Yet another look at omitted variable bias: A two-sample alternative to using instruments*

Masayuki Hirukawa† Irina Murtazashvili‡ Artem Prokhorov§
Ryukoku University Drexel University University of Sydney

June 2018

Abstract

When conducting regression analysis, econometricians often face the situation where some relevant regressors are unavailable in the data set at hand. One common solution is to look for some valid instruments for the regressors that are suspected to be endogenous due to their possible correlation with the omitted variables. Another solution is to look for a proxy. However, in many cases no such variable is available in the same data set. This paper shows how to combine the original data set with one containing the 'missing' regressors even when the two data sets do not have common observations. The use of additional data improves estimation efficiency and we propose a consistent semiparametric two-sample estimator of the parameters of interest. We explore the asymptotic properties of the estimator and show, using Monte Carlo simulations, that it dominates the solution involving instrumental variables, both in terms of bias and efficiency. An application to the PSID and NLS data indicates the importance of our estimation approach in empirical research.

Keywords: Data combination; endogeneity; instrumental variable estimation; omitted variable bias; two-sample estimation.

JEL Classification Codes: C13; C14; C31.

*Financial support through grants from Japan Society of the Promotion of Science (M. Hirukawa, Project No. 15K03405) and the Russian Science Foundation (A. Prokhorov, Project No. 16-18-10432) for various and non-overlapping parts of this research is gratefully acknowledged.

†e-mail: hirukawa@econ.ryukoku.ac.jp

‡e-mail: im99@drexel.edu.

§e-mail: artem.b.prokhorov@gmail.com.

1 Introduction

Omitted variable bias is a challenging problem in applied work. If the omitted variable is relevant and its correlation with the included regressors is strong, estimation bias is substantial. In the absence of a proxy, the most common solution is to look for an instrument.

The use of instrumental variables (IVs) has been associated with at least two problems. First, instruments that are not strongly correlated with the endogenous variables, i.e., weak instruments, lead to large inconsistencies in the IV estimates even if only a weak correlation exists between the instruments and the structural equation error, i.e., the instruments only slightly violate the validity assumption (see, e.g., Bound, Jaeger, and Baker, 1995). Even if the IVs are valid, the weak instrument problem leads to large asymptotic 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 estimates 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 properties.

A classic example of the missing regressor is an ability measure in Mincer's (1974) wage regression. Card (1995) argues that the estimation result suffers from the "ability bias" unless the regression includes a variable representing ability as a regressor. However, micro-level data sets such as the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID) do not 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., Zabalza and Arrufat, 1985; Black, Trainor, and Spencer, 1999). Work experience is an important regressor in the wage regression. Although the General Household Survey (GHS) and CPS contain wages and other predictors, they include no actual work experience. On the other hand, the National Longitudinal Survey (NLS) and PSID contain actual work experience along with other predictors.

Valid and strong instruments may not be available in the same sample, either, and ingenious methods have been proposed to combine data from more than one source in such cases (see, e.g., Angrist and Krueger, 1992, 1995; Murtazashvili, Liu, and Prokhorov, 2015). However, two-sample IV estimates inherit the same problems as their single-sample counterparts. In line with this notion, Choi, Gu and Shen (2018) extend inference based on the two-sample IV estimation to the framework of weak instruments.

This paper compares the two estimation strategies and argues for a two-sample approach using a proxy. In particular, we propose a new semiparametric estimator of regression parameters based on an imputed proxy, obtained using data from the second sample. Precisely speaking, our two-sample estimation procedure is to run ordinary least squares (OLS) after replacing the missing regressor with a nonparametric estimate of its conditional mean. Therefore, unlike Hirukawa and Prokhorov (2018), the estimator does not rely on the combined sample constructed via the nearest-neighbor matching (NNM). We explore large-sample properties of the estimator. Its finite-sample properties are examined through Monte Carlo simulations. It is numerically confirmed that the two-sample estimator dominates the IV-based alternative. In an empirical application, we are able to obtain data sets permitting estimation of the return to schooling using all the competing approaches, including IV estimation and proxy estimation based on both a sole data set and a two-sample procedure.

In addition to a comparison of IV and two-sample estimations within a unified

framework, nonparametric imputation of the missing regressor in our two-sample estimation contributes in two important areas. First, econometricians quite commonly impute the variables that are necessary for estimation but missing in the first sample, upon the availability of the second sample (see, e.g., Fang, Keane and Silverman, 2008; Flavin and Nakagawa, 2008). These papers impose some parametric form on the conditional mean of the missing regressor and proceed to imputation. However, their approaches have at least two problems. First, if a parametric model for imputation turns out to be misspecified, then consistency of parameter estimates in the model of interest is no longer guaranteed. Second, estimation errors associated to imputation are not explicitly considered. Rather, imputed values are treated as error-free, and it appears that no adjustments in standard errors of the parameter estimates are made. In contrast, our nonparametric imputation is by construction consistent, and thus consistency of the parameter estimates is also ensured.

Second, it is also worth mentioning the relation of our nonparametric imputation to nonparametrically generated regressors (see, e.g., Li and Wooldridge, 2002; Mammen, Rothe and Schienle, 2012, 2016; Hahn and Ridder, 2013). These papers explore statistical properties of semi- or nonparametric regression models when the regressors that cannot be observed directly (e.g., conditional means) are estimated by nonparametric methods, as Pagan (1984) does for fully parametric regression models. At first glance, our nonparametric imputation appears to be similar to their approaches. A crucial difference, however, is that the latter estimates the unobservable regressors within the same data set, whereas in our approach, the nonparametric imputation is made across two samples.

The remainder of this paper is organized as follows. Sections 2 describes the model and the two alternative estimation strategies in more detail. It is demonstrated that some stringent regularity conditions are required for consistency of IV estimates. Section 3 focuses on the new two-sample estimator and its convergence properties.

Section 4 presents results of Monte Carlo simulations to compare finite-sample properties of IV and two-sample estimators. As an empirical example, in Section 5, we apply the two-sample estimator to a version of Mincer's (1974) wage regression. Section 6 concludes with a few questions for future research. All proofs are given in the Appendix.[1]

The paper adopts the following notational conventions: $\|A\| = \{\mathrm{tr}\,(A'A)\}^{1/2}$ 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 The Model and Two Estimation Strategies

We consider the following standard 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 it is assumed that either β_1 or β_3 is the parameter of 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 the second data set which will be required for a two-sample estimation. We assume that $\mathcal{S}_1 = (Y, X_1, X_{3I}, X_{3E})$ and $\mathcal{S}_2 = (X_2, X_3) := (X_2, X_{3I}, X_{3E})$, where X_3 is the vector of common variables across the two samples that is partitioned into those included (X_{3I}) and those excluded (X_{3E}) from the regression in (1). Observe that the regressor X_2 is missing in $\mathcal{S}_1$

[1] GAUSS codes implementing the IV and two-sample estimators are available from the authors upon request.

although it is assumed to be relevant in the regression, i.e., $\beta_2 \neq 0$. Even though $\mathcal{S}_1$ and $\mathcal{S}_2$ contain common variables X_3 which we call *matching variables*, this does not mean they need to have common observations. 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.

In order to obtain a consistent estimator of the parameter in this setting, econometricians will typically follow one of two estimation strategies, namely, IV and two-sample estimations. In what follows, we discuss each strategy in further detail.

2.1 Strategy 1: IV estimation

Econometricians may attempt to complete the estimation using only $\mathcal{S}_1$, e.g., in light of a cost associated with searching for another data source $\mathcal{S}_2$. In this case, X_2 in (1) is treated as the vector of omitted variables. Accordingly, regression (1) becomes

$$Y = X_S'\beta_S + v, \tag{2}$$

where $X_S := (1, X_1', X_{3I}')'$, $\beta_S := (1, \beta_1', \beta_3')'$ and $v := u + X_2'\beta_2$ (the subscript "S" stands for a short regression).

OLS for (2) is inconsistent if (X_1, X_{3I}) and X_2 are correlated. Suppose that $\mathcal{S}_1$ additionally includes a vector of instruments $Z \in \mathbb{R}^{d_Z}$ that correlate with (X_1, X_{3I}) but not with X_2. Also suppose that $\mathcal{S}_1$ has n *iid* observations so that $\mathcal{S}_1 = \mathcal{S}_{1n} = \{(Y_i, X_{1i}, X_{3Ii}, X_{3Ei}, Z_i)\}_{i=1}^n$. Note that X_{3E} is not used for estimation as a set of variables that are neither instruments nor relevant regressors unless it is part of Z. Then, it is possible to estimate β_S consistently on the basis of the moment restriction

$$E(Z_S v) := E\left\{ \begin{bmatrix} 1 \\ Z \end{bmatrix} (u + X_2'\beta_2) \right\} = 0_{(d_Z+1)\times 1}.$$

Assuming $d_Z \geq d_1 + d_{3I}$, consistency of this estimator will be based on some restrictive assumptions such as $Z \perp\!\!\!\perp X_2$ and $E(X_2)'\beta_2 = 0$. The benefit is a simple

estimator which in the just-identified case ($d_Z = d_1 + d_{3I}$) can be written as

$$\hat{\beta}_{IV,S} := \hat{Q}_{ZX}^{-1}\hat{R}_{ZY} := \left(\frac{1}{n}\sum_{i=1}^{n} Z_{S,i}X_{S,i}'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} Z_{S,i}Y_i\right).$$

It is easy to show that it satisfies $\hat{\beta}_{IV,S} \xrightarrow{p} \beta_S$ and $\sqrt{n}\left(\hat{\beta}_{IV,S} - \beta_S\right) \xrightarrow{d} N\left(0_{(d_Z+1)\times 1}, V_{Z,S}\right)$, where

$$V_{Z,S} := Q_{ZX,S}^{-1}\Omega_{ZZ,S}Q_{ZX,S}'^{-1} := \left\{E\left(Z_S X_S'\right)\right\}^{-1} E\left\{\left(Z_S v\right)\left(Z_S v\right)'\right\}\left\{E\left(X_S Z_S'\right)\right\}^{-1}.$$

3 Strategy 2: Two-sample estimation with an imputed proxy of X_2

3.1 Additional notation

Suppose that instruments may be unavailable or there may be other reasons why strategy 1 may be avoided, e.g., weak instruments. Also suppose that econometricians have access to $\mathcal{S}_2 = \mathcal{S}_{2m} = \{(X_{2j}, X_{3Ij}, X_{3Ej})\}_{j=1}^{m}$.

We now introduce some additional notation that can help us define our new two-sample estimator and derive its asymptotic properties. First, it is assumed that the vector of matching variables $X_3 = (X_{3I}, X_{3E})$ consists 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. 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

$$g(X_3) = \begin{bmatrix} g_1(X_3) \\ g_2(X_3) \end{bmatrix} := \begin{bmatrix} E(X_1|X_3) \\ E(X_2|X_3) \end{bmatrix}, \text{ and}$$
$$\eta = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} := \begin{bmatrix} X_1 - g_1(X_3) \\ X_2 - g_2(X_3) \end{bmatrix}.$$

3.2 The estimators of β and $g_2(\cdot)$

Our two-sample estimator starts from reformulating the long regression (1) as

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

where $X := \left(1, X_1', g_2(X_3)', X_{3I}'\right)'$ and $\epsilon := u + \eta_2'\beta_2$. Therefore, the estimator fundamentally takes the form of an OLS-type regression of Y on X.

However, the conditional mean $g_2(\cdot)$ is unknown. In essence, we wish to obtain an estimator of $g_2(X_3)$ using $\mathcal{S}_2$ and use $\mathcal{S}_1$ in constructing a proxy of X_2. The proxy for X_2 will be based on the proximity of the common (or matching) variables X_3 between $\mathcal{S}_1$ and $\mathcal{S}_2$ but the estimator will not rely on nearest-neighbors or similar proximity measure. We will simply apply OLS after obtaining a nonparametric estimate of $g_2(X_3)$ and imputing it directly in place of X_2 in (1).

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

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

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

The estimator of β in the final form is

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

Because this estimator is the OLS with $\hat{g}_2(X_{3i})$ imputed in place of the missing regressor X_{2i}, we call it the *plug-in least squares* (PILS) estimator hereinafter. At first glance, this approach may be a variant of generated regressors studied by Pagan (1984). However, while Pagan (1984) exclusively considers the regressors generated parametrically within the same dataset, $\hat{g}_2(\cdot)$ is generated *nonparametrically from a different dataset*.

Our 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, from among all available nonparametric methods we choose the kernel regression smoother for mixed continuous and categorical data by Racine and Li (2004).

A key part of the estimator is the construction of a product kernel. Let $\mathcal{K}(t_{3C}; x_{3C}, \mathbf{h})$ and $\mathcal{L}(t_{3D}; x_{3D}, \lambda)$ denote the kernel smoothers for the continuous and discrete components of X_3, respectively. We provide details of both smoothers in Appendix A.1. Then, the product kernel for X_3 is

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

It follows that the nonparametric regression estimator of $g_2(\cdot)$ is defined as

$$\hat{g}_2(X_{3i}) := \frac{\sum_{j=1}^{m} X_{2j} \mathbb{K}(X_{3j}; X_{3i}, \mathbf{h}, \lambda)}{\sum_{j=1}^{m} \mathbb{K}(X_{3j}; X_{3i}, \mathbf{h}, \lambda)},\ i = 1, \ldots, n.$$

As discussed in Appendix A.1, there are options for the specific univariate kernel to use for continuous variables. In addition to the standard symmetric kernels, we consider the beta kernel by Chen (1999) as an attractive alternative because it 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. In particular, the latter implies that the amount of smoothing by this kernel changes in an adaptive manner.

3.3 Convergence properties of the PILS estimator

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) $E(\eta_1\eta_2') = 0$.

(iii) Second-order derivatives of $g_2(\cdot)$ with respect to X_{3C} are continuous and bounded on $\mathbb{X}_{3C}$.

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

Assumption 4. The univariate continuous kernel is either (a) a symmetric and bounded pdf that satisfies the first-order Lipschitz condition or (b) the beta kernel.

Assumption 5. For sequences of the smoothing parameters, one of the following holds as $m \to \infty$: (a) for a symmetric kernel, $(h_1, \ldots, h_{d_{3C}}) = (h_{1,m}, \ldots, h_{d_{3C},m})$ and $(\lambda_1, \ldots, \lambda_{d_{3D}}) = (\lambda_{1,m}, \ldots, \lambda_{d_{3D},m})$ satisfy

$$\sum_{s=1}^{d_{3C}} h_s + \sum_{s=1}^{d_{3D}} \lambda_s + \frac{\ln m}{m h_1 \cdots h_{d_{3C}}} \to 0;$$

and (b) for the beta kernel, $(b_1, \ldots, b_{d_{3C}}) = (b_{1,m}, \ldots, b_{d_{3C},m})$ and $(\lambda_1, \ldots, \lambda_{d_{3D}}) = (\lambda_{1,m}, \ldots, \lambda_{d_{3D},m})$ satisfy

$$\sum_{s=1}^{d_{3C}} b_s + \sum_{s=1}^{d_{3D}} \lambda_s + \frac{\ln m}{m \left(b_1 \cdots b_{d_{3C}}\right)^{1/2}} \to 0.$$

First three assumptions are basically the same as Assumptions 1-3 of Hirukawa and Prokhorov (2018). In particular, it follows from Assumption 3(i) that X_{3E} may be recognized as the variable that is necessary for model building but is not part of regression (1) due to an exclusion restriction. It is also worth emphasizing that nonlinearity of $g_2(\cdot)$ in Assumption 3(iv) is required only when all continuous matching variables are included as regressors in (1). Otherwise, excluded continuous matching variables introduce additional randomness in $g_2(\cdot)$, which helps in identifying the model parameter β. Assumptions 4 and 5 help establish uniform weak 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, Murtazashvili and Ullah (2013) for a symmetric kernel, and Bouezmarni and Rolin (2003) and Shi and Song (2016) for the beta kernel.

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 asymptotically negligible bias term due to kernel smoothing. Each theorem holds regardless of the number of matching variables and regardless of the divergence patterns in (n, m).

Theorem 1. *If Assumptions 1-5 hold, then* $\hat{\beta}_{PILS} \xrightarrow{p} \beta$ *as* $n, m \to \infty$.

Theorem 2. *If Assumptions 1-5 hold, then* $\sqrt{n}\left(\hat{\beta}_{PILS} - \beta - B_{g_2}\right) \xrightarrow{d} N\left(0_{(d+1)\times 1}, V_X\right)$ *as* $n, m \to \infty$, *where*

$$B_{g_2} := \hat{Q}_{\hat{X}\hat{X}}^{-1} \frac{1}{n} \sum_{i=1}^{n} \hat{X}_i \left\{ g_2\left(X_{3i}\right) - \hat{g}_2\left(X_{3i}\right) \right\}' \beta_2$$

and

$$V_X := Q_{XX}^{-1} \Omega_{XX} Q_{XX}^{-1} := \left\{ E\left(XX'\right) \right\}^{-1} E\left\{ \left(X\epsilon\right)\left(X\epsilon\right)' \right\} \left\{ E\left(XX'\right) \right\}^{-1}.$$

While the PILS estimator is consistent, its convergence rate is affected by the bias term 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. This can be recognized as the curse of dimensionality in continuous matching variables. It appears that this problem is unavoidable in a regression that uses a nonparametric component (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 this case, we modify Assumption 5 in order to control the bias and variance convergences of $\hat{g}_2\left(\cdot\right)$ more easily.

Assumption 5'. For sequences of smoothing parameters, one of the following holds as $m \to \infty$: (a) for a symmetric kernel, $(h_1, \ldots, h_{d_{3C}}) = (h_{1,n,m}, \ldots, h_{d_{3C},n,m})$ and $(\lambda_1, \ldots, \lambda_{d_{3D}}) = (\lambda_{1,n,m}, \ldots, \lambda_{d_{3D},n,m})$ satisfy

$$h_1, \ldots, h_{d_{3C}} \propto h,\ \lambda_1, \ldots, \lambda_{d_{3D}} \propto h^2, \text{ and } h + \frac{\ln m}{m h^{d_{3C}}} \to 0;$$

and (b) for the beta kernel, $(b_1, \ldots, b_{d_{3C}}) = (b_{1,n,m}, \ldots, b_{d_{3C},n,m})$ and $(\lambda_1, \ldots, \lambda_{d_{3D}}) = (\lambda_{1,n,m}, \ldots, \lambda_{d_{3D},n,m})$ satisfy

$$b_1, \ldots, b_{d_{3C}} \propto b,\ \lambda_1, \ldots, \lambda_{d_{3D}} \propto b, \text{ and } b + \frac{\ln m}{m b^{d_{3C}/2}} \to 0.$$

This allows us to obtained an unbiased estimator.

Corollary 1. Under Assumptions 1-4 and 5', if *(i)* $d_{3C} \leq 3$ *and* $n/m \to \kappa \in (0, \infty)$ *or (ii)* $d_{3C} \leq 4$ *and* $n/m \to 0$ *or (iii)* $d_{3C} \leq 3$ *and* $n/m \to \infty$ *. Then,* $\sqrt{n}\left(\hat{\beta}_{PILS} - \beta\right) \xrightarrow{d} N\left(0_{(d+1)\times 1}, V_X\right)$ *as* $n, m \to \infty$, where V_X *is defined in Theorem 2.*

Remark 1. (Smoothing parameter selection) We briefly refer to the problem of choosing the smoothing parameters h and b with a focus on the most realistic case $n/m \to \kappa$. Shrinkage rates of the smoothing parameters are $h \propto m^{-\alpha}$ and $b \propto m^{-2\alpha}$ for some $\alpha \in (1/4, 1/d_{3C})$; see the proof of Corollary 1 in Appendix A.4 for more details. Observe that these rates are faster than the optimal ones $h^* \propto m^{-1/(4+d_{3C})}$ and $b^* \propto m^{-2/(4+d_{3C})}$ that can balance 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. All we can do is to set the exponent α within the aforementioned range.

Remark 2. (Covariance estimation) Covariance estimation is essential for inference and is simple for PILS because it is a version of OLS. Define the PILS residual

as $\hat{\epsilon}_i := Y_i - \hat{X}_i'\hat{\beta}_{PILS}$. Then, it is straightforward to see that a consistent estimator of V_X is $\hat{V}_X := \hat{Q}_{\hat{X}\hat{X}}^{-1}\hat{\Omega}_{\hat{X}\hat{X}}\hat{Q}_{\hat{X}\hat{X}}^{-1}$, where $\hat{\Omega}_{\hat{X}\hat{X}} := (1/n)\sum_{i=1}^{n}\left(\hat{X}_i\hat{\epsilon}_i\right)\left(\hat{X}_i\hat{\epsilon}_i\right)'$ is the heteroskedasticity-robust covariance estimator by Eicker (1963) and White (1980).

Remark 3. (Measurement error in the missing regressor) It may be the case that the missing regressor X_2 can be observed with measurement error. This issue typically occurs if an ability measure is missing in the wage regression, for example. Suppose that we can observe $\tilde{X}_2 = X_2 + e$ at best, where e is the measurement error. As long as $E(e|X_3) = 0_{d_2\times 1}$, the effect of e is filtered out via kernel smoothing and

$$\tilde{g}_2(X_{3i}) := \frac{\sum_{j=1}^{m}\tilde{X}_{2j}\mathbb{K}(X_{3j}; X_{3i}, \mathbf{h}, \lambda)}{\sum_{j=1}^{m}\mathbb{K}(X_{3j}; X_{3i}, \mathbf{h}, \lambda)} = \hat{g}_2(X_{3i}) + o_p(1),\ i = 1, \ldots, n$$

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

Remark 4. (Testing for consistency of IV estimation) We recall 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} exists (i.e., some matching variables are included as regressors), then some elements of the instrument vector Z are either a part of X_{3I} or correlated with X_{3I} due to relevance. As a result Z and X_2 are also correlated, and IV estimation for the short regression (2) becomes inconsistent. This observation suggests that we can test the null of consistency of IV estimation 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 positive set on $\text{supp}(X_{3I})$. Several versions of the test of significance in nonparametric regression have been proposed in the literature: examples include Fan and Li (1996) and Racine (1997) for continuous regressors and Lavergne (2001) and Racine, Hart and Li (2006) for discrete or categorical regressors, to name a few.

3.4 A comparison with the MSII estimator

The matched-sample indirect inference (MSII) estimator was proposed by Hirukawa and Prokhorov (2018) to handle what is called *hot-deck imputation*, that is the prac-

tice of using imputed values for Census non-respondents without accounting for the imputation bias. We conclude this section by comparing PILS with MSII. Our focus is again on the case $n/m \to \kappa$.

The MSII estimation starts by obtaining a complete sample by means of NNM on the observed variables. As before, let $\mathcal{S}_1 = (Y, X_1, X_3)$ denote a sample for which we have all responses and let $\mathcal{S}_2 = (X_2, X_3)$ denote the sample which contains the variable with non-response as well as the observed variables to match on. Hot-deck imputation uses a distance measure between X_3 in $\mathcal{S}_1$ and in $\mathcal{S}_2$ to obtain a matched value of X_2 for $\mathcal{S}_1$. Specifically, for each observation of X_3 in $\mathcal{S}_1$, K-NNM picks out K closest matches of $X_2$from $\mathcal{S}_2$ through finding first to Kth closest matches of X_3 in S_2 with respect to some distance function such as the Mahalanobis distance.The resulting complete sample can be written as follows

$$\mathcal{S} = \left\{ \left(Y_i, X_{1i}, X_{2j_1(i)}, \ldots, X_{2j_K(i)}, X_{3Ii}, X_{3Ei} \right) \right\}_{i=1}^{n}.$$

A matched-sample OLS estimation (MSOLS) for the regression of Y_i on $X_{i,j(i)} := \left(1, X_{1i}', X_{2j(i)}', X_{3Ii}'\right)'$, where $X_{2j(i)} := (1/K)\sum_{k=1}^{K} X_{2j_k(i)}$, generates a non-vanishing, classical measurement error bias. The bias is attributed to using $X_{2j(i)}$ as a proxy for X_{2i}, and the source of attenuation is $\Sigma_2 = E\left(\eta_2 \eta_2'\right)$.

As is well known in the literature on errors-in-variables models, the bias cannot be corrected in general without imposing additional identification conditions. However, in the above two-sample setup it can be corrected analytically with no such extra conditions as $\mathcal{S}_2$ serves as repeated measurements. The MSII is a bias-corrected estimator defined as follows

$$\hat{\beta}_{II} := \left(\frac{1}{n}\sum_{i=1}^{n} X_{i,j(i)} X_{i,j(i)}' - \frac{1}{K}\hat{\Sigma} \right)^{-1} \frac{1}{n}\sum_{i=1}^{n} X_{i,j(i)} Y_i,$$

where $\hat{\Sigma} := \operatorname{diag}\left\{ 0_{(d_1+1)\times(d_1+1)}, \hat{\Sigma}_2, 0_{d_{3I}\times d_{3I}} \right\}$, and $\hat{\Sigma}_2$ is the difference-based variance estimator that can be obtained by reordering $\mathcal{S}_2$ with respect to X_3.

Notice that MSII can attain $\sqrt{n}$-asymptotic normality when the number of matching variables is only one. To overcome the curse of dimensionality in continuous matching variables, Hirukawa and Prokhorov (2018) propose a two-step estimator called the fully-modified MSII (MSII-FM) estimator. In its second step, MSII-FM eliminates the second-order bias due to the so-called *matching discrepancy* (Abadie and Imbens, 2006) asymptotically by means of a polynomial approximation similar to the one studied by Abadie and Imbens (2011). It is demonstrated that the estimator can achieve parametric convergence when the number of matching variables is four or less.

Both MSII(-FM) and PILS are $\sqrt{n}$-consistent and asymptotically normal two-sample estimators. Nonetheless, the two approaches to restoring consistency differ. Consistency of MSII(-FM) is established by imputing X_2 from $\mathcal{S}_2$ and then eliminating the non-vanishing bias caused by the imputation. The bias correction requires to estimate Σ_2, and the estimation error from $\hat{\Sigma}_2$ is $O_p\left(n^{-1/2}\right)$. 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 $O_p\left(n^{-1/2}\right)$ rate.

On the other hand, consistency of PILS comes from the fact that a consistent estimate of $E\left(X_2|\, X_3\right)$ is used in place of the missing X_2. Since PILS does not need bias correction, it involves only an asymptotically normal term, similarly to an OLS estimation. Therefore, the asymptotic variance of PILS does not much exceed that of OLS. We compare efficiency of MSII(-FM) and PILS numerically through Monte Carlo simulations in the next section.

PILS has three novel features relative to MSII(-FM). First, while imputation of the omitted variable is based on proximity of the variables that are common to both samples (the matching variables), the estimator does not rely on NNM. Essentially, we employ an estimate of the conditional mean of the omitted variable given the matching variables and so our estimator can be viewed as an analog to the kernel-

based matching estimator by Heckman, Ichimura and Todd (1998). Hence it requires no bias correction that is a key ingredient in MSII(-FM).

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

4 Finite-Sample Performance

4.1 Monte Carlo Setup

In this section we conduct Monte Carlo simulations to compare finite-sample properties of the IV, MSII and PILS estimators that are estimated from the same regression. Suppose that β_1 is the parameter of interest in the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u, \tag{3}$$

where X_1 and X_2 have reduced forms

$$X_1 = \pi_0 + \pi_1 Z + \pi_2 X_{3E_C} + \pi_3 X_{3E_D} + \eta_1, \text{ and} \tag{4}$$

$$X_2 = h\left(X_{3E_C}\right) + X_{3E_D} + \eta_2 \tag{5}$$

for some variables Z, X_{3E_C} and X_{3E_D} that will be described shortly. Throughout it is assumed that two samples, namely, $\mathcal{S}_1 = \{(Y_i, X_{1i}, X_{3E_C i}, X_{3E_D i}, Z_i)\}_{i=1}^n$ and $\mathcal{S}_2 = \{(X_{2j}, X_{3E_C j}, X_{3E_D j})\}_{j=1}^m$, are only observable. While Z is used exclusively as the instrument for X_1, X_{3E_C} and X_{3E_D} serve as excluded continuous and discrete matching variables that can be employed for imputing the missing regressor X_2. The complete sample $\mathcal{S}^* = \{(Y_i, X_{1i}, X_{2i}, X_{3E_C i}, X_{3E_D i}, Z_i)\}_{i=1}^n$ is the sample that would not be observed in practice.

There are three options to estimate β_1 consistently. In the first option only $\mathcal{S}_1$ is used. For this purpose the long regression (3) can be rewritten as the short regression

$$Y = \beta_0 + \beta_1 X_1 + (u + \beta_2 X_2) := \beta_0 + \beta_1 X_1 + v, \tag{6}$$

where X_2 is assumed to have a zero mean. Because both X_1 and X_2 depend on matching variables X_{3E_C} and X_{3E_D}, they are correlated and thus the OLS estimator from the regression (6) is inconsistent. On the other hand, the IV estimator with $(1, Z)$ chosen as instruments becomes consistent. The remaining two options rely on $\mathcal{S}_1$ and $\mathcal{S}_2$. In the second option, after constructing the matched sample $\mathcal{S}$ from $\mathcal{S}_1$ and $\mathcal{S}_2$, we run MSII for the regression (3) using $\mathcal{S}$. Alternatively, it is possible to run PILS as the third option using both $\mathcal{S}_1$ and $\mathcal{S}_2$.

The data are generated in the following manner. First, Z and X_{3E_C} are drawn independently from $U[-2, 2]$, and X_{3D} is generated as a Bernoulli-type binary variable that takes $\pm 1/2$ with equal probability. Second, given $(\xi_1, \xi_2)' \overset{iid}{\sim} N(0_{2\times 1}, I_2)$, X_1 is generated by putting $\pi_0 = \pi_1 = \pi_2 = 1$ in (4). Third, it follows from (5) that generating X_2 requires to specify the functional form of $h(\cdot)$. In this study, the following six functional forms are considered, where $E(X_2) = 0$ is ensured for each

specification:

$$h(x) = \begin{cases} x & \text{[Model A]} \\ \tanh(x) = \{\exp(x) - \exp(-x)\} / \{\exp(x) + \exp(-x)\} & \text{[Model B]} \\ \exp(x) - (1/4)\{\exp(2) - \exp(-2)\} & \text{[Model C]} \\ x^3 & \text{[Model D]} \\ 2\left(\sqrt{x+2} - 4/3\right) & \text{[Model E]} \\ x + (5/\tau)\,\phi(x/\tau) - (5/2)\{\Phi(2/\tau) - 1/2\},\ \tau = 3/4 & \text{[Model F]} \end{cases}.$$

Linearity of $h(\cdot)$ in Model A does not cause identification failure, because both of matching variables X_{3E_C} and X_{3E_D} are excluded from (3). Models B-E are also monotone, and Models C and E are globally convex and concave, respectively. Model F is a non-monotone function, where $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of $N(0,1)$, respectively. Inspired by the Monte Carlo design of Horowitz and Spokoiny (2001), it can be viewed as a linear function with a bump. Finally, Y is generated by setting $\beta_0 = \beta_1 = \beta_2 = 1$ and $u \overset{iid}{\sim} N(0,1)$ in (3).

The above procedure provides us with two observable samples $\mathcal{S}_1$ and $\mathcal{S}_2$, and one complete (but unobservable) sample $\mathcal{S}^*$. Finally, the matched sample

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

is constructed via the NNM with respect to (X_{3E_C}, X_{3E_D}). The NNM is based on the Mahalanobis metric and the number of matches $K \in \{1, 2, 4, 8\}$.

With regards to sample sizes, we choose $n \in \{500, 1000, 2000\}$ and put $m = n$ for simplicity. For each combination of sample sizes (n, m) and the functional form of $h_2(\cdot)$, 1000 Monte Carlo replications are drawn. The following seven estimators of β_1 are examined: (i) the infeasible OLS estimator from the long regression (3) using $\mathcal{S}^*$ [OLS*]; (ii) the OLS estimator from the short regression (6) using $\mathcal{S}_1$ only [OLS-S]; (iii) the IV estimator from the short regression (6) using $\mathcal{S}_1$ only [IV-S]; (iv) the MSOLS estimator from the long regression (3) using $\mathcal{S}$ [MSOLS]; (v) the MSII estimator from the long regression (3) using $\mathcal{S}$ [MSII]; (vi) the PILS estimator for the long regression (3) using $\mathcal{S}_1$, $\mathcal{S}_2$ and the Epanechnikov kernel $K(u) = (3/4)(1-u^2)\mathbf{1}\{|u| \leq 1\}$ [PILS-E]; and (vii) the PILS estimator for the long

regression (3) using $\mathcal{S}_1$, $\mathcal{S}_2$ and the beta kernel [PILS-B]. Notice that only OLS-S and MSOLS are inconsistent, whereas all others are consistent. Implementing two PILS estimators requires to choose the smoothing parameters. We put $\hat{h} = \hat{\sigma}_{X_{3E_C}} m^{-2/7}$ for the Epanechnikov kernel, $\hat{b} = \hat{\sigma}_{U_3} m^{-4/7}$ for the beta kernel, and $\hat{\lambda} = m^{-4/7}$ for the discrete kernel, where $\hat{\sigma}_{X_{3E_C}}$ and $\hat{\sigma}_{U_3}$ are sample standard deviations of the data in the original scale X_{3E_C} and in the transformed scale $U_3 := (X_{3E_C} + 2)/4 \in [0, 1]$, respectively.

For each estimator of β_1, 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). Since OLS-S and MSOLS are both inconsistent, their standard errors are not well defined. Accordingly, $\overline{SE}$ and CR are not computed for these estimators. The standard error for MSII follows Proposition 1 of Hirukawa and Prokhorov (2017), whereas heteroskedasticity-robust standard errors by Eicker (1963) and White (1980) are calculated for all other estimators. In addition, to indicate the degree of endogeneity in the short regression (6), we present simulation averages of sample correlation coefficients between X_1 and X_2, denoted as "$\hat{\rho}_{X_1X_2}$".

TABLE 1 ABOUT HERE

4.2 Results

Simulation results are summarized in Table 1. Only the results of MSII (and MSOLS) for the cases with $K = 1$ (i.e., the single-match cases) are reported, because the results for $K \in \{2, 4, 8\}$ are qualitatively similar.[2] Because of conditional homoskedasticity of the error term u, OLS* is the best linear unbiased estimator for the long regression

[2]In our preliminary Monte Carlo study larger values of matches (e.g., $K \in \{16, 32, 64, 128\}$) have been also investigated. However, the results look inferior.

(3). The results indeed suggest that it is unbiased and yields the smallest standard deviations. Normal approximation looks fine in the sense that in almost all cases coverage rates equal the nominal 95% level of confidence. However, OLS* is an infeasible, oracle estimator, and we should think of it merely as the best-case scenario. It is of more practical importance and interest to compare efficiency among the rest of consistent estimates of β_1.

We start from examining IV-S and MSII. Although both estimators are less efficient than OLS*, the reasons for their efficiency losses differ. The efficiency loss of IV-S is due to leaving the missing regressor X_2 omitted and relying on instruments, whereas that of MSII is attributed to imputing X_2 from another data source. For three out of six models, MSII exhibits smaller simulation standard deviations than IV-S. In other words, we may be often better off making effort to find another data source that contains the missing regressor than searching for valid instruments within the data set at hand. It is also noteworthy that coverage rates of both estimators are close to the nominal 95% level of confidence.

Now our focus shifts to two PILS estimators. It can be immediately found that both PILS estimators always exhibit smaller standard deviations than IV-S and MSII, and that efficiency gain from PILS is often substantial. An increase in the asymptotic variance of IV-S from that of PILS is attributed to the additional inclusion of $\beta_2^2 Var\left\{h\left(X_{3E_C}\right) + X_{3E_D}\right\}$ in the error variance. On the other hand, as described in the previous section, MSII is based on the matched data with a proxy of the missing regressor imputed, and it analytically corrects the attenuation generated by the measurement error in the proxy. The regression error and the bias-correction term serve as asymptotically normal terms with the same convergence rate, and as a consequence MSII tends to have a highly complicated (and often large) asymptotic variance. A closer inspection also reveals that PILS-B always outperforms PILS-E in terms of $RMSE$, although the margin is small. In addition, $\overline{SE}$ is reasonably

close to SD, which indicates that the (properly-scaled) covariance estimator $\hat{V}_X$ in Remark 2 yields good estimates of standard deviations of PILS. A concern, if any, is that sometimes CR indicates slight under-coverage.

We conclude this section by making a few remarks on a comparison between PILS and OLS*. First, unlike OLS*, PILS is not unbiased. However, it is nearly unbiased for large sample sizes. Second, standard deviations of the latter are always greater than those of the former. The relative efficiency loss can be thought of as the price to pay for identifying and estimating the regression using two samples jointly. In particular, SD and $\overline{SE}$ of PILS are second smallest. Judging from the fact that those of OLS* are smallest, PILS appears to be an efficient estimation strategy within the framework of missing regressors in regression analysis.

5 Application to Estimation of Return to Schooling

5.1 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 it to estimate the rate of return to additional schooling using alternative estimators available to researchers. This includes the traditional one-sample OLS and IV estimators as well as MSII and PSII.

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:

$$\log(earnings) = \beta_0 + \beta_1 education + \beta_2 ability + \beta_3 experience + \beta_4 experience^2 + \beta_5 married + \beta_6 black + \beta_7 south + \beta_8 urban + u, \quad (7)$$

where $\log(earnings)$ is the natural logarithm of the individual's total annual labor income, *education* is the person's completed years of education, *ability* is the individual's ability/skills (with zero mean), *experience* is work experience of the person, *married* is an indicator for whether the individual is married, *black* is an indicator for whether the person is black, *south* is an indicator for whether the person currently lives in the southern geographical region, *urban* is an indicator of the individual's urban residence while growing up[3], and u is an idiosyncratic error.

The main issue with estimating equation (7) is that researchers are typically unable to observe the individual's skills. Because of that, the (one-sample) OLS estimate of the return to schooling – β_1 – from equation (7) where *ability* is excluded from the vector of explanatory variables is likely to suffer from the so-called 'ability bias' (see, e.g., Card, 1995, for details).

Two textbook solutions to this problem are (1) to find within the same data set a valid proxy for the unobservable skills and use OLS estimation, and (2) 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 compare in practice the one- and two-sample approaches to the same applied task.

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

5.2 Constructing the samples

To estimate earnings and wage equations for the US, labor economists frequently use such micro-datasets as CPS, NLS and PSID. 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 usually collect such important wage determinants as actual work experience and ability. PSID 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.

More specifically, as our main dataset we employ the 1972 wave of PSID that has been ongoing since 1968. This longitudinal survey has been conducted annually through 1997, and biennially since then. We choose this PSID wave so we can estimate equation (7) using not only the IV and PILS but also using a proxy for the unobserved skills and ability contained in PSID. While the PSID generally does not contain any measures of unobserved ability, the 1972 wave is among the few PSID waves that does include an ability measure.

In the 1972 wave, the PSID respondents were administered a particular assessment of abstract thinking – the Lorge-Thorndike Intelligence Test. 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." The Lorge-Thorndike Test administered to the PSID respondents contained 13 questions, where each question received a point for the correct response. Thus, the score for the entire PSID 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.

To obtain actual labor market experience we utilize the longitudinal nature of PSID. Specifically, we use the 1974 and 1973 waves. The 1972-1973 waves did not ask respondents about their entire work experiences but only whether they worked positive hours in the previous year. However, in the 1974 wave, the respondents were also asked how many years they worked since they were 18 years old. Therefore, we first obtain actual years of work experience for the respondents as reported in 1974. We then subtract one or two from this total if individuals reported working positive hours in 1972 or in 1972-1973.

In addition to the information on the individual's ability and actual work experience discussed above, we also gather information on the individual's completed years of education, age, whether the person is black, married, and lives in the southern geographical region[4]. 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 the information on the completed years of education by the individual's father. The last variable – the father's educational level – is obtained so we can use it as an instrument for the individual's educational attainment when no measure of ability is available.

There are two main advantages of using family background variables for this purpose. First, they are available in many datasets including PSID. Second, they are usually highly correlated with the individual's educational level. While, generally, family background variables have been criticized for having a direct effect on the individual's income, violating the assumption of instrument's validity, they nevertheless have been used as instruments for education in many studies including those focusing

[4]We follow the 1979 National Longitudinal Survey 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.

on wage and earnings regressions (see, e.g., Blackburn and Neumark, 1992; Callan and Harmon, 1999; Parker and van Praag, 2006).

There is a literature that argues that once the individual's education is controlled for, (at least some) family background characteristics have no independent effect on income (see, e.g., Griliches, 1979). Moreover, recent research by Hoogerheide et al. (2012, p. 515) "provides confidence in the use of family background variables as instruments in income regressions" and shows (p. 519) that "using the father's education as an instrument in an income regression is a viable option for solving the endogeneity problem with regards to education." This further justifies using father's education as an instrument for individual's education.

TABLE 2 ABOUT HERE

Our main sample contains 2,430 men who reported positive labor income in 1971. Panel A of Table 2 reports summary statistics for the variables in our PSID sample. The sample correlation between individual's education and father's education is 0.434.

The PSID sample we construct allows us to estimate equation (7) using the one-sample OLS and IV approaches. However, there are several reasons why we might want to consider two-sample approaches. First, we might be concerned that the intelligence test reported in the PSID is not a good ability proxy. Second, the more likely scenario is that empirical researchers working with the other PSID waves (or other surveys) simply do not have any measures of ability in their samples.

To exploit the two-sample estimation approaches, we need to find a second sample where the ability measure is available and/or potentially more reliable. For this purpose, we employ the CARD dataset provided with Wooldridge (2010). This dataset contains observations from the 1977 wave of the Young Men Cohort of the National Longitudinal Survey (NLS). We exploit those observations that contain positive wages in 1976.

Our choice of the second data set is driven by three considerations. First, the CARD dataset is readily available with a popular textbook. Second, NLS collects information on several tests administered to the respondents during their interviews or in the secondary schools they attended. The results for some of these tests can be used to create a measure of unobserved individual ability. Third, some of the variables in the CARD dataset overlap with some of the variables in the PSID sample we construct.

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 1976. 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 equation (7).

Panels B and C of Table 2 report summary statistics for the NLS sample for two different sets of matching variables. Panel B provides summary statistics for $m = 1,102$ respondents when *education*, *married*, *black*, *south*, and *urban* are used as the included matching 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 matching variables, X_{3E}. Panel C contains summary statistics for $m = 197$ respondents when X_{3I} is the same as in Panel B but there are no excluded matching variables.

5.3 Estimation approaches

We report estimation results in Table 3. Columns (1)-(3) report the parametric estimation results based on just the PSID sample. Columns (1) reports the OLS estimates using the ability proxy available in the PSID 1972 waive. Column (2) reports the OLS results which does not account for unobserved ability. We note that

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 we use father's education as an instrumental variable for the individual's educational level. Columns (4)-(7) report two-sample semi-parametric estimation results based on both the PSID and NLS samples. Column (4) contains the OLS estimation results when the two samples are combined using hot-deck imputation and no bias correction is applied. This estimator is inconsistent. Columns (5) and (6) provide two sets of MSII-FM results that account for imputation biases using the second- or third-order polynomials as proposed by Hirukawa and Prokhorov (2018). Note that, in essence, the MSII approach is a bias-corrected version of the MSOLS estimation approach. Finally, column (7) reports the PILS results using the beta kernel, which we chose one the base of the results in the simulations section.

The two-sample estimation approaches use five variables common to both samples – *education*, *married*, *black*, *south*, and *urban* – as included matching variables and two variables – *age* and *south while growing up* – as excluded matching variables. We choose not to employ *experience* as a matching variable because experience in PSID represents actual experience, while experience in NLS represents potential experience. As a consequence, our entire set of matching variables is (*education*, *married*, *black*, *south*, *urban*, *age*, *south while growing up*), where *education* and *age* are treated as continuous.

The NNM approach for 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 of a respondent characterized by the corresponding values of the matching variables. As a consequence, $m = 1,102$ respondents remain in our $\mathcal{S}_2$.

The standard errors for MSOLS are the same as given in Table 2 of Hirukawa and

Prokhorov (2018). They should be interpreted with caution since MSOLS is inconsistent (and even its convergence rate is slower than parametric) and the standard errors for this method merely indicate measures of dispersion at the same scale as other estimates and are not intended for inference.

Our choice of kernel for PILS reflects the favorable Monte Carlo performance of the beta kernel in the previous section. For PILS-B, each continuous matching variable X is converted from its original scale to the variable U in the $[0,1]$ scale via

$$U := \frac{X - m_X}{M_X - m_X} \in [0,1]\,,$$

where M_X and m_X are maximal and minimal values of the pooled sample constructed from observations of X in S_1 and S_2. The reason for using the maximum and minimum of the pooled sample is differences in ranges of *education* and *age* between S_1 and S_2. Accordingly, $\hat{b} = \hat{\sigma}_U m^{-4/7}$ and $\hat{\lambda} = m^{-4/7}$ are chosen for the beta and discrete kernels, respectively, where $\hat{\sigma}_U$ is the sample standard deviation of a converted continuous matching variable U in S_2.

Table 4 provides parameter estimates for the same two-sample methods as in Table 3 but with no excluded matching variables. Parametric identification in this case is possible only when there is nonlinearity in conditional mean function of ability (X_2) given the matching variables (X_3). Columns (1)-(4) of Table 4 reproduce Columns (4)-(7) of Table 3 using the same included matching variables, kernel choices, etc., except that columns (2) and (3) in Table 4 report the initial (i.e. one-step) MSII results. In Table 4 we report the one-step MSII results and not the two-step MSII results as in Table 3 because the two sets of MSII results are nearly identical numerically suggesting that the second FM-step does not improve the quality of the estimates via bias correction. Since we use fewer matching variables now, only $m = 197$ respondents remain in our $\mathcal{S}_2$.

5.4 Empirical findings

As Table 3 suggests, the signs of coefficient estimates on all the regressors except for *ability* are as expected, and they are (mostly) highly significant. The MSOLS and MSII-FM estimator using the third-order polynomial for the power-series approximation in the second step yield negative (but insignificant) estimates of the ability effect, whereas the MSII-FM estimator using the second-order polynomial for the power-series approximation results in a positive (but also insignificant) estimate of the effect of ability on earnings. Interestingly, OLS* and PILS-B are the only two approaches that produce positive (as one would expect) and statistically significant estimates of the ability effect. Finally, note that the PILS-B standard errors tend to be smaller than those of MSII-FM, as predicted in the theoretical section.

TABLE 3 ABOUT HERE

Next, let us focus on the estimates of the rate of return to education. The estimation results reported in Table 3 provide some evidence that father's educational is unlikely to be a valid instrument for individual's education. This is so because if father's education is a valid instrument, the OLS-S and MSOLS estimates of β_1 – the return to education – are upward-inconsistent. However, we observe that the IV estimate of β_1 is the largest in magnitude. Furthermore, if the instrument is valid and there is no measurement error in the PSID ability measure then OLS* and IV approaches are both consistent. However, we see that OLS* and IV estimates are noticeably different, suggesting again that the two estimators are unlikely to be both consistent. We provide further discussion of such inconsistencies in Appendix A.5.

The infeasible OLS estimate of β_1 (OLS*) is the second smallest (conceding in magnitude only to PILS-B). 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 3. Indeed,

the feasible one-sample estimate of the return to education with omitted ability – OLS-S – is between IV and infeasible OLS. In fact, the MSOLS and both MSII-FM estimates of β_1 are also between the one-sample IV and infeasible OLS* estimates. Therefore, there is some 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-B. We note that PILS-B is also the closest (in absolute value) to the infeasible OLS estimate of β_1. The second closest to our benchmark estimate of β_1 is MSII-FM using the second-order polynomial in the second step. While this seems like a good outcome given that only MSII-FM and PILS are consistent approaches if the instrument is invalid, we cannot help but point out sensitivity of MSII-FM to the polynomial order. Indeed, the MSII-FM estimate of β_1 using the third-order polynomial is the second largest estimate (after the IV estimate). Furthermore, the possibility that PILS (not OLS*) is the closest to the true return to education is not out of the question. Based on our theoretical considerations in Appendix A.5 we know that if the PSID ability measure is error-ridden, OLS* is upward-inconsistent suggesting that the PILS estimate of β_1 is likely to be the closest to the true return to education.

TABLE 4 ABOUT HERE

Finally, Table 4 provides alternative MSOLS, MSII-FM and PILS estimates of the model parameters when fewer matching variables are used for estimation. Now we do not exploit the excluded matching variables available in the two samples. We produce the results in Table 4 to impose the nonlinearity in $E(X_2|X_3)$ restriction adopted in Hirukawa and Prokhorov (2018).

We note a substantial change in the performance of the MSII-FM estimator in this case. First, in contrast to Table 3, the MSII-FM approach is now insensitive to the order of the power-series approximation. Second, the MSII-FM and PILS estimates

are almost identical now. While the PILS estimate is further away from OLS* under the nonlinearity requirement, it is still the closest to OLS*.

Virtually no difference in numerical results between MSII-FM and PILS may be attributed to the omitted variable bias in estimating the conditional mean of the missing variable given matching variables. As indicated in Monte Carlo simulations, omitting truly relevant excluded matching variables yields numerically similar and heavily biased results among MSOLS, MSII-FM and PILS. Importantly, the results in Table 4 suggest caution in imposing the nonlinearity restriction: all the two-sample estimation approaches – MSOLS, MSII-FM, and PILS – are sensitive to the nonlinearity assumption.

6 Concluding Remarks

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

We have developed the PILS estimation procedure for models where endogenous regressors enter the model due to an omitted variable. The procedure first uses an auxiliary sample to obtain a nonparametric estimator of the conditional mean of the regressor missing in the first sample given matching variables available in both samples. The second step of the procedure is to simply run OLS after imputing the conditional mean estimate in place of the missing regressor.

We establish the asymptotic normality of PILS. The estimator turns out to be more efficient than existing two-sample estimators including the MSII(-FM) estimator recently proposed by Hirukawa and Prokhorov (2018). Attractive finite-sample

properties of PILS are confirmed in a Monte Carlo study. In particular, simulations have demonstrated numerically that PILS is more efficient than IV and MSII-FM and that its efficiency gain is often considerable.

However, in order for PILS to attain the parametric rate of convergence, the number of continuous matching variables must be four or less. The curse of dimensionality in continuous matching variables is still a concern, no different from the nearest-neighbor matching used for MSII-FM.

Therefore, as an extension we may adopt propensity score matching as a means of dimension reduction using multiple matching variables. This would involve using the observable variables to estimate a selection model for observations that are imputed, and obtaining the (imputation) propensity score.

In a related paper, Abadie and Imbens (2016) deliver asymptotic properties of the matching estimators of average treatment effects using an estimated propensity score as a plug-in. It may be worth pursuing a similar idea for matched-sample regression estimation.

Another research question that is worth pursuing is to derive efficiency bounds within the class of two-sample estimators under consideration. It is of importance and interest to examine whether PILS has attained the efficiency bound or there is still room to improve.

A Appendix: Technical Proofs

A.1 Product kernel construction for estimation of $g_2(\cdot)$

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

$X_{3C,s}$ is

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

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. Each of d_{3D} discrete variables $X_{3D,s}$, $s = 1, \ldots, d_{3D}$ is assumed to take $r_s\,(\geq 2)$ different values, i.e., $X_{3D,s} \in \{0, 1, \ldots, r_s - 1\}$. 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

$$l\left(t_{3D,s}; x_{3D,s}, \lambda_s\right) := \begin{cases} 1 & \text{if } t_{3D,s} = x_{3D,s} \\ \lambda_s & \text{if } t_{3D,s} \neq x_{3D,s} \end{cases},$$

where $t_{3D,s}$, $x_{3D,s}$ and $\lambda_s \in [0, 1]$ is the data point, the design point and the bandwidth, respectively. Given the same notations, the univariate discrete kernel for an *ordered* variable is in the form of

$$\ell\left(t_{3D,s}; x_{3D,s}, \lambda_s\right) := \begin{cases} 1 & \text{if } t_{3D,s} = x_{3D,s} \\ \lambda_s^{\left|t_{3D,s} - x_{3D,s}\right|} & \text{if } t_{3D,s} \neq x_{3D,s} \end{cases}.$$

If there are $p_1\,(\leq d_{3D})$ unordered discrete variables, then the product kernel for all d_{3D} discrete variables is given by

$$\begin{aligned} \mathcal{L}\left(t_{3D}; x_{3D}, \lambda\right) &= \left\{\prod_{s=1}^{p_1} l\left(t_{3D,s}; x_{3D,s}, \lambda_s\right)\right\}\left\{\prod_{s=p_1+1}^{d_{3D}} \ell\left(t_{3D,s}; x_{3D,s}, \lambda_s\right)\right\} \\ &= \left\{\prod_{s=1}^{p_1} \lambda_s^{\mathbf{1}\left\{t_{3D,s} \neq x_{3D,s}\right\}}\right\}\left\{\prod_{s=p_1+1}^{d_{3D}} \lambda_s^{\left|t_{3D,s} - x_{3D,s}\right|}\right\}, \end{aligned}$$

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

For the continuous component, an alternative kernel choice could be applied. Taking compactness of $\mathbb{X}_{3C}$ into account (see Assumption 2), we may employ the beta kernel by Chen (1999) in place of the univariate symmetric kernel. The beta kernel is defined as

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

for the design point $x \in [0,1]$ and the smoothing parameter b.

A.2 Proof of Theorem 1

To save space, we concentrate only on the case in which a symmetric kernel is employed for smoothing the continuous component in $\hat{g}_2(\cdot)$. A straightforward calculation yields

$$\hat{\beta}_{PILS} := \beta + B_{g_2} + \hat{Q}_{\hat{X}\hat{X}}^{-1} E_{R_X},$$

where

$$B_{g_2} = \hat{Q}_{\hat{X}\hat{X}}^{-1} \frac{1}{n} \sum_{i=1}^{n} \hat{X}_i \left\{ g_2(X_{3i}) - \hat{g}_2(X_{3i}) \right\}' \beta_2 \text{ and } E_{R_X} = \frac{1}{n} \sum_{i=1}^{n} \hat{X}_i \epsilon_i.$$

It follows from Theorem 2.1 of Li and Ouyang (2005) that

$$\begin{aligned} \|B_{g_2}\| \leq & \left\| \hat{Q}_{\hat{X}\hat{X}} \right\|^{-1} \|\beta_2\| \left\{ \frac{1}{n} \sum_{i=1}^{n} \left\| \hat{X}_i \right\| \|E\{\hat{g}_2(X_{3i})\} - g_2(X_{3i})\| \right. \\ & \left. + \frac{1}{n} \sum_{i=1}^{n} \left\| \hat{X}_i \right\| \|\hat{g}_2(X_{3i}) - E\{\hat{g}_2(X_{3i})\}\| \right\} \\ = & O_p \left(\sum_{s=1}^{d_{3C}} h_s^2 + \sum_{s=1}^{d_{3D}} \lambda_s \right) + o_p\left(n^{-1/2}\right). \end{aligned}$$

Likewise, by a central limit theorem,

$$\sqrt{n} E_{R_X} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \epsilon_i + o_p(1) \xrightarrow{d} N\left(0_{(d+1)\times 1}, \Omega_{XX}\right),$$

where $\Omega_{XX} := E\left\{ (X\epsilon)(X\epsilon)' \right\}$. Therefore,

$$\hat{\beta}_{PILS} = \beta + \left\{ O_p \left(\sum_{s=1}^{d_{3C}} h_s^2 + \sum_{s=1}^{d_{3D}} \lambda_s \right) + o_p\left(n^{-1/2}\right) \right\} + O_p\left(n^{-1/2}\right) \xrightarrow{p} \beta. \blacksquare$$

A.3 Proof of Theorem 2

This is obvious in light of the proof of Theorem 1. $\blacksquare$

A.4 Proof of Corollary 1

To save space, we again concentrate only on the case in which a symmetric kernel is employed for smoothing the continuous component in $\hat{g}_2(\cdot)$. We consider the most realistic case with $n/m \to \kappa$; other cases can be demonstrated in a similar manner. It follows from the proof of Theorem 1 that $B_{g_2} = O_p(h^2) + o_p\left(n^{-1/2}\right)$. Then, $\sqrt{n}\left(\hat{\beta}_{PILS} - \beta - B_{g_2}\right) = \sqrt{n}\left(\hat{\beta}_{PILS} - \beta\right) + o_p(1)$ if $nh^4 \to 0$. Combining this condition with Assumption 5' and $n \propto m$, we conclude that $h \propto m^{-\alpha}$ for some $\alpha \in (1/4, 1/d_{3C})$ and $1/4 < 1/d_{3C}$ must be the case. The latter is equivalent to $d_{3C} \leq 3$, which completes the proof. ■

A.5 Discussion of inconsistencies of the estimators

Here we discuss the estimators we consider in the empirical section in light of different assumptions applicable to the earnings equation and the PSID and NLS samples of data used for the empirical analysis. For simplicity, we focus on a reduced version of model (7) where we ignore all the covariates except for the ones of our main focus.

As a starting point, consider the following model:

$$\log(earnings) = \beta_0 + \beta_1 education + \beta_2 ability + u. \tag{A1}$$

The probability limit of the OLS estimate of β_1 from equation (A1) where *ability* is unobserved is

$$\operatorname*{plim}_{N\to\infty} \hat{\beta}_1^{OLS} = \beta_1 + \beta_2 \frac{cov(ability, education)}{var(education)}. \tag{A2}$$

If we assume that $cov(ability, education) > 0$ and $\beta_2 > 0$, $\hat{\beta}_1^{OLS}$ for equation (A1) with unobserved ability is upward-inconsistent.

Second, if there is no ability measure available in our sample, we can employ IV estimation of the wage equation (A1) and it will produce a consistent estimate of β_1 as long as the instrument used – father's education – is a valid instrument for *education*.

However, if our instrument has a direct effect on the individual's wage, the IV estimator of equation (A1) with unobservable ability will be inconsistent. Indeed, if the individual's wage is determined by:

$$\log(earnings) = \gamma_0 + \gamma_1 education + \gamma_2 father's\ education + u,$$

then the relation between β_1 (from equation (A1) without the ability variable) and γ_1 and γ_2 can be obtained following the omitted variable formula:

$$\beta_1 = \frac{cov(father's\ education, \log(earnings))}{cov(father's\ education, education)} = \gamma_1 + \gamma_2 \frac{\lambda \sigma_f}{\sigma_e},$$

where $\lambda = corr(father's\ education, education)$, $\sigma_f^2 = var(father's\ education)$, and $\sigma_e^2 = var(education)$. In this case, the probability limit of the IV estimator of β_1 from equation (A1) where *ability* is excluded from the equation is

$$\operatorname*{plim}_{N\to\infty} \hat{\beta}_1^{IV} = \frac{cov(father's\ education, \log(earnings))}{cov(father's\ education, education)} = \gamma_1 + \gamma_2 \frac{\sigma_f}{\lambda \sigma_e} = \beta_1 + \gamma_2 \frac{(1-\lambda^2)\sigma_f}{\lambda \sigma_e}. \tag{A3}$$

If we assume that $0 < \lambda < 1$ and $\gamma_2 > 0$, $\hat{\beta}_1^{IV}$ for equation (A1) with unobserved ability is upward-inconsistent.

Third, if there is measurement error in our measure of ability under the classical errors-in-variables assumption that the correlation between the true unobserved variable and its measurement error is zero, the OLS estimate of β_1 from equation (A1) with an error-ridden measure of ability will be inconsistent. Let

$$ability = ability^* + \epsilon,$$

where $ability^*$ is the unobserved error-free ability, *ability* is the observed mismeasured ability, and ϵ is the measurement error. We can show that

$$\operatorname*{plim}_{N\to\infty} \hat{\beta}_1^{OLS} = \beta_1 + \beta_2 \frac{\sigma_\epsilon^2 \sigma_{ea} - \sigma_a^2 \sigma_{e\epsilon}}{\sigma_e^2 \sigma_a^2 - \sigma_{ea}^2}, \tag{A4}$$

where $\sigma_\epsilon^2 = var(\epsilon)$, $\sigma_e^2 = var(education)$, $\sigma_a^2 = var(ability)$, $\sigma_{ea} = cov(education, ability)$, and $\sigma_{e\epsilon} = cov(education, \epsilon)$. If we assume that $\sigma_{ea} > 0$ and the correlation between

the individual's educational level and the measurement error in his ability is zero (i.e., $\sigma_{e\epsilon} = 0$), then

$$\underset{N\to\infty}{\text{plim}}\,\hat{\beta}_1^{OLS} = \beta_1 + \beta_2 \frac{\sigma_\epsilon^2 \sigma_{ea}}{\sigma_e^2 \sigma_a^2 (1 - \rho_{ea}^2)}, \tag{A5}$$

where $\rho_{ea} = corr(education, ability)$. Thus, $\hat{\beta}_1^{OLS}$ from equation (A1) with mismeasured ability is upward-inconsistent.

Finally, note that Hirukawa and Prokhorov (2018) provide a measurement error interpretation for the MSOLS estimator of the earnings equation when *ability* is unobserved in (A1). Thus, we expect the MSOLS estimate of β_1 from (A1) with missing *ability* to be also upward-inconsistent following our argument above for the case of mismeasured ability.

Based on the above calculations we can make the following conclusions. First, if we use a valid instrument for the endogenous regressor when using the IV approach to estimate β_1 in (A1) where *ability* is missing, then this IV estimator of β_1 is consistent while the OLS estimators of β_1 from the regression with mismeasured ability and the regression with omitted ability (as well as the MSOLS estimator) are upward-inconsistent.

Second, if we think there is no measurement error in some ability measure which is available to us, then the OLS* estimate of β_1 is consistent, while the OLS-S and MSOLS are still not.

Third, when the instrument used for endogenous education when ability is missing from equation (A1) is valid, generally, it is unclear which one of the three (four if we count MSOLS) upward inconsistencies considered above will be the largest and which the smallest.

References

[1] Abadie, A., and G. W. Imbens (2006): "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74, 235 - 267.

[2] Abadie, A., and G. W. Imbens (2011): "Bias-Corrected Matching Estimators for Average Treatment Effects," *Journal of Business & Economic Statistics*, 29, 1 - 11.

[3] Abadie, A., and G. W. Imbens (2016): "Matching on the Estimated Propensity Score," *Econometrica*, 84, 781 - 807.

[4] Angrist, J., and A. Krueger (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.

[5] Angrist, J., and A. Krueger (1995): "Split-sample instrumental variables estimates off the return to schooling," *Journal of Business and Economic Statistics*, 13(2), 225 - 235.

[6] Bekker, P. A. (1994): "Alternative Approximations to the Distributions of Instrumental Variable Estimators," *Econometrica*, 62(3), 657 - 681.

[7] Black, B., M. Trainor, and J. E. Spencer (1999): "Wage Protection Systems, Segregation and Gender Pay Inequalities: West Germany, the Netherlands and Great Britain," *Cambridge Journal of Economics*, 23, 449 - 464.

[8] Blackburn, M., and D. Neumark (1992): "Unobserved Ability, Efficiency Wages, and Interindustry Wage Differentials," *Quarterly Journal of Economics*, 1077, 1421 - 1436.

[9] Bouezmarni, T., and J.-M. Rolin (2003): "Consistency of the Beta Kernel Density Function Estimator," *Canadian Journal of Statistics*, 31, 89 - 98.

[10] Bound, J., D. A. Jaeger, and R. M. Baker (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.

[11] Callan, T., and C. Harmon (1999): "The Economic Returns to Schooling in Ireland," *Labour Economics*, 6, 543 - 550.

[12] Card, D. (1995): "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in L. N. Christophides, E. K. Grant, and R. Swidinsky (eds.), *Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp*. Toronto: University of Toronto Press, 201 - 222.

[13] Card, D. (2001): "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, 69, 1127 - 1160.

[14] Chao, J. C, and N. R. Swanson (2005): "Consistent estimation with a large number of weak instruments," *Econometrica*, 73(5), 1673 - 1692.

[15] Chen, S. X. (1999): "Beta Kernel Estimators for Density Functions," *Computational Statistics and Data Analysis*, 31, 131 - 145.

[16] Choi, J., J. Gu, and S. Shen (2018): "Weak-Instrument Robust Inference for Two-Sample Instrumental Variable Regression," *Journal of Applied Econometrics*, 33, 109 - 125.

[17] Eicker, F. (1963): "Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regressions," *Annals of Mathematica Statistics*, 34, 447 - 456.

[18] Fan, Y., and Q. Li (1996): "Consistent Model Specification Tests: Omitted Variables and Semiparametric Functional Forms," *Econometrica*, 64, 865 - 890.

[19] Fang, H., M. P. Keane, and D. Silverman (2008): "Sources of Advantageous Selection: Evidence from the Medigap Insurance Market," *Journal of Political Economy*, 116, 303 - 350.

[20] Flavin, M., and S. Nakagawa (2008): "A Model of Housing in the Presence of Adjustment Costs: A Structural Interpretation of Habit Persistence," *American Economic Review*, 98, 474 - 495.

[21] Flores-Lagunes, A. (2007): "Finite sample evidence of IV estimators under weak instruments," *Journal of Applied Econometrics*, 22(3), 677 - 694.

[22] Griliches, Z. (1977): "Estimating the Return to Schooling: Some Econometrics Problems," *Econometrica*, 45, 1 - 22.

[23] Griliches, Z. (1979): "Sibling Models and Data in Economics: Beginnings of a Survey" *Journal of Political Economy*, LXXXVII, Part 2, S37-64.

[24] Hahn, J., and G. Ridder (2013): "Asymptotic Variance of Semiparametric Estimators with Generated Regressors," *Econometrica*, 81, 315 - 340.

[25] Hansen, B. E. (2008): "Uniform Convergence Rates for Kernel Estimation with Dependent Data," *Econometric Theory*, 24, 726 - 748.

[26] Heckman, J. J., H. Ichimura, and P. Todd (1998): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261 - 294.

[27] Hirukawa, M., and A. Prokhorov (2018): "Consistent Estimation of Linear Regression Models Using Matched Data," *Journal of Econometrics*, 203(2), 344 - 358.

[28] Hoogerheide, L., J. Brock, and R. Thurik (2012): "Family Background Variables as Instruments for Education in Income Regressions: A Bayesian Analysis," *Economics of Education Review*, 31, 515 - 523.

[29] Horowitz, J. L., and V. G. Spokoiny (2001): "An Adaptive, Rate-Optimal Test of a Parametric Mean-Regression Model Against a Nonparametric Alternative," *Econometrica*, 69, 599 - 631.

[30] Lavergne, P. (2001): "An Equality Test across Nonparametric Regressions," *Journal of Econometrics*, 103, 307 - 344.

[31] Li, Q., and D. Ouyang (2005): "Uniform Convergence Rate of Kernel Estimation with Mixed Categorical and Continuous Data," *Economics Letters*, 86, 291 - 296.

[32] Li, Q., and J. M. Wooldridge (2002): "Semiparametric Estimation of Partially Linear Models for Dependent Data with Generated Regressors," *Econometric Theory*, 18, 625 - 645.

[33] Mammen, E., C. Rothe, and M. Schienle (2012): "Nonparametric Regression with Nonparametrically Generated Covariates," *Annals of Statistics*, 40, 1132 - 1170.

[34] Mammen, E., C. Rothe, and M. Schienle (2016): "Semiparametric Estimation with Generated Covariates," *Econometric Theory*, 32, 1140 - 1177.

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

[36] Murtazashvili, I., D. Liu, and A. Prokhorov (2005): "Two-sample nonparametric estimation of intergenerational income mobility in the United States and Sweden," *Canadian Journal of Economics*, 48(5), 1733 - 1761.

[37] Pagan, A. (1984): "Econometric Issues in the Analysis of Regressions with Generated Regressors," *International Economic Review*, 25, 221 - 247.

[38] Parker, S. C., and C. M. Van Praag (2006): "Schooling, Capital Constraints, and Entrepreneurial Performance: The Endogenous Triangle," *Journal of Business & Economic Statistics*, 24(4), 416 - 431.

[39] Racine, J. S. (1997): "Consistent Significance Testing for Nonparametric Regression," *Journal of Business & Economic Statistics*, 15, 369 - 378.

[40] Racine, J. S., J. Hart, and Q. Li (2006): "Testing the Significance of Categorical Predictor Variables in Nonparametric Regression Models," *Econometric Reviews*, 25, 523 - 544.

[41] Racine, J. S., and Q. Li (2004): "Nonparametric Estimation of Regression Functions with Both Categorical and Continuous Data," *Journal of Econometrics*, 119, 99 - 130.

[42] Shi, J., and W. Song (2016): "Asymptotic Results in Gamma Kernel Regression," *Communications in Statistics - Theory and Methods*, 45, 3489 - 3509.

[43] Su, L., I. Murtazashvili, and A. Ullah (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, 184 - 207.

[44] Veroff, J., L. McClelland, and K. Marquis (1971): "Measuring Intelligence and Achievement Motivation in Surveys," *Technical Series Paper #71-01, Survey Research Center-Institute for Social Research, University of Michigan, Ann Arbor.*

[45] White, H. (1980): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817 - 838.

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

[47] Zabalza, A., and J. L. Arrufat (1985): "The Extent of Sex Discrimination in Great Britain," in A. Zabalza and Z. Tzannatos (eds.), *Women and Equal Pay: The Effects of Legislation on Female Employment and Wages in Britain*, Cambridge, U.K.: Cambridge University Press, 70 - 96.

Table 1: **Monte Carlo Results**

$n\,(=m)$	$\hat{\rho}_{X_1X_2}$	Estimator	OLS*	OLS-S	IV-S	MSOLS	MSII	PILS-E	PILS-B
			Model A: $h(x) = x$						
500	0.4970	$Mean$	0.9980	1.4027	0.9935	1.2049	0.9868	1.0175	1.0122
		SD	0.0256	0.0362	0.0738	0.0423	0.0715	0.0415	0.0413
		$RMSE$	0.0257	0.4043	0.0741	0.2092	0.0727	0.0451	0.0431
		$\overline{SE}$	0.0260	–	0.0734	–	0.0744	0.0414	0.0414
		CR	96%	–	95%	–	97%	92%	94%
1000	0.4983	$Mean$	0.9999	1.4044	0.9996	1.2080	0.9963	1.0126	1.0095
		SD	0.0187	0.0269	0.0497	0.0296	0.0490	0.0284	0.0284
		$RMSE$	0.0187	0.4053	0.0497	0.2101	0.0491	0.0311	0.0299
		$\overline{SE}$	0.0184	–	0.0519	–	0.0509	0.0293	0.0292
		CR	95%	–	96%	–	96%	93%	94%
2000	0.4975	$Mean$	1.0008	1.4044	0.9996	1.2088	1.0000	1.0086	1.0067
		SD	0.0132	0.0186	0.0365	0.0211	0.0350	0.0206	0.0206
		$RMSE$	0.0133	0.4048	0.0365	0.2098	0.0350	0.0224	0.0216
		$\overline{SE}$	0.0130	–	0.0366	–	0.0355	0.0207	0.0207
		CR	95%	–	95%	–	95%	93%	94%
			Model B: $h(x) = \tanh(x) = \{\exp(x) - \exp(-x)\} / \{\exp(x) + \exp(-x)\}$						
500	0.4048	$Mean$	0.9983	1.2711	0.9942	1.1820	0.9766	1.0213	1.0144
		SD	0.0243	0.0340	0.0643	0.0384	0.0973	0.0405	0.0404
		$RMSE$	0.0243	0.2732	0.0645	0.1860	0.1001	0.0457	0.0429
		$\overline{SE}$	0.0247	–	0.0645	–	0.0933	0.0403	0.0404
		CR	96%	–	95%	–	97%	90%	93%
1000	0.4064	$Mean$	0.9998	1.2727	1.0000	1.1845	0.9917	1.0148	1.0107
		SD	0.0178	0.0250	0.0439	0.0269	0.0599	0.0281	0.0281
		$RMSE$	0.0178	0.2739	0.0439	0.1864	0.0605	0.0318	0.0300
		$\overline{SE}$	0.0175	–	0.0456	–	0.0596	0.0286	0.0286
		CR	95%	–	96%	–	97%	93%	94%
2000	0.4055	$Mean$	1.0007	1.2728	1.0000	1.1851	0.9983	1.0102	1.0075
		SD	0.0125	0.0174	0.0322	0.0191	0.0413	0.0202	0.0202
		$RMSE$	0.0125	0.2733	0.0322	0.1860	0.0413	0.0226	0.0215
		$\overline{SE}$	0.0124	–	0.0322	–	0.0403	0.0202	0.0202
		CR	95%	–	95%	–	94%	92%	94%

Continued on next page

Table 1 – continued from previous page

$n\,(=m)$	$\hat{\rho}_{X_1X_2}$	Estimator	OLS*	OLS-S	IV-S	MSOLS	MSII	PILS-E	PILS-B
		Model C: $h(x) = \exp(x) - (1/4)\{\exp(2) - \exp(-2)\}$							
500	0.5066	$Mean$	0.9982	1.5588	0.9917	1.1547	0.9925	1.0076	1.0032
		SD	0.0258	0.0463	0.0947	0.0447	0.0551	0.0402	0.0400
		$RMSE$	0.0259	0.5607	0.0950	0.1610	0.0557	0.0410	0.0401
		$\overline{SE}$	0.0262	–	0.0932	–	0.0568	0.0393	0.0392
		CR	95%	–	94%	–	96%	94%	95%
1000	0.5085	$Mean$	0.9995	1.5618	0.9989	1.1581	0.9980	1.0055	1.0030
		SD	0.0187	0.0335	0.0633	0.0313	0.0383	0.0275	0.0274
		$RMSE$	0.0187	0.5628	0.0633	0.1612	0.0384	0.0280	0.0276
		$\overline{SE}$	0.0185	–	0.0660	–	0.0398	0.0276	0.0276
		CR	95%	–	96%	–	95%	94%	94%
2000	0.5078	$Mean$	1.0006	1.5615	0.9999	1.1592	1.0003	1.0040	1.0024
		SD	0.0134	0.0238	0.0460	0.0225	0.0279	0.0201	0.0200
		$RMSE$	0.0134	0.5620	0.0460	0.1608	0.0279	0.0205	0.0201
		$\overline{SE}$	0.0131	–	0.0465	–	0.0280	0.0195	0.0195
		CR	95%	–	96%	–	95%	94%	95%
		Model D: $h(x) = x^3$							
500	0.5399	$Mean$	0.9980	1.8784	0.9914	1.1163	0.9920	0.9997	0.9966
		SD	0.0269	0.0637	0.1347	0.0474	0.0529	0.0413	0.0409
		$RMSE$	0.0270	0.8807	0.1349	0.1256	0.0535	0.0413	0.0410
		$\overline{SE}$	0.0268	–	0.1307	–	0.0543	0.0395	0.0393
		CR	95%	–	94%	–	96%	94%	94%
1000	0.5404	$Mean$	0.9995	1.8798	0.9971	1.1203	0.9990	1.0005	0.9991
		SD	0.0194	0.0467	0.0878	0.0329	0.0359	0.0276	0.0275
		$RMSE$	0.0194	0.8811	0.0878	0.1247	0.0359	0.0276	0.0275
		$\overline{SE}$	0.0190	–	0.0925	–	0.0379	0.0277	0.0276
		CR	95%	–	96%	–	96%	96%	96%
2000	0.5403	$Mean$	1.0005	1.8804	0.9989	1.1209	1.0001	1.0001	0.9995
		SD	0.0137	0.0318	0.0647	0.0242	0.0269	0.0202	0.0201
		$RMSE$	0.0137	0.8809	0.0647	0.1233	0.0269	0.0202	0.0201
		$\overline{SE}$	0.0134	–	0.0653	–	0.0267	0.0195	0.0195
		CR	95%	–	95%	–	94%	95%	95%

Continued on next page

Table 1 – continued from previous page

$n\,(=m)$	$\hat{\rho}_{X_1X_2}$	Estimator	OLS*	OLS-S	IV-S	MSOLS	MSII	PILS-E	PILS-B
			Model E: $h\,(x) = 2\left(\sqrt{x+2} - 4/3\right)$						
500	0.4543	$Mean$	0.9981	1.3349	0.9942	1.1956	0.9836	1.0184	1.0126
		SD	0.0250	0.0356	0.0690	0.0406	0.0781	0.0414	0.0412
		$RMSE$	0.0251	0.3368	0.0693	0.1997	0.0798	0.0453	0.0431
		$\overline{SE}$	0.0253	–	0.0687	–	0.0796	0.0409	0.0409
		CR	95%	–	95%	–	97%	92%	93%
1000	0.4554	$Mean$	0.9999	1.3361	0.9996	1.1980	0.9952	1.0132	1.0098
		SD	0.0184	0.0262	0.0465	0.0283	0.0522	0.0282	0.0282
		$RMSE$	0.0184	0.3371	0.0465	0.2000	0.0524	0.0312	0.0299
		$\overline{SE}$	0.0179	–	0.0485	–	0.0534	0.0289	0.0289
		CR	94%	–	96%	–	96%	93%	94%
2000	0.4547	$Mean$	1.0008	1.3364	0.9998	1.1988	0.9994	1.0087	1.0067
		SD	0.0129	0.0181	0.0344	0.0201	0.0371	0.0204	0.0203
		$RMSE$	0.0129	0.3368	0.0344	0.1998	0.0371	0.0222	0.0214
		$\overline{SE}$	0.0127	–	0.0343	–	0.0370	0.0204	0.0204
		CR	95%	–	94%	–	95%	93%	94%
			Model F: $h\,(x) = x + (5/\tau)\,\phi\,(x/\tau) - (5/2)\,\{\Phi\,(2/\tau) - 1/2\}\,,\ \tau = 3/4$						
500	0.4339	$Mean$	0.9983	1.4031	0.9945	1.1440	0.9926	1.0094	1.0038
		SD	0.0250	0.0417	0.0811	0.0424	0.0542	0.0394	0.0394
		$RMSE$	0.0251	0.4053	0.0813	0.1501	0.0547	0.0405	0.0396
		$\overline{SE}$	0.0250	–	0.0813	–	0.0547	0.0375	0.0376
		CR	94%	–	95%	–	96%	93%	93%
1000	0.4347	$Mean$	0.9999	1.4042	0.9998	1.1473	0.9995	1.0078	1.0043
		SD	0.0182	0.0311	0.0565	0.0299	0.0382	0.0275	0.0275
		$RMSE$	0.0182	0.4053	0.0565	0.1503	0.0382	0.0286	0.0278
		$\overline{SE}$	0.0177	–	0.0574	–	0.0381	0.0265	0.0265
		CR	95%	–	96%	–	95%	93%	94%
2000	0.4338	$Mean$	1.0007	1.4043	0.9991	1.1475	1.0006	1.0053	1.0029
		SD	0.0127	0.0214	0.0407	0.0212	0.0271	0.0195	0.0194
		$RMSE$	0.0128	0.4049	0.0407	0.1490	0.0271	0.0202	0.0197
		$\overline{SE}$	0.0125	–	0.0406	–	0.0268	0.0187	0.0187
		CR	95%	–	94%	–	94%	93%	94%

Notes: $Mean$ = simulation average of the parameter estimate; SD = simulation standard deviation of the parameter estimate; $RMSE$ = root mean-squared error of the parameter estimate; $\overline{SE}$ = simulation average of the standard error; and CR = coverage rate for the nominal 95% confidence interval.

Table 2: **Sample Characteristics**

Variable	Mean	Std. Dev.	Min.	Median	Max.
Panel A: The PSID Sample					
Earnings	8763.24	5968.39	30	7900	70000
ln (*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 with Excluded Matching Variables					
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
Panel C: The NLS sample with No Excluded Matching Variables					
Education	11.98	4.02	1	12	18
Married	0.54	0.50	0	1	1
Black	0.50	0.50	0	0	1
South	0.53	0.50	0	1	1
Urban	0.49	0.50	0	0	1
KWW score	29.59	7.95	4.00	30.70	43.43

Notes: $(n =)2,430$ individuals in the PSID sample. $(m =)1,102$ and $(m =)197$ distinct combinations of the variables are used from the NLS sample with and without excluded matching variables, respectively. We used the demeaned *IQ score* and *KWW score* when estimating the earnings equation.

Table 3: **Estimation Results with Excluded Matching Variables**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
					MSII-FM		
	OLS*	OLS-S	IV-S	MSOLS	2nd	3rd	PILS-B
Education	0.0635	0.0718	0.0953	0.0727	0.0685	0.0784	0.0598
	(0.0056)	(0.0053)	(0.0159)	(0.0059)	(0.0080)	(0.0080)	(0.0070)
Experience	0.0809	0.0818	0.0830	0.0826	0.0766	0.0546	0.0825
	(0.0043)	(0.0043)	(0.0045)	(0.0049)	(0.0065)	(0.0065)	(0.0043)
$Experience^2$	−0.0017	−0.0017	−0.0017	−0.0017	−0.0016	−0.0007	−0.0017
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Ability	0.0313	–	–	−0.0012	0.0029	−0.0008	0.0075
	(0.0078)	(–)	(–)	(0.0027)	(0.0081)	(0.0080)	(0.0029)
Married	0.3717	0.3793	0.3844	0.3799	0.3777	0.3770	0.3954
	(0.0539)	(0.0536)	(0.0536)	(0.0535)	(0.0535)	(0.0532)	(0.0541)
Black	−0.1302	−0.1741	−0.1249	−0.1849	−0.1504	−0.1869	−0.1749
	(0.0323)	(0.0316)	(0.0426)	(0.0393)	(0.0806)	(0.0784)	(0.0316)
South	−0.0921	−0.0983	−0.0814	−0.0979	−0.0989	−0.1059	−0.1036
	(0.0288)	(0.0287)	(0.0314)	(0.0286)	(0.0289)	(0.0285)	(0.0287)
Urban	0.1363	0.1499	0.1278	0.1538	0.1404	0.1355	0.1541
	(0.0282)	(0.0284)	(0.0332)	(0.0297)	(0.0390)	(0.0384)	(0.0281)
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 matching variables, and *age* and *south while growing up* are used as the excluded matching variables. The (demeaned) *IQ score* and *KWW score* variables are used as ability measures in the PSID and NLS samples, respectively.

Table 4: **Estimation Results without Excluded Matching Variables**

	(1)	(2)	(3)	(4)
		MSII-FM		
	MSOLS	2nd	3rd	PILS-B
Education	0.0814	0.0700	0.0700	0.0701
	(0.0109)	(0.0054)	(0.0054)	(0.0122)
Experience	0.0818	0.0818	0.0818	0.0818
	(0.0043)	(0.0043)	(0.0043)	(0.0043)
$Experience^2$	−0.0017	−0.0017	−0.0017	−0.0017
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Abilily	−0.0090	0.0016	0.0016	0.0012
	(0.0082)	(0.0015)	(0.0015)	(0.0082)
Married	0.3883	0.3775	0.3775	0.3776
	(0.0548)	(0.0535)	(0.0535)	(0.0562)
Black	−0.2413	−0.1620	−0.1619	−0.1721
	(0.0614)	(0.0362)	(0.0362)	(0.0344)
South	−0.1069	−0.0968	−0.0968	−0.0974
	(0.0311)	(0.0285)	(0.0285)	(0.0303)
Urban	0.1664	0.1468	0.1468	0.1501
	(0.0322)	(0.0285)	(0.0285)	(0.0285)
Data Combination?	Yes	Yes	Yes	Yes
Sample Size: n	2430	2430	2430	2430
m	197	197	197	197

Notes: The dependent variable is the log of total annual labor earnings. *Education*, *married*, *black*, *south*, and *urban* are used as the included matching variables, and no excluded matching variables are used. The (demeaned) *IQ score* and *KWW score* variables are used as ability measures in the PSID and NLS samples, respectively.