.9802	0.0450	1.9807	–	–	2.9807	0.0314	1.9809	–	–
IV1-S	0.7461	1.2942	1.3189	1.1668	92%	0.9019	0.6212	0.6289	0.6171	94%
IV2-S	0.7076	1.5997	1.6262	1.3316	93%	0.8921	0.6651	0.6738	0.6457	93%
2SLS-S	1.0207	0.9247	0.9249	0.9019	88%	1.0115	0.5629	0.5630	0.5744	92%
GMM-S	1.0209	0.9251	0.9253	0.9012	88%	1.0112	0.5631	0.5632	0.5741	92%
MSOLS	1.1265	0.0624	0.1411	–	–	1.0938	0.0434	0.1034	–	–
MSII	0.9058	0.0805	0.1239	–	–	0.9467	0.0504	0.0734	–	–
MSII-FM	0.8203	0.0827	0.1978	0.1066	67%	0.9052	0.0511	0.1077	0.0629	72%
PARA	0.9990	0.0612	0.0612	0.0611	95%	1.0009	0.0426	0.0426	0.0432	96%
PILS-E	1.0288	0.0720	0.0775	0.0773	95%	1.0024	0.0462	0.0463	0.0533	98%
PILS-B	1.0013	0.0637	0.0637	0.0804	99%	0.9963	0.0437	0.0439	0.0550	99%
				Model A ($\rho_1 = 0.4$)						
OLS*	0.9999	0.0294	0.0294	0.0293	94%	1.0012	0.0204	0.0205	0.0207	96%
OLS-S	1.4203	0.0451	0.4228	–	–	1.4204	0.0311	0.4216	–	–
IV1-S	0.9993	0.1163	0.1163	0.1165	96%	0.9970	0.0800	0.0801	0.0821	96%
IV2-S	0.9989	0.1198	0.1198	0.1202	95%	0.9965	0.0832	0.0832	0.0848	96%
2SLS-S	1.0018	0.1165	0.1165	0.1161	95%	0.9982	0.0798	0.0799	0.0820	96%
GMM-S	1.0019	0.1167	0.1168	0.1160	96%	0.9982	0.0798	0.0798	0.0819	96%
MSOLS	1.0166	0.0313	0.0354	–	–	1.0152	0.0219	0.0266	–	–
MSII	0.9952	0.0317	0.0321	–	–	0.9987	0.0221	0.0221	–	–
MSII-FM	0.9861	0.0317	0.0347	0.0308	92%	0.9938	0.0221	0.0229	0.0216	94%
PARA	0.9998	0.0310	0.0310	0.0307	95%	1.0010	0.0214	0.0214	0.0217	96%
PILS-E	1.0094	0.0354	0.0367	0.0327	91%	1.0035	0.0236	0.0239	0.0226	93%
PILS-B	1.0013	0.0318	0.0318	0.0315	95%	1.0012	0.0219	0.0219	0.0221	95%
				Model B ($\rho_1 = 0.1$)						
OLS*	0.9986	0.0602	0.0602	0.0613	95%	0.9996	0.0447	0.0447	0.0433	95%
OLS-S	1.4334	0.0288	0.4344	–	–	1.4343	0.0199	0.4347	–	–
IV1-S	0.9359	0.4259	0.4306	0.4207	97%	0.9820	0.2432	0.2439	0.2432	97%
IV2-S	0.9204	0.5361	0.5420	0.4801	96%	0.9794	0.2593	0.2602	0.2537	97%
2SLS-S	0.9978	0.3717	0.3717	0.3623	95%	1.0067	0.2299	0.2300	0.2330	96%
GMM-S	0.9974	0.3721	0.3721	0.3619	94%	1.0064	0.2299	0.2300	0.2329	96%
MSOLS	1.0907	0.0598	0.1087	–	–	1.0663	0.0440	0.0796	–	–
MSII	0.8611	0.1102	0.1773	–	–	0.9222	0.0616	0.0992	–	–
MSII-FM	0.8204	0.1185	0.2152	0.1309	81%	0.9079	0.0630	0.1116	0.0676	77%
PARA	1.4632	2.8122	2.8501	–	–	1.5260	4.1731	4.2061	–	–
PILS-E	1.0087	0.0702	0.0707	0.0751	95%	0.9926	0.0487	0.0492	0.0523	97%
PILS-B	0.9607	0.0727	0.0827	0.0896	98%	0.9636	0.0507	0.0624	0.0583	95%
				Model B ($\rho_1 = 0.4$)						
OLS*	0.9990	0.0371	0.0371	0.0373	96%	1.0005	0.0271	0.0271	0.0263	94%
OLS-S	1.1077	0.0283	0.1113	–	–	1.1084	0.0190	0.1101	–	–
IV1-S	0.9979	0.0687	0.0687	0.0703	96%	1.0004	0.0496	0.0496	0.0496	95%
IV2-S	0.9971	0.0709	0.0710	0.0726	95%	1.0002	0.0511	0.0511	0.0512	96%
2SLS-S	0.9987	0.0689	0.0689	0.0702	96%	1.0008	0.0494	0.0494	0.0495	95%
GMM-S	0.9987	0.0690	0.0691	0.0701	95%	1.0008	0.0494	0.0494	0.0495	95%
MSOLS	1.0404	0.0370	0.0548	–	–	1.0378	0.0265	0.0462	–	–
MSII	0.9854	0.0538	0.0557	–	–	0.9917	0.0362	0.0371	–	–
MSII-FM	0.9809	0.0553	0.0585	0.0534	93%	0.9897	0.0366	0.0380	0.0355	94%
PARA	1.1324	0.9097	0.9192	–	–	1.1148	2.0242	2.0275	–	–
PILS-E	1.0192	0.0414	0.0456	0.0487	94%	1.0110	0.0298	0.0318	0.0355	96%
PILS-B	0.9992	0.0417	0.0418	0.0640	98%	0.9992	0.0307	0.0307	0.0423	98%

0.4 in the original setup. Because the purpose is to check how the quality of two-sample estimates deteriorates as w shrinks, we only report *Mean*, *SD* and *RMSE* for each estimator.

3.2. Results

Table 1 presents the results from alternative specifications of X_2. It is immediately clear that OLS* outperforms all other estimators as would be expected. However, OLS* is infeasible, and its

Table 2. Estimates of β_3 with shrinking w.

	$w = 1.0$			$w = 0.1$		
Estimator	*Mean*	*SD*	*RMSE*	*Mean*	*SD*	*RMSE*
			Model A′ ($\rho_1 = 0.4$)			
OLS*	1.0012	0.0204	0.0205	1.0012	0.0435	0.0435
OLS-S	1.4204	0.0311	0.4216	1.4209	0.0194	0.4214
IV1-S	0.9970	0.0800	0.0801	0.9998	0.0548	0.0548
IV2-S	0.9965	0.0832	0.0832	0.9996	0.0569	0.0569
2SLS-S	0.9982	0.0798	0.0799	1.0011	0.0547	0.0547
GMM-S	0.9982	0.0798	0.0798	1.0011	0.0547	0.0547
MSOLS	1.0152	0.0219	0.0266	1.2922	0.0444	0.2955
MSII	0.9987	0.0221	0.0221	0.9460	0.1691	0.1775
MSII-FM	0.9938	0.0221	0.0229	0.9045	0.1872	0.2101
PARA	1.0010	0.0214	0.0214	1.0013	0.0819	0.0819
PILS-E	1.0035	0.0236	0.0239	1.1469	0.0625	0.1597
PILS-B	1.0012	0.0219	0.0219	1.0540	0.0708	0.0890
			Model B′ ($\rho_1 = 0.4$)			
OLS*	1.0005	0.0271	0.0271	1.0004	0.0317	0.0317
OLS-S	1.1084	0.0190	0.1101	1.1084	0.0190	0.1101
IV1-S	1.0004	0.0496	0.0496	1.0004	0.0495	0.0495
IV2-S	1.0002	0.0511	0.0511	1.0002	0.0511	0.0511
2SLS-S	1.0008	0.0494	0.0494	1.0008	0.0494	0.0494
GMM-S	1.0008	0.0494	0.0494	1.0008	0.0493	0.0493
MSOLS	1.0378	0.0265	0.0462	1.0629	0.0309	0.0701
MSII	0.9917	0.0362	0.0371	0.9647	0.0815	0.0888
MSII-FM	0.9897	0.0366	0.0380	0.9600	0.0840	0.0931
PARA	1.1148	2.0242	2.0275	1.0665	1.1649	1.1668
PILS-E	1.0110	0.0298	0.0318	1.0290	0.0375	0.0474
PILS-B	0.9992	0.0307	0.0307	1.0028	0.0424	0.0425

performance can be used only as a benchmark. Instead, realistic comparisons can be made between the other estimators.

Performance of IV1-S and IV2-S is largely governed by strengths of instruments, while huge biases and variability due to weak instruments are ameliorated by a larger sample size. While performances of 2SLS-S and GMM-S improve due to additional moment restrictions, their efficiencies do not surpass those of two PILS estimators.

The instrument strength also influences the performance of the two-sample estimators. MSII-FM is a prominent example. The estimator is severely biased and unstable for each of two weak instrument cases.

The performance of PARA depends heavily on specification of X_2. PARA performs exceptionally well in the more favorable case (i.e., Model A), as expected. On the other hand, the estimator generates substantial bias and variability for Model B. The poor performance may be attributed to identification failure due to difficulty in the linear projection of the product of periodic and binary random variables.

The two PILS estimators appear to be more robust against changes in the strength of instruments and/or specification of X_2. An inspection reveals that PILS-B outperforms PILS-E in terms of *RMSE*, except in the case of Model B with $\rho_1 = 0.1$.

Table 2 presents the results for the variation addressing near identification failure. Qualities of OLS*, OLS-S and all IV-based estimates remain unchanged throughout this table, as expected. In contrast, performances of two-sample estimators are influenced by both the functional form of $g_2(\cdot)$ and the value of w.

For Model A′, $g_2(\cdot)$ reduces to a linear function of the included matching variable X_{3I_C} as $w \to 0$, and identification failure is expected. Indeed, as w decreases, the finite-sample distribution of MSII-FM appears to be off-centered from the true value of 1 and highly dispersed. A potential explanation is that as w decreases, excluded matching variables (X_{3E_C}, X_{3E_D}) lose their importance in estimating correction terms for non-negligible and matching-discrepancy biases. Nonetheless, using the full set of $X_3 = (X_{3I_C}, X_{3E_C}, X_{3E_D})$ makes their estimates quite imprecise due to additional noises, whereas in reality it may suffice to estimate the terms by using X_{3I_C} only. Likewise, the finite-sample distribution of PARA for Model A$'$ also tends to be dispersed with w. This may be attributed to near multi-collinearity in the regression. The proxy of X_2 is almost a linear function of X_{3I_C} for a very small w, and imputing it as an extra regressor makes PARA unstable. Large biases in PILS-E and B are thought to be due to the same identification failure.

On the other hand, $g_2(\cdot)$ for Model B$'$ collapses to a cubic function of X_{3I_C} for $w \to 0$. In this case, the role of (X_{3E_C}, X_{3E_D}) is less important for identification, and the performance of all two-sample estimators is only marginally affected by w. In particular, PILS-B seems to be more robust against changes in w than any other two-sample estimator.

4. An application: Estimation of return to schooling

4.1. Earnings function

Estimation of the causal link between education and earnings has been a focus of labor economists over the last several decades. Card (2001) suggested that endogeneity of education in the earnings equation might at least in part be responsible for the continuing interest in uncovering the causal effect of education on labor market outcomes. As Griliches (1977) discusses in detail, empirical labor economists have long speculated that the primary reason why education is endogenous when estimating the returns to schooling is some omitted explanatory variable(s)—unobservable factors that influence education such as ability and motivation—that are likely to have a direct effect on individual earnings and wages. The aim of this section is to estimate the rate of return to additional schooling using alternative estimators available to us. These include the traditional one-sample OLS and linear and nonlinear IV estimators as well as MSII, PARA, and PILS. A more comprehensive comparison can be found in the online supplement.

Following the classical framework of the human capital earnings/wage function of Mincer (1974) we assume additivity and wish to use US data in order to estimate the causal effect of education on earnings using the following model:

$$\begin{aligned}\log(\mathit{earnings}) = \beta_0 + \beta_1 \mathit{education} + \beta_2 \mathit{ability} + \beta_3 \mathit{experience} + \beta_4 \mathit{experience}^2 \\ + \beta_5 \mathit{married} + \beta_6 \mathit{black} + \beta_7 \mathit{south} + \beta_8 \mathit{urban} + u,\end{aligned} \tag{2}$$

where $\log(\mathit{earnings})$ is the natural logarithm of the individual's total annual labor income, *education* is the individual's completed years of education, *ability* is the individual's ability/skills (with zero mean), *experience* is work experience of the individual, *married* is an indicator for whether the individual is married, *black* is an indicator for whether the individual is black, *south* is an indicator for whether the individual currently lives in the southern geographical region, *urban* is an indicator of the individual's urban residence while growing up,[7] and u is an idiosyncratic error.

The main issue with estimating Eq. (2) is that econometricians are typically unable to observe the individual's skills. Because of that, the (one-sample) OLS estimator of the return to schooling—β_1—from Eq. (2) where *ability* is excluded is likely to suffer from an "ability bias." Two textbook solutions to this problem are (i) to find within the same data set a valid proxy for the unobservable skills and use OLS estimation, and (ii) to find within the same data set a valid instrumental variable for the individual's educational level and use the IV approach. Since in addition to these two approaches, we advocate using two-sample estimation, we would like to be able to apply in practice the one- and two-sample approaches to the same applied task.

[7] The publicly available part of the PSID survey that we use does not provide information on current urban residence.

Table 3. Sample characteristics.

Variable	Mean	Std. Dev.	Min.	Median	Max.
		Panel A: The PSID sample			
Earnings	8763.24	5968.39	30	7900	70000
log (*Earnings*)	8.84	0.80	3.40	8.97	11.16
Education	12.15	3.09	5	12	17
Experience	19.57	13.17	0	18	68
Married	0.89	0.31	0	1	1
Black	0.26	0.44	0	0	1
South	0.41	0.49	0	0	1
Urban	0.28	0.45	0	0	1
Age	38.57	13.48	17	37	86
South while growing up	0.46	0.50	0	0	1
IQ score	9.49	2.28	0	10	13
Father's education	9.26	3.06	0	8	17
		Panel B: The NLS sample			
Education	12.98	3.20	1	13	18
Married	0.63	0.48	0	1	1
Black	0.34	0.48	0	0	1
South	0.53	0.50	0	1	1
Urban	0.58	0.49	0	1	1
Age	28.47	3.14	24	28	34
South while growing up	0.55	0.50	0	1	1
KWW score	32.48	8.42	4	33	56

Notes: $(n =)$2430 individuals in the PSID sample. $(m =)$1102 individuals in the NLS sample. Demeaned *IQ score* and *KWW score* are used.

4.2. Constructing the samples

To estimate earnings and wage equations for the US, labor economists frequently use such micro-data sets as CPS, PSID and the National Longitudinal Survey (NLS). While CPS has a larger sample and is more representative of the entire demographic composition of the US population than the other two surveys, it does not typically collect important wage determinants such as actual work experience and ability. The PSID data set does provide information on actual work experience but, similar to CPS, generally does not collect data on ability. NLS routinely reports ability measures and contains data on actual and potential work experience but it is less representative of the US labor force. Given the features of these surveys, PSID and NLS seem most suitable as our main and auxiliary data sets, respectively.

We employ the 1972 wave of PSID as our main data set. We choose this wave so that we can estimate Eq. (2) using not only IV and PILS but also using a proxy for unobserved skills and ability. While PSID generally does not contain any measures of unobserved ability, the 1972 wave is among the few PSID waves that do include an ability measure.

In the 1972 wave, the PSID respondents were administered a particular assessment of abstract thinking—the Lorge-Thorndike Intelligence Test (LTIT). In essence, this is a test of verbal skills—a sentence completion test—that Veroff et al. (1971, p.26) describe "as a feasible, reasonably valid assessment of what psychologists have labeled intelligence." Furthermore, Veroff et al. (1971, p.26) advise that this test "seems to correlate well with most different kinds of tests of intelligence, well enough to suggest using it singly without going to multiple measurement." LTIT administered to the PSID respondents contained 13 questions, where each question received a point for the correct response. Thus, the score for the entire sentence completion test can range from zero (the worst outcome) to 13 (the best outcome). We call the variable containing the total test score *IQ score*, and we demean it.

In addition to the information on the individual's ability discussed above, we also gather information on the individual's actual work experience, completed years of education, age, whether the person is black, married, and lives in the southern geographical region.[8] Furthermore, we include a dummy variable for whether the person grew up in an urban area and for whether the person grew up in the southern region. Finally, in our PSID sample, we also get information on the completed years of education by the individual's father. The last variable—father's educational level—is used as an instrument for the individual's educational attainment when no measure of ability is available.

Our main sample contains 2430 men who reported positive labor income in 1971. Panel A of Table 3 reports summary statistics for the variables in our PSID sample. The sample correlation between an individual's education and the father's education is 0.434.

To exploit the two-sample approaches, we need a second sample where the ability measure is available and/or potentially more reliable. We employ the `CARD` data set provided with Wooldridge (2013). This data set contains observations from the 1977 wave of the Young Men Cohort of NLS. We exploit observations with positive wages in 1976.

As an ability measure we employ the results of the "Knowledge of the World of Work" (*KWW score*) test that the NLS respondents were administered during their interviews in 1966. The KWW score from NLS is arguably a better measure of unobserved ability than the IQ score from the 1972 wave of PSID. A higher KWW score indicates higher intelligence. We demean this measure before using it to estimate Eq. (2).

Panel B of Table 3 provides summary statistics for the NLS sample of $m = 1102$ respondents when *education*, *married*, *black*, *south*, and *urban* are used as the included common variables, X_{3I}, and *age* (the individual's age, in years) and *south while growing up* (an indicator for whether the person resided in the southern region while growing up) are used as the excluded common variables, X_{3E}. Note that *experience* and *experience*^{2}, which are not observed in the NLS sample, represent X_1 in regression (2). The summary statistics for *experience* are provided in Panel A of Table 3.

4.3. Estimation approaches

We report estimation results in Table 4.[9] Columns (1)–(3) report the parametric estimation results based on just the PSID sample. Columns (1) reports the OLS estimates using the ability proxy from the 1972 waive. Column (2) reports the OLS results that do not account for unobserved ability. In most real-life settings, the results in column (1) are infeasible and those in column (2) are subject to the "ability bias." Column (3) provides the IV estimates where the father's education is used as an IV for an individual's educational level.

Columns (4)–(7) report two-sample estimates based on both the PSID and NLS samples. Columns (4), (5), and (7) contain the results based on semiparametric two-sample approaches, while column (6) reports the only estimates based on a fully parametric two-sample procedure. Specifically, column (4) contains the OLS estimates when the two samples are combined using hot-deck imputation and no bias correction is applied. As discussed by Hirukawa and Prokhorov (2018), this estimator is inconsistent. Column (5) provides the MSII-FM results that account for the imputation biases. In essence, the MSII approach is a bias-corrected version of the MSOLS approach. Column (6) exhibits the PARA results as yet another two-sample estimator. It is worth emphasizing that PARA is not guaranteed to be consistent in this case. Finally, column (7) reports the PILS results using the beta kernel, chosen on the basis of the simulation results in previous section.

[8] We follow the 1979 NLS of Youth definition of the southern region. The southern region includes Alabama, Arkansas, Delaware, District of Columbia, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, and West Virginia.

[9] GAUSS codes implementing the IV and two-sample estimators as well as the data sets are available at https://sites.google.com/site/artembprokhorovv0/papers/SimulationAndApplicationFiles.zip.

Table 4. Estimation results.

	(1) OLS*	(2) OLS-S	(3) IV-S	(4) MSOLS	(5) MSII-FM	(6) PARA	(7) PILS
Education	0.0635	0.0718	0.0953	0.0727	0.0685	−0.2850	0.0568
	(0.0056)	(0.0053)	(0.0159)	(0.0059)	(0.0080)	(0.2281)	(0.0087)
Experience	0.0809	0.0818	0.0830	0.0826	0.0766	0.1170	0.0818
	(0.0043)	(0.0043)	(0.0045)	(0.0049)	(0.0065)	(0.0095)	(0.0043)
Experience^2	−0.0017	−0.0017	−0.0017	−0.0017	−0.0016	−0.0017	−0.0017
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Ability	0.0313	–	–	−0.0012	0.0029	0.2444	0.0094
	(0.0078)	(–)	(–)	(0.0027)	(0.0081)	(0.1514)	(0.0045)
Married	0.3717	0.3793	0.3844	0.3799	0.3777	0.2226	0.3958
	(0.0539)	(0.0536)	(0.0536)	(0.0535)	(0.0535)	(0.1319)	(0.0543)
Black	−0.1302	−0.1741	−0.1249	−0.1849	−0.1504	−0.1819	−0.1776
	(0.0323)	(0.0316)	(0.0426)	(0.0393)	(0.0806)	(0.1146)	(0.0316)
South	−0.0921	−0.0983	−0.0814	−0.0979	−0.0989	−0.1372	−0.1042
	(0.0288)	(0.0287)	(0.0314)	(0.0286)	(0.0289)	(0.1135)	(0.0287)
Urban	0.1363	0.1499	0.1278	0.1538	0.1404	0.0138	0.1557
	(0.0282)	(0.0284)	(0.0332)	(0.0297)	(0.0390)	(0.1273)	(0.0282)
Data combination?	No	No	No	Yes	Yes	Yes	Yes
Sample size: n	2430	2430	2430	2430	2430	2430	2430
m	–	–	–	1102	1102	1102	1102

Notes: The dependent variable is the log of total annual labor earnings. *Education*, *married*, *black*, *south*, and *urban* are used as the included common variables, and *age* and *south while growing up* are used as the excluded common variables. The (demeaned) *IQ score* and *KWW score* variables are used as ability measures in the PSID and NLS samples, respectively.

The two-sample estimators use five variables common to both samples—*education*, *married*, *black*, *south*, and *urban*—as included variables and two variables—*age* and *south while growing up*—as excluded variables. Note that since in our case both the main (PSID) and auxiliary (NLS) samples contain information on the individual's educational level, we treat *education* as a part of X_3. We choose not to employ *experience* as a common variable because experience in PSID represents actual experience, while experience in NLS represents potential experience, which is quite different. As a consequence, our entire set of common variables contains *education*, *married*, *black*, *south*, *urban*, *age*, *south while growing up*, where *education* and *age* are treated as continuous.

In the context of Assumption 3(ii), the PILS results assume that *experience* and *ability* are related only through *education*, *married*, *black*, *south*, *urban*, *age*, and *south while growing up*. This seems plausible given the weak empirical link between experience and ability and the long list of observables we condition on. In the proxy literature, this is known as redundancy or ignorability of the proxy variable in the structural equation, akin to ignorability of selection in selectivity models (see, e.g., Wooldridge, 2010, p.67).

The NNM used by MSOLS and MSII-FM adopts a single match ($K = 1$) based on the Mahalanobis metric. We average the (demeaned) KWW score for ties in our second (NLS) sample and assign this average as a unique value of the ability measure to the respondent with the given values of our common variables. As a consequence, $m = 1102$ respondents remain in our $\mathcal{S}_2$. A second-order polynomial is used in MSII-FM.

Our choice of a kernel for PILS reflects the favorable performance of the beta kernel in simulations. For PILS, each continuous common variable X is converted from its original scale to a variable $U := (X - m_X)/(M_X - m_X) \in [0, 1]$, where M_X and m_X are maximal and minimal values in the pooled sample constructed from observations of X in $\mathcal{S}_1$ and $\mathcal{S}_2$. The reason for using the maximum and minimum of the pooled sample is that the ranges of *education* and *age* are quite different between $\mathcal{S}_1$ and $\mathcal{S}_2$. We use $\hat{b} = \hat{\sigma}_U(\log m/m)^{0.6}$ and $\hat{\lambda} = (\log m/m)^{0.6}$ for the beta and discrete kernels, respectively, where $\hat{\sigma}_U$ is the sample standard deviation of a converted continuous common variable U in $\mathcal{S}_2$.[10]

[10] One potential concern about PILS is how sensitive it is to the choice of smoothing parameter values in nonparametric estimation of $g_2(\cdot)$. It is confirmed that even after their original values are doubled or cut by half, the estimation results are qualitatively similar. Detailed results are available in the online supplement.

4.4. Empirical findings

It is immediately clear from Table 4 that the PARA estimation results substantially differ from the rest. First, contrary to the corresponding OLS* estimate, the estimate of the rate of return to education is negative. Second, the PARA estimate of the coefficient on *ability* is expectedly positive but extraordinarily large when compared to the OLS* estimate. Third, PARA estimates are quite unstable in that the standard error of each coefficient estimate is much larger than those of the corresponding estimates from other methods in general. Because it is hard to provide a meaningful interpretation to the PARA results, we exclude them from further discussion. One lesson from the PARA estimates is that parametric imputation methods for missing regressors should be used with caution. In this respect, MSII(-FM) and PILS adopt nonparametric imputation methods, and thus they are more robust than PARA.

Table 4 shows that the signs of coefficient estimates on all the regressors except *ability* are as expected and most of the estimates are significant. The MSOLS estimator yields negative and insignificant estimates of the ability effect, whereas the MSII-FM estimator results in a positive and insignificant estimate. Interestingly, OLS* and PILS are the only two estimators that produce positive (as one would expect) and statistically significant estimates of the ability effect. Finally, we note that the PILS standard errors tend to be smaller than those of MSII-FM, as predicted in a previous section.

Next we focus on the estimates of the rate of return to education. The estimates reported in Table 4 provide evidence that the father's education is unlikely to be a valid instrument for an individual's education. This is because if the father's education were a valid IV, the OLS-S and MSOLS estimators of β_1 would be upward-inconsistent. However, we observe that the IV estimate is the largest in magnitude. Furthermore, if this instrument is valid and there is no measurement error in the PSID ability measure, then OLS* and IV are both consistent. However, we see that the OLS* and IV estimates are noticeably different, suggesting again that the two estimators are unlikely to both be consistent.

The infeasible OLS – OLS*—is the second smallest (conceding in magnitude only to PILS). In addition, if there is no measurement error in the PSID ability measure then OLS* is consistent, but OLS-S and MSOLS are still upward-inconsistent. This possibility seems to fit well with our estimates in Table 4. Indeed, the feasible one-sample estimate of β_1 with omitted ability—OLS-S—is between IV and infeasible OLS. In fact, the MSOLS and MSII-FM estimates of β_1 are also between the one-sample IV and infeasible OLS* estimates. Therefore, there are grounds to view the infeasible OLS* as a benchmark result in our analysis.

Importantly, the only estimate of β_1 that is smaller than the infeasible OLS is PILS. We note that PILS is also the second closest (in absolute value) to the infeasible OLS estimate, while the first closest is MSII-FM. This would be expected given that only MSII-FM and PILS are the consistent approaches if the instrument is invalid. Furthermore, the possibility that PILS (and not OLS*) is closer to the true return to education is not unrealistic. If the PSID ability measure is error-ridden, OLS* is upward-inconsistent suggesting that PILS is likely to be the only estimator that delivers the closest estimate to the true population value.[11]

5. Concluding remarks

When some regressors are unavailable for a regression analysis, econometricians often resort to IV estimation. They typically make a considerable effort to find and justify valid instruments for the regressors that are suspected to be endogenous due to their possible correlation with the omitted regressors. In this article, we have explored two-sample alternatives to this conventional approach.

[11] The online supplement presents a detailed discussion on this claim.

We have developed the PILS estimation procedure for such models. The procedure first uses an auxiliary sample to obtain a nonparametric estimator of the conditional mean of the regressor which is missing in the main sample. Variables that are common in both samples are used in the conditioning set. In the second step of the procedure, we simply run OLS using a plug-in estimate of the conditional mean.

We establish asymptotic normality of PILS. Attractive finite-sample properties of PILS are confirmed in a Monte Carlo study. In particular, simulations demonstrate numerically that PILS is generally more efficient than IV and it dominates other available estimators, e.g., MSII(-FM), in terms of mean squared error. The results also indicate superior finite-sample properties of PILS using the beta kernel.

However, in order for PILS to attain the parametric rate of convergence, the number of continuous common variables must be three or less (provided that nonnegative kernels are employed). In this respect, the curse of dimensionality plays out similarly to the NNM-based estimation considered by Hirukawa and Prokhorov (2018).

A natural extension is to adopt a dimension reduction method when using multiple common variables. This can be done using a sparsity-based algorithm, e.g., lasso of the first step, or using a propensity score algorithm. We leave these ideas for future research.

Acknowledgments

The authors would like to thank the editor, an associate editor and two anonymous referees as well as the participants of CFE, IAAE, iCEBA, SETA, and a seminar at Durham University for their constructive comments and suggestions.

Funding

Financial support through grants from Japan Society for the Promotion of Science (M. Hirukawa, Project No. 19K01595) and the Russian Science Foundation (A. Prokhorov, Project No. 16-18-10432) for various and non-overlapping parts of this research is gratefully acknowledged.

References

Abadie, A., Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* 74(1):235–267. doi:10.1111/j.1468-0262.2006.00655.x

Abadie, A., Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics* 29(1):1–11. doi:10.1198/jbes.2009.07333

Aitchison, J., Aitken, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika* 63(3): 413–420. doi:10.1093/biomet/63.3.413

Angrist, J., Krueger, A. (1992). The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples. *Journal of the American Statistical Association* 87(418): 328–336. doi:10.1080/01621459.1992.10475212

Angrist, J., Krueger, A. (1995). Split-sample instrumental variables estimates off the return to schooling. *Journal of Business & Economic Statistics* 13(2):225–235. doi:10.2307/1392377

Arellano, M., Meghir, C. (1992). Female labour supply and on the job search: an empirical model estimated using complementary data sets. *Review of Economic Studies* 59(3):537–559. doi:10.2307/2297863

Bekker, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica* 62(3):657–681. doi:10.2307/2951662

Black, B., Trainor, M., Spencer, J. E. (1999). Wage protection systems, segregation and gender pay inequalities: West Germany, The Netherlands and Great Britain. *Cambridge Journal of Economics* 23(4):449–464. doi:10.1093/cje/23.4.449

Bound, J., Jaeger, D. A., Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 90(430):443–450. doi:10.2307/2291055

Card, D. (1995). Using geographic variation in college proximity to estimate the return to schooling. In: Christophides, L. N., Grant, E. K., Swidinsky, R., eds., *Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp*. Toronto: University of Toronto Press, pp. 201–222.

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica* 69(5):1127–1160. doi:10.1111/1468-0262.00237

Chao, J. C., Swanson, N. R. (2005). Consistent estimation with a large number of weak instruments. *Econometrica* 73(5):1673–1692. doi:10.1111/j.1468-0262.2005.00632.x

Chen, S. X. (1999). Beta kernel estimators for density functions. *Computational Statistics and Data Analysis* 31(2): 131–145. doi:10.1016/S0167-9473(99)00010-9

Chen, X., Hong, H., Tarozzi, A. (2008). Semiparametric efficiency in GMM models with auxiliary data. *Annals of Statistics* 36(2):808–843.

Choi, J., Gu, J., Shen, S. (2018). Weak-Instrument robust inference for two-sample instrumental variable regression. *Journal of Applied Econometrics* 33(1):109–125. doi:10.1002/jae.2580

Dieterle, S. G., Snell, A. (2016). A simple diagnostic to investigate instrument validity and heterogeneous effects when using a single instrument. *Labour Economics* 42:76–86. doi:10.1016/j.labeco.2016.08.002

Dominguez, M. A., Lobato, I. N. (2004). Consistent estimation of models defined by conditional moment restrictions. *Econometrica* 72(5):1601–1615. doi:10.1111/j.1468-0262.2004.00545.x

Eicker, F. (1963). Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *Annals of Mathematical Statistics* 34(2):447–456. doi:10.1214/aoms/1177704156

Fan, J., Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* 85(3):645–660. doi:10.1093/biomet/85.3.645

Fan, Y., Li, Q. (1996). Consistent model specification tests: omitted variables and semiparametric functional forms. *Econometrica* 64(4):865–890. doi:10.2307/2171848

Fang, H., Keane, M. P., Silverman, D. (2008). Sources of advantageous selection: evidence from the Medigap insurance market. *Journal of Political Economy* 116(2):303–350. doi:10.1086/587623

Flavin, M., Nakagawa, S. (2008). A model of housing in the presence of adjustment costs: a structural interpretation of habit persistence. *American Economic Review* 98(1):474–495. doi:10.1257/aer.98.1.474

Flores-Lagunes, A. (2007). Finite sample evidence of IV estimators under weak instruments. *Journal of Applied Econometrics* 22(3):677–694. doi:10.1002/jae.916

Graham, B. S., de Xavier Pinto, C. C., Egel, D. (2016). Efficient estimation of data combination models by the method of auxiliary-to-study tilting (AST). *Journal of Business & Economic Statistics* 34(2):288–301. doi:10.1080/07350015.2015.1038544

Griliches, Z. (1977). Estimating the return to schooling: some econometrics problems. *Econometrica* 45(1):1–22. doi:10.2307/1913285

Guerre, E., Perrigne, I., Vuong, Q. (2000). Optimal nonparametric estimation of first-price auctions. *Econometrica* 68(3):525–574. doi:10.1111/1468-0262.00123

Hansen, B. E. (2005). Exact mean integrated squared error of higher order kernel estimators. *Econometric Theory* 21(6):1031–1057. doi:10.1017/S0266466605050528

Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory* 24(3):726–748. doi:10.1017/S0266466608080304

Heckman, J. J., Ichimura, H., Todd, P. (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies* 65(2):261–294. doi:10.1111/1467-937X.00044

Hirukawa, M., Murtazashvili, I., Prokhorov, A. (2022). Uniform convergence rates for nonparametric estimators smoothed by the beta kernel. *Scandinavian Journal of Statistics* 49(3):1353–1382. doi:10.1111/sjos.12573

Hirukawa, M., Prokhorov, A. (2018). Consistent estimation of linear regression models using matched data. *Journal of Econometrics* 203(2):344–358. doi:10.1016/j.jeconom.2017.07.006

Hsu, Y.-C., Liao, J.-C., Lin, E. S. (2022). Two-step series estimation and specification testing of (partially) linear models with generated regressors, Econometric Reviews 41(9):985–1007. doi:10.1080/07474938.2022.2082169

Inoue, A., Solon, G. (2010). Two-sample instrumental variables estimators. *Review of Economics and Statistics* 92(3):557–561. doi:10.1162/REST_a_00011

Kitamura, Y., Tripathi, G., Ahn, H. (2004). Empirical likelihood-based inference in conditional moment restriction models. *Econometrica* 72(6):1667–1714. doi:10.1111/j.1468-0262.2004.00550.x

Klevmarken, A. (1982). Missing variables and two-stage least squares estimation from more than one data set, IFN Working Paper, No. 62, Stockholm: Research Institute of Industrial Economics (IFN).

Lavergne, P. (2001). An equality test across nonparametric regressions. *Journal of Econometrics* 103(1–2):307–344. doi:10.1016/S0304-4076(01)00046-X

Lavergne, P., Patilea, V. (2013). Smooth minimum distance estimation and testing with conditional estimating equations: uniform in bandwidth theory. *Journal of Econometrics* 177(1):47–59. doi:10.1016/j.jeconom.2013.05.006

Li, Q., Ouyang, D. (2005). Uniform convergence rate of kernel estimation with mixed categorical and continuous data. *Economics Letters* 86(2):291–296. doi:10.1016/j.econlet.2004.07.018

Mammen, E., Rothe, C., Schienle, M. (2016). Semiparametric estimation with generated covariates. *Econometric Theory* 32(5):1140–1177. doi:10.1017/S0266466615000134

Marron, J. S. (1994). Visual understanding of higher-order kernels. *Journal of Computational and Graphical Statistics* 3(4):447–458. doi:10.1080/10618600.1994.10474657

Marron, J. S., Wand, M. P. (1992). Exact mean integrated squared error. *Annals of Statistics* 20(2):712–736.

Mincer, J. A. (1974). *Schooling, Experience and Earnings*. New York: National Bureau of Economic Research.

Murtazashvili, I., Liu, D., Prokhorov, A. (2015). Two-sample nonparametric estimation of intergenerational income mobility in the United States and Sweden. *Canadian Journal of Economics* 48(5):1733–1761. doi:10.1111/caje.12178

Newey, W. K. (1994). Kernel estimation of partial means and a general variance estimator. *Econometric Theory* 10(2):1–21. doi:10.1017/S0266466600008409

Newey, W. K. (2009). Two-step series estimation of sample selection models. *Econometrics Journal* 12(S1): S217–S229. doi:10.1111/j.1368-423X.2008.00263.x

Pacini, D. (2019). Two-sample least squares projection. *Econometric Reviews* 38(1):95–123. doi:10.1080/07474938.2016.1222068

Pacini, D., Windmeijer, F. (2016). Robust inference for the two-sample 2SLS estimator. *Economics Letters* 146: 50–54. doi:10.1016/j.econlet.2016.06.033

Pagan, A. (1984). Econometric issues in the analysis of regressions with generated regressors. *International Economic Review* 25(1):221–247. doi:10.2307/2648877

Racine, J. S. (1997). Consistent significance testing for nonparametric regression. *Journal of Business & Economic Statistics* 15(3):369–378.

Racine, J. S., Hart, J., Li, Q. (2006). Testing the significance of categorical predictor variables in nonparametric regression models. *Econometric Reviews* 25(4):523–544. doi:10.1080/07474930600972590

Racine, J. S., Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics* 119(1):99–130. doi:10.1016/S0304-4076(03)00157-X

Rilstone, P. (1996). Nonparametric estimation of models with generated regressors. *International Economic Review* 37(2):299–313. doi:10.2307/2527325

Shah, R. D., Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *Annals of Statistics* 48(3):1514–1538.

Stengos, T., Yan, B. (2001). Double kernel nonparametric estimation in semiparametric econometric models. *Journal of Nonparametric Statistics* 13(6):883–906. doi:10.1080/10485250108832882

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics* 10(4): 1040–1053.

Su, L., Murtazashvili, I., Ullah, A. (2013). Local linear GMM estimation of functional coefficient IV models with an application to estimating the rate of return to schooling. *Journal of Business & Economic Statistics* 31(2): 184–207. doi:10.1080/07350015.2012.754314

Veroff, J., McClelland, L., Marquis, K. (1971). Measuring intelligence and achievement motivation in surveys. Technical Series Paper #71-01, Survey Research Center-Institute for Social Research, University of Michigan.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48(4):817–838. doi:10.2307/1912934

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. Cambridge, MA: MIT Press.

Wooldridge, J. M. (2013). *Introductory Econometrics: A Modern Approach*, 5th ed. Mason, OH: South-Western Cengage Learning.

Yatchew, A. (1997). An elementary estimator of the partial linear model. *Economics Letters* 57(2):135–143. doi:10.1016/S0165-1765(97)00218-8

Zabalza, A., Arrufat, J. L. (1985). The extent of sex discrimination in Great Britain. In: Zabalza, A., Tzannatos, Z., eds., *Women and Equal Pay: The Effects of Legislation on Female Employment and Wages in Britain*. Cambridge, UK: Cambridge University Press, pp. 70–96.

Appendix A. Technical proofs

A.1. Construction of product kernel for estimating $g_2(\cdot)$

We employ a continuous univariate kernel for each of d_{3C} continuous variables $X_{3C,p}$, $p = 1, \ldots, d_{3C}$. For simplicity, suppose that $X_{3C,p}$ is smoothed by a univariate symmetric kernel $K(\cdot)$ and bandwidth h_p. Then, the product kernel for $X_{3C,p}$ is

$$\mathcal{K}(t_{3C}; x_{3C}, \mathbf{h}) = \prod_{p=1}^{d_{3C}} \frac{1}{h_p} K\left(\frac{t_{3C,p} - x_{3C,p}}{h_p}\right),$$

where $t_{3C} := (t_{3C,1}, \ldots, t_{3C,d_{3C}})$, $x_{3C} := (x_{3C,1}, \ldots, x_{3C,d_{3C}})$ and $\mathbf{h} := (h_1, \ldots, h_{d_{3C}})$ are vectors of data points, design points and bandwidths, respectively.

Next, we construct a kernel for the discrete component. While a variety of discrete kernels can be applied for this component, our focus is on those given by Aitchison and Aitken (1976). In what follows a product of their discrete kernels is exclusively considered. Each of d_{3D} discrete variables $X_{3D,q}$, $q = 1, \ldots, d_{3D}$ is assumed to take $r_q(\geq 2)$ different values, i.e., $X_{3D,q} \in \{0, 1, \ldots, r_q - 1\}$. In addition, each discrete variable is classified into either unordered or ordered, because the kernels employed for the two types of categorical variables differ slightly. The univariate discrete kernel for an *unordered* variable is

$$l\left(t_{3D,q}; x_{3D,q}, \lambda_q\right) := \begin{cases} 1 & \text{if } t_{3D,q} = x_{3D,q} \\ \lambda_q/(r_q - 1) & \text{if } t_{3D,q} \neq x_{3D,q} \end{cases},$$

where $t_{3D,q}$, $x_{3D,q}$ and $\lambda_q \in (0, 1)$ are the data point, design point and bandwidth, respectively. The univariate discrete kernel for an *ordered* variable takes the form of

$$\ell\left(t_{3D,q}; x_{3D,q}, \lambda_q\right) := \begin{pmatrix} r_q \\ \left|t_{3D,q} - x_{3D,q}\right| \end{pmatrix} (1 - \lambda_q)^{r_q - \left|t_{3D,q} - x_{3D,q}\right|} \lambda_q^{\left|t_{3D,q} - x_{3D,q}\right|}.$$

If there are $q_1(\leq d_{3D})$ unordered discrete variables, then the product kernel for all d_{3D} discrete variables is given by

$$\mathcal{L}(t_{3D}; x_{3D}, \boldsymbol{\lambda}) = \left\{\prod_{q=1}^{q_1} l\left(t_{3D,q}; x_{3D,q}, \lambda_q\right)\right\}\left\{\prod_{q=q_1+1}^{d_{3D}} \ell\left(t_{3D,q}; x_{3D,q}, \lambda_q\right)\right\},$$

where $t_{3D} := (t_{3D,1}, \ldots, t_{3D,d_{3D}})$, $x_{3D} := (x_{3D,1}, \ldots, x_{3D,d_{3D}})$ and $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_{d_{3D}})$.

For the continuous component, we also apply an alternative nonstandard kernel. Taking compactness of $\mathbb{X}_{3C}$ into account (see Assumption 2), we employ the beta kernel by Chen (1999) in place of the univariate symmetric kernel. For a design point $x \in [0, 1]$ and the smoothing parameter b, the beta kernel is defined as

$$K_{B(x,b)}(t) = \frac{t^{x/b}(1 - t)^{(1-x)/b}}{B\{x/b + 1, (1 - x)/b + 1\}} \mathbf{1}\{t \in [0, 1]\}.$$

It turns out that convergence properties of nonparametric estimators based on this kernel family have not been established. In a companion paper (Hirukawa et al., 2022), we provide a careful proof of weak and strong uniform convergence with rates of nonparametric estimators smoothed by the beta kernel.

A.2. Proof of Theorem 1

The proof requires a lemma on uniform consistency of the kernel regression estimator $\hat{g}_2(\cdot)$. Let $g_{2k}(\cdot)$ be the kth element of $g_2(\cdot)$ for $k = 1, \ldots, d_2$. In order to differentiate the continuous kernels used in the nonparametric estimation of $g_{2k}(\cdot)$, we denote the estimates obtained using νth-order symmetric and beta kernels by $\hat{g}_{2k}^{S\nu}(\cdot)$ and $\hat{g}_{2k}^{B}(\cdot)$, respectively. Then, the lemma below establishes weak uniform consistency with rates of $\hat{g}_{2k}^{S\nu}(\cdot)$ and $\hat{g}_{2k}^{B}(\cdot)$.

Lemma A1. *If Assumptions 1–6 hold, then, for $k = 1, \ldots, d_2$, as $m \to \infty$,*

$$\sup_{x_3 \in \mathbb{X}_3} \left| E\left\{ \hat{g}_{2k}^{S\nu}(x_3) \right\} - g_{2k}(x_3) \right| = O\left(\sum_{p=1}^{d_{3C}} h_p^{\nu} + \sum_{q=1}^{d_{3D}} \lambda_q \right),$$

$$\sup_{x_3 \in \mathbb{X}_3} \left| \hat{g}_{2k}^{S\nu}(x_3) - E\left\{ \hat{g}_{2k}^{S\nu}(x_3) \right\} \right| = O_p\left(\sqrt{\frac{\log m}{m \prod_{p=1}^{d_{3C}} h_p}} \right),$$

$$\sup_{x_3 \in \mathbb{S}_{X_{3C}} \times \mathbb{X}_{3D}} \left| E\left\{ \hat{g}_{2k}^{B}(x_3) \right\} - g_{2k}(x_3) \right| = O\left(\sum_{p=1}^{d_{3C}} b_p + \sum_{q=1}^{d_{3D}} \lambda_q \right), \text{ and}$$

$$\sup_{x_3 \in \mathbb{S}_{X_{3C}} \times \mathbb{X}_{3D}} \left| \hat{g}_{2k}^{B}(x_3) - E\left\{ \hat{g}_{2k}^{B}(x_3) \right\} \right| = O_p\left(\sqrt{\frac{\log m}{m \sqrt{\prod_{p=1}^{d_{3C}} b_p \theta_p}}} \right).$$

A.2.1. Proof of Lemma A1

First two statements on $\hat{g}_{2k}^{S\nu}(\cdot)$ are implied by Theorem 2.1 of Li and Ouyang (2005). Remaining two statements on $\hat{g}_{2k}^{B}(\cdot)$ are established by Theorem 7 of Hirukawa et al. (2022) with $r_n = O(1)$ due to uniform boundedness from below of $f(\cdot)$. ■

A.2.2. Proof of Theorem 1

To save space, we concentrate only on the case with the beta estimator $\hat{g}_2^B(\cdot)$. The PILS estimator admits the expansion

$$\begin{aligned} \hat{\beta}_{PILS} &= \beta + S_{n,\hat{X},\hat{X}}^{-1} S_{n,\hat{X},\epsilon} + S_{n,\hat{X},\hat{X}}^{-1} S_{n,\hat{X},\left\{g_2(X_3) - \hat{g}_2^B(X_3)\right\}'\beta_2} \\ &= \beta + S_{n,\hat{X},\hat{X}}^{-1} S_{n,\hat{X},\left[g_2(X_3) - E\left\{\hat{g}_2^B(X_3)|X_3\right\}\right]'\beta_2} \\ &\quad + S_{n,\hat{X},\hat{X}}^{-1} S_{n,\hat{X},\epsilon} - S_{n,\hat{X},\hat{X}}^{-1} S_{n,\hat{X},\left[\hat{g}_2^B(X_3) - E\left\{\hat{g}_2^B(X_3)|X_3\right\}\right]'\beta_2}, \end{aligned} \tag{A1}$$

where the third and fourth terms on the right-hand side can be interpreted as those of the sampling error in regression (1) and the approximation error of g_2, respectively. It follows from Lemma A1 that $B_{g_2} := S_{n,\hat{X},\hat{X}}^{-1} S_{n,\hat{X},\left[g_2(X_3) - E\left\{\hat{g}_2^B(X_3)|X_3\right\}\right]'\beta_2} = O_p(\sum_{p=1}^{d_{3C}} b_p + \sum_{q=1}^{d_{3D}} \lambda_q)$. Each of the last two terms is $O_p(n^{-1/2})$ by the central limit theorem (CLT); see the proof of Theorem 2 for more details. Therefore,

$$\hat{\beta}_{PILS} = \beta + O_p\left(\sum_{p=1}^{d_{3C}} b_p + \sum_{q=1}^{d_{3D}} \lambda_q \right) + O_p(n^{-1/2}) + O_p(n^{-1/2}) \xrightarrow{p} \beta$$

as $n, m \to \infty$. ■

A.3. Proof of Theorem 2

We continue to focus on the case with the beta estimator $\hat{g}_2^B(\cdot)$. Short-hand notations such as $g_{2i} = g_2(X_{3i})$, $E_i(\hat{g}_{2i}^B) = E\{\hat{g}_2^B(X_{3i})|X_{3i}\}$, and $\mathcal{W}_{j.i} = \mathcal{W}(X_{3j}; X_{3i}, \mathbf{b}, \boldsymbol{\lambda})$ also apply. Then, by Lemma A1 and CLT,

$$S_{n,\hat{X},\hat{X}} = S_{n,X,X} + o_p(1) \xrightarrow{p} \Phi_{X,X} > 0, \text{ and} \tag{A2}$$

$$\sqrt{n} S_{n,\hat{X},\epsilon} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \epsilon_i + o_p(1) \xrightarrow{d} N\left(0_{(d+1)\times 1}, \Omega_1\right) \tag{A3}$$

for the sampling error part. Moreover,

$$S_{n,\hat{X},\left[\hat{g}_2^B(X_3) - E\left\{\hat{g}_2^B(X_3)|X_3\right\}\right]'\beta_2} = S_{n,X,\left[\hat{g}_2^B(X_3) - E\left\{\hat{g}_2^B(X_3)|X_3\right\}\right]'\beta_2} + S_{n,\hat{X}-X,\left[\hat{g}_2^B(X_3) - E\left\{\hat{g}_2^B(X_3)|X_3\right\}\right]'\beta_2}$$

for the approximation error part, where the second term is of a smaller order of magnitude than the first term. Also observe that $\hat{g}_{2i}^B - E_i(\hat{g}_{2i}^B) = \{\tilde{g}_{2i}^B - E_i(\tilde{g}_{2i}^B)\} + (\tilde{\eta}_{2i}^B - \eta_{2i}) + \eta_{2i}$, where $\tilde{\varphi}_i^B := \sum_{j=1}^{m} \varphi_j \mathcal{W}_{j.i} / \sum_{j=1}^{m} \mathcal{W}_{j.i}$ for $\varphi \in \{g_2, \eta_2\}$. Then, by a similar argument to Proposition 9 of Stengos and Yan (2001) and CLT,

$$\sqrt{n}S_{n,\hat{X},\left[\hat{g}_2^B(X_3)-E\left\{\hat{g}_2^B(X_3)|X_3\right\}\right]'\beta_2} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\eta_{2i}'\beta_2 + o_p(1) \xrightarrow{d} N\left(0_{(d+1)\times 1}, \Omega_2\right). \tag{A4}$$

Finally, it is straightforward to see that for every i and k,

$$E\left[\left(\hat{X}_i\epsilon_i\right)\left\{\hat{X}_k\left(\hat{g}_{2k}^B - E_k\left(\hat{g}_{2k}^B\right)\right)'\beta_2\right\}'\right] = 0_{(d+1)\times(d+1)}. \tag{A5}$$

The result is established by substituting (A2)–(A5) into (A1) and rearranging it. ■

A.4. Proof of Corollary 1

Again to save space, we provide the proof for the case with the beta estimator $\hat{g}_2^B(\cdot)$. To control the convergence rate of θ, let $\theta \propto b^{\delta/d_{3C}}$ for a sufficiently small $\delta > 0$ so that $b/\theta \to 0$ can be established. Because $B_{g_2} = O_p(b) + o_p(n^{-1/2})$, $\sqrt{n}(\hat{\beta}_{PILS} - \beta - B_{g_2}) = \sqrt{n}(\hat{\beta}_{PILS} - \beta) + o_p(1)$ holds if

$$nb^2 = O\left\{\left(\frac{n}{m}\right)(\log m)^{4\alpha} m^{1-4\alpha}\right\} \to 0.$$

This is the case if one of the following conditions is true: (i) $n/m \to \kappa$ and $\alpha > 1/4$; (ii) $n/m \to 0$ and $\alpha > 1/4$; or (iii) $n/m \to \infty$, $n/m^{4\alpha} \to 0$ at a polynomial rate and $\alpha > 1/4$. It also follows from Assumption 6′(b) that

$$\frac{\log m}{m(b\theta)^{d_{3C}/2}} = O\left\{\left(\frac{\log m}{m}\right)^{1-\alpha(d_{3C}+\delta)}\right\} \to 0$$

should be the case. This holds if $1 - \alpha(d_{3C} + \delta) > 0$ or $\alpha < 1/(d_{3C} + \delta) < 1/d_{3C}$. Therefore, for $\hat{\beta}_{PILS}$ to be $\sqrt{n}$-consistent, α must satisfy $1/4 < \alpha < 1/d_{3C}$. This range is non-empty if $d_{3C} \leq 3$. ■