The Stata Journal (*yyyy*) *vv*, Number *ii*, pp. 1–14

msreg: A Stata Command for Consistent Estimation of Linear Regression Models Using Matched Data

Masayuki Hirukawa
Ryukoku University

Di Liu
StataCorp

Artem Prokhorov
University of Sydney
& CEBA & CIREQ

Abstract. Economists often use matched samples, especially dealing with earning data where some observations are missing in one sample and need to be imputed from another sample. Hirukawa and Prokhorov (2018) show that the ordinary least squares estimator using matched samples is inconsistent and propose two consistent estimators. We describe a new Stata command, **msreg**, which implements these two consistent estimators based on two samples. The estimators attain the parametric convergence rate if the number of continuous matching variables is no greater than four.

Keywords: **msreg**, bias correction, linear regression, matching estimation

1 Introduction

Matching-based imputation is common in economic datasets. For example, the US Census uses a practice known as "hot-deck imputation" which is implemented when Census reports for non-responders values of important variables such as earnings and income borrowed from responders with a few similar characteristics. In some surveys the share of such imputed responses reaches 30%.

Hirukawa and Prokhorov (2018), abbreviated as "HP18" hereinafter, were concerned with this widely used but often ignored practice. The concern was that users of such data in applied econometrics are often unaware of the fact that these are imputed, rather than actual, observations and the resulting matching discrepancy leads to nonnegligible biases in the ordinary least squares estimator. They list a large number of other settings, usually involving more than one datasets, where matching is unavoidable and needs to be accounted for.

The goal of this paper is to facilitate the use of the consistent estimation approaches proposed by HP18 in its numerous applications. HP18 derive the imputation bias analytically and propose two bias-corrected estimators. In this paper, we introduce a new command, **msreg**, which implements both estimators in Stata.

Section 2 documents theoretical backgrounds for the **msreg** command. In Section 3 we discuss the **msreg** syntax and provide a numerical example. Section 4 contains a simulation study. Section 5 provides an empirical application by estimating the return to schooling as in HP18.

2 Setup and Estimators

2.1 Setup and Assumptions

Suppose that we are interested in estimating a linear regression model:

$$y = \beta_0 + \mathbf{x}_1'\beta_{\mathbf{1}} + \mathbf{x}_2'\beta_{\mathbf{2}} + \mathbf{z}'\gamma + u = \mathbf{w}'\theta + u, \qquad \mathbf{E}(u|\mathbf{w}) = 0,$$

where $\mathbf{x}_1 \in \mathbf{R}^{d_1}$, $\mathbf{x}_2 \in \mathbf{R}^{d_2}$, $\mathbf{z} \in \mathbf{R}^{d_z}$, $\mathbf{w} = (1, \mathbf{x}_1', \mathbf{x}_2', \mathbf{z}')$, and $\theta = (\beta_0, \beta_{\mathbf{1}}{}', \beta_{\mathbf{2}}{}', \gamma')'$.

If we can observe all the variables in one sample, OLS is a consistent estimator for θ. However, in reality, we often encounter a situation where the variables are taken from two different samples. To be precise, we need more notatons to distinguish between the two samples. The first sample is denoted by $\mathbf{S}_1 = \{(y_i, \mathbf{x}_{1i}, \mathbf{z}_i)\}_{i=1}^n$. The second sample is denoted by $\mathbf{S}_2 = \{(\mathbf{x}_{2j}, \mathbf{z}_j)\}_{j=1}^m$. For the purpose of inference, we also denote d_3 as the number of continuous common variables in $\mathbf{z}$ hereinafter, which is not always equal to d_z.

Estimation theories in HP18 are built on a set of assumptions, which are required for identification, consistency and asymptotic normality of their estimators. Some of them are quite common. For example, Assumption 2 imposes compactness of the support of continuous common variables. In our empirical analysis in Section 5, `educ`, `feduc` and `meduc` are such variables, and it is natural to think of their support as compact. On the other hand, there are more subtle assumptions in HP18 that may or may not hold in a given application. Examples include a common joint distribution for $\mathbf{S}_1$ and $\mathbf{S}_2$ (Assumption 1), $\mathbf{E}(\eta_1\eta_2') = 0$ and strict nonlinearity in $g_2(\cdot)$ (Assumption 3(ii)), where $\eta_\ell := \mathbf{x}_\ell - g_\ell(\mathbf{z})$ and $g_\ell(\mathbf{z}) := \mathbf{E}(\mathbf{x}_\ell|\mathbf{z})$ for $\ell = 1, 2$. It is difficult to test the validity of these assumptions because $\mathbf{x}_1$ and $\mathbf{x}_2$ belong to two distinct samples.

2.2 Nearest Neighbor Matching

The matched sample can be constructed via the nearest neighbor matching (NNM) using a vector of common variables $\mathbf{z}$ across two samples. It is worth emphasizing that $\mathbf{z}$ must contain at least one continuous variable for valid inference; inclusion of discrete common variables with a finite number of support points (e.g., binary variables) in $\mathbf{z}$ does not affect the asymptotic results that will be stated shortly.

To specify the NNM, we need to first define a matrix norm to measure the distance between two vectors. For a vector x and some symmetric and positive definite matrix $\mathbf{A}$, the vector norm is defined as $||x||_{\mathbf{A}} = (x'\mathbf{A}x)^{1/2}$. Following Abadie and Imbens (2011), we use the Mahalanobis metric $\mathbf{A}_M = \left\{(1/N)\sum_{i=1}^N(\mathbf{z}_i - \overline{\mathbf{z}})(\mathbf{z}_i - \overline{\mathbf{z}})'\right\}^{-1}$ and the normalized Euclidean metric $\mathbf{A}_{NE} = \mathbf{d}iag\left(\mathbf{A}_M^{-1}\right)^{-1}$, where $N = n + m$ and $\overline{\mathbf{z}} = (1/N)\sum_{i=1}^N \mathbf{z}_i$.

Let $j_k(i)$ be the index of the kth match in $\mathbf{S}_2$ to the unit i in $\mathbf{S}_1$, i.e., for each

$i \in \{1, \ldots, n\}$, $j_k(i)$ satisfies

$$\sum_{j=1}^{m} \mathbb{1}\left\{||\mathbf{z}_j - \mathbf{z}_i||_{\mathbf{A}} \leq ||\mathbf{z}_{j_k(i)} - \mathbf{z}_i||_{\mathbf{A}}\right\} = k.$$

In other words, $\mathbf{z}_{j_k(i)}$ is the kth nearest neighbor in $\mathbf{S}_2$ to the unit i in $\mathbf{S}_1$.

For each unit i, let $\mathbf{J}_K(i) = \{j_i(1), \ldots, j_i(K)\}$ denote the K matches from $\mathbf{S}_2$. The NNM based matched sample is

$$\mathbf{S} = \left\{(y_i, \mathbf{x}_{1i}, \mathbf{x}_{2j_1(i)}, \ldots, \mathbf{x}_{2j_K(i)}, \mathbf{z}_i)\right\}_{i=1}^{n}$$

We also write $\overline{\mathbf{x}_{2j(i)}} = (1/K)\sum_{j \in \mathbf{J}_K(i)} \mathbf{x}_{2j}$.

For the purpose of estimation, we use a transformation of the matched sample $\mathbf{S}$.

$$\mathbf{S}^* = \left\{(y_i, \mathbf{x}_{1i}, \overline{\mathbf{x}_{2j(i)}}, \mathbf{z}_i)\right\}_{i=1}^{n}.$$

Compared with the original matched sample $\mathbf{S}$, we replace the individual matched variable $\mathbf{x}_{2j_1(i)}, \ldots, \mathbf{x}_{2j_K(i)}$ by their mean $\overline{\mathbf{x}_{2j(i)}}$.

Throughout it is assumed that we estimate θ by regressing y_i on $\mathbf{w}_{i,j(i)} = (1, \mathbf{x}_{1i}', \overline{\mathbf{x}_{2j(i)}}', \mathbf{z}_i)'$

2.3 Inconsistency of Matched-Sample OLS

We start by using ordinary least squares estimator on the matched sample $\mathbf{S}^*$. The OLS estimator is

$$\widehat{\theta}_{\mathbf{MSOLS}} = \widehat{\mathbf{Q}_W}^{-1}\widehat{\mathbf{R}_W},$$

where $\widehat{\mathbf{Q}_W} = (1/n)\sum_{i=1}^{n} \mathbf{w}_{i,j(i)}\mathbf{w}_{i,j(i)}'$ and $\widehat{\mathbf{R}_W} = (1/n)\sum_{i=1}^{n} \mathbf{w}_{i,j(i)}y_i$. It is referred as the matched-sample OLS (**MSOLS**) estimator.

Theorem 1 *(HP18, Theorem 1) Under some regularity conditions,*

$$\widehat{\theta}_{\mathbf{MSOLS}} = \mathbf{Q}_W^{-1}\mathbf{P}_W\theta + O_p(m^{-1/d_3}) + O_p(n^{-1/2}),$$

where $\mathbf{Q}_W = \mathbf{E}(\mathbf{w}_{i,j(i)}\mathbf{w}_{i,j(i)}')$, $\mathbf{P}_W = \mathbf{Q}_W - (1/K)\Sigma$, Σ *is a* $(d+1) \times (d+1)$ *block-diagonal matrix of the form* $\Sigma = \mathbf{d}iag\{0_{(d_1+1)\times(d_1+1)}, \Sigma_2, 0_{d_z \times d_z}\}$, *and* $\Sigma_2 = \mathbf{E}(\eta_2\eta_2')$.

Theorem 1 implies that **MSOLS** is inconsistent in general. The inconsistency is attributed to correlation between the imputed regressor $\overline{\mathbf{x}_{2j(i)}}$ and $(1/K)\sum_{j \in \mathbf{J}_{K(i)}} \eta_{2j}$ in the composite error term $\epsilon_{i,j(i)}$. All asymptotic analyses in HP18 are based on letting n and m diverge while keeping K fixed. It is in principle possible to restore consistency by letting K diverge at a rate slower than n and m. However, a fixed K is what researchers are likely to do in practice; Abadie and Imbens (2006) also adopt this setup. Moreover, the $O_p\left(m^{-1/d_3}\right)$ term corresponds to the second-order bias term $\lambda_{i,j(i)}$ due to the matching discrepancy by Abadie and Imbens (2006). Observe that this term affects the convergence rate of $\widehat{\theta}_{\mathbf{MSOLS}}$ to its probability limit; see Remark 3 of HP18 for more details.

2.4 One-step bias-corrected estimator

The source of the inconsistency of **MSOLS** estimator is the fact that $\widehat{\mathbf{Q}_W} \xrightarrow{p} \mathbf{Q}_W$ whereas $\widehat{\mathbf{R}_W} \xrightarrow{p} \mathbf{P}_W\theta = \{\mathbf{Q}_W - (1/K)\Sigma\}\,\theta$. To eliminate the non-vanishing bias, the strategy is to replace the denominator $\widehat{\mathbf{Q}_W}$ by a consistent estimator of $\mathbf{P}_W$ with the numerator $\widehat{\mathbf{R}_W}$ left unchanged. Because this bias correction has an indirect inference interpretation, HP18 call this estimator the *matched-sample indirect inference* (**MSII**) estimator. Let $\widehat{\mathbf{P}_W}$ be some consistent estimator of $\mathbf{P}_W$. Then, the **MSII** estimator is defined as

$$\widehat{\theta}_{\mathbf{MSII}} = \widehat{\mathbf{P}_W}^{-1}\widehat{\mathbf{R}_W}$$

To consistently estimate $\mathbf{P}_W$, we need consistent estimators for $\mathbf{Q}_W$ and Σ. Apparently, $\widehat{\mathbf{Q}_W}$ is a natural estimator for $\mathbf{Q}_W$. Furthermore, it turns out that we can consistently estimate Σ without nonparametric estimation of $\mathbf{E}(\mathbf{x}_2|\mathbf{z})$. To do so, we first reorder $\mathbf{S}_2$ with respect to $\mathbf{z}$ in the ascending order.

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

2. For $j = 2, \ldots, m$, choose $(j) = \arg\min_{j\neq(1),\ldots,(j-1)} ||\mathbf{z}_j - \mathbf{z}_{(j-1)}||$, where the norm of a matrix $||A||$ is defined as $||A|| = \{tr(A'A)\}^{1/2}$.

Given the reordered sample $\mathbf{S}_2 = \{(\mathbf{x}_{2(j)}, \mathbf{z}_{(j)})\}_{j=1}^m$, Σ_2 can be consistently estimated by

$$\widehat{\Sigma_2} = \frac{1}{2(m-1)} \sum_{j=2}^{m} \Delta\mathbf{x}_{2(j)} \Delta\mathbf{x}'_{2(j)},$$

where $\Delta\mathbf{x}_{2(j)} = \mathbf{x}_{2(j)} - \mathbf{x}_{2(j-1)}$. This is known as the difference-based variance estimator; see von Neumann (1941), Yatchew (1997) and Horowitz and Spokoiny (2001) for references.

The estimator of $\mathbf{P}_W$ is given by

$$\widehat{\mathbf{P}_W} = \widehat{\mathbf{Q}_W} - \frac{1}{K}\widehat{\Sigma} = \widehat{\mathbf{Q}_W} - \frac{1}{K}\mathbf{d}iag\{0_{(d_1+1)\times(d_1+1)}, \widehat{\Sigma_2}, 0_{d_z\times d_z}\}$$

Theorem 2 below documents asymptotic normality of $\widehat{\theta}_{\mathbf{MSII}}$. The theorem applies only when the number of continuously distributed matching variables is so small that the second-order, matching discrepancy bias can be safely ignored. Observe that both the convergence rate of $\widehat{\theta}_{\mathbf{MSII}}$ and its asymptotic variance depend on the divergence pattern of (n, m).

Theorem 2 *(HP18, Corollary 1) Under some regularity conditions, as $n, m \to \infty$,*

$$\begin{cases} \sqrt{n}(\widehat{\theta}_{\mathbf{MSII}} - \theta) \xrightarrow{d} N(0, V_I) = N(0, \mathbf{P}_W^{-1}\Omega\mathbf{P}_W^{-1}), & \text{if } n/m \to \kappa \in (0,\infty) \text{ and } d_3 = 1. \\ \sqrt{n}(\widehat{\theta}_{\mathbf{MSII}} - \theta) \xrightarrow{d} N(0, V_{II}) = N(0, \mathbf{P}_W^{-1}\Omega_{11A}\mathbf{P}_W^{-1}), & \text{if } n/m \to 0 \text{ and } d_3 = 1, 2. \\ \sqrt{m}(\widehat{\theta}_{\mathbf{MSII}} - \theta) \xrightarrow{d} N(0, V_{III}) = N(0, \mathbf{P}_W^{-1}\Omega_{22}\mathbf{P}_W^{-1}), & \text{if } n/m \to \infty \text{ and } d_3 = 1, \end{cases}$$

where the definitions of Ω, Ω_{11A}, and Ω_{22} can be found in the Appendix, along with their consistent estimates.

As demonstrated in this theorem and Theorem 3 below, the bias-corrected estimators of HP18 attain the parametric convergence rate only when the number of continuous common variables is four or less. It may be tempting to include as many continuous common variables as possible in the NNM. However, this results in slowing down the convergence rate and we do not recommend it.

2.5 Two-step bias-corrected estimator

The one-step bias-corrected estimator can attain the parametric rate of convergence with at most two matching variables. To overcome this curse of dimensionality, we should eliminate the second order bias $\lambda_{i,j(i)}$. The entire procedure is reminiscent of the fully-modified least squares estimation for conintegrating regressions by Phillips and Hansen (1990). In this sense, HP18 call the estimator the fully-modified **MSII** (**MSII-FM**) estimator.

Estimating $\lambda_{i,j(i)}$ requires consistent estimates of θ and $g_2(\cdot)$. For θ, we can use the **MSII** estimate $\widehat{\theta}_{\mathbf{MSII}}$. For $g_2(\cdot)$, we use a nonparametric power-series estimation as in Abadie and Imbens (2011). Let $v = (v_1, \ldots, v_{d_z})$ be a multi-index of dimension d_z, which is a d_z-dimensional vector of nonnegative integers with $|v| = \sum_{l=1}^{d_z} v_l$. Also denote $z^v = \Pi_{l=1}^{d_z} z_l^{v_l}$. Consider a series $\{v_Q\}_{Q=1}^{\infty}$ containing distinct vectors such that $|v(Q)|$ is non-decreasing. Let $p_Q(z) = z^{v(Q)}$ and $p^Q(z) = (p_1(z), \ldots, p_Q(Z))'$. Then a nonparametric series estimator of the regression function $g_{2r}(z), r = 1, \ldots, d_2$ is

$$\widehat{g_{2r}} = p^{Q(m)}(\mathbf{z})' \left\{ \sum_{j=1}^{m} p^{Q(m)}(\mathbf{z}_j) p^{Q(m)}(\mathbf{z}_j)' \right\}^{-} \sum_{j=1}^{m} p^{Q(m)}(\mathbf{z}_j)\mathbf{x}_{2r,j},$$

where $\mathbf{x}_{2r,j}$ denotes the rth element of $\mathbf{x}_{2j}$ in $\mathbf{S}_2$, $(\cdot)^-$ denotes the generalized inverse, and $Q = Q(m)$ implies the dependence of Q on the sample size of $\mathbf{S}_2$.

The entire estimation procedure can be summarized in the following three steps:

1. Run **MSII** using the matched sample $\mathbf{S}^*$ to obtain $\widehat{\theta}_{\mathbf{MSII}} = (\widehat{\beta_{II,0}}', \widehat{\beta_{II,1}}', \widehat{\beta_{II,2}}', \widehat{\gamma_{II}}')'$

2. Construct adjusted dependent variables $\{y_i^+\}_{i=1}^n = \{y_i - \widehat{\lambda}_{i,j(i)}\}_{i=1}^n$, where

$$\widehat{\lambda}_{i,j(i)} = \left(\widehat{g}_2(\mathbf{z}_i) - \frac{1}{K} \sum_{j \in \mathbf{J}_K(i)} \widehat{g}_2(\mathbf{z}_j) \right)' \widehat{\beta_{II,2}}$$

3. Rerun **MSII** using the modified matched sample $\mathbf{S}^+ = \left\{ (y_i^+, \mathbf{x}_{1i}, \overline{\mathbf{x}_{2j(i)}}, \mathbf{z}_i) \right\}_{i=1}^n$ to obtain the final estimator

$$\widehat{\theta}_{\mathbf{MSII-FM}} = \widehat{\mathbf{P}_W}^{-1} \widehat{\mathbf{R}_W}^{+} = \widehat{\mathbf{P}_W}^{-1} \frac{1}{n} \sum_{i=1}^{n} \mathbf{w}'_{i,j(i)} y_i^+$$

Theorem 3 *(HP18, Theorem 4) Under some regularity conditions, as* $n, m \to \infty$*,*

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

where the definitions of V_I*,* V_{II} *and* V_{III} *are the same as in Theorem 2.*

In practice, the standard errors resulting from each of the three cases may be quite different. The relative magnitudes of n and m determine which case applies. HP18 did not provide any generic comparisons for the variance matrices. Besides the scaling factor, the differences can be attributed to the specific features of the datasets and model specification. In borderline cases it is advisable to use larger standard errors for conservative inference.

3 Syntax of msreg

msreg has the following syntax.

```
msreg depvar [ varlist_X1 ] (varlist_X2 = varlist_Z) using filename [ if ] [ in ]
      , [ vce(vce_spec) estimator(est_spec) nneighbor(#)
      metric(metric_spec) order(#) noconstant level(#)
      display_options coeflegend ]
```

3.1 Options

`vce(`*vce_spec*`)` specifies the type of variance-covariance matrix used in computation. *vce_spec* can be one of `vi`, `vii` or `viii`. The default is `vce(vi)`. The definition of `vi`, `vii` and `viii` can be found in Theorem 2.

`estimator(`*est_spec*`)` specifies the type of estimator. *est_spec* can either be `onestep` or `twostep`. `onestep` specifies to use the one-step bias-corrected estimator. `twostep` specifies to use the two-step bias-corrected estimator. The default is `estimator(twostep)`.[1]

`nneighbor(`#`)` specifies the number of matches per observation. The default is `nneighbor(1)`. The maximum allowed number of matches is 10. Each observation is matched with the mean of the specified number of observations from the other dataset.

`metric(`*metric_spec*`)` specifies the distance matrix used as the weight matrix in a quadratic form that transforms the multiple distances into a single distance measure. *metric_spec* can either be `mahalanobis` or `euclidean`. `metric(mahalanobis)` specifies to use the inverse of the sample covariance matrix of matching variables, which is the default. `metric(euclidean)` specifies to use the inverse of only diagonal elements of the sample covariance matrix of matching variables.

`order(`#`)` specifies the order of polynomial in the power-series approximation for **MSII-FM**. The default is `order(2)`. The maximum allowed number of order is 5.

`noconstant` suppresses the constant term.

`level(`#`)` specifies the level of significance for output table.

display_options: `noci`, `nopvalues`, `noomitted`, `vsquish`, `noemptycells`, `baselevels`, `allbaselevels`, `nofvlabel`, `fvwrap(`#`)`, `fvwrapon(style)`, `cformat(`*%fmt*`)`, `pformat(`*%fmt*`)`, `sformat(`*%fmt*`)`, and `nolstretch`; see [R] **Estimation options**.

`coeflegend` specifies that the legend of the coefficients and how to specify them in an expression be displayed rather than displaying the statistics for the coefficients.

3.2 Stored results

msreg stores the following results in `e()`.

1. We intentionally leave the option `estimator(msols)` undocumented. MSOLS is inconsistent and we includ it in the paper for the purpose of simulations only. We do not recommend users to use the option `estimator(msols)`.

Scalars

`e(N_1)`	number of observations in the first sample
`e(N_2)`	number of observations in the second sample
`e(nneighbor)`	number of nearest neighbors matched
`e(chi2)`	chi-squared
`e(p)`	p-value for test of variables
`e(df_m)`	degree-of-freedom in the model
`e(order)`	order of polynomial in the power-series estimation

Macros

`e(cmd)`	msreg
`e(footnote)`	footnote displayed under the output table
`e(chi2type)`	Wald
`e(vce)`	vce specified in `vce()`
`e(vcetype)`	title used to label Std. Err.
`e(common)`	common variables that exist in both samples
`e(matched)`	matched variables
`e(metric)`	type of distance matrix
`e(estimator)`	type of estimator
`e(title)`	title in estimation output
`e(properties)`	`b V`

Matrices

`e(b)`	coefficient vector
`e(V)`	variance-covariance matrix of the estimators

Functions

`e(sample)`	marks estimation sample

3.3 A numerical example

We illustrate the use of **msreg** with a numerical example.

For the purpose of illustration, we simulate two datasets: `s1.dta` and `s2.dta`. The first sample `s1.dta` contains the dependent variable `y`, and some independent variables `x11`, `x12`, `z1` and `z2`. The second sample `s2.dta` contains some other dependent variables `x21`, `x22`, `z1` and `z2`. Notice that variables `z1` and `z2` exist in both samples. In contrast, the variables `x21` and `x22` only exist in the second sample `s2.dta`. The data generating process is described in Section 4.

We want to fit the following regression model

$$y = \beta_0 + \beta_{11}x_{11} + \beta_{12}x_{12} + \beta_{21}x_{21} + \beta_{22}x_{22} + \gamma_1 z_1 + \gamma_2 z_2 + u \tag{1}$$

The true values of all the coefficients are set to be 1.

Apparently, we cannot estimate regression (1) using just the first sample `s1.dta`, because the variables `x21` and `x22` are missing in this dataset. Instead, we want to use the common variables that exist in both samples, i.e. `z1` and `z2`, to construct matched variables `x22` and `x21` from the second sample `s2.dta`.

We now use **msreg** to estimate the coefficients in regression (1). We can use the default two-step bias-corrected estimator and the default `vi` type variance if we assume n/m converges to a nonzero constant.

Here are some comments on the syntax.

- The `(x21 x22 = z1 z2)` specifies that the variables `x21` and `x22` are the variables to be matched, and the variables `z1` and `z2` are the common variables that exists in both samples.
- The `using s2` specifies that the variables `x21` and `x22` come from `s2.dta`.
- We use the default two-step bias corrected estimator.
- The default option `vce(vi)` specifies to use the `vi` type variance matrix as specified in Theorem 2, because we assume the sample size ratio between the two samples converges to a nonzero constant and there are only two continuous common variables used for matching. Actually, the footnote states a similar explanation about `vce(vi)`.
- Option `nneighbor(2)` specifies to pick out 2 matches via the NNM.
- Option `order(3)` specifies to fit a third order polynomial in the power-series approximation for **MSII-FM**.
- The output shows the point estimates of coefficients and their standard errors, and they can be interpreted as in a regular linear regression framework.

4 A simulation study

4.1 Simulation design

We conduct a Monte Carlo simulation study for two purposes: first, we want to see the finite-sample performance of **MSII-FM** in contrast to **MSOLS**; second, the simulation results can serve as a verification of the numerical implementation of our command **msreg**. The simulation study replicates that of HP18.

The model considered throughout is

$$y = \beta_0 + \mathbf{x}_1'\beta_{\mathbf{1}} + \mathbf{x}_2'\beta_{\mathbf{2}} + \mathbf{z}'\gamma + u \tag{2}$$

where $\mathbf{x}_1 = (x_{11}, x_{12})'$, $\beta_{\mathbf{1}} = (\beta_{11}, \beta_{12})' \in \mathbf{R}^2$, $\mathbf{x}_2 = (x_{21}, x_{22})'$, $\beta_{\mathbf{2}} = (\beta_{21}, \beta_{22})' \in \mathbf{R}^2$, and $\mathbf{z} = (z_1, \ldots, z_{d_3})'$, $\gamma = (\gamma_1, \ldots, \gamma_{d_3})' \in \mathbf{R}^{d_3}$ for $d_3 = 1, 2, 3$. It is assumed that two samples, namely, $\mathbf{S}_1 = \{(y_i, \mathbf{x}_{1i}, \mathbf{z}_i)\}_{i=1}^n$ and $\mathbf{S}_2 = \{(\mathbf{x}_{2j}, \mathbf{z}_j)\}_{j=1}^m$ are observable in practice.

Here is how to generate the data. First, $\mathbf{z}^* = (z_1^*, z_2^*, z_3^*)$ is generated by

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

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

Second, $\mathbf{x}_1 = (x_{11}, x_{12})'$ is generated by $x_{1q} = \sum_{p=1}^{d_3} z_p + \eta_q$ $(q = 1, 2)$, where $\eta_q \sim N(0, 1)$. Third, $\mathbf{x}_2 = (x_{21}, x_{22})'$ 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 $\eta_{2r} \sim N(0, 1)$. Specifically, $g_{21}(z) = z + (5/\tau)\,\phi(z/\tau), \tau = 0.25$, where $\phi(\cdot)$ is the pdf of $N(0, 1)$, and $g_{22}(z) = 4\sqrt{|z/2|(1 - |z/2|)}$ $\sin\{2\pi(1 + \epsilon)/(|z/2| + \epsilon)\}$, with $\epsilon = 0.05$.

Finally, y is generated by setting all coefficients equal to 1 with $u \overset{\text{iid}}{\sim} N(0, 1)$. The sample sizes are set to be $(n, m) = (1000, 1000)$. The number of replications is 1000.

We focus on the finite-sample properties of estimators of β_{22} and γ_1. For each estimator, the following performance measures are computed: (i) Bias (1 - Mean), where Mean is the simulation average of the parameter estimate (ii) SD (simulation standard deviation of the parameter estimate) (iii) $\overline{SE}$ (simulation average of the standard error) and (iv) Rej. rate (rejection rate for the test of parameter estimate equal to its true value 1 against the nominal 5% level of significance)

For $d_3 = 1, 2, 3$, we estimate the coefficients in regression (2) using both **MSOLS** and **MSII-FM**. For **MSOLS** the number of matches is $K = 1, 2, 4, 8$. For **MSII-FM**, the number of matches K is fixed at 1, and orders of polynomials in the power-series approximation are $2, 3$ and 4. For a more complete simulation study, see Section 4 of HP18.

4.2 Results

The simulation results are summarized in Table (1) and (2) for **MSOLS** and **MSII-FM**, respectively.

(a) Table (1) shows that regardless of the number of matches, there is a big bias of β_{22} and there is a large rejection rate, which indicate inconsistency of **MSOLS** as implied in Theorem 1.

(b) Table (2) shows that (1) the bias is small, i.e., the mean of the point estimates is very close to its true value, (2) the standard deviation of the point estimate is very close to the mean of the standard errors, and (3) the overall rejection rate is close to the nominal 5% level. Notice that for the case $d_3 = 2$, the rejection rate is a little bit off for β_{22}. However, based on results for larger samples, it seems that the over-rejection rate is due to the finite-sample bias of **MSII-FM** (reported in the Supplement of HP18). The simulation result shows that **MSII-FM** performs well in finite sample as predicted by Theorem 3, and it also numerically verifies the implementation of **msreg**.

Table 1: Monte Carlo results for **MSOLS**

	β_{22}				γ_1			
K	1	2	4	8	1	2	4	8
$d_3 = 1$								
Bias	0.4486	0.2866	0.1634	0.0773	-0.1680	-0.0919	-0.0442	-0.0216
SD	0.0426	0.0455	0.0474	0.0492	0.1122	0.1052	0.1005	0.0980
$\overline{SE}$	0.0507	0.0523	0.0546	0.0603	0.1174	0.1141	0.1119	0.1109
Rej. rate	1.0000	1.0000	0.9110	0.3800	0.3210	0.1550	0.1070	0.0910
$d_3 = 2$								
Bias	0.5280	0.3724	0.2239	0.0828	-0.1289	-0.0723	-0.0372	0.0093
SD	0.0408	0.0460	0.0523	0.0619	0.1468	0.1398	0.1369	0.1379
$\overline{SE}$	0.0462	0.0529	0.0627	0.0754	0.1548	0.1502	0.1534	0.1667
Rej. rate	1.0000	1.0000	0.9660	0.3110	0.1640	0.1030	0.0880	0.1100
$d_3 = 3$								
Bias	0.7532	0.6149	0.4472	0.2305	-0.2206	-0.1306	-0.0475	0.0281
SD	0.0472	0.0575	0.0707	0.0893	0.1993	0.1906	0.1881	0.1922
$\overline{SE}$	0.0514	0.0648	0.0794	0.1082	0.2015	0.1990	0.2078	0.2270
Rej. rate	1.0000	1.0000	1.0000	0.6860	0.1980	0.1210	0.0860	0.0920

Table 2: Simulation results for **MSII-FM**

	β_{22}			γ_1		
Order	2	3	4	2	3	4
$d_3 = 1$						
Bias	-0.0305	-0.0305	-0.0323	0.0110	0.0110	0.0119
SD	0.1049	0.1049	0.1047	0.1257	0.1259	0.1258
$\overline{SE}$	0.1130	0.1148	0.1149	0.1353	0.1361	0.1359
Rej. rate	0.0620	0.0610	0.0640	0.0730	0.0730	0.0720
$d_3 = 2$						
Bias	-0.1740	-0.1735	-0.1641	0.0318	0.0382	0.0401
SD	0.1539	0.1541	0.1540	0.1750	0.1754	0.1765
$\overline{SE}$	0.1637	0.1636	0.1643	0.1912	0.1941	0.1930
Rej. rate	0.1400	0.1380	0.1270	0.0820	0.0840	0.0760
$d_3 = 3$						
Bias	-0.0948	-0.0904	-0.0866	0.0372	0.0408	0.0526
SD	0.2884	0.2925	0.2904	0.2481	0.2499	0.2495
$\overline{SE}$	0.3041	0.3152	0.3132	0.2680	0.2749	0.2736
Rej. rate	0.0370	0.0350	0.0370	0.0610	0.0690	0.0650

5 An empirical application: returns to schooling

We now apply **msreg** to a version of Mincer (1974) wage equation. We consider the following wage regression

$$\begin{aligned}\texttt{log(wage)} = \beta_0 + \beta_1\texttt{expr} + \beta_2\texttt{expr}^2 + \beta_3\texttt{kww} + \beta_4\texttt{educ} + \beta_5\texttt{feduc} + \beta_6\texttt{meduc} \\ + \beta_7\texttt{smsa} + \beta_8\texttt{south} + \beta_9\texttt{black} + u, \end{aligned} \tag{3}$$

where `expr` is years of experience, `educ` is years of education, `kww` is Knowledge of World of Work test score, `feduc` and `meduc` are years of father's and mother's education, `black`, `smsa` and `south` are dummy variables to indicate if an individual is black, lives in an urban area and in the south, respectively.

We can estimate regression (3) using only the dataset `card.dta` from Card (1995), as in the benchmark OLS result below.

The estimation result is stored as `ols`.

Nonetheless, we pretend that the variable `kww` is missing in this dataset. In accordance with this scenario, we employ yet another dataset `wage2.dta` from Blackburn and Neumark (1992). The dataset contains six variables `educ`, `feduc`, `meduc`, `smsa`, `south`, and `black` other than `kww`. All six variables are used as matching variables to impute the missing `kww`, where `educ`, `feduc` and `meduc` are assumed to be continuous. Our aim is to see how the estimation result of (3) changes if `kww` is imputed from `wage2.dta`.

We use the default `vi` type covariance estimation assuming the sample size ratio between the two datasets converge to a nonzero constant. We use a third order polynomial in power-series estimation to remove the second order bias. The estimation result is stored as `twostep_vi`.

We can now compare these two estimation results.

The first column shows the benchmark OLS results. The signs of `expr`, $\texttt{expr}^2$, `kww`, `educ` are as expected and the estimates are significant at the 5% level.

The second column shows the results from the two-step bias-corrected estimator with the default `vi` type covariance. All the point estimates have the same sign as in the OLS benchmark. However, the coefficient on `kww` is insigificant due to a large stardard error.

6 Conclusion

In this paper, we describe a new Stata command, **msreg**, which implements two estimators proposed in HP18. The command allows practitioners to obtain consistent

estimators of linear regression models after imputing missing regressors via the NNM. We illustrate the use of **msreg** through a numerical example and an empirical application.

7 References

Abadie, A., and G. W. Imbens. 2006. Large sample properties of matching estimators for average treatment effects. *Econometrica* 74(1): 235–267.

———. 2011. Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics* 29(1): 1–11.

Blackburn, M., and D. Neumark. 1992. Unobserved Ability, Efficiency Wages, and Interindustry Wage Differentials. *Quarterly Journal of Economics* 107(4): 1421–1436.

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

Hirukawa, M., and A. Prokhorov. 2018. Consistent estimation of linear regression models using matched data. *Journal of Econometrics* 203(2): 344–358.

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(3): 599–631.

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

von Neumann, J. 1941. Distribution of the ratio of the mean square successive difference to the variance. *Annals of Mathematical Statistics* 12(4): 367–395.

Phillips, P. C., and B. E. Hansen. 1990. Statistical inference in instrumental variables regression with I (1) processes. *Review of Economic Studies* 57(1): 99–125.

Yatchew, A. 1997. An elementary estimator of the partial linear model. *Economics Letters* 57(2): 135–143.

Appendix

Theorems 2 and 3 give the asymptotic distributions of $\widehat{\theta}_{\mathbf{MSII}}$ and $\widehat{\theta}_{\mathbf{MSII-FM}}$, respectively. The covariance matrix V_I, and V_{II} and V_{III} depend on the definitions of Ω, Ω_{11A}, and Ω_{22}. We define Ω, Ω_{11A}, and Ω_{22} as follows.

$$\Omega = \Omega_{11A} + \kappa \left[(\beta_2' \Sigma_2 \beta_2) \mathbf{E}(\mathbf{w}) \mathbf{E}(\mathbf{w})' + \frac{1}{K^2} \mathbf{d}iag \left\{ 0_{(d_1+1)\times(d_1+1)}, (\beta_2' \Sigma_2 \beta_2) V_{g2} + \frac{1}{2}\Psi, 0_{d_z \times d_z} \right\} \right],$$

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

$$\Omega_{22} = \frac{1}{K^2} \mathbf{d}iag \left\{ 0_{(d_1+1)\times(d_1+1)}, \Xi + \frac{1}{2}\Psi, 0_{d_z \times d_z} \right\},$$

$$V_{g_2} = Var\left\{ g_2(z) \right\},$$

$$\Xi = \mathbf{E}\left\{ \left(\eta_2 \eta_2' - \Sigma_2 \right) \beta_2 \beta_2' \left(\eta_2 \eta_2' - \Sigma_2 \right) \right\}, and$$

$$\Psi = (\beta_2' \Sigma_2 \beta_2) \Sigma_2 + \Sigma_2 \beta_2 \beta_2' \Sigma_2.$$

We present consistent estimators of Ω_{11A}, Ω_{22} and Ω for **MSII** below. Because **MSII-FM** is first-order asymptotically equivalent to **MSII** as documented in Theorem 3, simply replacing $\widehat{\theta}_{\mathbf{MSII}}$ in $\widehat{\Omega}_{11A}$, $\widehat{\Omega}_{22}$ and $\widehat{\Omega}$ with $\widehat{\theta}_{\mathbf{MSII-FM}}$ yields the estimators for **MSII-FM**.

$$\widehat{\Omega}_{11A} = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{w}_{i,j(i)} \widehat{\epsilon_{i,j(i)}} + \frac{1}{K} \widehat{\Sigma} \widehat{\theta}_{\mathbf{MSII}} \right) \left(\mathbf{w}_{i,j(i)} \widehat{\epsilon_{i,j(i)}} + \frac{1}{K} \widehat{\Sigma} \widehat{\theta}_{\mathbf{MSII}} \right)',$$

$$\widehat{\Omega}_{22} = \frac{1}{K^2} \mathbf{d}iag \left\{ 0_{(d_1+1)\times(d_1+1)}, \widehat{\Gamma}(-1) + \widehat{\Gamma}(0) + \widehat{\Gamma}(1), 0_{d_z \times d_z} \right\}, \text{and}$$

$$\widehat{\Omega} = \widehat{\Omega}_{11A} + \frac{n}{m} \Bigg\{ \left(\widehat{\beta_{2,\mathbf{MSII}}}' \widehat{\Sigma}_2 \widehat{\beta_{2,\mathbf{MSII}}} \right) \bar{\mathbf{w}} \bar{\mathbf{w}}'$$

$$+ \frac{1}{K^2} \mathbf{d}iag\{ 0_{(d_1+1)\times(d_1+1)}, \widehat{\beta_{2,\mathbf{MSII}}}' \widehat{\Sigma}_2 \widehat{\beta_{2,\mathbf{MSII}}} \widehat{V}_{g_2} + \widehat{\Gamma}(0) - \left[\widehat{\Gamma}(-1) + \widehat{\Gamma}(1) \right], 0_{d_z \times d_z} \} \Bigg\},$$

where $\widehat{\epsilon_{i,j(i)}} = y_i - \mathbf{w}_{i,j(i)}' \widehat{\theta}_{\mathbf{MSII}}$, $\widehat{\beta_{2,II}}$ is the **MSII** estimate of β_2, $\widehat{\Gamma}(l)$ is the lth sample auto-covariance of $\left\{ \Delta \mathbf{x}_{2j} \Delta \mathbf{x}_{2j}'/2 - \widehat{\Sigma} \right\}$, i.e.,

$$\widehat{\Gamma}(l) = \frac{1}{m-1} \sum_{j=max\{2,2+l\}}^{min\{m,m+l\}} \left(\frac{\Delta \mathbf{x}_{2j} \Delta \mathbf{x}_{2j}'}{2} - \widehat{\Sigma}_2 \right) \widehat{\beta_{2,\mathbf{MSII}}} \widehat{\beta_{2,\mathbf{MSII}}}' \left(\frac{\Delta \mathbf{x}_{2j-l} \Delta \mathbf{x}_{2j-l}'}{2} - \widehat{\Sigma}_2 \right),$$

$$\bar{\mathbf{w}} = \begin{bmatrix} 1 \\ \bar{\mathbf{x}}_1 \\ \bar{\mathbf{x}}_2 \\ \bar{\mathbf{z}} \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_{1i} \\ \frac{1}{m} \sum_{j=1}^{m} \mathbf{x}_{2j} \\ \frac{1}{N} \sum_{i=1}^{N} \mathbf{z}_i \end{bmatrix}, \text{and}$$

$$\widehat{V}_{g_2} = \frac{1}{m-1} \sum_{j=1}^{m} (\mathbf{x}_{2j} - \bar{\mathbf{x}}_{2j})(\mathbf{x}_{2j} - \bar{\mathbf{x}}_{2j})' - \widehat{\Sigma}_2.$$

About the authors

Masayuki Hirukawa is Professor in the Faculty of Economics of Ryukoku University in Japan. Research for this paper was supported by a grant from Japan Society for the Promotion of Science (Project No. 19K01595).

Di Liu is Senior Econometrician in StataCorp in the United States.

Artem Prokhorov is Professor in the Displine of Business Analytics of the University of Sydney in Australia. Research for this paper was supported by a grant from the Russian Science Foundation (Project No. 20-18-00365).