# Consistent Estimation of Linear Regression Models Using Matched Data*

Masayuki Hirukawa† Setsunan University

Artem Prokhorov‡ University of Sydney

14 April 2017

**Abstract**

Economists often use matched samples, especially when dealing with earnings data where a number of missing observations need to be imputed. In this paper, we demonstrate that the ordinary least squares estimator of the linear regression model using matched samples is inconsistent and has a non-standard convergence rate to its probability limit. If only a few variables are used to impute the missing data, then it is possible to correct for the bias. We propose two semiparametric bias-corrected estimators and explore their asymptotic properties. The estimators have an indirect-inference interpretation, and they attain the parametric convergence rate when the number of matching variables is no greater than four. Monte Carlo simulations confirm that the bias correction works very well in such cases.

**Keywords:** Bias correction; indirect inference; linear regression; matching estimation; measurement error bias.

**JEL Classification Codes:** C13; C14; C31.

---

*The first author gratefully acknowledges financial support from Japan Society of the Promotion of Science (grant numbers 23530259 and 15K03405).

†Faculty of Economics, Setsunan University, 17-8 Ikeda Nakamachi, Neyagawa, Osaka 572-8508, Japan; phone: (+81)72-839-8095; fax: (+81)72-839-8138; e-mail: hirukawa@econ.setsunan.ac.jp.

‡Discipline of Business Analytics, Business School, University of Sydney, H04-499 Merewether Building, Sydney, NSW 2006, Australia; phone: (+61)2-9351-6584; fax: (+61)2-9351-6409; e-mail: artem.prokhorov@sydney.edu.au.

# 1 Introduction

Suppose that we are interested in estimating a linear regression model

$$Y = \beta_0 + X_1'\beta_1 + X_2'\beta_2 + Z'\gamma + u := W'\theta + u,\ E(u|W) = 0, \tag{1}$$

using a random sample, where $X_1 \in \mathbb{R}^{d_1}$, $X_2 \in \mathbb{R}^{d_2}$ and $Z \in \mathbb{R}^{d_3}$. The reason for distinguishing between the regressors $X_1$, $X_2$ and $Z$ will become clear shortly. In addition, while $d_1 = 0$ is allowed, $d_2, d_3 > 0$ must be the case in our setup. When $W = (1, X_1', X_2', Z')' \in \mathbb{R}^{d+1}$, where $d := d_1 + d_2 + d_3$, is exogenous and a single random sample of $(Y, X_1, X_2, Z)$ can be obtained, the ordinary least squares (OLS) estimator of $\theta = (\beta_0, \beta_1', \beta_2', \gamma')'$ is consistent.

In reality, however, we often face the problem that $(Y, X_1, X_2, Z)$ cannot be taken from a single data source. It is not uncommon that economists who use survey data for empirical analysis must collect all necessary variables from more than one source. Examples include Lusardi (1996), Björklund and Jäntti (1997), Currie and Yelowitz (2000), Dee and Evans (2003), Borjas (2004), Bover (2005), Fujii (2008), Bostic *et al.* (2009), and Murtazashvili *et al.* (2015), to name a few. Ridder and Moffitt (2007) provide an excellent survey. This is the setting in which we are interested. Specifically, suppose that instead of observing a complete data set $(Y, X_1, X_2, Z)$, we have the following two overlapping subsets of data, $(Y, X_1, Z)$ and $(X_2, Z)$, i.e., some of the regressors are not available in the initial data set, where the initial data set is the one containing observations on the dependent variable along with a few other regressors. In such a setting, it is natural to construct a matched data set via exploiting the proximity of the common regressor(s) $Z$ across the two samples. This is often called "probabilistic record linkage". Here are two examples of the setting.

**Example 1. (Earnings data)** Matching is currently used for imputing missing records of earnings in important economic data sets. For example, the U.S. Cur-

rent Population Survey (CPS) files use the so-called "hot deck imputation" procedure of the Census (see, e.g., Little and Rubin, 2002; Hirsch and Schumacher, 2004; Bollinger and Hirsch, 2006), which allocates to nonrespondents the reported earnings of a matched respondent who has similar recorded attributes.[1] The share of imputed values is as high as 30%. The resulting earnings data have been used to uncover much of what is known about the labor market dynamics and outcomes.

**Example 2. (Returns to schooling)** Let $Y$ denote (the logarithm of) earnings, $X_1$ individual characteristics, $X_2$ ability measured by test scores, and $Z$ education. Although $(Y, X_1, Z)$ is available in the Panel Study of Income Dynamics (PSID), for instance, it is often the case that $(X_2, Z)$ can be found only in a different, psychometric data set. Utilizing the proximity of the common variable $Z$, we must construct a matched data set of $(Y, X_1, X_2, Z)$.

There are many algorithms that can be used to construct matched data sets (see, e.g., Smith and Todd, 2005; Ridder and Moffitt, 2007). We focus on the nearest neighbor matching (NNM) because of its simplicity and wide use. Abadie and Imbens (2006, 2012) use it in the context of the average treatment effect (ATE) estimation. Chen and Shao (2001) and Shao and Wang (2008) study the problem of variance estimation after a nearest neighbor-based imputation. The NNM can be used as a building block in construction of more complicated matching algorithms, most notably the single index or propensity score matching, but we do not pursue these extensions here.

We demonstrate that the OLS estimator from the regression (1) using NNM-based matched samples is inconsistent. The source of the inconsistency is a non-vanishing

[1] The distinction between hot and cold deck imputation seems to primarily refer to which sample (of punch cards) to use for matching, a current sample (hot) or an earlier sample (cold). Hence, hot deck imputation often means imputation of missing *values* of an existing variable, whereas cold deck imputation means imputation of entire missing variables. In this respect, this paper may be closer to cold rather than hot deck imputation.

bias term, which can be viewed as a measurement error bias stemming from replacing unobservable $X_2$ with a proxy in the matched data. In this sense, the paper is related to the literature on the classical problem of generated regressors and missing data (see, e.g., Pagan, 1984; Prokhorov and Schmidt, 2009). Moreover, we show that the rate of convergence to the probability limit of OLS depends on the number of common, matching variables and the divergence patterns of two sample sizes.

In line with these findings, we propose two semiparametric bias-corrected estimators. The first, one-step estimator is shown to attain the parametric convergence rate for the cases with at most two matching variables. The second estimator attempts to remedy the curse of dimensionality with respect to the number of matching variables. It is a two-step estimator, and in the second step it eliminates the second-order bias due to the so-called *matching discrepancy* (Abadie and Imbens, 2006) asymptotically in a similar manner 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 estimators can be also interpreted as indirect inference estimators (Gouriéroux, Monfort and Renault, 1993; Smith, 1993) in the sense that they can be obtained by taking the probability limit of the OLS estimator from the regression (1) as the "binding" function.

The paper contributes to three important areas. First, we provide new asymptotic results for regression analysis using matched data. In particular, we explicitly handle the issue of biases due to matching errors, which has been often ignored in the literature as if there were no mismatches; see Ridder and Moffitt (2007, p.5480) for a discussion and Bover (2005) and Bostic *et al.* (2009) for regression analysis using matched data. Available results are limited to the case of matching in the ATE estimation. For example, Abadie and Imbens (2006) show that when there is only one matching covariate, the bias in NNM-based matching estimators of the ATE may be asymptotically ignored; they attain the parametric convergence rate in that case.

To the best of our knowledge, bias-corrected estimation using matched data and the convergence properties of estimators within the framework of regression analysis have not been explored in the literature before.

Second, the estimation theory we develop provides guidance on repeated survey sampling when some covariates are found to be completely or partially missing after the initial survey. Our theory suggests (approximately) how many observations should be collected in a follow-up survey and how to estimate the linear regression model of interest consistently using the matched data from two surveys.

Finally, the paper offers an alternative to some well-known estimation methods based on two samples. A number of such methods have been designed within the framework of instrumental variables (IV) or generalized method of moments (GMM) estimation, where we can construct required moments from the two samples individually so that no matching is required (e.g., Angrist and Krueger, 1992, 1995; Arellano and Meghir, 1992; Inoue and Solon, 2010; Murtazashvili *et al.*, 2015). These approaches are not applicable to the setting of a linear regression where some regressors are missing and two-sample moment based estimation is infeasible.

Throughout we assume that the two samples *jointly* identify the regression models. There are other two-sample estimators that cover the cases where the first sample *alone* identifies the models and the second sample is used for efficiency gains (see, e.g., Imbens and Lancaster, 1994; Hellerstein and Imbens, 1999). These are not the settings we consider.

The remainder of this paper is organized as follows. Section 2 shows inconsistency of the OLS estimation from the regression (1) using matched samples. Section 3 proposes two bias-corrected estimators and explores their convergence properties. We also discuss consistent estimation of their asymptotic covariance matrices. Section 4 contains selected Monte Carlo simulations, examining how the bias correction works in finite samples. As an empirical example, in Section 5, we apply the bias-corrected

two-sample estimation 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.[2]

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 $q = 1$; and the symbol $>$ applied to matrices means positive definiteness.

# 2 Inconsistency of OLS Estimation Using Matched Samples

## 2.1 Setup

In order to explain how a matched sample is constructed, we need more notations. Denote the two random samples by $\mathcal{S}_1$ and $\mathcal{S}_2$. Also let $n$ and $m$ be sample sizes of $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively. Then, the two samples can be expressed as $\mathcal{S}_1 = \mathcal{S}_{1n} = \{(Y_i, X_{1i}, Z_i)\}_{i=1}^n$ and $\mathcal{S}_2 = \mathcal{S}_{2m} = \{(X_{2j}, Z_j)\}_{j=1}^m$. A natural way of matching based on $Z$ is to use the NNM based on some metric. For a vector $x$ and some symmetric matrix $A > 0$, a vector norm is denoted by $\|x\|_A = (x'Ax)^{1/2}$. While there may be numerous choices of $A$, following Abadie and Imbens (2011), we adopt the Mahalanobis metric $A_M = \left\{(1/N)\sum_{i=1}^N \left(Z_i - \bar{Z}\right)\left(Z_i - \bar{Z}\right)'\right\}^{-1}$ and the normalized Euclidean metric $A_{NE} = \mathrm{diag}\left(A_M^{-1}\right)^{-1}$, where $N := n + m$ and $\bar{Z} = (1/N)\sum_{i=1}^N Z_i$.

Furthermore, let $j_k(i)$ be the index of the $k$th match in $\mathcal{S}_2$ to the unit $i$ in $\mathcal{S}_1$, i.e., for each $i \in \{1, \ldots, n\}$, $j_k(i)$ satisfies

$$\sum_{j=1}^{m} \mathbf{1}\left\{\|Z_j - Z_i\|_A \leq \left\|Z_{j_k(i)} - Z_i\right\|_A\right\} = k.$$

Also let $\mathcal{J}_K(i) = \{j_1(i), \ldots, j_K(i)\}$ denote the set of indices for the first $K$ matches

[2] An online Supplement contains some lengthy derivations and comprehensive simulation results. GAUSS codes implementing the estimators are also available from the authors upon request.

for the unit $i$. The NNM constructs the matched data set

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

We also define $X_{2j(i)} := (1/K) \sum_{j \in \mathcal{J}_K(i)} X_{2j}$ and $Z_{j(i)} := (1/K) \sum_{j \in \mathcal{J}_K(i)} Z_j$.

It is worth noting that $X_2$ is missing entirely but only from the first sample. When the problem is considered in the context of the imputed sample, only the values corresponding to the first sample are missing. Thus formally, the problem can be viewed as both value imputation and variable imputation. However, in what follows we view the problem as missing variable (rather than missing values) imputation.

In our NNM, the number of matches $K$ remains *fixed*, as in Abadie and Imbens (2006). While it is possible to achieve consistency as in the $K$-nearest neighbor ($K$-NN) method by letting $K$ diverge at a slower rate than $n$ and $m$, there are two reasons why we keep $K$ fixed. First, this is what is done in practice. In many applications, the NNM is implemented with small values of $K$, and $K = 1$ (i.e., NNM with a single match) is often chosen even for large $n$ and $m$. Second, it can be argued that the analysis with fixed $K$ provides a better approximation to the finite sample behavior of the estimator than under $K \to \infty$.

A few additional remarks on NNM are in order. First, matching is made with replacement, and each element of the matching vector $Z$ is assumed to be continuous. Hence, our setting can be viewed as a foundation for more complicated methods of kernel-based matching (see, e.g., Busso, DiNardo and McCrary, 2014). Second, matching with replacement, allowing each unit to be used as a match more than once, seems to be standard in the econometric literature. Third, inclusion of discrete matching variables with a finite number of support points does not affect the subsequent asymptotic results but raises the question of how to treat ties. For simplicity, we ignore ties in the NNM, which happen with probability zero as long as $Z$ is continuous.

Throughout it is assumed that we estimate $\theta$ by regressing $Y_i$ on $W_{i,j(i)} :=$

$\left(1, X_{1i}', X_{2j(i)}', Z_i'\right)'$. It is possible to use $Z_{j(i)}$ in place of $Z_i$ and run the regression of $Y_i$ on $W_{i,j(i)}^{\dagger} := \left(1, X_{1i}', X_{2j(i)}', Z_{j(i)}'\right)'$. However, we focus exclusively on the former scenario because of the following two reasons. First, the two scenarios yield first-order asymptotically equivalent results. To see this, observe that $W_{i,j(i)}^{\dagger} = W_{i,j(i)} + \left[\ 0_{1\times(d_1+d_2+1)} \quad \left(Z_{j(i)} - Z_i\right)'\ \right]' = W_{i,j(i)} + O_p\left(m^{-1/d_3}\right)$ by Lemma A1, i.e., the second term serves merely as an extra second-order bias term. It is noteworthy that the identification condition is derived from the latter scenario. Second, as illustrated in Section 4, bias-corrected estimators based on $W_{i,j(i)}$ exhibits better finite-sample properties.

We start our analysis from running OLS for the regression of $Y_i$ on $W_{i,j(i)}$. The OLS estimator

$$\hat{\theta}_{OLS} := \hat{Q}_W^{-1}\hat{R}_W := \left(\frac{1}{n}\sum_{i=1}^{n} W_{i,j(i)}W_{i,j(i)}'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n} W_{i,j(i)}Y_i$$

is referred to as the *matched-sample OLS* (MSOLS) estimator hereinafter.

## 2.2 Regularity Conditions

In what follows, we develop the asymptotic theory of estimation of $\theta$ in the regression (1) as $n$ and $m$ diverge while $K$ is fixed. All of the estimation theory, including the bias-corrected estimation methods and their convergence properties, is new to the literature.

All of the results in the paper, including the inconsistency proof of the MSOLS estimator and the consistency proof of the new bias-corrected estimators of $\theta$, require the following conditions.

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

**Assumption 2.** The matching variable $Z$ is continuously distributed with a convex and compact support $\mathbb{Z}$, and its density is bounded and bounded away from zero on $\mathbb{Z}$.

**Assumption 3.**

**(i)** The regression error $u$ satisfies $E(u|W) = 0$ and $\sigma_u^2(W) := E(u^2|W) \in (0, \infty)$.

**(ii)** Let $g(Z) := \left[ g_1(Z)' \quad g_2(Z)' \right]' := \left[ E(X_1|Z)' \quad E(X_2|Z)' \right]'$ and let $\eta := \left[ \eta_1' \quad \eta_2' \right]' := \left[ X_1' - g_1(Z)' \quad X_2' - g_2(Z)' \right]'$. Then, $\Sigma_1 := E(\eta_1\eta_1') > 0$, $\Sigma_2 := E(\eta_2\eta_2') > 0$, $E(\eta_1\eta_2') = 0_{d_1 \times d_2}$, and $g_2(\cdot)$ is a first-order Lipschitz continuous, strictly nonlinear function on $\mathbb{Z}$.

These regularity conditions are largely inspired by those in the literature on semi-parametric, partial linear regression models (e.g., Robinson, 1988; Yatchew, 1997), matching estimators for the ATE (e.g., Abadie and Imbens, 2006), and regression estimation based on two samples (e.g., Angrist and Krueger, 1992; Inoue and Solon, 2010). In particular, equivalents to Assumption 1 (the common distribution assumption) are often imposed in the literature (e.g., Assumption 3 of Abadie and Imbens, 2006; Assumption a of Inoue and Solon, 2010). This is a strong assumption which simplifies the subsequent derivations considerably. It implies that the matched sample $\mathcal{S}$ behaves as a pseudo-population, from which the two samples are drawn. Assumption 2 plays a key role in controlling the order of magnitude in the matching discrepancy. Nonlinearity of $g_2(\cdot)$ in Assumption 3(ii) will be discussed in Remark 1 below in relation to identification. Zero correlation between $\eta_1$ and $\eta_2$ in Assumption 3(ii) is also a key assumption. Because we observe $X_1$ and $X_2$ in two independent samples, there seems to be no way to estimate $E(\eta_1\eta_2')$ consistently. Therefore we assume uncorrelatedness between $\eta_1$ and $\eta_2$.

## 2.3 Inconsistency of MSOLS

Our asymptotic analysis is built on reformulating the regression (1) in a 'partial linear'-like format. A straightforward calculation yields

$$Y_i := W'_{i,j(i)}\theta + \lambda_{i,j(i)} + \epsilon_{i,j(i)},\ i = 1, \ldots, n, \tag{2}$$

where

$$\lambda_{i,j(i)} = \lambda\left(Z_i, Z_{j(i)}\right) = \left\{ g_2\left(Z_i\right) - \frac{1}{K}\sum_{j\in\mathcal{J}_K(i)} g_2\left(Z_j\right) \right\}' \beta_2, \text{ and}$$

$$\epsilon_{i,j(i)} = u_i + \left( \eta_{2i} - \frac{1}{K}\sum_{j\in\mathcal{J}_K(i)} \eta_{2j} \right)' \beta_2 := u_i + \left(\eta_{2i} - \eta_{2j(i)}\right)' \beta_2.$$

The reason why this is not exactly a partial linear model is that there is a common regressor $Z_{j(i)}$ included in $W_{i,j(i)}$ and $\lambda_{i,j(i)}$. In this formulation, $W_{i,j(i)}$ is employed as the regressor of the fully parametric part $W'_{i,j(i)}\theta$. On the other hand, the semi-parametric part $\lambda_{i,j(i)}$ generates the second-order bias that will be discussed shortly, and thus it could be viewed as an analogue to the conditional bias discussed in Abadie and Imbens (2006). A key difference from the partial linear regression models studied in Robinson (1988) and Yatchew (1997) is that the matched regressor $X_{2j(i)}$ is endogenous, i.e., $X_{2j(i)}$ and the composite error $\epsilon_{i,j(i)}$ are correlated. The theorem below is established for the model in (2); it provides the probability limit of $\hat{\theta}_{OLS}$ and its associated rate of convergence.

**Theorem 1.** *If Assumptions 1-3 hold, then* $\hat{\theta}_{OLS} = Q_W^{-1}P_W\theta + O\left(m^{-1/d_3}\right) + O_p\left(n^{-1/2}\right)$ *as* $n, m \to \infty$, *where* $Q_W := E\left(W_{i,j(i)}W'_{i,j(i)}\right)$, $P_W := Q_W - (1/K)\,\Sigma$, *and* $\Sigma$ *is a* $(d+1)\times(d+1)$ *block-diagonal matrix of the form* $\Sigma := \text{diag}\left\{0_{(d_1+1)\times(d_1+1)}, \Sigma_2, 0_{d_3\times d_3}\right\}$.

**Remark 1.** Basic identification assumptions for MSOLS follow from those of the standard OLS. Fundamentally, they require that $\eta_1$ and $\eta_2$ are not in the linear span of each other and that $X_1$ and $X_2$ are not in the linear span of $Z$. As in the standard

OLS, we need $E(WW')$ to be of full rank. In our setting, the additional issue is whether $\hat{Q}_W$ and $Q_W$ are invertible. While we implicitly assume non-singularity of the former, the invertibility of the latter can be examined explicitly.

For simplicity and concreteness, consider the regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \gamma_1 Z + u,$$

where $X_1, X_2, Z \in \mathbb{R}$. The identification condition in question can be derived from the regression of $Y_i$ on $W^{\dagger}_{i,j(i)} = \left(1, X_{1i}, X_{2j(i)}, Z_{j(i)}\right)'$. The same condition is valid for the regression of $Y_i$ on $W_{i,j(i)} = \left(1, X_{1i}, X_{2j(i)}, Z_i\right)'$, because $\hat{Q}_{W^{\dagger}} := (1/n)\sum_{i=1}^{n} W^{\dagger}_{i,j(i)} W^{\dagger\prime}_{i,j(i)}$ and $\hat{Q}_W$ are first-order asymptotically equivalent in that $\hat{Q}_W = \hat{Q}_{W^{\dagger}} + O_p\left(m^{-1/d_3}\right)$ by Lemma A1. Let $Q_{W^{\dagger}} := E\left(W^{\dagger}_{i,j(i)} W^{\dagger\prime}_{i,j(i)}\right)$. Then,

$$Q_{W^{\dagger}} = \begin{bmatrix} 1 & E(X_1) & E(X_2) & E(Z) \\ E(X_1) & E(X_1^2) & E(X_1)E(X_2) & E(X_1)E(Z) \\ E(X_2) & E(X_1)E(X_2) & E^2(X_2) + Var(X_2)/K & E(X_2)E(Z) + Cov(X_2, Z)/K \\ E(Z) & E(X_1)E(Z) & E(X_2)E(Z) + Cov(X_2, Z)/K & E^2(Z) + Var(Z)/K \end{bmatrix},$$

and $\det(Q_{W^{\dagger}}) = Var(X_1)Var(X_2)Var(Z)\{1 - Corr^2(X_2, Z)\}/K^2 > 0$ with no additional restrictions. Hence, $Q_{W^{\dagger}}$ is invertible. Furthermore, the identification of bias-corrected estimators that will be proposed in the next section requires us to ensure non-singularity of $P_{W^{\dagger}} := Q_{W^{\dagger}} - (1/K)\Sigma$. It is easy to obtain $\det(P_{W^{\dagger}}) = Var(X_1)Var\{g_2(Z)\}Var(Z)[1 - Corr^2\{g_2(Z), Z\}]/K^2$, and $\det(P_{W^{\dagger}}) > 0$ if and only if $g_2(\cdot)$ is strictly nonlinear, as assumed in Assumption 3(ii).

So far we have maintained the assumption that the vector of common variables $Z$ is employed for both matching and estimation. It is possible that at least one common variable is used exclusively for matching (and thus not included in the regression (1)).[3] In this case the variable can be used to form yet another identification condition, which would allow us to relax somewhat our identification restrictions and/or improve

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

efficiency. For example, in the presence of an outside matching variable, $g_2(\cdot)$ can be allowed to be linear. But we do not pursue this point here.

**Remark 2.** Theorem 1 implies that MSOLS is inconsistent in general. The term $(1/K)\,\Sigma$ in $P_W$, which is the source of inconsistency, is generated by misspecifying the regression of $Y_i$ on $W_i$ as the one of $Y_i$ on $W_{i,j(i)}$, or equivalently, employing $X_{2j(i)}$ as a proxy of the latent variable $X_{2i}$. Therefore, the non-vanishing bias in MSOLS can be thought of as a measurement error bias. The measurement error interpretation is revisited in Section 2.4 below.

A straightforward calculation shows that the MSOLS estimator of $\beta_2$ is biased toward zero (an attenuation bias) in the limit and that all other parameter estimates are also biased. Perhaps surprisingly, the entire set of MSOLS estimates is inconsistent. This is contrary to the textbook regression with proxy variables where only a part of the OLS estimates is biased (see, e.g., Wooldridge, 2013, Ch.9). It is also easy to demonstrate that the bias in $\hat{\theta}_{OLS}$ is a function of $\beta_2$, $\Sigma_2$ and $K$. Therefore, the estimator would be consistent if either (i) $\beta_2 = 0$, i.e., $X_2$ were irrelevant in the correctly specified model; or (ii) $\Sigma_2 = 0$, i.e., $X_2$ were a deterministic function of $Z$.

**Remark 3.** The convergence rate of $\hat{\theta}_{OLS}$ is affected by the $O_p\left(m^{-1/d_3}\right)$ term, which corresponds to the second-order bias term $\lambda_{i,j(i)}$ due to the matching discrepancy. The rate can be determined by three different divergence patterns of $(n, m)$, namely, $n/m \to \kappa \in (0, \infty)$, $n/m \to 0$, and $n/m \to \infty$ as $n, m \to \infty$, and there exists a *curse of dimensionality* with respect to the matching variable $Z$ for each divergence pattern.

When $n/m \to \kappa$, $\hat{\theta}_{OLS} = Q_W^{-1} P_W \theta + O_p\left(n^{-\min\{1/2, 1/d_3\}}\right)$. For $d_3 = 1$, a central limit theorem (CLT) implies that $\sqrt{n}\left(\hat{\theta}_{OLS} - Q_W^{-1} P_W \theta\right)$ has a normal limit. For $d_3 = 2$, $\hat{\theta}_{OLS}$ is still $\sqrt{n}$-convergent, but we can only demonstrate asymptotic normality of $\hat{\theta}_{OLS}$ after subtracting the second-order bias term, i.e., the best we can do in this case

is to apply the CLT to $\sqrt{n}\left(\hat{\theta}_{OLS} - Q_W^{-1}P_W\theta - B_{OLS2}\right)$, where

$$B_{OLS2} := \hat{Q}_W^{-1}B_{R_W 2} := \hat{Q}_W^{-1}\frac{1}{n}\sum_{i=1}^{n} W_{i,j(i)}\lambda_{i,j(i)}.$$

These limiting distributions would reduce to the usual one of OLS if a complete data set of $(Y, X_1, X_2, Z)$ were available. For $d_3 \geq 3$, the convergence rate of $\hat{\theta}_{OLS}$ is slower than the parametric one, and it becomes slower as $d_3$ increases.

When $n/m \to 0$, $m^{-1/d_3} = o\left(n^{-1/2}\right)$ for $d_3 \leq 2$. Hence, $\hat{\theta}_{OLS} = Q_W^{-1}P_W\theta + O_p\left(n^{-1/2}\right)$, and $\sqrt{n}\left(\hat{\theta}_{OLS} - Q_W^{-1}P_W\theta\right)$ has a normal limit in this case. However, for $d_3 \geq 3$, the convergence rate of $\hat{\theta}_{OLS}$ can be determined only if an extra divergence pattern of $(n, m)$ is imposed. For instance, when $d_3 = 3$, $\hat{\theta}_{OLS}$ is $\sqrt{n}$-convergent if $n^3 = O\left(m^2\right)$ and its convergence rate is a nonparametric one if $n^3/m^2 \to \infty$.

When $n/m \to \infty$, a $\sqrt{n}$-convergent of $\hat{\theta}_{OLS}$ can be attained only if $d_3 = 1$ and $n = O\left(m^2\right)$. Moreover, $\sqrt{n}\left(\hat{\theta}_{OLS} - Q_W^{-1}P_W\theta\right)$ has a normal limit when $d_3 = 1$ and $n/m^2 \to 0$. On the other hand, if $d_3 = 1$ and $n/m^2 \to \infty$ or if $d_3 \geq 2$, then $\hat{\theta}_{OLS} = Q_W^{-1}P_W\theta + O_p\left(m^{-1/d_3}\right)$, and the convergence bound $m^{1/d_3}$ is slower than $\sqrt{n}$, regardless of whether or not the bound is sharp.

## 2.4 A Measurement Error Interpretation

Before moving to our proposal for bias-corrected estimation, it is helpful to consider the problem of imputation as a measurement error problem arising from using a proxy.[4] Write the model in (1) as

$$Y = \beta_0 + X_1'\beta_1 + g_2\left(Z\right)'\beta_2 + Z'\gamma + e,$$

where $e := \left\{X_2 - g_2\left(Z\right)\right\}'\beta_2 + u$. Then, $g_2\left(Z\right)$ can be viewed as a proxy for $X_2$ and if we could observe $g_2\left(Z\right)$ then the model could be estimated by OLS as long as $X_1$ is uncorrelated with $\left\{X_2 - g_2\left(Z\right)\right\}$ and $g_2\left(Z\right)$ is not in the linear span of $Z$.

[4] We thank an anonymous referee for suggesting this interpretation to us.

However, $g_2(Z)$ is not observed and needs to be estimated. There are two complications here. One is that we need to use an estimator $\hat{g}_2(Z)$ based on another sample. The other is that the estimator uses matched values of $X_2$ obtained using nearest-neighbors of $Z$ from the other sample, not the $Z$ itself. Suppose that $\hat{g}_2(Z)$ is the estimate via the $K$-NN method for the moment.[5] Rewriting the model as

$$Y = \beta_0 + X_1'\beta_1 + \hat{g}_2(Z)'\beta_2 + Z'\gamma + v,$$

where $v := \{X_2 - \hat{g}_2(Z)\}'\beta_2 + u$, we attempt to estimate this regression by OLS. If $\hat{g}_2(Z)$ were estimated from the same sample, then the correlation between $\hat{g}_2(Z)$ and $\{X_2 - \hat{g}_2(Z)\}$ would be near zero because of orthogonality of $g_2(Z)$ and $\{X_2 - g_2(Z)\}$. We actually employ a different sample to estimate (or impute) $g_2(Z)$, and thus the correlation does not equal zero, which causes a non-negligible bias in the OLS estimator. This can be interpreted as a classical measurement error problem.

As is well known in the literature on measurement error problems, the bias of OLS can be corrected if the variance of the measurement error can be obtained analytically, given that the matching discrepancy from $K$-NN is bounded. Our bias correction methods in the next section basically follow this idea, although the nearest-neighbor algorithm that we use is intended only to find $K$ closest matches to $Z$ and not to estimate $g_2(Z)$.

# 3 Bias-Corrected Estimation

## 3.1 One-Step Bias-Corrected Estimator

This section develops bias-corrected estimation of $\theta$. As suggested by the proof of Theorem 1, inconsistency of MSOLS comes from the fact that $\hat{Q}_W \xrightarrow{p} Q_W$ whereas $\hat{R}_W \xrightarrow{p} P_W\theta = \{Q_W - (1/K)\Sigma\}\theta$. Therefore, the non-vanishing bias in MSOLS can be eliminated if either

[5] We adopt a power-series approximation to estimate $g_2(Z)$; see Section 3.3 for details.

**(1a)** the denominator $\hat{Q}_W$ is replaced by a consistent estimator of $P_W$ with the numerator $\hat{R}_W$ left unchanged; or

**(1b)** an extra term consistent for $(1/K)\,\Sigma\theta$ is added to $\hat{R}_W$ with $\hat{Q}_W$ held as it is.

Bias correction in each strategy is semiparametric in that a consistent estimate of $\Sigma_2$ (covariance matrix of the nonparametric regression error $\eta_2$) is required. Moreover, implementing (1b) inevitably leads to a two-step estimation with an initial consistent estimate of $\theta$ plugged in. However, if the plug-in estimator is the one using strategy (1a), then the two-step estimation will produce a numerically identical result. To see why, let an initial estimator of $\theta$ using strategy (1a) be $\hat{\theta}_{(1a)} = \hat{P}_W^{-1}\hat{R}_W$, where $\hat{P}_W \xrightarrow{p} P_W$. Given $\hat{\theta}_{(1a)}$, we obtain the second-step estimator as

$$\hat{\theta}_{(1b)} := \hat{Q}_W^{-1}\left(\hat{R}_W + \frac{1}{K}\hat{\Sigma}\hat{\theta}_{(1a)}\right) = \hat{Q}_W^{-1}\left(I_{d+1} + \frac{1}{K}\hat{\Sigma}\hat{P}_W^{-1}\right)\hat{R}_W, \tag{3}$$

where $\hat{\Sigma}$ is a consistent estimate of $\Sigma$. Post-multiplying both sides of $\hat{P}_W + (1/K)\,\hat{\Sigma} = \hat{Q}_W$ by $\hat{P}_W^{-1}$ yields $I_{d+1} + (1/K)\,\hat{\Sigma}\hat{P}_W^{-1} = \hat{Q}_W\hat{P}_W^{-1}$. Substituting this into the right-hand side of (3) immediately establishes that $\hat{\theta}_{(1b)} = \hat{\theta}_{(1a)}$. Therefore, there is no point in pursuing strategy (1b) separately; strategy (1b) is interesting only if an alternative consistent estimator of $\theta$ (other than $\hat{\theta}_{(1a)}$) is chosen.

Now we turn to the bias correction based on strategy (1a). The idea behind the strategy comes from indirect inference (II) estimation by Gouriéroux, Monfort and Renault (1993) and Smith (1993). Take the probability limit of $\hat{\theta}_{OLS}$ as the binding function $b(\theta)$, i.e., $b(\theta) = Q_W^{-1}P_W\theta$.[6] Because $P_W^{-1}$ exists as discussed in Remark 1, the II estimator can be built on the inverse mapping of $\hat{\theta}_{OLS} = b(\theta)$, i.e., $\theta = P_W^{-1}Q_W\hat{\theta}_{OLS}$. The interpretation then follows from replacing $P_W$ with its $\sqrt{n}$-consistent estimator $\hat{P}_W$ and regarding $\hat{R}_W$ as a 'sample analog' of $Q_W\hat{\theta}_{OLS}$.

[6] Typically the binding function is unknown, and it must be approximated via simulations. However, when the function has a closed form, there is no need for simulations; see Carrasco and Florens (2002) for another example. As suggested by a referee, this also gives the estimator the interpretation of a classical minimum distance estimation based on the distance $b(\theta) - \hat{\theta}_{OLS}$ but we do not pursue this point further.

Accordingly, we call this estimation method *the matched-sample indirect inference* (MSII) estimation. We formally define the MSII estimator as

$$\hat{\theta}_{II} := \hat{P}_W^{-1} \hat{R}_W,$$

which has been called $\hat{\theta}_{(1a)}$ before.[7]

Our remaining task is to deliver a consistent estimator of $P_W$. Obviously, $\hat{Q}_W$ is a natural estimator of $Q_W$. Furthermore, it turns out that when estimating $\Sigma = \text{diag}\left\{0_{(d_1+1)\times(d_1+1)}, \Sigma_2, 0_{d_3\times d_3}\right\}$, we can do without a nonparametric estimation of $g_2(\cdot)$. To do so, we first reorder $\mathcal{S}_2$ with respect to $Z$ by the following recursion:

1. Define $Z_{(1)}$ as the observation that has the smallest first element, i.e., $(1) = \arg\min_{1\le j\le m} Z_{j1}$.

2. For $j = 2, \ldots, m$, choose $(j) = \arg\min_{j\neq(1),\ldots,(j-1)} \left\|Z_j - Z_{(j-1)}\right\|$.[8]

Given the reordered sample $\mathcal{S}_2 = \left\{X_{2(j)}, Z_{(j)}\right\}_{j=1}^m$, $\Sigma_2$ can be consistently estimated by

$$\hat{\Sigma}_2 = \frac{1}{2(m-1)} \sum_{j=2}^{m} \Delta X_{2(j)} \Delta X_{2(j)}', \tag{4}$$

where $\Delta X_{2(j)} := X_{2(j)} - X_{2(j-1)}$. This is known as the difference-based variance estimator; see von Neumann (1941) and Rice (1984) for univariate and Yatchew (1997) and Horowitz and Spokoiny (2001) for multivariate cases. In the end, the estimator of $P_W$ is given by

$$\hat{P}_W := \hat{Q}_W - \frac{1}{K}\hat{\Sigma} = \hat{Q}_W - \frac{1}{K}\text{diag}\left\{0_{(d_1+1)\times(d_1+1)}, \hat{\Sigma}_2, 0_{d_3\times d_3}\right\}.$$

[7] The estimator $\hat{\theta}_{II}$ also has a method-of-moment interpretation, where the moment is

$$E\left(W_{i,j(i)}\epsilon_{i,j(i)}\right) = -\frac{1}{K}\Sigma\theta.$$

From the viewpoint of likelihood-based methods MSII may leave some information (or moment restrictions) unused, and thus there may be room for efficiency improvement. But pursuing this point is beyond the scope of this paper.

[8] If $Z$ is a scalar, then the recursion reduces to rearranging $\{Z_j\}_{j=1}^m$ in an ascending order $Z_{(1)} \le \ldots \le Z_{(m)}$.

## 3.2 Convergence Properties of the MSII Estimator

We first provide a consistency result for the MSII estimator $\hat{\theta}_{II}$. The result holds regardless of the number of matching variables $d_3$ and of the divergence patterns of $(n, m)$.

**Theorem 2.** *If Assumptions 1-3 hold, then* $\hat{\theta}_{II} \xrightarrow{p} \theta$ *as* $n, m \to \infty$.

Next we establish asymptotic normality of $\hat{\theta}_{II}$ after correcting for the two bias terms, namely, $B_{MD}$ (second-order bias term due to the matching discrepancy) and $B_\Sigma$ (third-order bias term due to the nonparametric estimation of $\Sigma$). Before proceeding, we make an additional assumption. Like Assumption c of Inoue and Solon (2010), Assumption 4 makes derivations of asymptotic variances in the limiting distributions of $\hat{\theta}_{II}$ easier.

**Assumption 4.** In the nonparametric regression $X_2 = g_2(Z) + \eta_2$, $g_2(Z)$ and $\eta_2$ are independent, and third-order moments of $\eta_2$ are zeros.

The next theorem establishes the limiting distributions of $\hat{\theta}_{II}$ under a variety of divergence patterns of $(n, m)$. Again, these results hold regardless of the number of matching variables $d_3$.

**Theorem 3.** *If Assumptions 1-4 hold, then, as* $n, m \to \infty$,

$$\begin{cases} \sqrt{n}\left(\hat{\theta}_{II} - \theta - B_{MD} - B_\Sigma\right) \xrightarrow{d} N(0, V_I) := N\left(0, P_W^{-1}\Omega P_W^{-1}\right) & \text{if } n/m \to \kappa \in (0, \infty) \\ \sqrt{n}\left(\hat{\theta}_{II} - \theta - B_{MD} - B_\Sigma\right) \xrightarrow{d} N(0, V_{II}) := N\left(0, P_W^{-1}\Omega_{11A} P_W^{-1}\right) & \text{if } n/m \to 0 \\ \sqrt{m}\left(\hat{\theta}_{II} - \theta - B_{MD} - B_\Sigma\right) \xrightarrow{d} N(0, V_{III}) := N\left(0, P_W^{-1}\Omega_{22} P_W^{-1}\right) & \text{if } n/m \to \infty \end{cases},$$

*where* $B_{MD} = \frac{1}{n}\hat{P}_W^{-1}\sum_{i=1}^{n} W_{i,j(i)}\lambda_{i,j(i)}$ *and*

$$B_\Sigma = \hat{P}_W^{-1}\frac{1}{K}\operatorname{diag}\left\{0_{(d_1+1)\times(d_1+1)}, \frac{1}{2(m-1)}\sum_{j=2}^{m}\Delta g_2\left(Z_{(j)}\right)\Delta g_2\left(Z_{(j)}\right)', 0_{d_3\times d_3}\right\}\theta$$

*are the two bias terms,*

$$\Omega := \Omega_{11} + \sqrt{\kappa}\left(\Omega_{12} + \Omega_{12}'\right) + \kappa\Omega_{22} := \left(\Omega_{11A} + \Omega_{11B}\right) + \sqrt{\kappa}\left(\Omega_{12} + \Omega_{12}'\right) + \kappa\Omega_{22},$$

$$\Omega_{11A} := E\left\{\left(W_{i,j(i)}\epsilon_{i,j(i)} + \frac{1}{K}\Sigma\theta\right)\left(W_{i,j(i)}\epsilon_{i,j(i)} + \frac{1}{K}\Sigma\theta\right)'\right\},$$

$$\Omega_{11B} := \kappa\left[\left(\beta_2'\Sigma_2\beta_2\right)E\left(W\right)E\left(W\right)' + \frac{1}{K^2}\operatorname{diag}\left\{0_{(d_1+1)\times(d_1+1)}, \left(\beta_2'\Sigma_2\beta_2\right)V_{g2} + \Xi, 0_{d_3\times d_3}\right\}\right],$$

$$\Omega_{12} := -\frac{\sqrt{\kappa}}{K^2}\operatorname{diag}\left\{0_{(d_1+1)\times(d_1+1)}, \Xi, 0_{d_3\times d_3}\right\},$$

$$\Omega_{22} := \frac{1}{K^2}\operatorname{diag}\left\{0_{(d_1+1)\times(d_1+1)}, \Xi + \frac{1}{2}\Psi, 0_{d_3\times d_3}\right\},$$

$$V_{g_2} := Var\left\{g_2\left(Z\right)\right\}, \ \Xi := E\left\{\left(\eta_2\eta_2' - \Sigma_2\right)\beta_2\beta_2'\left(\eta_2\eta_2' - \Sigma_2\right)\right\}, \ \textit{and}$$

$$\Psi := \left(\beta_2'\Sigma_2\beta_2\right)\Sigma_2 + \Sigma_2\beta_2\beta_2'\Sigma_2.$$

Observe that $\Omega$ in $V_I$ can be simplified as

$$\Omega = E\left\{\left(W_{i,j(i)}\epsilon_{i,j(i)} + \frac{1}{K}\Sigma\theta\right)\left(W_{i,j(i)}\epsilon_{i,j(i)} + \frac{1}{K}\Sigma\theta\right)'\right\}$$
$$+ \kappa\left[\left(\beta_2'\Sigma_2\beta_2\right)E\left(W\right)E\left(W\right)' + \frac{1}{K^2}\operatorname{diag}\left\{0_{(d_1+1)\times(d_1+1)}, \left(\beta_2'\Sigma_2\beta_2\right)V_{g_2} + \frac{1}{2}\Psi, 0_{d_3\times d_3}\right\}\right].$$

This suggests that $V_I$ is greater than $V_{II}$ in the positive definite sense, i.e., amassing more information in the form of a quicker growing $m$ provides asymptotic efficiency gains.

Another relative efficiency implication is that when we include an extra matching variable, i.e., $d_3$ increases, the additional variability captured by $P_W$ leads to a smaller $V_{III}$ (in the positive definite sense) since the non-zero elements of $\Omega_{22}$ remain unchanged. This is similar to the effect of adding a regressor to a regression. The effect of a larger $d_3$ on $V_I$ and $V_{II}$ is unclear as $\Omega_{11A}$ and $\Omega$ will be also inflated by the inclusion.

More importantly, the limiting distributions of $\hat{\theta}_{II}$ vary across divergence patterns of $(n, m)$. The convergence rate of $\hat{\theta}_{II}$ is determined by the sample size of the smaller sample. The rate is $\sqrt{n}$ when $n/m \to \kappa$, where $\kappa \in [0, \infty)$, i.e., $\mathcal{S}_2$ is no smaller than $\mathcal{S}_1$ in order of magnitude. In contrast, when $n/m \to \infty$ or $\mathcal{S}_1$ is much larger than $\mathcal{S}_2$,

the best possible rate slows down to $\sqrt{m} = o(\sqrt{n})$. The slow rate can be thought of as the price paid for estimating $\theta$ using a considerably smaller sample $\mathcal{S}_2$ via the NNM. As a result of the bias correction, the order of magnitude in the estimation error of $\Sigma_2$ dominates.

While the MSII estimator is consistent, its convergence is affected by the asymptotic bias terms. For a large $d_3$ the second-order bias term $B_{MD}$ dominates and the convergence rate becomes inferior. This is the curse of dimensionality of NNM that is commonly observed in other applications. With regards to the ATE estimation, Abadie and Imbens (2006, Corollary 1), for instance, show that the matching discrepancy bias can be safely ignored only when matching is made on a single variable.

A similar result applies in our setting for small values of $d_3$. The next corollary illustrates such cases.

**Corollary 1.** *If Assumptions 1-4 hold, then, as* $n, m \to \infty$,

$$\begin{cases} \sqrt{n}\left(\hat{\theta}_{II} - \theta\right) \xrightarrow{d} N\left(0, V_I\right) & \text{if } n/m \to \kappa \in (0, \infty) \text{ and } d_3 = 1 \\ \sqrt{n}\left(\hat{\theta}_{II} - \theta\right) \xrightarrow{d} N\left(0, V_{II}\right) & \text{if } n/m \to 0 \text{ and } d_3 = 1, 2 \\ \sqrt{m}\left(\hat{\theta}_{II} - \theta\right) \xrightarrow{d} N\left(0, V_{III}\right) & \text{if } n/m \to \infty \text{ and } d_3 = 1 \end{cases},$$

*where* $V_I$, $V_{II}$ *and* $V_{III}$ *are defined in Theorem 3.*

In the cases covered by Corollary 1, the bias terms $B_{MD}$ as $B_\Sigma$ are already of smaller order than $n^{-1/2}$ or $m^{-1/2}$, and thus we obtain parametric convergence rates without bias correction.

## 3.3 Two-Step Bias-Corrected Estimator

If a parametric rate of convergence cannot be obtained without bias correction as in Corollary 1, then we need to find a way of eliminating the second-order bias $B_{MD}$, or equivalently, the effect of $\lambda_{i,j(i)}$ asymptotically from (2). There are two possible strategies, namely:

**Strategy 1:** taking the first-order difference of (2); and

**Strategy 2:** subtracting a consistent estimate of $\lambda_{i,j(i)}$ from the dependent variable $Y_i$.

Yatchew (1997) advocates Strategy 1 in semiparametric regression estimation, whereas Robinson (1988) and Abadie and Imbens (2011) adopt a similar strategy to Strategy 2 in semiparametric regression and ATE estimations, respectively. In our settings, we have found that Strategy 1 has a few disadvantages. First, differencing (2) leaves $\beta_0$ and $\gamma$ unidentified. Second, our preliminary Monte Carlo study suggests that MSII estimates from the differenced regression are numerically quite unstable. For these reasons we focus on Strategy 2.

Estimating $\lambda_{i,j(i)}$ requires consistent estimates of $\theta$ and $g_2(\cdot)$. For the former, it suffices to employ the MSII estimate $\hat{\theta}_{II}$. For the latter, as in Abadie and Imbens (2011), we adopt a nonparametric power-series estimation. Let $\upsilon = (\upsilon_1, \ldots, \upsilon_{d_3})$ be a multi-index of dimension $d_3$, which is a $d_3$-dimensional vector of nonnegative integers with $|\upsilon| = \sum_{l=1}^{d_3} \upsilon_l$. Also denote $z^{\upsilon} = \prod_{l=1}^{d_3} z_l^{\upsilon_l}$, where $z_l$ is the $l$th element of $z$. Consider a series $\{\upsilon(\mathcal{K})\}_{\mathcal{K}=1}^{\infty}$ containing distinct vectors such that $|\upsilon(\mathcal{K})|$ is non-decreasing. Let $p_{\mathcal{K}}(z) = z^{\upsilon(\mathcal{K})}$ and $p^{\mathcal{K}}(z) = (p_1(z), \ldots, p_{\mathcal{K}}(z))'$. Then, a nonparametric series estimator of the regression function $g_{2r}(z)$, $r = 1, \ldots, d_2$, is given by

$$\hat{g}_{2r}(z) := p^{\mathcal{K}(m)}(z)' \left\{ \sum_{j=1}^{m} p^{\mathcal{K}(m)}(Z_j) p^{\mathcal{K}(m)}(Z_j)' \right\}^{-} \sum_{j=1}^{m} p^{\mathcal{K}(m)}(Z_j) X_{2r,j},$$

where $X_{2r,j}$ is the $r$th element of $X_{2j}$ in $\mathcal{S}_2$, $(\cdot)^{-}$ denotes the generalized inverse, and $\mathcal{K} = \mathcal{K}(m)$ signifies the dependence of $\mathcal{K}$ on the sample size of $\mathcal{S}_2$.

The entire estimation procedure based on Strategy 2 can be summarized in the following two steps:

1. Run MSII using the original matched sample $\mathcal{S}$ to obtain the initial estimate $\hat{\theta}_{II}^{(1)} = \left(\hat{\beta}_{II,0}^{(1)}, \hat{\beta}_{II,1}^{(1)\prime}, \hat{\beta}_{II,2}^{(1)\prime}, \hat{\gamma}_{II}^{(1)\prime}\right)'$.

2. Construct adjusted dependent variables $\left\{Y_i^+\right\}_{i=1}^n := \left\{Y_i - \hat{\lambda}_{i,j(i)}\right\}_{i=1}^n$, where

$$\hat{\lambda}_{i,j(i)} = \left\{\hat{g}_2\left(Z_i\right) - \frac{1}{K}\sum_{j\in\mathcal{J}_K(i)}\hat{g}_2\left(Z_j\right)\right\}'\hat{\beta}_{II,2}^{(1)}$$

and $\hat{g}_2\left(z\right) = \left(\hat{g}_{21}\left(z\right),\ldots,\hat{g}_{2d_2}\left(z\right)\right)'$, and rerun MSII using the modified matched sample $\mathcal{S}^+ = \mathcal{S}_n^+ := \left\{\left(Y_i^+, X_{1i}, X_{2j_1(i)},\ldots,X_{2j_K(i)}, Z_i, Z_{j_1(i)},\ldots,Z_{j_K(i)}\right)\right\}_{i=1}^n$ to obtain the final estimator

$$\hat{\theta}_{II-FM} := \hat{P}_W^{-1}\hat{R}_W^+ := \hat{P}_W^{-1}\frac{1}{n}\sum_{i=1}^n W_{i,j(i)}Y_i^+.$$

The idea behind the above procedure is as follows. The initial MSII estimate $\hat{\theta}_{II}^{(1)}$ is consistent but inefficient, because the slow convergence rate $m^{1/d_3}$ of the second-order bias term $B_{MD}$ dominates. Then, in the second step, we (asymptotically) eliminate the source of the inferior rate by subtracting $\hat{\lambda}_{i,j(i)}$ from the dependent variable and reestimate $\theta$ by MSII using the bias-adjusted data to ensure $\sqrt{n}$- or $\sqrt{m}$-consistency. The entire procedure is reminiscent of the fully-modified least squares estimation for cointegrating regressions by Phillips and Hansen (1990). In this sense, we call the estimator the *fully-modified MSII* (MSII-FM) estimator hereinafter.

In order to deliver convergence results for $\hat{\theta}_{II-FM}$, we must additionally impose the following regularity conditions. These are analogous to conditions (i)-(iii) in Theorem 2 of Abadie and Imbens (2011).

**Assumption 5.** $\mathbb{Z}$ is a Cartesian product of compact intervals.

**Assumption 6.** $\mathcal{K}\left(m\right) \asymp m^{\nu}$ for some constant $\nu \in \left(0, \min\left\{2/\left(4d_3+3\right), 2/\left(4d_3^2-d_3\right)\right\}\right)$.

**Assumption 7.** There is a constant $C$ such that for each multi-index $\upsilon$, the $\upsilon$th partial derivative of $g_2\left(z\right)$ exists and its norm is bounded by $C^{|\upsilon|}$.

The theorem below refers to the limiting distributions of $\hat{\theta}_{II-FM}$ under a variety of divergence patterns of $(n, m)$. It is worth emphasizing that asymptotic variances

of the MSII-FM estimator take the same forms as in Theorem 3 and Corollary 1, i.e., the FM procedure asymptotically removes the second- and third-order bias terms without inflating the variance.

**Theorem 4.** *If Assumptions 1-7 hold, then, as* $n, m \to \infty$,

$$\begin{cases} \sqrt{n}\left(\hat{\theta}_{II-FM} - \theta\right) \xrightarrow{d} N\left(0, V_I\right) & \text{if } n/m \to \kappa \in (0, \infty) \text{ and } d_3 = 2, 3 \\ \sqrt{n}\left(\hat{\theta}_{II-FM} - \theta\right) \xrightarrow{d} N\left(0, V_{II}\right) & \text{if } n/m \to 0 \text{ and } d_3 = 3, 4 \\ \sqrt{m}\left(\hat{\theta}_{II-FM} - \theta\right) \xrightarrow{d} N\left(0, V_{III}\right) & \text{if } n/m \to \infty \text{ and } d_3 = 2, 3 \end{cases},$$

*where* $V_I$, $V_{II}$ *and* $V_{III}$ *are defined in Theorem 3.*

An important practical question when implementing MSII-FM is how to choose $\mathcal{K}(m)$, the number of terms in the series approximation. We will address this issue in Section 4 and in Supplement C.

## 3.4 Covariance Estimation

We conclude this section by discussing covariance estimation, which is essential for inference. Our focus is on the MSII estimator of Corollary 1 and the MSII-FM estimator of Theorem 4, which are first-order asymptotically equivalent. Because $\hat{P}_W$ is consistent for $P_W$, the problem of estimating $V_I$, $V_{II}$ and $V_{III}$ consistently is boiled down to proposing consistent estimators of $\Omega$, $\Omega_{11A}$ and $\Omega_{22}$.

The consistent estimators are presented in the proposition below. Notice that the proposition is built on the assumption that $\hat{\theta}_{II}$ is employed as a consistent estimator for $\theta$; it is easy to see that the result equally holds after it is replaced by $\hat{\theta}_{II-FM}$.

**Proposition 1.** *Let the estimators of $\Omega_{11A}$, $\Omega_{22}$ and $\Omega$ be*

$$\hat{\Omega}_{11A} = \frac{1}{n}\sum_{i=1}^{n}\left(W_{i,j(i)}\hat{\epsilon}_{i,j(i)} + \frac{1}{K}\hat{\Sigma}\hat{\theta}_{II}\right)\left(W_{i,j(i)}\hat{\epsilon}_{i,j(i)} + \frac{1}{K}\hat{\Sigma}\hat{\theta}_{II}\right)',$$

$$\hat{\Omega}_{22} = \frac{1}{K^2}\,\mathrm{diag}\left\{0_{(d_1+1)\times(d_1+1)}, \hat{\Gamma}(-1) + \hat{\Gamma}(0) + \hat{\Gamma}(1), 0_{d_3\times d_3}\right\}, \text{ and}$$

$$\hat{\Omega} = \hat{\Omega}_{11A} + \frac{n}{m}\left[\left(\hat{\beta}'_{2,II}\hat{\Sigma}_2\hat{\beta}_{2,II}\right)\bar{W}\bar{W}'\right.$$
$$\left. + \frac{1}{K^2}\,\mathrm{diag}\left\{0_{(d_1+1)\times(d_1+1)}, \left(\hat{\beta}'_{2,II}\hat{\Sigma}_2\hat{\beta}_{2,II}\right)\hat{V}_{g_2} + \hat{\Gamma}(0) - \left\{\hat{\Gamma}(-1) + \hat{\Gamma}(1)\right\}, 0_{d_3\times d_3}\right\}\right],$$

*where $\hat{\epsilon}_{i,j(i)} = Y_i - W'_{i,j(i)}\hat{\theta}_{II}$ is the MSII residual, $\hat{\beta}_{2,II}$ is the MSII estimator of $\beta_2$, $\hat{\Gamma}(\ell)$ is the $\ell$th sample autocovariance of $\left\{\left(\Delta X_{2j}\Delta X'_{2j}/2\right) - \hat{\Sigma}_2\right\}\hat{\beta}_{2,II}$, i.e.,*

$$\hat{\Gamma}(\ell) = \frac{1}{m-1}\sum_{j=\max\{2,2+\ell\}}^{\min\{m,m+\ell\}}\left(\frac{\Delta X_{2j}\Delta X'_{2j}}{2} - \hat{\Sigma}_2\right)\hat{\beta}_{2,II}\hat{\beta}'_{2,II}\left(\frac{\Delta X_{2j-\ell}\Delta X'_{2j-\ell}}{2} - \hat{\Sigma}_2\right),$$

$$\bar{W} = \begin{bmatrix} 1 \\ \bar{X}_1 \\ \bar{X}_2 \\ \bar{Z} \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{1}{n}\sum_{i=1}^{n} X_{1i} \\ \frac{1}{m}\sum_{j=1}^{m} X_{2j} \\ \frac{1}{N}\sum_{i=1}^{N} Z_i \end{bmatrix}, \text{ and}$$

$$\hat{V}_{g_2} = \frac{1}{m-1}\sum_{j=1}^{m}\left(X_{2j} - \bar{X}_2\right)\left(X_{2j} - \bar{X}_2\right)' - \hat{\Sigma}_2.$$

*Then, under Assumptions 1-4, $\hat{\Omega}_{11A} \xrightarrow{p} \Omega_{11A}$, $\hat{\Omega}_{22} \xrightarrow{p} \Omega_{22}$ and $\hat{\Omega} \xrightarrow{p} \Omega$ as $n, m \to \infty$.*

# 4 Finite-Sample Performance

## 4.1 Monte Carlo Setup

In this section we conduct Monte Carlo simulations to examine finite-sample properties of proposed bias-corrected estimators. The simulation study takes a unified approach in the sense that the same regression model is employed regardless of the number of matching variables $d_3$. The model considered throughout is

$$Y = \beta_0 + X_1'\beta_1 + X_2'\beta_2 + Z'\gamma + u, \tag{5}$$

where $X_1 = (X_{11}, X_{12})', \beta_1 = (\beta_{11}, \beta_{12})' \in \mathbb{R}^2$, $X_2 = (X_{21}, X_{22})', \beta_2 = (\beta_{21}, \beta_{22})' \in \mathbb{R}^2$, and $Z = (Z_1, \ldots, Z_{d_3})', \gamma = (\gamma_1, \ldots, \gamma_{d_3})' \in \mathbb{R}^{d_3}$ for $d_3 = 1, 2, 3$. It is assumed

that two samples, namely, $\mathcal{S}_1 = \{(Y_i, X_{1i}, Z_i)\}_{i=1}^n$ and $\mathcal{S}_2 = \{(X_{2j}, Z_j)\}_{j=1}^m$, are only observable. The complete sample $\mathcal{S}^* = \{(Y_i, X_{1i}, X_{2i}, Z_i)\}_{i=1}^n$ is the sample that would not be observed in practice.

The data are generated in the following manner. First, $Z^* = (Z_1^*, Z_2^*, Z_3^*)'$ is generated by

$$Z^* \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 1/\sqrt{2} & 1/\sqrt{3} \\ 1/\sqrt{2} & 1 & \sqrt{2}/\sqrt{3} \\ 1/\sqrt{3} & \sqrt{2}/\sqrt{3} & 1 \end{bmatrix}\right).$$

Each $Z_p^*$ $(p = 1, 2, 3)$ is transformed to $Z_p = 4\Phi\left(Z_p^*\right) - 2$, where $\Phi(\cdot)$ is the cdf of $N(0, 1)$. Observe that the $Z_p$ are mutually correlated $U[-2, 2]$ random variables. Then, for a given $d_3$, the $Z_p$ $(p \leq d_3)$ are used as matching variables.

Second, $X_1 = (X_{11}, X_{12})'$ is generated by $X_{1q} = \sum_{p=1}^{d_3} Z_p + \eta_{1q}$ $(q = 1, 2)$, where $\eta_1 = (\eta_{11}, \eta_{12})' \overset{iid}{\sim} N(0_{2\times 1}, I_2)$. Third, $X_2 = (X_{21}, X_{22})'$ is generated by $X_{2r} = \sum_{p=1}^{d_3} g_{2r}(Z_p) + \eta_{2r}$ $(r = 1, 2)$ for some nonlinear function $g_{2r}(\cdot)$, where $\eta_2 = (\eta_{21}, \eta_{22})' \overset{iid}{\sim} N(0_{2\times 1}, I_2)$. While $g_{21}(z) = z + (5/\tau)\,\phi(z/\tau)$, $\tau = 0.25$ is employed throughout, one of the following three functional forms is chosen as $g_{22}(z)$:

$$g_{22}(z) = \begin{cases} z + (5/\tau)\,\phi(z/\tau),\ \tau = 0.75 & \text{[Model A]} \\ 2|z| & \text{[Model B]} \\ 4\sqrt{|z/2|\,(1 - |z/2|)}\sin\{2\pi(1+\epsilon)/(|z/2| + \epsilon)\},\ \epsilon = 0.05 & \text{[Model C]} \end{cases}.$$

Both $g_{21}(\cdot)$ and Model A, which are inspired by the Monte Carlo design of Horowitz and Spokoiny (2001), can be viewed as a linear function with a bump. Model A is a smooth function, whereas Models B and C have a kink at the origin. Strictly speaking, these models violate the smoothness condition given in Assumption 7. Nonetheless we investigate them to see how the violation affects finite-sample properties of MSII-FM. In addition, Model C is (a mirror image of) the Doppler function, which is a rapidly oscillating, spatially inhomogeneous function, as illustrated in Figure 1 of Donoho and Johnstone (1994). Therefore, the model may be thought of as the most difficult case among the three. This is the model for which we report the results

here.[9]

Finally, $Y$ is generated by setting all coefficients in (5) equal to 1 with $u \overset{iid}{\sim} N(0,1)$.

The above procedure provides us with two observable samples $\mathcal{S}_1 = \{(Y_i, X_{1i}, Z_i)\}_{i=1}^n$ and $\mathcal{S}_2 = \{(X_{2j}, Z_j)\}_{j=1}^m$, and one complete sample $\mathcal{S}^*$. Finally, the matched sample $\mathcal{S} = \left\{\left(Y_i, X_{1i}, X_{2j_1(i)}, \ldots, X_{2j_K(i)}, Z_i, Z_{j_1(i)}, \ldots, Z_{j_K(i)}\right)\right\}_{i=1}^n$ is constructed via the NNM with respect to $Z$, where the NNM is based on the Mahalanobis metric. We focus only on small numbers of matches and examine $K \in \{1, 2, 4, 8\}$.[10]

With regards to sample sizes, for each of $n \in \{1000, 2000\}$, $m$ is chosen as one of $m \in \{n/2, n, 2n\}$ so that the values of $\kappa$ are $\kappa = 2, 1$ and $1/2$, respectively. For each combination of sample sizes $(n, m)$ and the functional form of $g_{22}(z)$, we generate 1000 Monte Carlo samples. The following five estimators are examined: (i) the infeasible OLS estimator using the complete sample $\mathcal{S}^*$ [OLS*]; (ii) the MSOLS estimator using the matched sample $\mathcal{S}$ and $W_{i,j(i)}$ [MSOLS-A]; (iii) the MSOLS estimator using the matched sample $\mathcal{S}$ and $W_{i,j(i)}^\dagger$ [MSOLS-B]; (iv) the MSII(-FM) estimator using the matched sample $\mathcal{S}$ and $W_{i,j(i)}$ [MSII(-FM)-A]; and (v) the MSII(-FM) estimator using the matched sample $\mathcal{S}$ and $W_{i,j(i)}^\dagger$ [MSII(-FM)-B]. Second-, third- and fourth-order polynomials are investigated in the power-series approximation for MSII-FM, and these specifications are denoted as "*2nd*", "*3rd*" and "*4th*" in the row "*Poly.*", respectively. Results on the initial MSII are also available as "*initial*" for reference. Moreover, the consistent estimator of the second-order bias term for MSII-FM-B is $\hat{\lambda}_{i,j(i)}^\dagger = \hat{\lambda}_{i,j(i)} + \left\{Z_i - (1/K)\sum_{j \in \mathcal{J}_K(i)} Z_j\right\}' \hat{\gamma}_{II}^{(1)\prime}$.

We focus on finite-sample properties of estimators of $\beta_{22}$ and $\gamma_1$. For each estimator, the following performance measures are computed: (i) $Mean$ (simulation average of the parameter estimate); (ii) $SD$ (simulation average of the parameter estimate);

[9] Comprehensive simulation results, available in the online Supplement, are even more favorable for Models A and B.

[10] In our preliminary Monte Carlo study larger values of matches (e.g., $K = 16, 32, 64, 128$) have been also investigated. However, the results are quite poor.

(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 MSOLS is inconsistent and limiting distributions of the initial MSII for $d_3 = 2, 3$ are not available, their standard errors are not well defined. Accordingly, $\overline{SE}$ and $CR$ are not computed for these estimators.

TABLE 1 ABOUT HERE

## 4.2 Results

Simulation results are summarized in Table 1. To save space, we present only the results from the most difficult case (Model C) for $(n, m) = (1000, 1000)$ and $(2000, 2000)$.

**(a) For $d_3 = 1$:** Panel (a) reports the results for a single matching variable. Because of conditional homoskedasticity of the error term $u$, OLS* is the best linear unbiased estimator. The results indicate that it is unbiased and yields small standard deviations. However, OLS* is an infeasible, oracle estimator. Instead, we should make a realistic comparison between MSOLS and MSII and use OLS* as the benchmark to measure the efficiency loss when all variables cannot be taken from a single data source.

For MSOLS, whether $W_{i,j(i)}$ or $W^{\dagger}_{i,j(i)}$ is used as the regressor has almost no difference; this reflects the fact that the extra second-order bias induced by replacing $Z_i$ with $Z_{j(i)}$ is $O_p\left(n^{-1}\right) = o_p\left(n^{-1/2}\right)$. As predicted in Theorem 1, the bias of the MSOLS estimate decreases with the number of matches $K$. However, it is inconsistent in that its bias does not vanish with the sample size $n$. Also observe that the standard deviation of each MSOLS estimate shrinks with $n$, as Theorem 1 suggests.

Now we turn to MSII. At a glance, we can find that the proposed bias-correction method works remarkably well, and that the choice of the regressor again does not change the results. However, unlike MSOLS, increasing $K$ has little effect at best,

which suggests that MSII works well across small values of $K$. The results also confirm consistency of MSII; as $n$ increases, the simulation average of each MSII estimate gets closer to the truth and its standard deviation shrinks. In addition, $\overline{SE}$ is reasonably close to $SD$, which indicates that the (properly-scaled) covariance estimator $\hat{\Omega}$ yields good estimates of standard deviations of MSII. Coverage rates are also close to the nominal level of confidence, and the single match case appears to have advantage from the viewpoint of coverage accuracy.

Comparing MSII with OLS*, we have the following two findings. First, unlike OLS*, MSII 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. It is worth noting that while standard deviations of MSOLS are greater than those of OLS*, they are smaller than those of MSII. This can be explained by the fact that the asymptotic variance of $\sqrt{n}\left(\hat{\theta}_{OLS} - Q_W^{-1}P_W\theta - B_{OLS2}\right)$ is $Q_W^{-1}\Omega_{11}Q_W^{-1}$, which tends to be smaller (in the matrix sense) than $P_W^{-1}\Omega P_W^{-1}$.

**(b) For $d_3 = 2$:** Next, we look into Panel (b), which presents the results from two matching variables. Only results of MSII-FM for $K = 1$ are provided, because those for $K \geq 2$ are quite poor. As in the case for $d_3 = 1$, employing $W_{i,j(i)}$ or $W^{\dagger}_{i,j(i)}$ has little effect on MSOLS or MSII-FM; although the extra second-order bias generated by switching $Z_i$ to $Z_{j(i)}$ is $O_p\left(n^{-1/2}\right)$, its effect appears to be minor at best.

Even after the number of matching variables increases, the general tendency remains unchanged. Performance of MSOLS varies with $K$. MSII-FM successfully corrects the bias generated by MSOLS, at the expense of precision in estimation. Standard deviations of MSII-FM are close to that of the initial MSII, which reflects that the FM procedure corrects the second-order bias of MSII without inflating the

variance. However, FM works only for $K = 1$. The rationale could be that FM requires both the initial MSII and second-order bias estimates to be of good quality. This requirement is unlikely to be satisfied with many matches, which include poor ones and thus inevitably affect the performance of MSII-FM. In terms of the power-series approximation, results from the second- and third-order polynomials look similar, and those from the fourth-order polynomial differ slightly. Coverage accuracy in estimates of $\beta_{22}$ may be a concern. However, based on the results for larger samples (reported in the Supplement), it seems that the under-coverage is due to finite-sample bias of MSII-FM.

**(c) For $d_3 = 3$:** In Panel (c), only results of MSII-FM for $K = 1$ are provided again in view of quality. An apparent difference is that once the number of matching variables increases to three, results from using $W_{i,j(i)}$ or $W^{\dagger}_{i,j(i)}$ differ substantially for each of MSOLS and MSII-FM. Observe that MSII-FM using $W_{i,j(i)}$ exhibits much better finite-sample properties. In contrast, MSII-FM based on $W^{\dagger}_{i,j(i)}$ generates considerable biases in estimates of $\gamma$. The extra second-order bias when $Z_{j(i)}$ is used in place of $Z_i$ becomes as slow as $O_p\left(n^{-1/3}\right)$, and its adverse effect is no longer negligible in finite samples. Coverage rates of MSII-FM are improved from those for $d_3 = 2$. In terms of the series approximation, results from the second-and third-order polynomials are again similar. However, those from the fourth-order polynomial look inferior in the presence of non-smoothness in $g_{22}(\cdot)$, in particular, for Model B.

**(d) Summary:** Simulation results confirm that the bias-corrected estimation proposed in this paper works reasonably well. Simulation averages of MSII(-FM) for $d_3 = 1$ ($d_3 = 2, 3$) tend to be closer to the truths as $n$ increases, even in the most difficult case. Judging from the Monte Carlo evidence, we recommend setting $K = 1$, employing $W_{i,j(i)}$ as the regressor, and applying the second- or third-order polynomials for the series approximation in MSII-FM. It follows that making MSOLS consistent

by use of $K$-NN method (i.e., by letting $K$ diverge at a slower rate than $n$ and $m$) does not appear to be a solution in the setting of matched sample estimation. Rather, it looks promising to pursue the strategy of constructing a matched sample based on a single match and then correcting the non-negligible bias of the estimate analytically.

# 5 An Empirical Application: Returns to Schooling

We now apply our proposed estimation methods to a version of Mincer's (1974) wage regression. As argued in Card (1995), the estimation result may suffer from the "ability bias" unless it includes a variable representing ability as a regressor. Therefore, we consider the following wage regression

$$\begin{aligned}\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 abil \\ + \beta_5 feduc + \beta_6 meduc + \beta_7 black + \beta_8 smsa + \beta_9 south + u, \end{aligned} \tag{6}$$

where *educ* is years of education, *exper* is work experience, *abil* is an ability measure, *feduc* and *meduc* are years of father's and mother's education, and *black*, *smsa* and *south* are indicator variables that take one if the individual is black, lives in the urban area and south, respectively.

We estimate regression (6) using three data sets, namely, those used in Card (1995), Blackburn and Neumark (1992), and Heckman, Tobias and Vytlacil (2000). The data sets are available under the names "`card`", "`wage2`" and "`htv`", respectively, as supplemental materials for Wooldridge (2013). Each of the three data sets is drawn from the National Longitudinal Survey (NLS) and contains some ability measure; to be precise, while both `card` and `wage2` include scores of IQ and Knowledge of the World of Work (*kww*) tests, `htv` has the "*g*" measure constructed from 10 component tests of the Armed Services Vocational Aptitude Battery.

We conduct two exercises that address the following questions:

**(Q1)** How would the estimation result change if *kww* in `card` were missing and

instead taken from `wage2`?

**(Q2)** What would happen if *kww* in `card` were replaced by $g$ from `htv`?

For these exercises, the OLS result using 2191 male observations in `card` with *kww* chosen as *abil* can be viewed as the benchmark result from the infeasible OLS*. Because each of Q1 and Q2 requires a matched sample, we regard `card` as $\mathcal{S}_1$ and `wage2` or `htv` as $\mathcal{S}_2$. The NNM is made in the following manner. When `wage2` is employed as $\mathcal{S}_2$, $(educ, feduc, meduc, black, smsa, south)$ are chosen as matching variables, where the first three variables are treated as continuous. On the other hand, `htv` contains only white-male observations. Accordingly, when using it as $\mathcal{S}_2$, we choose five matching variables excluding *black*. Not surprisingly, there are several ties of the matching variables in $\mathcal{S}_2$. Then, we take an average of *kww* or $g$ within ties and assign the average as the unique value of the ability measure to each combination of matching variables. As a consequence, 466 and 589 distinct combinations of matching variables remain in male samples of `wage2` and `htv`, respectively. In both cases, the NNM is based on the Mahalanobis metric, and we set the number of matches $K = 1$ (single match) based on the simulation results.

Given the matched sample, we estimate (6) by MSOLS and MSII-FM. Specifically, MSOLS-A and MSII-FM-A (i.e., estimators with $W_{i,j(i)}$ used as the regressor) are chosen, and the third-order polynomial is applied for the power-series approximation of MSII-FM, again based on the simulation results; estimation results from second- and fourth-order polynomials are qualitatively similar.

TABLE 2 ABOUT HERE

Table 2 presents estimation results and standard errors (in parentheses). White's (1980) heteroskedasticity-robust standard errors are computed for OLS*, whereas 'standard errors' for MSOLS are square-roots of diagonal elements of $\hat{Q}_W^{-1}\hat{\Omega}_{11}\hat{Q}_W^{-1}/n :=$

$\hat{Q}_W^{-1}\left(\hat{\Omega}_{11A}+\hat{\Omega}_{11B}\right)\hat{Q}_W^{-1}/n$, where

$$\hat{\Omega}_{11B}=\frac{n}{m}\left[\left(\hat{\beta}_{2,II}'\hat{\Sigma}_2\hat{\beta}_{2,II}\right)\bar{W}\bar{W}'\right.$$
$$\left.+\frac{1}{K^2}\,\mathrm{diag}\left\{0_{(d_1+1)\times(d_1+1)},\left(\hat{\beta}_{2,II}'\hat{\Sigma}_2\hat{\beta}_{2,II}\right)\hat{V}_{g_2}+2\left\{\hat{\Gamma}\left(-1\right)+\hat{\Gamma}\left(1\right)\right\},0_{d_3\times d_3}\right\}\right]$$

for $(n,m)$ given in the corresponding column of Table 2. The latter should be interpreted with caution; because $\hat{\theta}_{OLS}$ is inconsistent (and even its convergence rate is slower than the parametric one), the numbers merely indicate measures of dispersion at the same scale as other estimates and are not intended for inference.

The benchmark OLS* result using `card` is provided in the first column. Signs of the coefficient estimates on *educ*, *exper*, $exper^2$, and $abil\,(=kww)$ are as expected, and they are significant at the 5% level. To answer Q1, we run MSOLS and MSII-FM using the matched sample with `wage2`. The results are reported in columns 2 and 3. Signs of the coefficient estimates by MSII-FM are the same as those by OLS*. On the other hand, MSOLS overestimates returns to schooling due to failure to correct for matching results. It also yields a negative estimate of the ability effect, whereas the one from MSII-FM is positive (but insignificant due to the large standard error).

Furthermore, to answer Q2, we replace the ability measure with $g$ by constructing the matched sample with `htv`. Results from MSOLS and MSII-FM using this sample are presented in columns 4 and 5. There is still the tendency that MSII-FM estimates are closer to those of OLS*. MSOLS again tends to inflate returns to schooling. The estimated ability effect turns positive, but its magnitude is much smaller than the one from MSII-FM.

# 6 Conclusion

Regression estimation using samples constructed via the NNM from two sources is not uncommon in applied economics. This paper has demonstrated that such OLS estimators are generally inconsistent, and thus an appropriate bias correction is re-

quired. It has also been shown that the convergence rate to the probability limit of the OLS depends on the number of matching variables and the divergence pattern of two sample sizes.

Two versions of bias-corrected estimators have been proposed, and each can be interpreted as a variant of indirect inference estimators. The MSII estimator attains the parametric convergence rate for the cases with at most two matching variables, whereas the MSII-FM estimator achieves the parametric convergence rate when the number of matching variables does not exceed four. Monte Carlo results suggest that a small number of matches work well in practice, and in particular, we should consider the single match when the number of matching variables is two or three.

The paper aims at providing corrections for an established practice, which is to run (parametric) OLS ignoring imputation. In particular, our proposal for MSII-FM is based on a nonparametric series estimation of $g_2(Z) = E(X_2|Z)$. The nonparametric estimator is employed only when the curse of dimensionality in matching variables prevents MSII from attaining parametric convergence. Alternatively, it is possible to use a nonparametric estimate of $g_2(Z)$ as a (generated) regressor in place of $X_2$ in regression (1) from the beginning. As illustrated in Section 2.4, this would result in a partially linear semiparametric model with a measurement error problem. There could be several different (nonparametric) estimators available for the model of this class. However, such estimators are not as widely used in practice as the OLS, and hence we leave the development of the estimators for future work.

Several other extensions would be fruitful. First, 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 closely 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.

Second, combining our matched-sample estimation theory with IV/GMM estimation would be also of interest in the presence of endogeneity in regressors. This is particularly relevant to empirical studies using earnings data, which are thought to include measurement errors and imputation biases.

Finally, the estimation theory may be extended to kernel estimation of varying coefficient models using matched samples. It is not difficult to see that kernel estimators of the varying coefficients are also inconsistent, and appropriate bias-correction methods similar to those proposed in this paper are worth investigating.

# A Appendix: Technical Proofs

## A.1 A Useful Lemma

Before proceeding, we present a lemma about the error bounds from NNM, which is repeatedly applied in the technical proofs below. To do so, we provide the formal definition of the matching discrepancy from Abadie and Imbens (2006).

Let $z \in \mathbb{Z}$ be a fixed value of the matching variable $Z$, where, in practice, $z$ is one of $\{Z_i\}_{i=1}^n$ in $\mathcal{S}_1$. Then, the $k$th closest matching discrepancy $U_k = U_k(z)$, $k = 1, \ldots, K$ is defined as $U_k := Z_{j_k(z)} - z$ if $Z_{j_k(z)}$ is the $k$th closest match to $z$ among all $\{Z_j\}_{j=1}^m$ in $\mathcal{S}_2$. The following lemma states uniform moment bounds of the matching discrepancy.

**Lemma A1. (Abadie and Imbens, 2006, Lemma 2)** *If Assumptions 1-2 hold, then all the moments of $m^{1/d_3}\|U_k\|$ are uniformly bounded in $m$ and $z \in \mathbb{Z}$.*

## A.2 Proof of Theorem 1

The proof requires the following lemma.

**Lemma A2.** *If Assumptions 1-3 hold, then $B_{R_W 2} = O\left(m^{-1/d_3}\right) + o_p\left(n^{-1/2}\right)$.*

### A.2.1 Proof of Lemma A2

Consider the identity $B_{R_W 2} \equiv E\left(B_{R_W 2}\right) + \left\{B_{R_W 2} - E\left(B_{R_W 2}\right)\right\}$. It follows from Assumption 1, Lipschitz continuity of $g_2$, the Cauchy-Schwarz inequality, and Lemma A1 that $\left\|E\left(m^{1/d_3} B_{R_W 2}\right)\right\| < \infty$, and thus $E\left(B_{R_W 2}\right) = O\left(m^{-1/d_3}\right)$. Similarly, we have $Var\left(n^{1/2} m^{1/d_3} \left\|B_{R_W 2}\right\|\right) < \infty$ so that

$$B_{R_W 2} = O\left(m^{-1/d_3}\right) + O_p\left(n^{-1/2} m^{-1/d_3}\right) = O\left(m^{-1/d_3}\right) + o_p\left(n^{-1/2}\right). \blacksquare$$

### A.2.2 Proof of Theorem 1

It is easy to see from (2) that $\hat{R}_W := \hat{Q}_W \theta + B_{R_W 1} + B_{R_W 2} + E_{R_W}$, where

$$\begin{aligned}
B_{R_W 1} &= E\left(W_{i,j(i)} \epsilon_{i,j(i)}\right), \\
B_{R_W 2} &= \frac{1}{n} \sum_{i=1}^{n} W_{i,j(i)} \lambda_{i,j(i)}, \text{ and} \\
E_{R_W} &= \frac{1}{n} \sum_{i=1}^{n} \left\{W_{i,j(i)} \epsilon_{i,j(i)} - E\left(W_{i,j(i)} \epsilon_{i,j(i)}\right)\right\}.
\end{aligned}$$

It follows that $\hat{\theta}_{OLS} := \theta + B_{OLS1} + B_{OLS2} + E_{OLS}$, where $B_{OLS1} = \hat{Q}_W^{-1} B_{R_W 1}$, $B_{OLS2} = \hat{Q}_W^{-1} B_{R_W 2}$ and $E_{OLS} = \hat{Q}_W^{-1} E_{R_W}$ correspond to the first-order (or leading) bias, the second-order bias due to the matching discrepancy and the weighted average of errors, respectively.

We begin with evaluating $B_{OLS1}$. First note that $E\left(X_{1i} \eta_{2i}'\right) = E\left\{g_1\left(Z\right) \eta_2'\right\} + E\left(\eta_1 \eta_2'\right) = 0_{d_1 \times d_2}$, $E\left(X_{2j(i)} \eta_{2j(i)}'\right) = \left(1/K\right) \Sigma_2$, and that the $i$th and $j_k\left(i\right)$th observations are independent. Then,

$$B_{R_W 1} = \left[\begin{array}{c} 0_{(d_1+1)\times 1} \\ -\left(1/K\right) \Sigma_2 \beta_2 \\ 0_{d_3 \times 1} \end{array}\right] = -\frac{1}{K} \operatorname{diag}\left\{0_{(d_1+1)\times(d_1+1)}, \Sigma_2, 0_{d_3 \times d_3}\right\} \theta := -\frac{1}{K} \Sigma \theta.$$

Because $\hat{Q}_W = Q_W + O_p\left(n^{-1/2}\right)$, we obtain $B_{OLS1} = -\left(1/K\right) Q_W^{-1} \Sigma \theta + O_p\left(n^{-1/2}\right)$. Next, Lemma A2 implies that $B_{OLS2} = O\left(m^{-1/d_3}\right) + o_p\left(n^{-1/2}\right)$. Finally, $E_{R_W} = O_p\left(n^{-1/2}\right)$ by CLT, and thus $E_{OLS} = O_p\left(n^{-1/2}\right)$. Therefore, $\hat{\theta}_{OLS} = Q_W^{-1} P_W \theta + O\left(m^{-1/d_3}\right) + O_p\left(n^{-1/2}\right)$ by denoting $P_W := Q_W - \left(1/K\right) \Sigma$. $\blacksquare$

## A.3 Proof of Theorem 2

The proof requires the following lemma.

**Lemma A3.** *If Assumptions 1-3 hold, then* $\hat{\Sigma}_2$ *and* $B_{\Sigma_2}$ *admit the expansions* $\hat{\Sigma}_2 - B_{\Sigma_2} = \Sigma_2 + O_p\left(m^{-1/2}\right)$ *and* $B_{\Sigma_2} = O\left(m^{-2/d_3}\right) + o_p\left(m^{-1/2}\right)$.

### A.3.1 Proof of Lemma A3

Observe that

$$\begin{aligned}
\hat{\Sigma}_2 &= \frac{1}{2(m-1)} \sum_{j=2}^{m} \Delta X_{2(j)} \Delta X_{2(j)}' \\
&= \frac{1}{2(m-1)} \sum_{j=2}^{m} \Delta \eta_{2(j)} \Delta \eta_{2(j)}' + \frac{1}{2(m-1)} \sum_{j=2}^{m} \Delta \eta_{2(j)} \Delta g_{2(j)}' \\
&+ \frac{1}{2(m-1)} \sum_{j=2}^{m} \Delta g_{2(j)} \Delta \eta_{2(j)}' + \frac{1}{2(m-1)} \sum_{j=2}^{m} \Delta g_{2(j)} \Delta g_{2(j)}' \\
&= S_1 + S_2 + S_3 + S_4 \text{ (say)}.
\end{aligned}$$

As in the proof of Lemma A2, consider the identity $S_i \equiv E(S_i) + \{S_i - E(S_i)\}$ for $i = 1, 2, 3, 4$. First, $S_1 = E(S_1) + \{S_1 - E(S_1)\} = \Sigma_2 + O_p\left(m^{-1/2}\right)$. Second, for $S_4$, it follows from Lipschitz continuity of $g_2(\cdot)$ in Assumption 3(ii) that the order of magnitude in $\|E(S_4)\|$ is determined by $E\left\{\sum_{j=2}^{m} \left\|\Delta Z_{(j)}\right\|^2\right\} / \{2(m-1)\}$. Because the re-ordered sample $\left\{Z_{(j)}\right\}_{j=1}^{m}$ is constructed by the nearest-neighbor algorithm in accordance with Yatchew (1997), we may apply the moment approximation in Theorem 5.4 of Evans, Jones and Schmidt (2002) to obtain $E\left\{\sum_{j=2}^{m} \left\|\Delta Z_{(j)}\right\|^2\right\} / \{2(m-1)\} = O\left(m^{-2/d_3}\right)$ so that $E(S_4) = O\left(m^{-2/d_3}\right)$. Also $Var\left(m^{1/2+2/d_3} \|S_4\|\right) < \infty$ by the theorem, and thus $S_4 = O\left(m^{-2/d_3}\right) + O_p\left(m^{-(1/2+2/d_3)}\right) = O\left(m^{-2/d_3}\right) + o_p\left(m^{-1/2}\right)$. Third, for $S_2$ and $S_3$, we have $E(S_2) = E(S_3) = 0_{d_2 \times d_2}$. Because, by a similar argument to above, $Var\left(m^{1/2+1/d_3} \|S_3\|\right), Var\left(m^{1/2+1/d_3} \|S_3\|\right) < \infty$, each of $S_2$ and $S_3$ is $O_p\left(m^{-(1/2+1/d_3)}\right) = o_p\left(m^{-1/2}\right)$. Finally, rewriting $S_4$ as $B_{\Sigma_2}$ yields the stated result. ■

### A.3.2 Proof of Theorem 2

By the proof of Theorem 1,

$$\hat{R}_W = \left(\hat{Q}_W - \frac{1}{K}\Sigma\right)\theta + B_{R_W 2} + E_{R_W} = \hat{P}_W\theta + B_{R_W 2} + \frac{1}{K}\left(\hat{\Sigma} - \Sigma\right)\theta + E_{R_W}.$$

It also follows from Lemma A3 that

$$\begin{aligned}\frac{1}{K}\left(\hat{\Sigma} - \Sigma\right)\theta &= \frac{1}{K}\,\mathrm{diag}\left\{0_{(d_1+1)\times(d_1+1)}, \left(\hat{\Sigma}_2 - B_{\Sigma_2}\right) - \Sigma_2, 0_{d_3\times d_3}\right\}\theta \\ &\quad + \frac{1}{K}\,\mathrm{diag}\left\{0_{(d_1+1)\times(d_1+1)}, B_{\Sigma_2}, 0_{d_3\times d_3}\right\}\theta \\ &= O_p\left(m^{-1/2}\right) + \left\{O\left(m^{-2/d_3}\right) + o_p\left(m^{-1/2}\right)\right\},\end{aligned}$$

where the first term on the right-hand side is denoted as $E_{\Sigma_2}$ hereinafter. Therefore,

$$\begin{aligned}\hat{\theta}_{II} &= \hat{P}_W^{-1}\hat{R}_W \\ &= \theta + B_{MD} + B_\Sigma + \hat{P}_W^{-1}E_{R_W} + \hat{P}_W^{-1}E_{\Sigma_2} \\ &= \theta + \left\{O\left(m^{-1/d_3}\right) + o_p\left(n^{-1/2}\right)\right\} + \left\{O\left(m^{-2/d_3}\right) + o_p\left(m^{-1/2}\right)\right\} \\ &\quad + O_p\left(n^{-1/2}\right) + O_p\left(m^{-1/2}\right) \xrightarrow{p} \theta. \blacksquare\end{aligned} \tag{A1}$$

## A.4 Proof of Theorem 3

To save space, we only demonstrate asymptotic normality results depending on the divergence pattern of $(n, m)$; see the Supplement for detailed derivation of asymptotic variances $V_I$, $V_{II}$ and $V_{III}$. It follows from (A1) that if $n/m \to \kappa$, then $\sqrt{n}\left(\hat{\theta}_{II} - \theta - B_{MD} - B_\Sigma\right) \xrightarrow{d} N\left(0, V_I\right)$, where

$$V_I := P_W^{-1} \lim_{\substack{n,m\to\infty \\ n/m\to\kappa}} Var\left\{\sqrt{n}\left(E_{R_W} + E_{\Sigma_2}\right)\right\} P_W^{-1}.$$

Alternatively, if $n/m \to 0$, then $\sqrt{n}\left(\hat{\theta}_{II} - \theta - B_{MD} - B_\Sigma\right) \xrightarrow{d} N\left(0, V_{II}\right)$, where

$$V_{II} := P_W^{-1} \lim_{\substack{n,m\to\infty \\ n/m\to 0}} Var\left(\sqrt{n}E_{R_W}\right) P_W^{-1}.$$

Finally, if $n/m \to \infty$, then $\sqrt{m}\left(\hat{\theta}_{II} - \theta - B_{MD} - B_\Sigma\right) \xrightarrow{d} N\left(0, V_{III}\right)$, where

$$V_{III} := P_W^{-1} \lim_{m\to\infty} Var\left(\sqrt{m}E_{\Sigma_2}\right) P_W^{-1}. \blacksquare$$

## A.5 Proof of Corollary 1

Because the corollary is a straightforward outcome of Theorem 3, its proof is omitted. ■

## A.6 Proof of Theorem 4

The proof requires the following lemma.

**Lemma A4.** *If Assumptions 1-7 hold, then*

$$\max_{1\leq i\leq n}\left|\hat{\lambda}_{i,j(i)}-\lambda_{i,j(i)}\right| = \begin{cases} o_p\left(n^{-1/2}\right) & \text{if } n/m\to\kappa \text{ and } d_3=2,3 \text{ or if } n/m\to 0 \text{ and } d_3=3,4 \\ o_p\left(m^{-1/2}\right) & \text{if } n/m\to\infty \text{ and } d_3=2,3 \end{cases}.$$

### A.6.1 Proof of Lemma A4

It is easy to see that $\hat{\lambda}_{i,j(i)} := R_{1i}+R_{2i}+R_{3i}+\lambda_{i,j(i)}$, where

$$\begin{aligned} R_{1i} &= \frac{1}{K}\sum_{j\in\mathcal{J}_K(i)}\left[\{\hat{g}_2(Z_i)-g_2(Z_i)\}-\{\hat{g}_2(Z_j)-g_2(Z_j)\}\right]'\left(\hat{\beta}^{(1)}_{II,2}-\beta_2\right), \\ R_{2i} &= \frac{1}{K}\sum_{j\in\mathcal{J}_K(i)}\left[\{\hat{g}_2(Z_i)-g_2(Z_i)\}-\{\hat{g}_2(Z_j)-g_2(Z_j)\}\right]'\beta_2, \text{ and} \\ R_{3i} &= \frac{1}{K}\sum_{j\in\mathcal{J}_K(i)}\{g_2(Z_i)-g_2(Z_j)\}'\left(\hat{\beta}^{(1)}_{II,2}-\beta_2\right). \end{aligned}$$

Hence, the proof is boiled down to demonstrating that each of $\max_{1\leq i\leq n}|R_{\ell i}|$, $\ell=1,2,3$ is bounded by either $o_p\left(n^{-1/2}\right)$ or $o_p\left(m^{-1/2}\right)$, depending on the divergence pattern of $(n,m)$ and $d_3$.

We first work on $R_{3i}$. To derive the bounds for $R_{1i}$ and $R_{3i}$, we may apply (A1) to obtain

$$\hat{\theta}^{(1)}_{II}=\theta+\begin{cases} O\left(n^{-1/d_3}\right)+O_p\left(n^{-1/2}\right) & \text{if } n/m\to\kappa \text{ and } d_3=2,3 \\ O\left(m^{-1/d_3}\right)+O_p\left(n^{-1/2}\right) & \text{if } n/m\to 0 \text{ and } d_3=3,4 \\ O\left(m^{-1/d_3}\right)+O_p\left(m^{-1/2}\right) & \text{if } n/m\to\infty \text{ and } d_3=2,3 \end{cases}.$$

It follows from Lemma A1 and Lipschitz continuity of $g_2(\cdot)$ that $\max_{1\leq i\leq n}|R_{3i}|$ is bounded by

$$\begin{cases} O_p\left(m^{-1/d_3}\right)\left\{O\left(n^{-1/d_3}\right)+O_p\left(n^{-1/2}\right)\right\}=o_p\left(n^{-1/2}\right) & \text{if } n/m\to\kappa \text{ and } d_3=2,3 \\ O_p\left(m^{-1/d_3}\right)\left\{O\left(m^{-1/d_3}\right)+O_p\left(n^{-1/2}\right)\right\}=o_p\left(n^{-1/2}\right) & \text{if } n/m\to 0 \text{ and } d_3=3,4 \\ O_p\left(m^{-1/d_3}\right)\left\{O\left(m^{-1/d_3}\right)+O_p\left(m^{-1/2}\right)\right\}=o_p\left(m^{-1/2}\right) & \text{if } n/m\to\infty \text{ and } d_3=2,3 \end{cases}.$$

The remaining task is to demonstrate that for $k=1,\ldots,K$,

$$\begin{aligned} &\max_{1\leq i\leq n}\left\|\left\{\hat{g}_2(Z_i)-g_2(Z_i)\right\}-\left\{\hat{g}_2\left(Z_{j_k(i)}\right)-g_2\left(Z_{j_k(i)}\right)\right\}\right\| \\ &=\begin{cases} o_p\left(n^{-1/2}\right) & \text{if } n/m\to\kappa \text{ and } d_3=2,3 \text{ or if } n/m\to 0 \text{ and } d_3=3,4 \\ o_p\left(m^{-1/2}\right) & \text{if } n/m\to\infty \text{ and } d_3=2,3 \end{cases}. \end{aligned} \quad \text{(A2)}$$

However, Lemma A.2 of Abadie and Imbens (2011) holds under Assumptions 1-7. Therefore,

$$\max_{1\leq i\leq n}\left|\left\{\hat{g}_{2r}(Z_i)-\hat{g}_{2r}\left(Z_{j_k(i)}\right)\right\}-\left\{g_{2r}(Z_i)-g_{2r}\left(Z_{j_k(i)}\right)\right\}\right|=o_p\left(m^{-1/2}\right),\ r=1,\ldots,d_2,$$

and thus (A2) immediately follows. Then, each of $\max_{1\leq i\leq n}|R_{1i}|$ and $\max_{1\leq i\leq n}|R_{2i}|$ is also bounded by either $o_p\left(n^{-1/2}\right)$ or $o_p\left(m^{-1/2}\right)$. This completes the proof. ■

### A.6.2 Proof of Theorem 4

To save space, we focus on the case in which $n/m\to\kappa\in(0,\infty)$ and $d_3=2,3$. Observe that $\hat{\theta}_{II-FM}:=\hat{\theta}_{II}^{(1)}-\hat{B}_{MD}:=\hat{\theta}_{II}^{(1)}-(1/n)\sum_{i=1}^{n}W_{i,j(i)}\hat{\lambda}_{i,j(i)}$. By Theorem 3, $\sqrt{n}\left(\hat{\theta}_{II}^{(1)}-\theta-B_{MD}-B_{\Sigma}\right)\xrightarrow{d}N(0,V_I)$. Lemma A4 implies that $\left\|B_{MD}-\hat{B}_{MD}\right\|=o_p\left(n^{-1/2}\right)$. Then, $\sqrt{n}\left\{\left(B_{MD}-\hat{B}_{MD}\right)+B_{\Sigma}\right\}=o_p(1)$, and thus the result immediately follows. ■

## A.7 Proof of Proposition 1

The proof is obvious in light of Section A of the Supplement and thus omitted. ■

# 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 (2012): "A Martingale Representation for Matching Estimators," *Journal of the American Statistical Association*, 107, 833 - 843.

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

[5] Angrist, J. D., and A. B. 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, 328 - 336.

[6] Angrist, J. D., and A. B. Krueger (1995): "Split-Sample Instrumental Variables Estimates of the Return to Schooling," *Journal of Business & Economic Statistics*, 13, 225 - 235.

[7] Arellano, M., and C. Meghir (1992): "Female Labour Supply and on the Job Search: An Empirical Model Estimated Using Complementary Data Sets," *Review of Economic Studies*, 59, 537 - 559.

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

[9] Björklund, J., and M. Jäntti (1997): "Intergenerational Income Mobility in Sweden Compared to the United States," *American Economic Review*, 87, 1009 - 1018.

[10] Bollinger, C. R., and B. T. Hirsch (2006): "Match Bias from Earnings Imputation in the Current Population Survey: The Case of Imperfect Matching," *Journal of Labor Economics*, 24, 483 - 519.

[11] Borjas, G. J. (2004): "Food Insecurity and Public Assistance," *Journal of Public Economics*, 88, 1421 - 1443.

[12] Bostic, R., S. Gabriel, and G. Painter (2009): "Housing Wealth, Financial Wealth, and Consumption: New Evidence from Micro Data," *Regional Science and Urban Economics*, 39, 79 - 89.

[13] Bover, O. (2005): "Wealth Effects on Consumption: Microeconometric Estimates from the Spanish Survey of Household Finances," Documentos de Trabajo No.0522, Banco de España.

[14] Busso, M., J. DiNardo, and J. McCrary (2014): "New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators," *Review of Economics and Statistics*, 96, 885 - 897.

[15] 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.

[16] Carrasco, M., and J.-P. Florens (2002): "Simulation-Based Method of Moments and Efficiency," *Journal of Business & Economic Statistics*, 20, 482 - 492.

[17] Chen, J., and H. Shao (2001): "Jackknife Variance Estimation for Nearest-Neighbor Imputation," *Journal of the American Statistical Association*, 96, 260 - 269.

[18] Currie, J., and A. Yelowitz (2000): "Are Public Housing Projects Good for Kids?" *Journal of Public Economics*, 75, 99 - 124.

[19] Dee, T. S., and W. N. Evans (2003): "Teen Drinking and Educational Attainment: Evidence from Two-Sample Instrumental Variables Estimates," *Journal of Labor Economics*, 21, 178 - 209.

[20] Donoho, D. L., and I. M. Johnstone (1994): "Ideal Spacial Adaptation by Wavelet Shrinkage," *Biometrika*, 81, 425 - 455.

[21] Evans, D., A. J. Jones, and W. M. Schmidt (2002): "Asymptotic Moments of Near-Neighbour Distance Distributions," *Proceedings of the Royal Society A*, 458, 2839 - 2849.

[22] Fujii, T. (2008): "Two-Sample Estimation of Poverty Rates for Disabled People: An Application to Tanzania," Singapore Management University Economics & Statistics Working Paper No.02-2008.

[23] Gouriéroux, C., A. Monfort, and E. Renault (1993): "Indirect Inference," *Journal of Applied Econometrics*, 8, S85 - S118.

[24] Heckman, J., J. L. Tobias, and E. Vytlacil (2000): "Simple Estimators for Treatment Parameters in a Latent Variable Framework with an Application to Estimating the Return to Schooling," NBER Working Paper No.7950.

[25] Hellerstein, J. K., and G. W. Imbens (1999): "Imposing Moment Restrictions from Auxiliary Data by Weighting," *Review of Economics and Statistics*, 81, 1 - 14.

[26] Hirsch, B. T., and E. J. Schumacher (2004): "Match Bias in Wage Gap Estimates due to Earnings Imputation," *Journal of Labor Economics*, 22, 689 - 722.

[27] 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.

[28] Imbens, G. W., and T. Lancaster (1994): "Combining Micro and Macro Data in Microeconometric Models," *Review of Economic Studies*, 61, 655 - 680.

[29] Inoue, A., and G. Solon (2010): "Two-Sample Instrumental Variables Estimators," *Review of Economics and Statistics*, 92, 557 - 561.

[30] Little, R. J. A., and D. B. Rubin (2002): *Statistical Analysis with Missing Data*, Second Edition. New York: John Wiley & Sons.

[31] Lusardi, A. (1996): "Permanent Income, Current Income, and Consumption: Evidence from Two Panel Data Sets," *Journal of Business & Economic Statistics*, 14, 81 - 90.

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

[33] Murtazashvili, I., D. Liu, and A. Prokhorov (2015): "Two-Sample Nonparametric Estimation of Intergenerational Income Mobility in the United States and Sweden," *Canadian Journal of Economics*, 48, 1733 - 1761.

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

[35] Phillips, P. C. B., and B. E. Hansen (1990): "Statistical Inference in Instrumental Variables Regression with I(1) Processes," *Review of Economic Studies*, 57, 99 - 125.

[36] Prokhorov, A., and P. Schmidt (2009): "GMM Redundancy Results for General Missing Data Problems," *Journal of Econometrics*, 151, 47 - 55.

[37] Rice, J. (1984): "Bandwidth Choice for Nonparametric Regression," *Annals of Statistics*, 12, 1215 - 1230.

[38] Ridder, G., and R. Moffitt (2007): "The Econometrics of Data Combination," in J. J. Heckman and E. E. Leamer (eds.), *Handbook of Econometrics*, Vol. 6, Part B. Amsterdam: Elsevier, Chapter 75, 5469 - 5547.

[39] Robinson, P. M. (1988): "Root-$N$-Consistent Semiparametric Regression," *Econometrica*, 56, 931 - 954.

[40] Shao, J., and H. Wang (2008): "Confidence Intervals Based on Survey Data with Nearest Neighbor Imputation," *Statistica Sinica*, 18, 281 - 297.

[41] Smith, A. A., Jr. (1993): "Estimating Nonlinear Time-Series Models Using Simulated Vector Autoregressions," *Journal of Applied Econometrics*, 8, S63 - S84.

[42] Smith, J. A., and P. E. Todd (2005): "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators," *Journal of Econometrics*, 125, 305 - 353.

[43] von Neumann, J. (1941): "Distribution of the Ratio of the Mean Square Successive Difference to the Variance," *Annals of Mathematical Statistics*, 12, 367 - 395.

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

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

[46] Yatchew, A. (1997): "An Elementary Estimator of the Partial Linear Model," *Economics Letters*, 57, 135 - 143.

**Table 1:** Monte Carlo Results for Model C

**Panel (a):** $d_3 = 1$

| $(n,m)$ | Estimator | | $\beta_{22}$ | | | | $\gamma_1$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| (1000, 1000) | OLS* | *Mean* | 1.0003 | | | | 0.9970 | | | |
| | | *SD* | 0.0202 | | | | 0.0529 | | | |
| | | *RMSE* | 0.0202 | | | | 0.0529 | | | |
| | | $\overline{SE}$ | 0.0207 | | | | 0.0527 | | | |
| | | *CR* | 96% | | | | 95% | | | |
| | | *K* | *1* | *2* | *4* | *8* | *1* | *2* | *4* | *8* |
| | MSOLS-A | *Mean* | 0.5556 | 0.7148 | 0.8355 | 0.9203 | 1.0513 | 1.0272 | 1.0145 | 1.0091 |
| | | *SD* | 0.0512 | 0.0546 | 0.0582 | 0.0611 | 0.1134 | 0.1052 | 0.1019 | 0.1008 |
| | | *RMSE* | 0.4474 | 0.2903 | 0.1745 | 0.1004 | 0.1245 | 0.1087 | 0.1029 | 0.1012 |
| | MSOLS-B | *Mean* | 0.5556 | 0.7148 | 0.8355 | 0.9203 | 1.0513 | 1.0271 | 1.0145 | 1.0092 |
| | | *SD* | 0.0512 | 0.0546 | 0.0582 | 0.0611 | 0.1135 | 0.1052 | 0.1020 | 0.1008 |
| | | *RMSE* | 0.4474 | 0.2903 | 0.1745 | 0.1005 | 0.1245 | 0.1087 | 0.1030 | 0.1012 |
| | | *K* | *1* | *2* | *4* | *8* | *1* | *2* | *4* | *8* |
| | MSII-A | *Mean* | 1.0251 | 1.0142 | 1.0126 | 1.0221 | 0.9970 | 0.9980 | 0.9993 | 1.0013 |
| | | *SD* | 0.1141 | 0.0906 | 0.0774 | 0.0711 | 0.1231 | 0.1098 | 0.1040 | 0.1019 |
| | | *RMSE* | 0.1168 | 0.0917 | 0.0784 | 0.0744 | 0.1231 | 0.1098 | 0.1040 | 0.1019 |
| | | $\overline{SE}$ | 0.1040 | 0.0740 | 0.0633 | 0.0609 | 0.1199 | 0.1064 | 0.0994 | 0.0961 |
| | | *CR* | 94% | 89% | 88% | 90% | 95% | 94% | 93% | 93% |
| | MSII-B | *Mean* | 1.0251 | 1.0142 | 1.0126 | 1.0221 | 0.9970 | 0.9979 | 0.9993 | 1.0013 |
| | | *SD* | 0.1141 | 0.0906 | 0.0774 | 0.0711 | 0.1231 | 0.1098 | 0.1040 | 0.1019 |
| | | *RMSE* | 0.1168 | 0.0917 | 0.0784 | 0.0744 | 0.1231 | 0.1099 | 0.1040 | 0.1019 |
| | | $\overline{SE}$ | 0.1040 | 0.0740 | 0.0633 | 0.0609 | 0.1199 | 0.1064 | 0.0994 | 0.0962 |
| | | *CR* | 94% | 89% | 88% | 89% | 95% | 94% | 93% | 93% |
| (2000, 2000) | OLS* | *Mean* | 0.9995 | | | | 0.9988 | | | |
| | | *SD* | 0.0145 | | | | 0.0372 | | | |
| | | *RMSE* | 0.0145 | | | | 0.0372 | | | |
| | | $\overline{SE}$ | 0.0147 | | | | 0.0374 | | | |
| | | *CR* | 96% | | | | 94% | | | |
| | | *K* | *1* | *2* | *4* | *8* | *1* | *2* | *4* | *8* |
| | MSOLS-A | *Mean* | 0.5602 | 0.7204 | 0.8406 | 0.9191 | 1.0502 | 1.0263 | 1.0137 | 1.0063 |
| | | *SD* | 0.0359 | 0.0380 | 0.0399 | 0.0416 | 0.0814 | 0.0758 | 0.0729 | 0.0716 |
| | | *RMSE* | 0.4413 | 0.2821 | 0.1643 | 0.0909 | 0.0956 | 0.0803 | 0.0742 | 0.0719 |
| | MSOLS-B | *Mean* | 0.5602 | 0.7204 | 0.8406 | 0.9191 | 1.0502 | 1.0263 | 1.0137 | 1.0063 |
| | | *SD* | 0.0359 | 0.0380 | 0.0399 | 0.0416 | 0.0814 | 0.0758 | 0.0729 | 0.0716 |
| | | *RMSE* | 0.4413 | 0.2821 | 0.1643 | 0.0909 | 0.0956 | 0.0803 | 0.0742 | 0.0719 |
| | | *K* | *1* | *2* | *4* | *8* | *1* | *2* | *4* | *8* |
| | MSII-A | *Mean* | 1.0144 | 1.0100 | 1.0099 | 1.0135 | 0.9961 | 0.9975 | 0.9986 | 0.9985 |
| | | *SD* | 0.0745 | 0.0614 | 0.0519 | 0.0478 | 0.0879 | 0.0790 | 0.0744 | 0.0724 |
| | | *RMSE* | 0.0758 | 0.0622 | 0.0528 | 0.0496 | 0.0880 | 0.0790 | 0.0745 | 0.0724 |
| | | $\overline{SE}$ | 0.0712 | 0.0512 | 0.0436 | 0.0413 | 0.0843 | 0.0750 | 0.0700 | 0.0676 |
| | | *CR* | 94% | 90% | 90% | 91% | 94% | 94% | 94% | 93% |
| | MSII-B | *Mean* | 1.0144 | 1.0100 | 1.0099 | 1.0135 | 0.9961 | 0.9975 | 0.9987 | 0.9985 |
| | | *SD* | 0.0745 | 0.0614 | 0.0519 | 0.0478 | 0.0879 | 0.0790 | 0.0744 | 0.0724 |
| | | *RMSE* | 0.0758 | 0.0622 | 0.0528 | 0.0496 | 0.0880 | 0.0790 | 0.0745 | 0.0724 |
| | | $\overline{SE}$ | 0.0712 | 0.0512 | 0.0436 | 0.0413 | 0.0843 | 0.0750 | 0.0700 | 0.0676 |
| | | *CR* | 94% | 90% | 90% | 91% | 95% | 94% | 94% | 93% |

**Table 1:** *Continued*

**Panel (b):** $d_3 = 2$

| $(n,m)$ | Estimator | | $\beta_{22}$ | | | | $\gamma_1$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| (1000, 1000) | OLS* | $Mean$ | 0.9986 | | | | 0.9977 | | | |
| | | $SD$ | 0.0165 | | | | 0.0571 | | | |
| | | $RMSE$ | 0.0166 | | | | 0.0572 | | | |
| | | $\overline{SE}$ | 0.0164 | | | | 0.0588 | | | |
| | | $CR$ | 95% | | | | 95% | | | |
| | | $K$ | *1* | *2* | *4* | *8* | *1* | *2* | *4* | *8* |
| | MSOLS-A | $Mean$ | 0.4733 | 0.6337 | 0.7856 | 0.9459 | 1.0597 | 1.0291 | 1.0100 | 0.9780 |
| | | $SD$ | 0.0528 | 0.0571 | 0.0662 | 0.0847 | 0.1767 | 0.1725 | 0.1766 | 0.1967 |
| | | $RMSE$ | 0.5294 | 0.3707 | 0.2243 | 0.1005 | 0.1865 | 0.1749 | 0.1769 | 0.1979 |
| | MSOLS-B | $Mean$ | 0.4735 | 0.6340 | 0.7858 | 0.9461 | 1.0123 | 0.9931 | 0.9795 | 0.9457 |
| | | $SD$ | 0.0529 | 0.0573 | 0.0664 | 0.0850 | 0.1782 | 0.1731 | 0.1786 | 0.1991 |
| | | $RMSE$ | 0.5292 | 0.3705 | 0.2242 | 0.1006 | 0.1786 | 0.1732 | 0.1798 | 0.2064 |
| | | $Poly.$ | *(initial)* | *2nd* | *3rd* | *4th* | *(initial)* | *2nd* | *3rd* | *4th* |
| | MSII-FM-A | $Mean$ | 1.1785 | 1.1803 | 1.1805 | 1.1588 | 0.9740 | 0.9723 | 0.9725 | 0.9667 |
| | | $SD$ | 0.1768 | 0.1772 | 0.1773 | 0.1750 | 0.2100 | 0.2123 | 0.2133 | 0.2165 |
| | | $RMSE$ | 0.2512 | 0.2528 | 0.2530 | 0.2363 | 0.2116 | 0.2141 | 0.2150 | 0.2191 |
| | | $\overline{SE}$ | – | 0.1688 | 0.1689 | 0.1679 | – | 0.1869 | 0.1871 | 0.1891 |
| | | $CR$ | – | 87% | 87% | 90% | – | 92% | 92% | 92% |
| | MSII-FM-B | $Mean$ | 1.1791 | 1.1803 | 1.1805 | 1.1587 | 0.9272 | 0.9710 | 0.9714 | 0.9654 |
| | | $SD$ | 0.1770 | 0.1772 | 0.1773 | 0.1750 | 0.2114 | 0.2153 | 0.2160 | 0.2185 |
| | | $RMSE$ | 0.2518 | 0.2528 | 0.2530 | 0.2363 | 0.2236 | 0.2173 | 0.2179 | 0.2212 |
| | | $\overline{SE}$ | – | 0.1688 | 0.1689 | 0.1679 | – | 0.1866 | 0.1868 | 0.1887 |
| | | $CR$ | – | 87% | 87% | 90% | – | 90% | 90% | 91% |
| (2000, 2000) | OLS* | $Mean$ | 0.9997 | | | | 1.0009 | | | |
| | | $SD$ | 0.0116 | | | | 0.0405 | | | |
| | | $RMSE$ | 0.0116 | | | | 0.0406 | | | |
| | | $\overline{SE}$ | 0.0116 | | | | 0.0415 | | | |
| | | $CR$ | 95% | | | | 95% | | | |
| | | $K$ | *1* | *2* | *4* | *8* | *1* | *2* | *4* | *8* |
| | MSOLS-A | $Mean$ | 0.5365 | 0.6953 | 0.8374 | 0.9698 | 1.0429 | 1.0200 | 1.0000 | 0.9811 |
| | | $SD$ | 0.0350 | 0.0372 | 0.0421 | 0.0502 | 0.1095 | 0.1049 | 0.1071 | 0.1160 |
| | | $RMSE$ | 0.4648 | 0.3070 | 0.1680 | 0.0586 | 0.1176 | 0.1068 | 0.1071 | 0.1175 |
| | MSOLS-B | $Mean$ | 0.5365 | 0.6953 | 0.8374 | 0.9699 | 1.0192 | 1.0020 | 0.9844 | 0.9656 |
| | | $SD$ | 0.0351 | 0.0372 | 0.0421 | 0.0503 | 0.1105 | 0.1055 | 0.1077 | 0.1171 |
| | | $RMSE$ | 0.4648 | 0.3069 | 0.1679 | 0.0587 | 0.1121 | 0.1055 | 0.1089 | 0.1220 |
| | | $Poly.$ | *(initial)* | *2nd* | *3rd* | *4th* | *(initial)* | *2nd* | *3rd* | *4th* |
| | MSII-FM-A | $Mean$ | 1.1229 | 1.1242 | 1.1243 | 1.1132 | 0.9787 | 0.9778 | 0.9776 | 0.9752 |
| | | $SD$ | 0.0932 | 0.0933 | 0.0933 | 0.0924 | 0.1250 | 0.1254 | 0.1256 | 0.1274 |
| | | $RMSE$ | 0.1543 | 0.1553 | 0.1554 | 0.1461 | 0.1268 | 0.1274 | 0.1276 | 0.1298 |
| | | $\overline{SE}$ | – | 0.0894 | 0.0894 | 0.0892 | – | 0.1183 | 0.1183 | 0.1198 |
| | | $CR$ | – | 63% | 63% | 69% | – | 87% | 87% | 88% |
| | MSII-FM-B | $Mean$ | 1.1230 | 1.1242 | 1.1243 | 1.1131 | 0.9548 | 0.9774 | 0.9772 | 0.9753 |
| | | $SD$ | 0.0933 | 0.0933 | 0.0933 | 0.0924 | 0.1258 | 0.1267 | 0.1269 | 0.1288 |
| | | $RMSE$ | 0.1544 | 0.1553 | 0.1554 | 0.1461 | 0.1337 | 0.1287 | 0.1289 | 0.1312 |
| | | $\overline{SE}$ | – | 0.0894 | 0.0894 | 0.0892 | – | 0.1182 | 0.1182 | 0.1196 |
| | | $CR$ | – | 63% | 63% | 69% | – | 86% | 86% | 87% |

**Table 1:** *Continued*

**Panel (c):** $d_3 = 3$

| $(n, m)$ | Estimator | | $\beta_{22}$ | | | | $\gamma_1$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| (1000, 1000) | OLS* | *Mean* | 0.9994 | | | | 0.9997 | | | |
| | | *SD* | 0.0135 | | | | 0.0580 | | | |
| | | *RMSE* | 0.0135 | | | | 0.0580 | | | |
| | | $\overline{SE}$ | 0.0139 | | | | 0.0585 | | | |
| | | *CR* | 96% | | | | 96% | | | |
| | | *K* | *1* | *2* | *4* | *8* | *1* | *2* | *4* | *8* |
| | MSOLS-A | *Mean* | 0.2193 | 0.3687 | 0.5758 | 0.8942 | 1.1498 | 1.0658 | 0.9978 | 0.9333 |
| | | *SD* | 0.0748 | 0.0887 | 0.1163 | 0.1528 | 0.3103 | 0.3050 | 0.3246 | 0.3601 |
| | | *RMSE* | 0.7843 | 0.6375 | 0.4398 | 0.1859 | 0.3445 | 0.3121 | 0.3246 | 0.3663 |
| | MSOLS-B | *Mean* | 0.2205 | 0.3703 | 0.5788 | 0.8994 | 0.6439 | 0.6835 | 0.6775 | 0.6299 |
| | | *SD* | 0.0755 | 0.0895 | 0.1168 | 0.1542 | 0.3176 | 0.3146 | 0.3406 | 0.3837 |
| | | *RMSE* | 0.7832 | 0.6360 | 0.4371 | 0.1842 | 0.4772 | 0.4463 | 0.4691 | 0.5332 |
| | | *Poly.* | *(initial)* | *2nd* | *3rd* | *4th* | *(initial)* | *2nd* | *3rd* | *4th* |
| | MSII-FM-A | *Mean* | 1.1151 | 1.0889 | 1.0901 | 1.0651 | 0.9763 | 0.9550 | 0.9534 | 0.9404 |
| | | *SD* | 0.4064 | 0.4009 | 0.4005 | 0.3953 | 0.3698 | 0.3751 | 0.3770 | 0.3712 |
| | | *RMSE* | 0.4224 | 0.4106 | 0.4105 | 0.4007 | 0.3705 | 0.3778 | 0.3799 | 0.3760 |
| | | $\overline{SE}$ | – | 0.3718 | 0.3726 | 0.3669 | – | 0.3288 | 0.3304 | 0.3245 |
| | | *CR* | – | 92% | 92% | 91% | – | 85% | 85% | 86% |
| | MSII-FM-B | *Mean* | 1.1217 | 1.0890 | 1.0903 | 1.0649 | 0.4709 | 0.8358 | 0.8318 | 0.8200 |
| | | *SD* | 0.4099 | 0.4012 | 0.4009 | 0.3954 | 0.3777 | 0.4210 | 0.4249 | 0.4136 |
| | | *RMSE* | 0.4275 | 0.4110 | 0.4110 | 0.4007 | 0.6501 | 0.4519 | 0.4570 | 0.4511 |
| | | $\overline{SE}$ | – | 0.3722 | 0.3730 | 0.3673 | – | 0.3273 | 0.3290 | 0.3200 |
| | | *CR* | – | 92% | 92% | 91% | – | 77% | 77% | 76% |
| (2000, 2000) | OLS* | *Mean* | 1.0002 | | | | 0.9991 | | | |
| | | *SD* | 0.0096 | | | | 0.0419 | | | |
| | | *RMSE* | 0.0096 | | | | 0.0419 | | | |
| | | $\overline{SE}$ | 0.0099 | | | | 0.0415 | | | |
| | | *CR* | 96% | | | | 94% | | | |
| | | *K* | *1* | *2* | *4* | *8* | *1* | *2* | *4* | *8* |
| | MSOLS-A | *Mean* | 0.2994 | 0.4653 | 0.6632 | 0.9149 | 1.1037 | 1.0492 | 0.9910 | 0.9347 |
| | | *SD* | 0.0454 | 0.0541 | 0.0657 | 0.0877 | 0.2007 | 0.1904 | 0.1946 | 0.2169 |
| | | *RMSE* | 0.7021 | 0.5374 | 0.3432 | 0.1222 | 0.2259 | 0.1967 | 0.1948 | 0.2265 |
| | MSOLS-B | *Mean* | 0.3003 | 0.4664 | 0.6648 | 0.9175 | 0.7534 | 0.7911 | 0.7804 | 0.7366 |
| | | *SD* | 0.0459 | 0.0546 | 0.0664 | 0.0886 | 0.2033 | 0.1931 | 0.1977 | 0.2235 |
| | | *RMSE* | 0.7012 | 0.5364 | 0.3417 | 0.1211 | 0.3195 | 0.2845 | 0.2955 | 0.3454 |
| | | *Poly.* | *(initial)* | *2nd* | *3rd* | *4th* | *(initial)* | *2nd* | *3rd* | *4th* |
| | MSII-FM-A | *Mean* | 1.0576 | 1.0477 | 1.0481 | 1.0191 | 0.9800 | 0.9667 | 0.9663 | 0.9617 |
| | | *SD* | 0.1826 | 0.1816 | 0.1818 | 0.1787 | 0.2277 | 0.2305 | 0.2307 | 0.2239 |
| | | *RMSE* | 0.1915 | 0.1877 | 0.1881 | 0.1797 | 0.2286 | 0.2328 | 0.2332 | 0.2271 |
| | | $\overline{SE}$ | – | 0.1800 | 0.1802 | 0.1771 | – | 0.1985 | 0.1989 | 0.1961 |
| | | *CR* | – | 97% | 97% | 96% | – | 90% | 90% | 91% |
| | MSII-FM-B | *Mean* | 1.0608 | 1.0477 | 1.0481 | 1.0189 | 0.6300 | 0.9094 | 0.9094 | 0.9032 |
| | | *SD* | 0.1840 | 0.1817 | 0.1819 | 0.1788 | 0.2328 | 0.2526 | 0.2530 | 0.2438 |
| | | *RMSE* | 0.1938 | 0.1879 | 0.1882 | 0.1798 | 0.4371 | 0.2683 | 0.2687 | 0.2623 |
| | | $\overline{SE}$ | – | 0.1801 | 0.1802 | 0.1772 | – | 0.1996 | 0.2000 | 0.1955 |
| | | *CR* | – | 96% | 96% | 96% | – | 84% | 85% | 85% |

***Note:*** $Mean$ = simulation average of the parameter estimate; $SD$ = simulation average 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:** Estimation Results of Wage Regressions with Ability Measures

**Dependent Variable:** $\log(wage)$

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| **Regressors** | OLS* | MSOLS | MSII-FM | MSOLS | MSII-FM |
| *educ* | 0.0612 | 0.0736 | 0.0690 | 0.0724 | 0.0693 |
| | (0.0054) | (0.0050) | (0.0074) | (0.0050) | (0.0165) |
| *exper* | 0.0787 | 0.0875 | 0.0847 | 0.0876 | 0.0876 |
| | (0.0084) | (0.0082) | (0.0083) | (0.0081) | (0.0082) |
| $exper^2$ | −0.0022 | −0.0023 | −0.0022 | −0.0023 | −0.0023 |
| | (0.0004) | (0.0004) | (0.0004) | (0.0004) | (0.0004) |
| *abil* | 0.0056 | −0.0007 | 0.0016 | 0.0006 | 0.0070 |
| | (0.0013) | (0.0014) | (0.0046) | (0.0049) | (0.0356) |
| *feduc* | −0.0018 | −0.0006 | −0.0007 | −0.0007 | −0.0010 |
| | (0.0031) | (0.0031) | (0.0031) | (0.0032) | (0.0038) |
| *meduc* | 0.0071 | 0.0080 | 0.0073 | 0.0079 | 0.0072 |
| | (0.0037) | (0.0037) | (0.0039) | (0.0037) | (0.0041) |
| *black* | −0.1321 | −0.1664 | −0.1559 | −0.1630 | −0.1607 |
| | (0.0258) | (0.0259) | (0.0331) | (0.0249) | (0.0283) |
| *smsa* | 0.1517 | 0.1602 | 0.1576 | 0.1595 | 0.1612 |
| | (0.0179) | (0.0183) | (0.0186) | (0.0181) | (0.0198) |
| *south* | −0.1111 | −0.1126 | −0.1125 | −0.1125 | −0.1104 |
| | (0.0178) | (0.0179) | (0.0178) | (0.0180) | (0.0216) |
| *intercept* | 4.6861 | 4.6491 | 4.6540 | 4.6425 | 4.6818 |
| | (0.0841) | (0.0861) | (0.1107) | (0.0849) | (0.1945) |
| *abil*? | *kww* | *kww* | *kww* | *g* | *g* |
| Matching? | No | Yes | Yes | Yes | Yes |
| $(n, m)$ | $(2191, -)$ | $(2191, 466)$ | $(2191, 466)$ | $(2191, 589)$ | $(2191, 589)$ |

***Note:*** Numbers in parentheses are standard errors. White's (1980) heteroskedasticity-robust standard errors are calculated for OLS*, whereas 'standard errors' for MSOLS are square-roots of diagonal elements of $\hat{Q}_W^{-1}\hat{\Omega}_{11}\hat{Q}_W^{-1}/n$.