Consistent Estimation of Linear Regression Models Using Matched Data[*]

Masayuki Hirukawa[†] Artem Prokhorov[‡]
Setsunan University University of Sydney

June 2015

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 nonstandard 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 if the number of matching variables is no greater than three. 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 number 23530259).

[†] 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

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

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.) While $d_1 = 0$ is allowed, the intercept term is assumed to be included as a component of X_2 so that $d_2 \geq 2$ must be the case. When $W = (X_1', X_2', Z')' \in \mathbb{R}^d$, where $d := d_1 + d_2 + d_3$, is exogenous and a single random sample of (Y, W) can be obtained, the ordinary least squares (OLS) estimator of $\theta = (\beta_1', \beta_2', \gamma')'$ is consistent and even best linear unbiased when the error term u is conditionally homoskedastic.

In reality, however, we often face the problem that (Y, W) cannot be taken from a single data source. It is not uncommon in labor and public economics to collect the variables necessary for regression analysis from more than one source. Examples include Lusardi (1996), Björklund and Jäntti (1997), Currie and Yelowitz (2000), Dee and Evans (2003), Borjas (2004), Fujii (2008), 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, W), we have the following two overlapping subsets of the data, (Y, X_1, Z) and (X_2, Z). That is, 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. Current 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. 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. (Return 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, W).

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 in the treatment effect literature (see, e.g., Abadie and Imbens, 2006, 2012a). 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 here.

We demonstrate that the OLS estimator of (1) using NNM-based matched samples is inconsistent. The source of the inconsistency is a non-vanishing nonparametric 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 matching variables. In particular, the parametric rate is attained only when $d_3 = 1$, i.e. when there is only one matching variable.

In line with these findings, we propose two semiparametric bias-corrected estimators. The first, one-step estimator is designed exclusively for the case with $d_3 = 1$. On the other hand, the second one 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 effect of the second-order bias asymptotically in a similar manner to the one studied by Abadie and Imbens (2011). It is demonstrated that this estimator attains the parametric convergence rate as long as $d_3 \leq 3$. Both estimators can be 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 regressions involving 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, e.g., Ridder and Moffitt, 2007, p.5480, for a discussion). Available results are limited to the case of matching in average treatment effect (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.

Second, the estimation theory we develop provides guidance on repeated survey sampling when some covariates are found to be completely 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 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 in 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 of the regression model (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 conducts Monte Carlo simulations and examines 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. Gauss codes implementing the estimators are available from the authors upon request.

The paper adopts the following notational conventions: $\|A\| = \{\mathrm{tr}\,(A'A)\}^{1/2}$ is the Euclidean norm of matrix A; $0_{p\times q}$ signifies the $p \times q$ zero matrix, where the subscript may be suppressed if $q = 1$; and the symbol $>$ ($\geq$) applied to matrices means positive (semi-) definiteness.

2 Inconsistency of OLS Estimation Using Matched Samples

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. Specifically, 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 $A_{NE} = \operatorname{diag}\left(A_M^{-1}\right)^{-1}$, where $N = n + m$ and $\bar{Z} = (1/N)\sum_{i=1}^{N} Z_i$. Then, the NNM chooses the index

$$j\left(i\right) := \arg\min_{1\le j\le m} \|Z_j - Z_i\|_A$$

for each Z_i $(1 \le i \le n)$.[1]

Our NNM is based on a single match, where matching is done with replacement, and each element of the matching vector Z is assumed to be continuous. So our setting can be viewed as a foundation for more complicated methods of kernel-based matching (see, e.g., Busso, DiNardo and McCrary, 2014; Abadie and Imbens, 2006). Matching with replacement, allowing each unit to be used as a match more than once, seems to be standard in the econometric literature, while inclusion of discrete matching variables with a finite number of support points does not affect the subsequent asymptotic results. Finally, for simplicity, we ignore ties in the NNM, which happen with probability zero as long as Z is continuous.

Applying the NNM we obtain a matched data set $\mathcal{S} = \left\{\left(Y_i, X_{1i}, X_{2j(i)}, Z_i, Z_{j(i)}\right)\right\}_{i=1}^{n}$, where $X_{2j(i)}$ is the observation paired with $Z_{j(i)}$ in $\mathcal{S}_2$. Throughout, it is assumed

[1] When Z is a scalar, the NNM collapses to $j\left(i\right) = \arg\min_{1\le j\le m}|Z_j - Z_i|$.

that we estimate θ by regressing Y_i on $W_{i,j(i)} := \left(X'_{1i}, X'_{2j(i)}, Z'_{j(i)} \right)'$. Alternatively, we could use Z_i in place of $Z_{j(i)}$. However, both alternatives are first-order asymptotically equivalent and thus we concentrate exclusively on the former case.

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. It will be shown shortly that the MSOLS estimator is inconsistent. Demonstrating this result and deriving the bias-corrected, consistent estimators of θ require the following assumptions.

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, W) with finite fourth-order moments, where the two sample sizes satisfy $n/m \to \kappa \in (0, \infty)$ as $n, m \to \infty$.

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

Assumption 3.

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

(ii) Let $g(Z) := \left[\begin{array}{cc} g_1(Z)' & g_2(Z)' \end{array} \right]' := \left[\begin{array}{cc} E\left(X_1 \middle| Z\right)' & E\left(X_2 \middle| Z\right)' \end{array} \right]'$ and let $\eta := \left[\begin{array}{cc} \eta'_1 & \eta'_2 \end{array} \right]' := \left[\begin{array}{cc} X'_1 - g_1(Z)' & X'_2 - g_2(Z)' \end{array} \right]'$. Then, $\Sigma_1 := E(\eta_1 \eta'_1) > 0$, $\Sigma_2 := E(\eta_2 \eta'_2) \geq 0$ with $\operatorname{rank}(\Sigma_2) = d_2 - 1$, $E(\eta_1 \eta'_2) = 0_{d_1 \times d_2}$, and $g_2(\cdot)$ is first-order Lipschitz continuous on $\mathbb{Z}$.

These regularity conditions are largely inspired by those in the literature on semiparametric, 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). Assumption 1 refers to the same divergence rate of the two sample sizes. This condition can be commonly found in the literature on two-sample regression estimation. Assumption 2 plays a key role in controlling the order of magnitude in the so called *matching discrepancy* (Abadie and Imbens, 2006). The rank condition in Assumption 3(ii) comes from the fact that the row and column of Σ_2 corresponding to the intercept are identically zero. While uncorrelatedness between η_1 and η_2 in this assumption appears to be restrictive, the condition simplifies subsequent analysis considerably. We could have relaxed this assumption at the expense of having to estimate the matrix and using the estimate in our asymptotic derivations.

We start our asymptotic analysis from rewriting Y_i 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

$$\begin{aligned}\lambda_{i,j(i)} &= \lambda\left(Z_i, Z_{j(i)}\right) = \left\{g_2\left(Z_i\right) - g_2\left(Z_{j(i)}\right)\right\}'\beta_2 + \left(Z_i - Z_{j(i)}\right)'\gamma, \text{ and} \\ \epsilon_{i,j(i)} &= u_i + \left(\eta_{2i} - \eta_{2j(i)}\right)'\beta_2.\end{aligned}$$

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 semiparametric part $\lambda_{i,j(i)}$ generates a second-order bias that will be discussed shortly and so 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 a rate of convergence.

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

The theorem states that MSOLS is inconsistent in general. The term Σ 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. We can also find from a straightforward calculation that the OLS estimator of β_2, which is the coefficient vector on the matched regressor X_2, is biased toward zero in the limit.

A quick inspection also reveals that $\hat{\theta}_{OLS}$ 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 *nonlinear* deterministic function of Z. However, even if either one of the above conditions were true, $\hat{\theta}_{OLS}$ might not be $\sqrt{n}$-consistent. Clearly, there exists a *curse of dimensionality* with respect to the matching variable Z. The proof of Theorem 1, which is provided in the Appendix, suggests that when $d_3 = 1$, $\sqrt{n}\left(\hat{\theta}_{OLS} - Q_W^{-1}P_W\theta\right)$ has a normal limit. When $d_3 = 2$, $\hat{\theta}_{OLS}$ is still $\sqrt{n}$-consistent, but we could only demonstrate asymptotic normality of $\hat{\theta}_{OLS}$ after subtracting the bias term due to the matching discrepancy, i.e. the best we can do in this case is to apply the central limit theorem (CLT) to $\sqrt{n}\left(\hat{\theta}_{OLS} - Q_W^{-1}P_W\theta - B_{OLS2}\right)$, where

$$B_{OLS2} := \hat{Q}_W^{-1}B_{R_W2} := \hat{Q}_W^{-1}\frac{1}{n}\sum_{i=1}^{n} W_{i,j(i)}\lambda_{i,j(i)},$$

and its numerator $B_{R_W2} = O_p\left(n^{-1/d_3}\right)$ is the source of the second-order bias due to the matching discrepancy. These limiting distributions would reduce to the usual one of OLS if a complete data set of (Y, W) were available. When $d_3 \geq 3$, the convergence rate of $\hat{\theta}_{OLS}$ is slower than the parametric one, and it becomes slower as d_3 increases.

3 Bias-Corrected Estimation of the Parameter

3.1 Estimation Strategies for a Single Matching Variable

This section develops bias-corrected, $\sqrt{n}$-consistent estimation of θ. Taking into account the order of magnitude of the bias term, we consider the estimation problem for the cases with $d_3 = 1$ and $d_3 \geq 2$ separately. Clearly the second-order bias must be removed explicitly in the latter case, whereas it is not required in the former case.

Our analysis starts with the case of a single matching variable. As suggested by the proof of Theorem 1 in the Appendix, inconsistency of MSOLS comes from the fact that $\hat{Q}_W \xrightarrow{p} Q_W$ whereas $\hat{R}_W \xrightarrow{p} P_W \theta = (Q_W - \Sigma)\,\theta$. Therefore, the non-vanishing bias in MSOLS can be eliminated if either

(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 $\Sigma\theta$ 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 Σ_2 (covariance matrix of the nonparametric regression error η_2) is required. Moreover, implementing (1b) requires a two-step estimation with an initial consistent plug-in estimate of θ. However, as we show next if the plug-in we use is the estimator using strategy (1a) than the two step estimation will produce a numerially identical result. So there is no point in pursuing strategy (1b) separately.

Suppose we have an initial estimator $\hat{\theta}$ using strategy (1a), i.e. $\hat{\theta} = \hat{P}_W^{-1} \hat{R}_W$, where $\hat{P}_W \xrightarrow{p} P_W$. And suppose we want to employ strategy (1b), that is, we wish to estimate a second-step estimator $\tilde{\theta}$ as follows

$$\tilde{\theta} := \hat{Q}_W^{-1} \left(\hat{R}_W + \hat{\Sigma}\hat{\theta} \right),$$

where $\hat{\Sigma}$ is a consistent estimate of Σ. Then, we can write

$$\tilde{\theta} = \hat{Q}_W^{-1} \left(I_d + \hat{\Sigma} \hat{P}_W^{-1} \right) \hat{R}_W.$$

On the other hand, post-multiplying both sides of $\hat{P}_W + \hat{\Sigma} = \hat{Q}_W$ by $\hat{P}_W^{-1}$ yields $I_d + \hat{\Sigma} \hat{P}_W^{-1} = \hat{Q}_W \hat{P}_W^{-1}$. Substituting this into the right-hand side of (3) immediately establishes that $\tilde{\theta} = \hat{\theta}$. Therefore, strategy (1b) is interesting only in terms of using alternative consistent estimators of θ (other than $\hat{\theta}$).

3.2 One-Step Bias Correction

In this section we obtain the consistent estimator of θ based on strategy (1a). We adopt the idea of 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. $b(\theta) = Q_W^{-1} P_W \theta$.[2] Provided that P_W^{-1} exists, 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 $Q_W \hat{\theta}_{OLS}$. Accordingly, we call this estimation method *the matched-sample indirect inference* (MSII) estimation. We formally define the MSII estimator as

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

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 $\Sigma = \operatorname{diag}\{0_{d_1 \times d_1}, \Sigma_2, 0_{d_3 \times d_3}\}$, 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. $(1) = \arg\min_{1 \leq j \leq m} Z_{j1}$.

[2] 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.

2. For $j = 2, \ldots, m$, choose $(j) = \arg\min_{j \neq (1), \ldots, (j-1)} \left\| Z_j - Z_{(j-1)} \right\|$.[3]

Given the reordered sample $\mathcal{S}_2 = \left\{ X_{2(j)}, Z_{(j)} \right\}_{j=1}^{m}$, Σ_2 can be consistently estimated by

$$\hat{\Sigma}_2 = \frac{1}{2(m-1)} \sum_{j=2}^{m} \Delta X_{2(j)} \Delta X'_{2(j)}, \tag{3}$$

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. It follows from Lemma of Yatchew (1997) that as long as Assumptions 1 and 2 hold and $d_3 \leq 3$, we have $\hat{\Sigma}_2 = \Sigma_2 + O_p\left(m^{-1/2}\right)$. In the end, the estimator of P_W is given by

$$\hat{P}_W := \hat{Q}_W - \hat{\Sigma} = \hat{Q}_W - \operatorname{diag}\left\{ 0_{d_1 \times d_1}, \hat{\Sigma}_2, 0_{d_3 \times d_3} \right\}.$$

The next theorem establishes $\sqrt{n}$-consistency of $\hat{\theta}_{II}$ and derives its limiting distribution.

Theorem 2. *If Assumptions 1-3 hold, $d_3 = 1$ and P_W^{-1} exists, then $\hat{\theta}_{II} \xrightarrow{p} \theta$ and $\sqrt{n}\left(\hat{\theta}_{II} - \theta\right) \xrightarrow{d} N\left(0, V_W\right)$, where $V_W := P_W^{-1} \Omega_W P_W^{-1}$,*

$$\begin{aligned}
\Omega_W &= \Omega_{W,11} + \sqrt{\kappa}\left(\Omega_{W,12} + \Omega'_{W,12}\right) + \kappa \Omega_{W,22}, \\
\Omega_{W,11} &= E\left(\phi_{i,j(i)} \phi'_{i,j(i)}\right), \\
\Omega_{W,12} &= \left[\begin{array}{ccc} 0_{d \times d_1} & E\left(\phi_{i,j(i)} \omega'_{j(i)}\right) & 0_{d \times d_3} \end{array}\right], \\
\Omega_{W,22} &= \operatorname{diag}\left\{ 0_{d_1 \times d_1}, E\left(\psi_j \psi'_{j-1}\right) + E\left(\psi_j \psi'_j\right) + E\left(\psi_j \psi'_{j+1}\right), 0_{d_3 \times d_3} \right\}, \\
\phi_{i,j(i)} &= W_{i,j(i)} \epsilon_{i,j(i)} + \Sigma \theta, \\
\omega_{j(i)} &= \left(\eta_{2j(i)} \eta'_{2j(i)} - \Sigma_2\right) \beta_2, \ \text{and} \\
\psi_j &= \left(\frac{\Delta \eta_{2j} \Delta \eta'_{2j}}{2} - \Sigma_2\right) \beta_2.
\end{aligned}$$

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

Remark 1. There are two important observations here. First, some nonlinearity in all elements of $g_2(\cdot)$ other than the intercept is necessary for a non-singular P_W and thus for identification of θ. To see why, observe that the lower-right block of P_W collapses to

$$\left[\begin{array}{cc} E\left(X_{2j(i)}X'_{2j(i)}\right)-\Sigma_2 & E\left(X_{2j(i)}Z'_{j(i)}\right) \\ E\left(Z_{j(i)}X'_{2j(i)}\right) & E\left(Z_{j(i)}Z'_{j(i)}\right) \end{array}\right]=\left[\begin{array}{cc} E\left\{g_2(Z)\,g_2(Z)'\right\} & E\left\{g_2(Z)\,Z'\right\} \\ E\left\{Zg_2(Z)'\right\} & E\left(ZZ'\right) \end{array}\right],$$

which becomes singular if one or more elements of $g_2(\cdot)$ other than the intercept are linear. Second, Theorem 2 suggests that in the special case where $n=o(m)$, Ω_W reduces to $\Omega_{W,11}=Var\left(\phi_{i,j(i)}\right)$.

3.3 Consistent Estimation for Two or More Matching Variables

While MSII yields a consistent estimate of θ, its apparent deficiency is that it can attain the parametric rate of convergence only for the case with a single matching variable. The curse of dimensionality in the NNM can be 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 done on a single variable.

To overcome the curse of dimensionality, we should find a way of eliminating the second-order bias, or equivalently, the effect of $\lambda_{i,j(i)}$ asymptotically from (2). There are two possible strategies, namely,

(2a) taking the first-order difference of (2); and

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

Yatchew (1997) advocates (2a) in semiparametric regression estimation, whereas Robinson (1988) and Abadie and Imbens (2011) adopt a similar strategy to (2b) in semiparametric regression and ATE estimations, respectively. In our settings, we have found that the strategy (2a) has a few disadvantages. First, differencing (2) makes γ as well as the intercept 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 (2b).

Estimating $\lambda_{i,j(i)}$ requires consistent estimates of θ and $g_2(\cdot)$. For the former, we 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 $|\upsilon| = \sum_{\ell=1}^{d_3} \upsilon_\ell$. Also denote $z^\upsilon = \prod_{\ell=1}^{d_3} z_\ell^{\upsilon_\ell}$, where z_ℓ is the ℓth element of z. Consider a series $\{\upsilon(K)\}_{K=1}^{\infty}$ containing distinct vectors such that $|\upsilon(K)|$ is nondecreasing. Let $p_K(z) = z^{\upsilon(K)}$ and $p^K(z) = (p_1(z), \ldots, p_K(z))'$. Then, a nonparametric series estimator of the regression function $g_{2l}(z)$, $l = 1, \ldots, d_2$, is given by

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

where $X_{2l,j}$ is the lth element of X_{2j} in $\mathcal{S}_2$, and $(\cdot)^-$ signifies the generalized inverse.

The entire estimation procedure based on the strategy (2b) can be summarized in the following two steps:

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

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

$$\hat{\lambda}_{i,j(i)} = \left\{\hat{g}_2(Z_i) - \hat{g}_2\left(Z_{j(i)}\right)\right\}' \hat{\beta}_{II,2}^{(1)} + \left(Z_i - Z_{j(i)}\right)' \hat{\gamma}_{II}^{(1)}$$

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

The idea behind the above procedure is as follows. The initial MSII estimate $\hat{\theta}_{II}^{(1)}$ is consistent but inefficient in general, because its convergence rate is n^{1/d_3} for $d_3 \geq 2$.

Then, in the second step, we (asymptotically) eliminate the source of the second-order bias by subtracting $\hat{\lambda}_{i,j(i)}$ from the dependent variable and reestimate θ by MSII using the bias-adjusted data to obtain a $\sqrt{n}$-consistent estimate. 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 4. $\mathbb{Z}$ is a Cartesian product of compact intervals.

Assumption 5. $K(m) = O(m^{\nu})$ for some $\nu \in (0, \min\{2/(4d_3+3), 2/(4d_3^2 - d_3)\})$.

Assumption 6. There is a constant C such that for each multi-index υ, the υth partial derivative of $g_2(z)$ exists and its norm is bounded by $C^{|\upsilon|}$.

Below we present the asymptotic theory for $\hat{\theta}_{II-FM}$. Because of Lemma A2 and the asymptotic properties of $\hat{\Sigma}_2$, the MSII-FM estimator is $\sqrt{n}$-consistent only for $d_3 = 2, 3$. It is also worth remarking that the asymptotic variance of $\sqrt{n}\left(\hat{\theta}_{II-FM} - \theta\right)$ takes the same form as the one for $\sqrt{n}\left(\hat{\theta}_{II} - \theta - \hat{Q}_W^{-1} B_{R_W 2}\right)$, i.e. the FM procedure removes the bias without inflating the variance.

Theorem 3. *If Assumptions 1-6 hold, $d_3 = 2, 3$ and P_W^{-1} exists, then $\hat{\theta}_{II-FM} \xrightarrow{p} \theta$ and $\sqrt{n}\left(\hat{\theta}_{II-FM} - \theta\right) \xrightarrow{d} N(0, V_W)$, where V_W is the same as in Theorem 2.*

Remark 2. An important practical question on implementing MSII-FM is how to choose the order of polynomials in the power series approximation. Because

$$\min\left\{\frac{2}{4d_3+3}, \frac{2}{4d_3^2 - d_3}\right\} = \left\{\begin{array}{ll} 1/7 & \text{for } d_3 = 2 \\ 2/33 & \text{for } d_3 = 3 \end{array}\right.,$$

we first set the number of terms $K(m) = O(m^{\bar{\nu}})$ with $\bar{\nu} = 1/8, 3/50$ for $d_3 = 2, 3$, respectively. Also observe that the total number of terms in the qth-order polynomial constructed from d_3 variables can be expressed as $(1/d_3!)\prod_{k=1}^{d_3}(q+k)$. Because this number must diverge at the rate $O(m^{\bar{\nu}})$, we choose the order q^* so that

$$\frac{1}{d_3!}\prod_{k=1}^{d_3}(q^*+k) \leq 4m^{1/8} < \frac{1}{d_3!}\prod_{k=1}^{d_3}(q^*+1+k) \text{ for } d_3 = 2, \text{ and}$$
$$\frac{1}{d_3!}\prod_{k=1}^{d_3}(q^*+k) \leq 8m^{3/50} < \frac{1}{d_3!}\prod_{k=1}^{d_3}(q^*+1+k) \text{ for } d_3 = 3.$$

Under such q^*, $K^* = (1/d_3!)\prod_{k=1}^{d_3}(q^*+k)$ indeed satisfies $K^* = O(m^{\bar{\nu}})$.

3.4 Covariance Estimation

We conclude this section by discussing covariance estimation, which is essential for inference. Theorems 2 and 3 indicate that asymptotic variances of $\sqrt{n}\left(\hat{\theta}_{II} - \theta\right)$ and $\sqrt{n}\left(\hat{\theta}_{II-FM} - \theta\right)$ take the same form. Hence, the problem is boiled down to estimating $V_W = P_W^{-1}\Omega_W P_W^{-1}$ consistently. Because $\hat{P}_W$ is consistent for P_W, it suffices to propose a consistent estimator of $\Omega_W = \Omega_{W,11} + \sqrt{\kappa}\left(\Omega_{W,12} + \Omega_{W,12}'\right) + \kappa\Omega_{W,22}$. Below we assume $d_3 = 1$ so that the MSII estimator is employed as a consistent estimator for θ; it is easy to see that the result equally holds after it is replaced by the MSII-FM estimator for $d_3 = 2, 3$.

Let the MSII residual be $\hat{\epsilon}_{i,j(i)} := Y_i - W_{i,j(i)}'\hat{\theta}_{II}$. Also denote the MSII estimator of β_2 as $\hat{\beta}_{2,II}$. Moreover, define $\hat{\phi}_{i,j(i)} := W_{i,j(i)}\hat{\epsilon}_{i,j(i)} + \hat{\Sigma}\hat{\theta}_{II}$ and $\hat{\psi}_j := \left\{\left(\Delta X_{2j}\Delta X_{2j}'/2\right) - \hat{\Sigma}_2\right\}\hat{\beta}_{2,II}$. Then, a natural estimator of $\Omega_{W,11}$ is

$$\hat{\Omega}_{W,11} = \frac{1}{n}\sum_{i=1}^{n}\hat{\phi}_{i,j(i)}\hat{\phi}_{i,j(i)}'.$$

In addition, $E\left(\psi_j\psi'_{j-k}\right)$, $k=\pm1,0$, can be consistently estimated as

$$\begin{aligned}\hat{E}\left(\psi_j\psi'_{j-k}\right) &= \frac{1}{m-1}\sum_{j=\max\{2,2+k\}}^{\min\{m,m+k\}}\hat{\psi}_j\hat{\psi}'_{j-k}\\ &= \frac{1}{m-1}\sum_{j=\max\{2,2+k\}}^{\min\{m,m+k\}}\left(\frac{\Delta\eta_{2j}\Delta\eta'_{2j}}{2}-\hat{\Sigma}_2\right)\hat{\beta}_{2,II}\hat{\beta}'_{2,II}\left(\frac{\Delta\eta_{2j-k}\Delta\eta'_{2j-k}}{2}-\hat{\Sigma}_2\right)\\ &\quad+o_p\left(m^{-1/2}\right),\end{aligned}$$

where the second equality holds for $d_3\le3$. Hence, a natural estimator of $\Omega_{W,22}$ is given by

$$\hat{\Omega}_{W,22}=\operatorname{diag}\left\{0_{d_1\times d_1},\hat{E}\left(\psi_j\psi'_{j-1}\right)+\hat{E}\left(\psi_j\psi'_j\right)+\hat{E}\left(\psi_j\psi'_{j+1}\right),0_{d_3\times d_3}\right\}.$$

Moreover, it follows from the proof of Theorem 2 that $E\left(\phi_{i,j(i)}\omega'_{j(i)}\right)=2E\left(\phi_{i,j(i)}\psi'_{j(i)}\right)$. Then, a consistent estimator of $E\left(\phi_{i,j(i)}\omega'_{j(i)}\right)$ takes the form of

$$\begin{aligned}\hat{E}\left(\phi_{i,j(i)}\omega'_{j(i)}\right) &= 2\hat{E}\left(\phi_{i,j(i)}\psi'_{j(i)}\right)\\ &= \frac{2}{n-1}\sum_{i=2}^{n}\hat{\phi}_{i,j(i)}\hat{\psi}'_{j(i)}\\ &= \frac{2}{n-1}\sum_{i=2}^{n}\left(W_{i,j(i)}\hat{\epsilon}_{i,j(i)}+\hat{\Sigma}\hat{\theta}_{II}\right)\hat{\beta}'_{2,II}\left(\frac{\Delta\eta_{2j(i)}\Delta\eta'_{2j(i)}}{2}-\hat{\Sigma}_2\right)+o_p\left(n^{-1/2}\right)\end{aligned}$$

as before. Therefore, $\Omega_{W,12}$ can be estimated as

$$\hat{\Omega}_{W,12}=\left[\begin{array}{ccc}0_{d\times d_1} & \hat{E}\left(\phi_{i,j(i)}\omega'_{j(i)}\right) & 0_{d\times d_3}\end{array}\right].$$

Since $n/m=\kappa+o(1)$, we finally obtain an estimator of V_W as

$$\hat{V}_W=\hat{P}_W^{-1}\hat{\Omega}_W\hat{P}_W^{-1}=\hat{P}_W^{-1}\left\{\hat{\Omega}_{W,11}+\sqrt{\frac{n}{m}}\left(\hat{\Omega}_{W,12}+\hat{\Omega}'_{W,12}\right)+\left(\frac{n}{m}\right)\hat{\Omega}_{W,22}\right\}\hat{P}_W^{-1}.$$

The following proposition refers to consistency of the covariance estimator. This proposition can be established by the techniques employed for the proofs of Theorems 1-3, and thus the proof is omitted.

Proposition 1. *If Assumptions 1-3 (Assumptions 1-6) hold for* $d_3=1$ *(*$d_3=2,3$*) and* P_W^{-1} *exists, then* $\hat{V}_W\xrightarrow{p}V_W$.

4 Finite-Sample Performance

4.1 Case 1: $d_3 = 1$

4.1.1 Monte Carlo Setup

We conduct Monte Carlo simulations to examine finite-sample properties of the MSII estimation by changing the number of matching variables. We start from the case with $d_3 = 1$. Consider the regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \gamma_1 Z + u, \tag{4}$$

where the 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 assumed to be 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, we draw $Z \overset{iid}{\sim} U[-2, 2]$. Second, X_1 and X_2 are generated by $1 + Z + \eta_1$ and $g_2(Z) + \eta_2$, respectively, where $\eta_1, \eta_2 \overset{iid}{\sim} N(0, 1)$ and $g_2(z)$ takes one of the following four functional forms:

$$\begin{aligned}
&\text{A} : g(z) = z^2 \\
&\text{B} : g(z) = \exp(z) \\
&\text{C} : g(z) = z + (5/\tau)\,\phi(z/\tau), \tau = 0.9 \\
&\text{D} : g(z) = z + (5/\tau)\,\phi(z/\tau), \tau = 0.3
\end{aligned}$$

with $\phi(\cdot)$ being the pdf of $N(0, 1)$. Models A and B are convex and monotone increasing, respectively. While these functions are purely nonlinear, the remaining two models may be thought of as 'intermediate' cases between linear and nonlinear functions in that each of these is constructed as a linear combination of linear and nonlinear functions. These models are inspired by the Monte Carlo design of Horowitz and Spokoiny (2001). Third, using (X_1, X_2, Z), we generate Y by setting all coefficients in (4) equal to 1 with $u \overset{iid}{\sim} N(0, 1)$. This 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 $\mathcal{S} = \left\{\left(Y_i, X_{1i}, X_{2j(i)}, Z_i, Z_{j(i)}\right)\right\}_{i=1}^n$ is constructed via the NNM with respect to Z.

With regards to sample sizes, for each of $n \in \{500, 1000, 2000\}$, m is chosen as one of $m \in \{n/2, n, 2n\}$ so that the values of κ are $\kappa = 2, 1$ and $1/2$, respectively. For each combination of sample sizes (n, m) and the functional form of $g_2(z)$, we generate 1000 Monte Carlo samples. The following three estimators are examined: (i) the infeasible OLS (OLS*) estimator using the complete sample $\mathcal{S}^*$; (ii) the MSOLS estimator using the matched sample $\mathcal{S}$; and (iii) the MSII estimator using the matched sample $\mathcal{S}$. For each estimator, averages, standard deviations (in parentheses) and root-mean squared errors (RMSEs) [in brackets] of estimates of β_2 and γ_1 over 1000 replications are reported.

TABLE 1 ABOUT HERE

4.1.2 Results

In Table 1, we report the results for Models C and D.[4] 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. As expected, the standard deviations tend to be smaller as n increases.

However, OLS* is an infeasible, oracle estimator. Instead, we should focus on the 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. Table 1 illustrates that MSOLS is inconsistent in that its bias does not vanish with the sample size. In addition, the MSOLS estimate of β_2 is biased toward zero, as predicted. However, the magnitude of the bias depends on the specification of $g_2(\cdot)$. It can be also seen that their standard deviations shrink with n, as Theorem 1 suggests.

Now we turn to MSII. At a glance, we can observe that our bias correction works remarkably well. Simulation averages of these estimators get closer to the truth as

[4] Comprehensive simulation results covering Cases 1 and 2 are available as a supplement to this paper, posted on the authors' web pages.

n increases, which confirms their consistency. At a closer look, we find that the performance of MSII depends on the degree of nonlinearity in $g_2(\cdot)$. For Model C, which is somewhat close to a linear function, the bias and standard deviation tend to be large. In contrast, for Model D, which is highly nonlinear, they are quite small even when $n = 500$. It can be also seen that standard deviations for each n tend to be smaller as κ decreases, as Theorem 1 suggests.

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. Also observe that while standard deviations of MSOLS are also 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 $Q_W^{-1}\Omega_{W,11}Q_W^{-1}$, which tends to be smaller (in the matrix sense) than V_W.

4.2 Case 2: $d_3 = 2$

4.2.1 Monte Carlo Setup

Now we increase the number of matching variables to two. Accordingly, the following regression is examined:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \gamma_1 Z_1 + \gamma_2 Z_2 + u, \tag{5}$$

where $\mathcal{S}_1 = \{(Y_i, X_{1i}, Z_{1i}, Z_{2i})\}_{i=1}^n$ and $\mathcal{S}_2 = \{(X_{2j}, Z_{1j}, Z_{2j})\}_{j=1}^m$ are assumed to be observable. There is a slight modification of the data generation, with all other setup details left unchanged. To generate $Z = (Z_1, Z_2)'$, we first draw

$$Z^* \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right).$$

Then, $Z^* = (Z_1^*, Z_2^*)'$ is transformed into $Z = (Z_1, Z_2)' = (2\Phi(Z_1^*) - 1, 4\Phi(Z_2^*) - 2)'$, where $\Phi(\cdot)$ is the cdf of $N(0,1)$. Observe that $Z_1 \sim U[-1,1]$, $Z_2 \sim U[-2,2]$ and these are correlated. Subsequently, X_1 and X_2 are generated by $1+Z_1+Z_2+\eta_1$ and $g_{21}(Z_1)+g_{22}(Z_2)+\eta_2$, respectively, where $\eta_1, \eta_2 \overset{iid}{\sim} N(0,1)$, $g_{21}(z)$ is either Model A or B, and $g_{22}(z)$ is either Model C or D. Finally, Y is generated by setting all coefficients in (5) equal to 1 with $u \overset{iid}{\sim} N(0,1)$. Besides, the NNM is based on the Mahalanobis metric.

We compare four estimators, namely, OLS*, MSOLS, MSII, and the MSII-FM estimator using the matched sample $\mathcal{S}$. Table 2 reports the results for Models AC and AD (see the supplement for other models). The order of polynomials in the power series approximation for MSII-FM is determined by the method in Remark 2. Specifically, $q^*=2$ (or $K^*=6$) if $m=250,500,1000$ and $q^*=3$ (or $K^*=10$) if $m=2000,4000$. Finally, for each estimator, averages, standard deviations (in parentheses) and root-mean squared errors (RMSEs) [in brackets] of estimates of β_2 and γ_2 over 1000 replications are reported.

TABLE 2 ABOUT HERE

4.2.2 Results

As shown in Table 2, even after the number of matching variables increases, the general tendency remains unchanged. The two MSII estimators successfully correct the bias generated by MSOLS, at the expense of precision in estimation. Furthermore, asymptotic results suggest that MSII-FM is more efficient than MSII because the effect of the second-order bias in the latter is eliminated in the former, although they are both $\sqrt{n}$-consistent. While the efficiency gain in MSII-FM is confirmed for Model AC, it is not obvious for Model AD.

To sum up, simulation results indicate that the bias-corrected estimation proposed in this paper works remarkably well. Simulation averages of MSII for Case 1 and MSII-FM for Case 2 tend to be closer to the truths as n increases, regardless of the specification in $g_2(\cdot)$. While the degree of variability is also subject to the functional form of $g_2(\cdot)$, it is likely to be fairly small for large samples even when nonlinearity in $g_2(\cdot)$ is not very pronounced.

5 An Empirical Application: Returns to Schooling

We now apply the proposed estimation method 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

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

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 *smsa* and *south* are indicator variables that take one if the individual 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 1846 white-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 based on the Mahalanobis metric of five matching variables $(educ, feduc, meduc, smsa, south)$, where the first three variables are treated as continuous. 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, 396 and 589 distinct combinations of matching variables remain in white-male samples of `wage2` and `htv`, respectively. After constructing a matched sample, we estimate (6) by MSOLS and MSII-FM, where the order of polynomials in the power series approximation for MSII-FM is again determined by the method in Remark 2. Specifically, $q^* = 2$ (or $K^* = 10$) is chosen for three continuous matching variables in $\mathcal{S}_2$.

TABLE 3 ABOUT HERE

Table 3 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 $\hat{Q}_W^{-1}\hat{\Omega}_{W,11}\hat{Q}_W^{-1}/n$, where $n = 1846$. 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 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. While MSII-FM estimates are closer to OLS* ones, the signs of the coefficient estimates by both estimation methods remain unchanged from OLS*. However, due to large standard errors, the MSII-FM estimate of the coefficient on *abil* becomes insignificant.

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*. Failure to correct for matching results in overestimates of the return to education. It is also worth remarking that the MSOLS estimate of the coefficient on *abil* turns negative, whereas the one from MSII-FM remains positive (but insignificant).

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 required. It has also been shown that the convergence rate to the probability limit of the OLS depends on the number of matching variables. Two versions of bias-corrected estimators have been proposed – each can be interpreted as a variant of II estimators. The MSII estimator attains the parametric convergence rate for the cases with only one matching variable, whereas the MSII-FM estimator achieves the parametric convergence rate in cases with no more than three matching variables.

Several research extensions would be fruitful. First, we may adopt propensity score matching as a means of dimension reduction using multiple matching variables. In a closely related paper, Abadie and Imbens (2012b) deliver asymptotic properties of the matching estimators of average treatment effects using an estimated propensity score as a plug-in. It is 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. Third, 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 closest matching discrepancy $U = U(z)$ is defined as $U := Z_{j(z)} - z$ if $Z_{j(z)}$ is the 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 with the number of closest neighbors M set equal to 1 (see Lemma 2 of Abadie and Imbens, 2006).

Lemma A1. *Under Assumptions 1-2, all the moments of $n^{1/d_3} \|U\|$ are uniformly bounded in n and $z \in \mathbb{Z}$.*

A.2 Proof of Theorem 1

It is easy to see from (2) that $\hat{R}_W := \hat{Q}_W\theta + B_{R_W1} + B_{R_W2} + E_{R_W}$, where

$$\begin{aligned} B_{R_W1} &= E\left(W_{i,j(i)}\epsilon_{i,j(i)}\right), \\ B_{R_W2} &= \frac{1}{n}\sum_{i=1}^{n} W_{i,j(i)}\lambda_{i,j(i)}, \text{ and} \\ E_{R_W} &= \frac{1}{n}\sum_{i=1}^{n}\left\{W_{i,j(i)}\epsilon_{i,j(i)} - E\left(W_{i,j(i)}\epsilon_{i,j(i)}\right)\right\}. \end{aligned}$$

It follows that $\hat{\theta}_{OLS} := \theta + B_{OLS1} + B_{OLS2} + E_{OLS}$, where $B_{OLS1} = \hat{Q}_W^{-1}B_{R_W1}$, $B_{OLS2} = \hat{Q}_W^{-1}B_{R_W2}$ 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 $E\left(X_{1i}\eta'_{2_i}\right) = E\left\{g_1(Z)\eta'_2\right\} + E\left(\eta_1\eta'_2\right) = 0_{d_1\times d_2}$, $E\left(X_{2j(i)}\eta'_{2_{j(i)}}\right) = \Sigma_2$, and that the ith and $j(i)$th observations are independent. Then, $B_{R_W1} = \left[\begin{array}{ccc} 0_{1\times d_1} & (-\Sigma_2\beta_2)' & 0_{1\times d_3}\end{array}\right]' = -\operatorname{diag}\left\{0_{d_1\times d_1}, \Sigma_2, 0_{d_3\times d_3}\right\}\theta := -\Sigma\theta$. Because $\hat{Q}_W = Q_W + O_p\left(n^{-1/2}\right)$, we obtain $B_{OLS1} = -Q_W^{-1}\Sigma\theta + O_p\left(n^{-1/2}\right)$. Next, Lemma A1 implies that $\max_{1\le i\le n}\left\|Z_{j(i)} - Z_i\right\| = O_p\left(n^{-1/d_3}\right)$. Then, by the Cauchy-Schwarz inequality and Lipschitz continuity of g_2, $\|B_{R_W2}\|$ is bounded by $O_p\left(n^{-1/d_3}\right)$. Hence, $B_{OLS2} = O_p\left(n^{-1/d_3}\right)$. 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, $\hat{\theta}_{OLS} = \theta - Q_W^{-1}\Sigma\theta + O_p\left(n^{-1/d_3}\right) + O_p\left(n^{-1/2}\right) = Q_W^{-1}P_W\theta + O_p\left(n^{-\min\{1/2,1/d_3\}}\right)$ by denoting $P_W := Q_W - \Sigma$. ■

A.3 Proof of Theorem 2

Consistency of $\hat{\theta}_{II}$ can be established in line with the proof of Theorem 1. To derive the asymptotic distribution of $\sqrt{n}\left(\hat{\theta}_{II} - \theta\right)$, we first obtain

$$\hat{R}_W = \left(\hat{Q}_W - \Sigma\right)\theta + B_{R_W2} + E_{R_W} = \hat{P}_W\theta + \left(\hat{\Sigma} - \Sigma\right)\theta + B_{R_W2} + E_{R_W} \qquad \text{(A1)}$$

by the proof of Theorem 1. Substituting this into $\sqrt{n}\left(\hat{\theta}_{II} - \theta\right)$ yields

$$\sqrt{n}\left(\hat{\theta}_{II} - \theta\right) = \hat{P}_W^{-1}\left\{\sqrt{n}\left(\hat{\Sigma} - \Sigma\right)\theta + \sqrt{n}B_{R_W2} + \sqrt{n}E_{R_W}\right\}. \qquad \text{(A2)}$$

When $d_3 = 1$, $B_{R_W 2} = O_p\left(n^{-1}\right)$ and thus $\sqrt{n} B_{R_W 2} = O_p\left(n^{-1/2}\right) = o_p(1)$. Using $\hat{P}_W^{-1} = P_W^{-1} + o_p(1)$, $\sqrt{n}\left(\hat{\Sigma} - \Sigma\right)\theta = \sqrt{n}\left(\hat{\Sigma}_2 - \Sigma_2\right)\beta_2$, and $n/m = \kappa + o(1)$, we finally have

$$\sqrt{n}\left(\hat{\theta}_{II} - \theta\right) = P_W^{-1}\left\{\sqrt{n} E_{R_W} + \sqrt{\kappa}\sqrt{m}\left(\hat{\Sigma}_2 - \Sigma_2\right)\beta_2\right\} + o_p(1),$$

where, by CLT, each of $\sqrt{n} E_{R_W}$ and $\sqrt{m}\left(\hat{\Sigma}_2 - \Sigma_2\right)\beta_2$ is asymptotically normal. Therefore, $\sqrt{n}\left(\hat{\theta}_{II} - \theta\right) \xrightarrow{d} N\left(0, P_W^{-1}\Omega_W P_W^{-1}\right)$, where Ω_W is some $d \times d$ long-run variance matrix implied by two summands $W_{i,j(i)}\epsilon_{i,j(i)} - E\left(W_{i,j(i)}\epsilon_{i,j(i)}\right) = W_{i,j(i)}\epsilon_{i,j(i)} + \Sigma\theta$ and $\left\{\left(\Delta X_{2j}\Delta X_{2j}{}'/2\right) - \Sigma_2\right\}\beta_2$ in E_{R_W} and $\left(\hat{\Sigma}_2 - \Sigma_2\right)\beta_2$, respectively.

The remaining task is to provide the analytical expression of Ω_W. Observe that Ω_W may be rewritten as

$$\Omega_W := \Omega_{W,11} + \sqrt{\kappa}\left(\Omega_{W,12} + \Omega_{W,12}'\right) + \kappa\Omega_{W,22},$$

where $\Omega_{W,11}$ and $\Omega_{W,22}$ are the long-run variance matrices of $W_{i,j(i)}\epsilon_{i,j(i)} + \Sigma\theta$ and $\left\{\left(\Delta X_{2j}\Delta X_{2j}{}'/2\right) - \Sigma_2\right\}\beta_2$, respectively, and $\Omega_{W,12}$ is their long-run covariance. Now, let $\phi_{i,j(i)} := W_{i,j(i)}\epsilon_{i,j(i)} + \Sigma\theta$. Clearly, this has no serial dependence, and thus

$$\Omega_{W,11} = Var\left(\phi_{i,j(i)}\right) = E\left(\phi_{i,j(i)}\phi_{i,j(i)}'\right).$$

Next, $\hat{\Sigma}_2 = (m-1)^{-1}\sum_{j=2}^{m}\left(\Delta\eta_{2j}\Delta\eta_{2j}'/2\right) + o_p\left(m^{-1/2}\right)$ as long as $d_3 \leq 3$, where $E\left(\Delta\eta_{2j}\Delta\eta_{2j}'/2\right) = \Sigma_2$ (see, e.g., Yatchew,1997). Because $\psi_j := \left\{\left(\Delta\eta_{2j}\Delta\eta_{2j}'/2\right) - \Sigma_2\right\}\beta_2$ is one-dependent, its long-run variance is $E\left(\psi_j\psi_{j-1}'\right) + E\left(\psi_j\psi_j'\right) + E\left(\psi_j\psi_{j+1}'\right)$ and thus

$$\Omega_{W,22} = \operatorname{diag}\left\{0_{d_1 \times d_1}, E\left(\psi_j\psi_{j-1}'\right) + E\left(\psi_j\psi_j'\right) + E\left(\psi_j\psi_{j+1}'\right), 0_{d_3 \times d_3}\right\}.$$

Lastly, for each i, $\eta_{2j(i)}$ in $\phi_{i,j(i)}$ is correlated with $\psi_{j(i)}$ and $\psi_{j(i+1)}$. Now we decompose $\psi_{j(i)}$ as follows

$$\begin{aligned}\psi_{j(i)} = &\left\{\left(\eta_{2j(i)}\eta_{2j(i)}' - \Sigma_2\right)/2 + \left(\eta_{2j(i-1)}\eta_{2j(i-1)}' - \Sigma_2\right)/2\right.\\ &\left.-\eta_{2j(i)}\eta_{2j(i-1)}' - \eta_{2j(i-1)}\eta_{2j(i)}'\right\}\beta_2\end{aligned}$$

Because ith and $j(i)$th observations are independent, only the expectation between $\phi_{i,j(i)}$ and the first term is non-zero. Defining $\omega_{j(i)} := \left(\eta_{2j(i)}\eta'_{2j(i)} - \Sigma_2\right)\beta_2$, we have $E\left(\phi_{i,j(i)}\psi'_{j(i)}\right) = E\left(\phi_{i,j(i)}\omega'_{j(i)}\right)/2$. Similarly, $E\left(\phi_{i,j(i)}\psi'_{j(i+1)}\right) = E\left(\phi_{i,j(i)}\omega'_{j(i)}\right)/2$ also holds. Therefore,

$$\begin{aligned}\Omega_{W,12} &= \left[\begin{array}{ccc} 0_{d\times d_1} & E\left(\phi_{i,j(i)}\psi'_{j(i)}\right) + E\left(\phi_{i,j(i)}\psi'_{j(i+1)}\right) & 0_{d\times d_3}\end{array}\right] \\ &= \left[\begin{array}{ccc} 0_{d\times d_1} & E\left(\phi_{i,j(i)}\omega'_{j(i)}\right) & 0_{d\times d_3}\end{array}\right],\end{aligned}$$

which completes the proof. ■

A.4 Proof of Theorem 3

The proof requires the following lemma.

Lemma A2. *If Assumptions 1-6 holds and $d_3 = 2, 3$, then*

$$\max_{1\le i\le n}\left|\hat{\lambda}_{i,j(i)} - \lambda_{i,j(i)}\right| = o_p\left(n^{-1/2}\right).$$

A.4.1 Proof of Lemma A2

It is easy to see that $\hat{\lambda}_{i,j(i)} := R_{1i} + R_{2i} + R_{3i} + R_{4i} + \lambda_{i,j(i)}$, where

$$\begin{aligned}R_{1i} &= \left[\left\{\hat{g}_2(Z_i) - \hat{g}_2\left(Z_{j(i)}\right)\right\} - \left\{g_2(Z_i) - g_2\left(Z_{j(i)}\right)\right\}\right]'\left(\hat{\beta}^{(1)}_{II,2} - \beta_2\right), \\ R_{2i} &= \left[\left\{\hat{g}_2(Z_i) - \hat{g}_2\left(Z_{j(i)}\right)\right\} - \left\{g_2(Z_i) - g_2\left(Z_{j(i)}\right)\right\}\right]'\beta_2, \\ R_{3i} &= \left\{g_2(Z_i) - g_2\left(Z_{j(i)}\right)\right\}'\left(\hat{\beta}^{(1)}_{II,2} - \beta_2\right), \text{ and} \\ R_{4i} &= \left(Z_i - Z_{j(i)}\right)'\left(\hat{\gamma}^{(1)}_{II} - \gamma\right).\end{aligned}$$

Hence, the proof is boiled down to demonstrating that $\max_{1\le i\le n}|R_{ki}| = o_p\left(n^{-1/2}\right)$ for $k = 1, 2, 3, 4$.

We first work on R_{3i} and R_{4i}. It follows from Lemma A1, Lipschitz continuity of $g_2(\cdot)$ and $\hat{\theta}^{(1)}_{II} = \theta + O_p\left(n^{-1/d_3}\right)$ that each of $\max_{1\le i\le n}|R_{3i}|$ and $\max_{1\le i\le n}|R_{4i}|$ is bounded by $O_p\left(n^{-2/d_3}\right)$, which becomes $o_p\left(n^{-1/2}\right)$ if $d_3 \le 3$.

The remaining task is to demonstrate that

$$\max_{1\leq i\leq n}\left\|\left\{\hat{g}_2\left(Z_i\right)-\hat{g}_2\left(Z_{j(i)}\right)\right\}-\left\{g_2\left(Z_i\right)-g_2\left(Z_{j(i)}\right)\right\}\right\|=o_p\left(n^{-1/2}\right). \tag{A3}$$

However, Lemma A.2 of Abadie and Imbens (2011) holds under Assumptions 1-6. Therefore, by $m=O(n)$, we have

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

and thus (A3) 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 $o_p\left(n^{-1/2}\right)$. This completes the proof. ■

A.4.2 Proof of Theorem 3

Observe that $\hat{\theta}_{II-FM}:=\hat{P}_W^{-1}\hat{R}_W^{+}$, where $\hat{R}_W^{+}=(1/n)\sum_{i=1}^{n}W_{i,j(i)}Y_i^{+}$. Consistency of the estimator can be shown as before. For asymptotic normality, it follows from $Y_i^{+}=W_{i,j(i)}'\theta+\epsilon_{i,j(i)}+\left(\lambda_{i,j(i)}-\hat{\lambda}_{i,j(i)}\right)$ and Lemma A2 that

$$\hat{R}_W^{+}=\hat{P}_W\theta+\left(\hat{\Sigma}-\Sigma\right)\theta+E_{R_W}+o_p\left(n^{-1/2}\right),$$

just as in (A1). Then,

$$\sqrt{n}\left(\hat{\theta}_{II-FM}-\theta\right)=\hat{P}_W^{-1}\left\{\sqrt{n}\left(\hat{\Sigma}-\Sigma\right)\theta+\sqrt{n}E_{R_W}\right\}+o_p(1).$$

The asymptotic normality of $\sqrt{n}\left(\hat{\theta}_{II-FM}-\theta\right)$ with its asymptotic variance can be established in the same manner as in the proof of Theorem 2. ■

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

[4] Abadie, A., and G. W. Imbens (2012b): "Matching on the Estimated Propensity Score," Working Paper, John F. Kennedy School of Government, Harvard University.

[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] 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.

[13] 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.

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

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

[16] 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.

[17] 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.

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

[19] 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.

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

[21] 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.

[22] 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.

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

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

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

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

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

[28] 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*, forthcoming.

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

[30] 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.

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

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

[33] 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.

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

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

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

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

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

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

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

Table 1: Monte Carlo Results for $d_3 = 1$

Model C: $g_2\left(z\right) = z + \left(5/\tau\right)\phi\left(z/\tau\right), \tau = 0.9$

		$m = n/2$ $(\kappa = 2)$		$m = n$ $(\kappa = 1)$		$m = 2n$ $(\kappa = 1/2)$	
n	Estimator	β_2	γ_1	β_2	γ_1	β_2	γ_1
500	OLS*	1.0008	0.9979	1.0003	0.9983	1.0019	0.9978
		(0.0377)	(0.0707)	(0.0362)	(0.0704)	(0.0360)	(0.0677)
		[0.0377]	[0.0707]	[0.0362]	[0.0704]	[0.0360]	[0.0678]
	MSOLS	0.3510	1.6480	0.3500	1.6455	0.3479	1.6515
		(0.0593)	(0.1075)	(0.0573)	(0.1085)	(0.0560)	(0.1071)
		[0.6517]	[0.6568]	[0.6526]	[0.6546]	[0.6545]	[0.6602]
	MSII	1.1800	0.8230	1.0999	0.8928	1.0712	0.9287
		(0.6010)	(0.6118)	(0.3322)	(0.3528)	(0.2724)	(0.2926)
		[0.6274]	[0.6369]	[0.3469]	[0.3687]	[0.2816]	[0.3012]
1000	OLS*	0.9986	1.0028	0.9996	1.0009	1.0000	1.0003
		(0.0252)	(0.0490)	(0.0254)	(0.0487)	(0.0254)	(0.0496)
		[0.0253]	[0.0491]	[0.0254]	[0.0487]	[0.0254]	[0.0496]
	MSOLS	0.3455	1.6513	0.3461	1.6535	0.3462	1.6536
		(0.0413)	(0.0773)	(0.0420)	(0.0765)	(0.0389)	(0.0759)
		[0.6558]	[0.6559]	[0.6553]	[0.6580]	[0.6549]	[0.6580]
	MSII	1.0644	0.9336	1.0462	0.9542	1.0340	0.9666
		(0.2661)	(0.2754)	(0.2129)	(0.2243)	(0.1743)	(0.1933)
		[0.2738]	[0.2833]	[0.2178]	[0.2290]	[0.1776]	[0.1961]
2000	OLS*	1.0007	1.0004	1.0003	1.0001	0.9996	1.0001
		(0.0177)	(0.0349)	(0.0182)	(0.0340)	(0.0174)	(0.0344)
		[0.0178]	[0.0349]	[0.0182]	[0.0340]	[0.0174]	[0.0344]
	MSOLS	0.3470	1.6542	0.3465	1.6539	0.3447	1.6522
		(0.0302)	(0.0565)	(0.0287)	(0.0539)	(0.0277)	(0.0546)
		[0.6537]	[0.6567]	[0.6541]	[0.6561]	[0.6559]	[0.6545]
	MSII	1.0307	0.9705	1.0169	0.9838	1.0058	0.9900
		(0.1557)	(0.1679)	(0.1272)	(0.1373)	(0.1118)	(0.1260)
		[0.1587]	[0.1704]	[0.1283]	[0.1383]	[0.1119]	[0.1264]

Table 1: *Continued*

Model D: $g_2(z) = z + (5/\tau)\,\phi(z/\tau), \tau = 0.3$

		$m = n/2$ $(\kappa = 2)$		$m = n$ $(\kappa = 1)$		$m = 2n$ $(\kappa = 1/2)$	
n	Estimator	β_2	γ_1	β_2	γ_1	β_2	γ_1
500	OLS*	1.0002	0.9985	1.0001	0.9984	1.0001	0.9996
		(0.0122)	(0.0605)	(0.0115)	(0.0610)	(0.0114)	(0.0588)
		[0.0122]	[0.0606]	[0.0115]	[0.0610]	[0.0114]	[0.0588]
	MSOLS	0.9368	1.0635	0.9365	1.0572	0.9356	1.0639
		(0.0260)	(0.1185)	(0.0233)	(0.1111)	(0.0218)	(0.1061)
		[0.0684]	[0.1344]	[0.0677]	[0.1249]	[0.0680]	[0.1239]
	MSII	1.0053	0.9956	1.0026	0.9903	1.0019	0.9976
		(0.0295)	(0.1229)	(0.0260)	(0.1139)	(0.0243)	(0.1097)
		[0.0299]	[0.1230]	[0.0261]	[0.1144]	[0.0244]	[0.1097]
1000	OLS*	0.9993	1.0021	1.0001	1.0004	1.0005	0.9997
		(0.0081)	(0.0432)	(0.0080)	(0.0428)	(0.0081)	(0.0422)
		[0.0081]	[0.0433]	[0.0080]	[0.0428]	[0.0081]	[0.0422]
	MSOLS	0.9352	1.0619	0.9353	1.0650	0.9359	1.0649
		(0.0190)	(0.0869)	(0.0166)	(0.0790)	(0.0157)	(0.0745)
		[0.0676]	[0.1067]	[0.0668]	[0.1023]	[0.0660]	[0.0988]
	MSII	1.0013	0.9962	1.0007	0.9995	1.0014	0.9995
		(0.0221)	(0.0901)	(0.0185)	(0.0812)	(0.0172)	(0.0757)
		[0.0221]	[0.0902]	[0.0185]	[0.0812]	[0.0173]	[0.0757]
2000	OLS*	1.0000	1.0010	1.0000	1.0004	1.0000	0.9997
		(0.0055)	(0.0300)	(0.0056)	(0.0288)	(0.0057)	(0.0304)
		[0.0055]	[0.0300]	[0.0056]	[0.0288]	[0.0057]	[0.0304]
	MSOLS	0.9356	1.0656	0.9352	1.0656	0.9346	1.0612
		(0.0132)	(0.0617)	(0.0114)	(0.0551)	(0.0106)	(0.0543)
		[0.0657]	[0.0900]	[0.0658]	[0.0857]	[0.0663]	[0.0818]
	MSII	1.0011	1.0002	1.0003	1.0005	0.9997	0.9961
		(0.0150)	(0.0637)	(0.0126)	(0.0564)	(0.0115)	(0.0554)
		[0.0150]	[0.0637]	[0.0126]	[0.0564]	[0.0115]	[0.0555]

Note: For each estimator, simulation averages of estimates, simulation standard deviations (in parentheses) and RMSEs [in brackets] are presented.

Table 2: Monte Carlo Results for $d_3 = 2$

Model AC: $(g_{21}(z), g_{22}(z)) = (z^2, z + (5/\tau)\phi(z/\tau)), \tau = 0.9$

		$m = n/2$ ($\kappa = 2$)		$m = n$ ($\kappa = 1$)		$m = 2n$ ($\kappa = 1/2$)	
n	Estimator	β_2	γ_2	β_2	γ_2	β_2	γ_2
500	OLS*	1.0004	0.9995	1.0004	1.0002	0.9996	1.0001
		(0.0350)	(0.0732)	(0.0372)	(0.0732)	(0.0349)	(0.0749)
		[0.0350]	[0.0732]	[0.0372]	[0.0732]	[0.0349]	[0.0749]
	MSOLS	0.3457	1.6167	0.3487	1.6306	0.3511	1.6409
		(0.0603)	(0.1206)	(0.0581)	(0.1202)	(0.0573)	(0.1165)
		[0.6571]	[0.6284]	[0.6539]	[0.6420]	[0.6514]	[0.6513]
	MSII	1.2812	0.6772	1.1538	0.8231	1.1080	0.8862
		(0.9132)	(1.0065)	(0.3684)	(0.3995)	(0.2898)	(0.3136)
		[0.9555]	[1.0570]	[0.3992]	[0.4369]	[0.3092]	[0.3336]
	MSII-FM	1.2909	0.7008	1.1536	0.8394	1.1075	0.8942
		(0.8881)	(0.9668)	(0.3653)	(0.3975)	(0.2920)	(0.3159)
		[0.9346]	[1.0121]	[0.3963]	[0.4287]	[0.3112]	[0.3332]
1000	OLS*	0.9991	1.0013	1.0008	0.9999	0.9989	0.9997
		(0.0249)	(0.0505)	(0.0256)	(0.0527)	(0.0258)	(0.0503)
		[0.0249]	[0.0506]	[0.0256]	[0.0527]	[0.0258]	[0.0503]
	MSOLS	0.3490	1.6390	0.3501	1.6420	0.3492	1.6470
		(0.0430)	(0.0831)	(0.0399)	(0.0828)	(0.0402)	(0.0782)
		[0.6524]	[0.6443]	[0.6511]	[0.6474]	[0.6520]	[0.6517]
	MSII	1.1105	0.8823	1.0594	0.9341	1.0439	0.9516
		(0.2588)	(0.2720)	(0.1927)	(0.2121)	(0.1734)	(0.1935)
		[0.2814]	[0.2964]	[0.2017]	[0.2221]	[0.1789]	[0.1994]
	MSII-FM	1.1104	0.8993	1.0589	0.9426	1.0439	0.9558
		(0.2574)	(0.2713)	(0.1918)	(0.2112)	(0.1726)	(0.1932)
		[0.2801]	[0.2894]	[0.2007]	[0.2189]	[0.1781]	[0.1982]
2000	OLS*	1.0004	1.0004	0.9988	1.0026	0.9996	1.0011
		(0.0178)	(0.0357)	(0.0179)	(0.0353)	(0.0187)	(0.0359)
		[0.0178]	[0.0357]	[0.0180]	[0.0354]	[0.0187]	[0.0359]
	MSOLS	0.3484	1.6446	0.3492	1.6508	0.3503	1.6489
		(0.0302)	(0.0568)	(0.0272)	(0.0562)	(0.0273)	(0.0563)
		[0.6523]	[0.6471]	[0.6514]	[0.6532]	[0.6503]	[0.6513]
	MSII	1.0564	0.9387	1.0314	0.9693	1.0208	0.9783
		(0.1574)	(0.1685)	(0.1260)	(0.1403)	(0.1112)	(0.1236)
		[0.1672]	[0.1793]	[0.1298]	[0.1436]	[0.1131]	[0.1255]
	MSII-FM	1.0560	0.9475	1.0308	0.9737	1.0205	0.9804
		(0.1567)	(0.1677)	(0.1259)	(0.1404)	(0.1114)	(0.1238)
		[0.1664]	[0.1758]	[0.1296]	[0.1428]	[0.1133]	[0.1253]

Table 2: *Continued*

Model AD: $(g_{21}(z), g_{22}(z)) = (z^2, z + (5/\tau)\,\phi(z/\tau)), \tau = 0.3$

		$m = n/2$ ($\kappa = 2$)		$m = n$ ($\kappa = 1$)		$m = 2n$ ($\kappa = 1/2$)	
n	Estimator	β_2	γ_2	β_2	γ_2	β_2	γ_2
500	OLS*	1.0000	0.9999	1.0001	1.0004	1.0003	0.9994
		(0.0111)	(0.0650)	(0.0114)	(0.0640)	(0.0117)	(0.0666)
		[0.0111]	[0.0650]	[0.0114]	[0.0640]	[0.0117]	[0.0666]
	MSOLS	0.8969	1.0683	0.9151	1.0631	0.9255	1.0686
		(0.0339)	(0.1463)	(0.0269)	(0.1343)	(0.0234)	(0.1198)
		[0.1085]	[0.1614]	[0.0890]	[0.1484]	[0.0780]	[0.1380]
	MSII	1.0359	0.9286	1.0225	0.9559	1.0149	0.9802
		(0.0499)	(0.1597)	(0.0342)	(0.1401)	(0.0277)	(0.1238)
		[0.0615]	[0.1749]	[0.0409]	[0.1469]	[0.0315]	[0.1254]
	MSII-FM	1.0366	0.9624	1.0223	0.9725	1.0147	0.9879
		(0.0490)	(0.1625)	(0.0339)	(0.1424)	(0.0277)	(0.1256)
		[0.0611]	[0.1668]	[0.0406]	[0.1451]	[0.0314]	[0.1262]
1000	OLS*	1.0000	1.0004	1.0004	1.0002	0.9998	0.9988
		(0.0079)	(0.0445)	(0.0079)	(0.0462)	(0.0082)	(0.0443)
		[0.0079]	[0.0445]	[0.0079]	[0.0462]	[0.0082]	[0.0443]
	MSOLS	0.9145	1.0768	0.9245	1.0682	0.9291	1.0661
		(0.0219)	(0.0980)	(0.0177)	(0.0904)	(0.0158)	(0.0821)
		[0.0882]	[0.1245]	[0.0775]	[0.1133]	[0.0727]	[0.1054]
	MSII	1.0218	0.9705	1.0124	0.9808	1.0072	0.9879
		(0.0295)	(0.1048)	(0.0213)	(0.0943)	(0.0179)	(0.0855)
		[0.0367]	[0.1089]	[0.0246]	[0.0962]	[0.0193]	[0.0863]
	MSII-FM	1.0216	0.9893	1.0123	0.9898	1.0072	0.9925
		(0.0292)	(0.1077)	(0.0212)	(0.0946)	(0.0178)	(0.0866)
		[0.0363]	[0.1082]	[0.0244]	[0.0952]	[0.0192]	[0.0870]
2000	OLS*	0.9999	1.0008	0.9995	1.0020	0.9999	1.0007
		(0.0057)	(0.0315)	(0.0059)	(0.0313)	(0.0058)	(0.0315)
		[0.0057]	[0.0315]	[0.0059]	[0.0313]	[0.0058]	[0.0315]
	MSOLS	0.9234	1.0717	0.9296	1.0716	0.9322	1.0671
		(0.0140)	(0.0633)	(0.0120)	(0.0591)	(0.0110)	(0.0572)
		[0.0779]	[0.0956]	[0.0714]	[0.0929]	[0.0687]	[0.0882]
	MSII	1.0113	0.9841	1.0076	0.9934	1.0045	0.9949
		(0.0175)	(0.0663)	(0.0139)	(0.0606)	(0.0123)	(0.0582)
		[0.0208]	[0.0682]	[0.0158]	[0.0610]	[0.0131]	[0.0584]
	MSII-FM	1.0112	0.9929	1.0075	0.9972	1.0044	0.9969
		(0.0174)	(0.0667)	(0.0139)	(0.0611)	(0.0123)	(0.0584)
		[0.0207]	[0.0670]	[0.0158]	[0.0612]	[0.0131]	[0.0585]

Note: For each estimator, simulation averages of estimates, simulation standard deviations (in parentheses) and RMSEs [in brackets] are presented.

Table 3: Returns to Schooling with Ability Measures Using Matched Samples

Dependent Variable: $\log(wage)$

Regressors	(1) OLS*	(2) MSOLS	(3) MSII-FM	(4) MSOLS	(5) MSII-FM
educ	0.0594	0.0722	0.0699	0.0664	0.0600
	(0.0057)	(0.0058)	(0.0094)	(0.0059)	(0.0290)
exper	0.0827	0.0992	0.0913	0.0898	0.0887
	(0.0094)	(0.0092)	(0.0092)	(0.0090)	(0.0094)
$exper^2$	−0.0022	−0.0030	−0.0024	−0.0026	−0.0023
	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0005)
abil	0.0048	0.0002	0.0014	−0.0005	0.0160
	(0.0015)	(0.0017)	(0.0055)	(0.0059)	(0.0581)
feduc	−0.0025	−0.0012	−0.0005	−0.0016	−0.0006
	(0.0034)	(0.0035)	(0.0035)	(0.0037)	(0.0044)
meduc	0.0099	0.0111	0.0099	0.0110	0.0071
	(0.0041)	(0.0043)	(0.0045)	(0.0046)	(0.0074)
smsa	0.1494	0.1525	0.1524	0.1430	0.1580
	(0.0193)	(0.0195)	(0.0197)	(0.0197)	(0.0251)
south	−0.1037	−0.1069	−0.1030	−0.1052	−0.1000
	(0.0191)	(0.0192)	(0.0192)	(0.0194)	(0.0262)
intercept	4.6863	4.5702	4.5762	4.7224	4.7776
	(0.0898)	(0.0962)	(0.1161)	(0.0988)	(0.3715)
abil?	*kww*	*kww*	*kww*	*g*	*g*
Matching?	No	Yes	Yes	Yes	Yes
(n, m)	$(1846, -)$	$(1846, 396)$	$(1846, 396)$	$(1846, 589)$	$(1846, 589)$

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 $\hat{Q}_W^{-1}\hat{\Omega}_{W,11}\hat{Q}_W^{-1}/n$.