Estimation and inference using imperfectly matched data

Rachel S. Anderson* Bo E. Honoré† Adriana Lleras-Muney‡

At Most one More

August 13, 2019

Abstract

This paper studies estimation and inference in standard econometric models that use matched data sets with multiple matches of the dependent variable. We show that it is straightforward to use multiple matches provided that the true match is included among the potential matches. On the other hand, identification is generally not possible if the true match is not included. We also investigate the possibility of using information about the quality of a match. The main result here is that if the probability of a correct match is only approximate then using the match quality will lead to inconsistent estimators although we also illustrate that bias-variance tradeoffs can justify using approximate match information informations.

1 Introduction

The recent availability of large administrative data sets has increased the use of matched data in recent years in economic applications (Chetty 2012 not in references). When matching data sets, there will be three possible outcomes, the consequences of

*Mailing Address: Department of Economics, Julis Romo Rabinowitz Building, Princeton, NJ 08544. Email: rachelsa@Princeton.edu.

†Mailing Address: Department of Economics, Julis Romo Rabinowitz Building, Princeton, NJ 08544. Email: honore@Princeton.edu.

‡Mailing Address: UCLA.

which we need to account for in the analysis. Some observations will not be matched to an outcome and thus will result in missing data (recipients for whom no outcomes can be found). Among those that are matched, it is not always possible to find a unique match, thus for some individuals we will have multiple matches/outcomes. Finally there might measurement error in matching: even when find a unique match per individual, it might not correspond to the correct match.

Many studies by economic historians and historical demographers have employed record linkage techniques. For example, **?** link Norwegian-born from the Norwegian census of 1865 to the 1900 U.S. Census of Population and the 1900 Norwegian census, finding only 26% of those observed in the base year. **?** link males born 1895-1900 from the 5% IPUMS sample of the 1900 U.S. Census of Population to the Social Security Death Master File, and successfully link 29% of those sought. **?** are able to link 21% of black, Southern-born males from the 1% IPUMS sample of the 1910 U.S. Census of Population to the 1930 U.S. Census of Population.

There are several prominent recent examples that match administrative data bases to do program evaluation. The Moving to Opportunity experiment, offering vouchers to poor families to move to low-poverty areas, was evaluated by matching data on participants to many administrative data sets including state UI data, state AFDC/TANF/Food stamps data, juvenile and criminal justice administrative data, National Student Clearinghouse data on college going, school data among others (**?**, **?**). **?** look at the long-run effects of smaller classes and better teachers, by matching the original Tennessee Project STAR experimental data to later IRS administrative tax records of the children when they are adults. **?** match 2.5M NYC public school records to IRS data to look at the long-run impacts of teacher value added. **?** evaluate the effect of the Harlem Children's Zone by matching the HCZ data to the National Student Clearinghouse college enrollment data.

There are also many publicly available and commonly used datasets that are constructed by matching administrative data sets, for instance the Linked Birth and Infant Death data (matches birth certificates and death certificates from population databases), the IPUMS Linked Representative Samples, 1850-1930 (matches indi-

viduals from census to census to create a panel data), the National Longitudinal Mortality Survey (matches CPS data to death certificates from the National Death Index, or NDI), the National Health Interview survey Linked Mortality File (NHIS survey linked to NDI), the General Social Survey-National Death Index (GSS-NDI) and National Health and Nutrition Examination Survey Linked Mortality Files (I, II, and III). Even in these high quality data sets, containing social security numbers, there are a non-trivial number of multiple matches and measurement error.[1]

The standard approach in the existing literature that uses matched data consists of dropping any observation without a match or with multiple matches, and to treat unique matches as correct. Inference and estimation then proceed by treating the data as if it came from a single source. In this paper we investigate how to best use matched data by a) incorporating multiple matches and b) allowing for measurement error in matching.

2 General problem

We are interested in estimating a model of the general form:

$$E_0 \left[m \left(y_i, x_i, z_i; \theta_0 \right) \right] = 0 \tag{1}$$

where y_i, x_i and z_i are vectors or scalars of data for an individual i. We will typically think of y_i as the dependent variable (such as earnings, disability or age at death), and (x_i, z_i) as a set of individual-level covariates or instruments. The function $m \left(y_i, x_i, z_i; \theta_0 \right)$ is known. θ_0 is the parameter of interest. The expectation E_0 is taken with respect to the joint distribution of the data $f_0(y, x, z)$.

The complication of the estimation problem we consider arises because the variables y_i, x_i and z_i are not observed in a single data set. Instead we have access to two different data sets. The first (the y–dataset) contains y_i and another vector of characteristics, w_i. The other (the x–dataset) contains x_i, z_i and w_i. The vector w_i

[1] need a reference

can be used to match observations across data sets, but not always uniquely – in other words w_i is not a unique person identifier. Specifically, for each observation, i, in the x–dataset there will be L_i "matching" observations in the y–dataset with identical w_i. In the applications that we have in mind, the y–dataset will in principle be the population. This implies that one of the L_i matched y's will correspond to observation i and the remaining $(L_i - 1)$ y's are mismatches. This will be our leading case, but we will also discuss extension to the case where it is possible that none of the matches is correct.

For example **?** match individuals in the experiment to hospital discharge records, credit reports and mortality to look at the impact of offering Medicaid on health care use, bankrupcy, and health. In this case the x–dataset contains then list of individuals that were randomly assigned to Medicaid. The y–dataset is one of the administrative data sets, for instance the hospital discharge records. The data sets are matched based on name and date of birth which are known in both data (w_i). In this example (and in all others cited above), the administrative data contain the information for the population. In principle all individuals in the x–dataset should be found in the y–dataset, but they are not, because name and date of birth do not uniquely identify individuals, and because of record errors (mispelled last names for instance).

3 GMM

Consider an individual from the x–dataset with L_i matches in the y–dataset. The data for that observation is $\left(x_i, z_i, \{y_{i\ell}\}_{\ell=1}^{L_i}, w_i\right)$. We start by considering the case where the correct match is in the set of matches and where every element in that set is equally likely to be the correct match. For instance consider the case where we observe an individual "Joe Smith" in the x–dataset born on January 3, 1984. In the y–dataset we find two male individuals born on the same date with names "Joe A. Smith" and "Joseph B. Smith". In the absence of additional information we assume that each of the two matches is equally likely to be the correct match. For a fixed

matching criteria, the number of matches L_i is a random variable.

Since we are matching observations on the basis of w_i, we will assume that the y–dataset and the x–dataset are random samples conditional on w_i (and L_i?) More formally, we make the following assumptions.

Assumption 1. The observed (x_i, z_i, w_i) is a random sample drawnfrom the marginal distribution $f_0(x, z|w)$. The $\{y_{i\ell}\}_{\ell=1}^{L_i}$ is a random sample drawn from $f_0(y|w)$.

Assumption 1 rules out unobserved sample selection, in the sense that all individuals with the same identifying information have equal probability of appearing in the sample. This assumption would be violated if, for example, higher income individuals have a greater probability of appearing in the sample, unless w_i includes income.

Assumption 2. There is exactly one $y_{i\ell}$ that is drawn from $f_0(y|x_i, z_i)$for all i. That is, we assume that the y–dataset contains the true outcome for each observation in the x–dataset.

This assumption implies we can write $y_i = \sum_{\ell=1}^{L_i} s_{i\ell} y_{i\ell}$, where$s_{i\ell}$ is an unobserved latent variable that equals 1 if $(y_{i\ell}, x_i, z_i)$ is drawn from $f_0(y, x, z)$, and that equals 0 otherwise. Since $\sum_{\ell=1}^{L_i} s_{i\ell} = 1$ for all i, we can rewrite (1) as,

$$E_0\left[m(y_i, x_i, z_i, \theta_0)\right] = 0 \iff E\left[m(y_{i\ell}, x_i, z_i; \theta_0)|s_{i\ell} = 1\right] = 0 \tag{2}$$

which is an expectation with respect to the DGP that produces the x–dataset.

Assumption 3. The identifying variables w_i in (x_i, z_i, w_i) are such that the researcher behaves as if

$$P(s_{i\ell} = 1|w_i, L_i) = \frac{1}{L_i}$$

hence, all matches are equally likely to be drawn from $f_0(y|x, z)$.

Dropping z_i for ease of exposition, observe that by the Law of Total Probability

and Assumption 3,

$$E\left[\sum_{\ell=1}^{L_i} m\left(y_{i\ell}, x_i; \theta\right) \middle| w_i, L_i, \right] = \sum_{\ell=1}^{L_i} \Big\{ E[m(y_{i\ell}, x_i; \theta)|w_i, L_i, s_{i\ell} = 1] P(s_{i\ell} = 1|w_i, L_i) + E[m(y_{i\ell}, x_i; \theta)|w_i, L_i, s_{i\ell} = 0] P(s_{i\ell} = 0|w_i, L_i) \Big\}$$
$$= \frac{1}{L_i} \sum_{\ell=1}^{L_i} E[m(y_{i\ell}, x_i; \theta)|w_i, L_i, s_{i\ell} = 1] + \frac{L_i - 1}{L_i} E[m(y_{i\ell}, x_i; \theta)|w_i, L_i, s_{i\ell} = 0]$$

Under random sampling, the expectations are equal for all i,

$$E\left[\sum_{\ell=1}^{L_i} m\left(y_{i\ell}, x_i; \theta\right) \middle| w_i, L_i, \right] = E\left[m(y_{i\ell}, x_i; \theta)|w_i, L_i, s_{i\ell} = 1\right] + (L_i - 1) E\left[m(y_{i\ell}, x_i; \theta)|w_i, L_i, s_{i\ell} = 0\right]$$

Rearranging terms,

$$E[m(y_{i\ell}, x_i; \theta)|w_i, L_i, s_{i\ell} = 1] = E\left[\sum_{\ell=1}^{L_i} m(y_{i\ell}, x_i; \theta) \middle| w_i, L_i\right] - (L_i - 1) E[m(y_{i\ell}, x_i; \theta)|w_i, L_i, s_{i\ell} = 0]$$

Note that $E[m(y_{i\ell}, x_i; \theta)|w_i, L_i, s_{i\ell} = 0]$ is an expectation with respect to the distribution of false matches. In this case, $y_{i\ell}$ is a draw from $f_0(y|w)$ that is independent of x_i, so that

$$E[m(y_{i\ell}, x_i; \theta)|w_i, L_i, s_{i\ell} = 0] = E[m(y_{i\ell}, x_i; \theta)|w_i, L_i, x_i, s_{i\ell} = 0] \equiv g(w_i, L_i, x_i; \theta)$$

which is equal to $\int m(y, x_i, \theta) f_0(y|w, L)$ and can be approximated numerically by replacing f_0 with a kernel or sieve estimator. Finally, by the Law of Iterated Expectations,

$$E[m(y_{i\ell}, x_i; \theta)|s_{i\ell} = 1] = E\left[\sum_{\ell=1}^{L_i} m(y_{i\ell}, x_i; \theta)\right] - E[(L_i - 1) g(w_i, L_i, x_i, \theta)] \quad (3)$$

where the left hand side is equal to the moment conditions in (1). This result suggests

that we can estimate θ_0 in (1) by replacing the moment conditions with an estimate (3). Indeed, if we use

$$\hat{E}\left[m(y_{i\ell}, x_i; \theta)|s_{i\ell} = 1\right] = \frac{1}{n}\sum_{i=1}^{n}\sum_{\ell=1}^{L_i} m(y_{i\ell}, x_i; \theta) - \frac{1}{n}\sum_{i=1}^{n}(L_i - 1)\hat{g}(w_i, L_i, x_i; \theta) \quad (4)$$

where n is the number of observations in the (x_i, z_i, w_i) file, and $\hat{g}$ is a parametric or nonparametric estimate of $g(\cdot)$, then the usual GMM estimator is consistent (see Appendix REF HERE).

[remark: this case covers multiple xs in addition to many ys–compare to WP by Mahajan or Poirer)

3.1 Linear Instrumental Variables Estimation

Consider the text–book linear instrumental variables model

$$y_i = x_i'\beta + \varepsilon_i \qquad E\left[z_i\varepsilon_i\right] = 0$$

or

$$E\left[z_i\left(y_i - x_i'\beta\right)\right] = 0 \quad (5)$$

or

$$E\left[z_i y_i\right] = E\left[z_i x_i'\right]\beta$$

In this case (**??**) becomes

$$\begin{aligned} E\left[z_i\left(y_i - x_i'b\right)\right] &= \sum_{\ell} E\left[z_i\left(y_{i\ell} - x_i'b\right)\right] - E\left[\left(L_i - 1\right)z_i\left(g\left(w_i\right) - x_i'b\right)\right] \\ &= E\left[z_i\left(\left(\sum_{\ell} y_{i\ell} - \left(L_i - 1\right)g\left(w_i\right)\right) - x_i'b\right)\right] \end{aligned}$$

where $g\left(w_i\right) = E\left[y_i|\, w_i\right]$. In other words, β satisfies the moment condition

$$E\left[z_i\left(\left(\sum_{\ell} y_{i\ell} - \left(L_i - 1\right)g\left(w_i\right)\right) - x_i'\beta\right)\right] = 0 \quad (6)$$

When the model is just–identified this gives the explicit expression for β

$$\beta = E\left[z_i x_i'\right]^{-1} E\left[z_i\left(\sum_{\ell} y_{i\ell} - (L_i - 1)\, g\left(w_i\right)\right)\right] \tag{7}$$

When $z_i = x_i$ and $L_i = 1$, this is the OLS estimator in a regression of y_i on x_i.

When the model is over–identified ($\dim\left(z_i\right) = m > k = \dim\left(x_i\right)$), we have

$$\beta = \left(PE\left[z_i x_i'\right]\right)^{-1} PE\left[z_i\left(\sum_{\ell} y_{i\ell} - (L_i - 1)\, g\left(w_i\right)\right)\right] \tag{8}$$

for any $k \times m$-matrix such that $E\left[Pz_i x_i'\right]^{-1}$exists. The choice of P is equivalent to the choice of weighting matrix in GMM. For example $P = E\left[x_i z_i'\right]\, E\left[z_i z_i'\right]^{-1}$ yields the usual 2SLS estimator when $L_i = 1$ for all i.

We now turn to the problem of converting (7) and (8) into estimators. The first complication in this is that $g\left(\cdot\right)$ must be estimated. In many applications, the y–dataset will be much, much larger than the x–dataset and it is then reasonable to treat $g\left(\cdot\right)$ as if it is known. In other cases, we will explicitly think of $g\left(\cdot\right)$ as an object to be estimated. The second complication is that (5) is often thought of as an implication of the conditional moment condition

$$E\left[\left(y_i - x_i'\beta\right)\middle|\, z_i\right] = 0. \tag{9}$$

In this case there is room for improving efficiency by weighing different observations differently when forming sample analogues to expectations.

To fix ideas, consider first the case where $g\left(\cdot\right)$ is known and the starting point is (5). In that case the optimal GMM estimator is

$$\left(\left(\sum x_i z_i'\right) W \left(\sum x_i z_i'\right)'\right)^{-1} \left(\sum x_i z_i'\right) W \left(\sum x_i y_i\right)$$

where

$$W = E\left[\left(\left(\sum_{\ell} y_{i\ell} - (L_i - 1)\, g\left(w_i\right)\right) - x_i'b\right)^2 z_i z_i'\right]^{-1}$$

Let $\nu_{i\ell} = y_{i\ell} - g\left(w_i\right)$ then W can be written as

$$W = E\left[\left(\varepsilon_i + \sum \nu_{i\ell}\right)^2 z_i z_i'\right]^{-1}$$

where the sum if over the $L_i - 1$ incorrect matches.

If $E\left[\nu_{i\ell}^2 \middle| z_i, L_i\right] = \sigma_\nu^2$ and $E\left[\varepsilon_i^2 \middle| z_i, L_i\right] = \sigma_\varepsilon^2$ then (assuming independence)

$$W = E\left[\left(\sigma_\varepsilon^2 + (L_i - 1)\,\sigma_\nu^2\right) E\left[z_i z_i' \middle| L_i\right]\right]^{-1}$$

If in addition $E\left[\varepsilon_i \middle| z_i, L_i\right] = 0$ then

$$\begin{aligned}\left(\sum_{\ell} y_{i\ell} - (L_i - 1)\, g\left(w_i\right)\right) &= x_i'b + \varepsilon_i + \sum \nu_{i\ell} \\ &= x_i'b + u_i\end{aligned}$$

with $E\left[u_i \middle| z_i, L_i\right] = 0$ and $V\left[u_i \middle| z_i, L_i\right] = (\sigma_\varepsilon^2 + (L_i - 1)\,\sigma_\nu^2)$ and the efficient estimator (check and reference) is... weighted 2sls... When $z_i = x_i$ the optimal estimator of β then is the weighted least squares estimator

$$\left(\sum_i \frac{1}{(\sigma_\varepsilon^2 + (L_i - 1)\,\sigma_\nu^2)} x_i x_i'\right)^{-1} \left(\sum_i \frac{1}{(\sigma_\varepsilon^2 + (L_i - 1)\,\sigma_\nu^2)} x_i \left(\sum_{\ell=1}^{L_i} y_{i\ell} - (L_i - 1)\, g\left(w_i\right)\right)\right)$$

where σ_ε^2 and σ_ν^2 can be replaced by consistent estimators.

We next consider the case where $g\left(\cdot\right)$ is parameterized, so we write $g\left(w_i\right) = g\left(w_i; \alpha\right)$ and estimate α by some standard estimator that can be written as a solution to a moment condition. For simplicity, we first assume that it is estimated from the sample of matches $\left\{\left\{y_{i\ell}\right\}_{\ell=1}^{L_i}, w_i\right\}_{i=1}^{n}$, but it could also be estimated from a larger sample.

There are then two cases to consider. One where we first estimate α and then

subsequently β, and one where α and β are potentially estimated jointly. In the former case, we stack the sample moment conditions determining $\widehat{\alpha}$ and the moment conditions that determine $\widehat{\beta}$:

$$\begin{pmatrix} \frac{1}{n}\sum_i \left(\sum_\ell \rho\left(y_{i\ell}, w_i, \widehat{\alpha}\right)\right) \\ \frac{1}{n}\sum_i Pz_i \left(\left(\sum_\ell y_{i\ell} - (L_i - 1)\, g\left(w_i; \widehat{\alpha}\right)\right) - x_i'\widehat{\beta}\right) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

These deliver the asymptotic distribution of $\left(\widehat{\alpha}, \widehat{\beta}\right)$. If one allows α and β to be estimated jointly, then one can simply consider a GMM estimator of (α, β) based on the moment conditions

$$\begin{pmatrix} E\left[\sum_\ell \rho\left(y_{i\ell}, w_i, \alpha\right)\right] \\ E\left[z_i\left(\left(\sum_\ell y_{i\ell} - (L_i - 1)\, g\left(w_i; \alpha\right)\right) - x_i'\beta\right)\right] \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

When $g(\cdot)$ is estimated nonparametrically for a (much) larger sample than the x–data set, the standard erros for $\widehat{\beta}$ can be calculated as in **?**. See also **?**

3.2 Maximum Likelihood Estimation

We next turn to estimation of parametric nonlinear model which would typically be estimated by maximum likelihood estimation if there were no multiple matches.

One way to approach this is to think of the first order condition for maximum likelihood as a moment condition and then proceed as above. Alternatively, one might think of maximum likelihood estimation in the presence of multiple matches.

Using the same notation as earlier, assume that an observation consists of $\left(x_i, \{y_{i\ell}\}_{\ell=1}^{L_i}, w_i\right)$. For $\ell = 1, ..., L$, there is probability $\frac{1}{L_i}$ that $y_{i\ell}$ is drawn from the parametric model, $f\left(y; \theta \middle| x_i\right)$; in this case, y_{ik} is drawn from some pre-estimated "reduced form" $g\left(y \middle| w_i\right)$ for $k \neq \ell$.

This gives the likelihood function

$$\sum_{\ell}\left\{\frac{1}{L_i}f\left(y_{i\ell};\theta|\,x_i\right)\prod_{k\neq\ell}\,g\left(y_{ik}|\,w_i\right)\right\}$$
$$=\prod_{k}\,g\left(y_{ik}|\,w_i\right)\left(\sum_{\ell}\frac{1}{L_i}\frac{f\left(y_{i\ell};\theta|\,x_i\right)}{g\left(y_{i\ell}|\,w_i\right)}\right)$$

So except for a constant, the (pseudo) log–likelihood function is

$$\sum_{i}\ln\left(\sum_{\ell}\frac{1}{L_i}\frac{f\left(y_{i\ell};\theta|\,x_i\right)}{g\left(y_{i\ell}|\,w_i\right)}\right) \tag{10}$$

The trick is coming up with a good g. This will have to be application–specific.

The asymptotic properties of the estimator defined by maximizing (10) will again depend on how one thinks of g. In some cases, the "y"-dataset will be so large relative to the "x"–dataset that it is reasonable to consider g known. In that case the maximizer of (10) is a standard maximum likelihood estimator. In other cases, one will want to account for the fact that g has been estimated (parametrically or nonparametrically).

3.2.1 Relationship to methods of moments (optional)

The first order condition for maximizing (10) is

$$\sum_{i}\left(\sum_{\ell}\frac{f\left(y_{i\ell};\theta|\,x_i\right)}{g\left(y_{i\ell}|\,w_i\right)}\right)^{-1}\sum_{\ell}\frac{f'\left(y_{i\ell};\theta|\,x_i\right)}{g\left(y_{i\ell}|\,w_i\right)}=0. \tag{11}$$

By comparison, the estimator defined by (**??**) with $m()$ the derivative of f with respect to θ would solve

$$\sum_{i}\left(\sum_{\ell}f'\left(y_{i\ell};\theta|\,x_i\right)-\left(L_i-1\right)E\left[f'\left(y;\theta|\,x_i\right)|\,w\right]\right)=0 \tag{12}$$

It seems that this is different from (11).

4 Possible Generalizations

4.1 Using information about the match quality

When presented with multiple matches, the researcher will often have information about which of the matches is most likely to be correct. In this section, we argue that using this information can lead to biases unless it is possible to consistently estimate the probability that each match is correct. We make this pooint by considering a very simple example.

Suppose we want to estimate a mean, $\mu = E[X]$. For each i, we have two observations, X_{1i} and X_{2i}. One is drawn from the correct distribution which has mean μ and variance σ^2 and one is drawn from a known incorrect distribution with *known* mean κ and variance ω^2.

Suppose that the probability that the first is drawn from the correct correct distribution is π. Then

$$\begin{aligned} E[X_{1i}] &= \pi\mu + (1-\pi)\kappa \\ E[X_{2i}] &= \pi\kappa + (1-\pi)\mu \end{aligned}$$

so

$$E[X_{1i} + X_{2i}] - \kappa = \pi(\mu + \kappa) + (1-\pi)(\kappa + \mu) - \kappa = \mu$$

It therefore follows that

$$\frac{1}{n}\sum_{i=1}^{n}(X_{1i} + X_{2i}) - \kappa$$

is a consistent estimator of μ.

More generally consider an estimator of the form

$$\widehat{\mu} = \frac{a_1}{n}\sum_{i=1}^{n} X_{1i} + \frac{a_2}{n}\sum_{i=1}^{n} X_{2i} - a_3\kappa \tag{13}$$

Its mean would be

$$E\left[\widehat{\mu}\right] = \pi\left(a_1\mu + a_2\kappa\right) + \left(1-\pi\right)\left(a_1\kappa + a_2\mu\right) - a_3\kappa \tag{14}$$

$$= \left(\pi a_1 + \left(1-\pi\right)a_2\right)\mu + \left(\pi a_2 + \left(1-\pi\right)a_1 - a_3\right)\kappa$$

For unbiasedness, we then need

$$\left(\pi a_1 + \left(1-\pi\right)a_2\right) = 1 \tag{15}$$

or

$$a_2 = \frac{1-\pi a_1}{1-\pi} = \frac{1}{1-\pi} - \frac{\pi}{1-\pi}a_1$$

and

$$\left(\pi a_2 + \left(1-\pi\right)a_1 - a_3\right) = 0 \tag{16}$$

The only way to do this without knowing π is to set $a_1 = a_2$. But in that case (16) inplies that $a_1 = a_2 = 1$.

If we know π then (15 and (16) can be solved for a_2 and a_3 as a function of a_1:

[here insert numerical example]

REMARK: By Oct 13, have a numerical example.

Also think about this: if we postulate a π_1 , calculate the bias and variance of various estimators when $\pi_1 \neq \pi$.

If possible do this for OLS as well

4.2 Mean-Variance Trade-Off

Simple formulas and numerical illustration

Perhaps also try to do explicit calculations for OLS

5 Allowing for the correct match to not be included

5.1 This section will argue that identification is not possible in this case

We next consider the case where there is some probability that none of the observations is drawn from the distribution of interest. In the motivation setup, this corresponds to the case where none of the matches is the correct one.

Suppose again that we want to estimate a mean, $\mu = E[X]$. For each i, we have two observations, X_{1i} and X_{2i}. With probability π_j, X_{ji} is drawn from the correct distribution which has mean μ and variance σ^2 and the other is drawn from a known incorrect distribution with *known* mean κ and variance ω^2. With probability $1-\pi_1-\pi_2$, X_{1i} and X_{2i} are drawn independently from the known incorrect distribution with *known* mean κ and variance ω^2.

Consider an estimator of the (natural) form

$$\widehat{\mu} = \frac{a_1}{n}\sum_{i=1}^{n} X_{1i} + \frac{a_2}{n}\sum_{i=1}^{n} X_{2i} - a_3\kappa \tag{17}$$

Its mean would be

$$\begin{aligned} & a_1(\pi_1\mu + (1-\pi_1)\kappa) + a_2(\pi_2\mu + (1-\pi_2)\kappa) - a_3\kappa \\ = \; & (a_1\pi_1 + a_2\pi_2)\mu + (a_1(1-\pi_1) + a_2(1-\pi_2) - a_3)\kappa \end{aligned} \tag{18}$$

Without knowledge of π_1 and π_2 is is impossible to choose a_1 and a_2 such that this is always equal to μ. This would be true even if we knew that $\pi_1 = \pi_2$.

On the other hand, if we know π_1 and π_2 then we could construct a class of unbiasedness estimators (only) by setting

$$a_2 = \frac{1 - a_1\pi_1}{\pi_2} \qquad \text{and} \qquad a_3 = a_1(1-\pi_1) + a_2(1-\pi_2)$$

(I realize that I am going off on a tangent here, but...) We can restate (18) as

$$E\left[a_1\overline{X}_1 + a_2\overline{X}_2\right] = (a_1\pi_1 + a_2\pi_2)\,\mu + (a_1\,(1-\pi_1) + a_2\,(1-\pi_2))\,\kappa$$

or if we assume that $\pi_1 = \pi_2$ and $a_1 = a_2 = \frac{1}{2}$ (the scale of the a's is irrelevant and at this point it seems meaningless to distinguish between $\overline{X}_1$ and $\overline{X}_2$)

$$E\left[\overline{X}\right] - (1-\pi)\,\kappa = \pi\mu$$

or

$$\mu = \frac{E\left[\overline{X}\right] - (1-\pi)\,\kappa}{\pi}$$

The derivative of this with respect to π is

$$\frac{\pi\kappa - E\left[\overline{X}\right] + (1-\pi)\,\kappa}{\pi^2} = \frac{\kappa - E\left[\overline{X}\right]}{\pi^2}$$

This is monotone, so if we know that $\pi \in [\pi^o, 1]$, then we know that μ must be between $E\left[\overline{X}\right]$ and $\frac{E\left[\overline{X}\right]-(1-\pi^o)\kappa}{\pi^o}$.

Note that

- Allowing different a_1 and a_2 should not provide more information.
- There seems to be no scope for identifying π

We next turn to the case where different observations have different number of matches. In some (unrealistic) cases, this may sharpen the bound above...

Suppuse that for $\ell = 1, .., L$, each of $X_{i\ell}$ has the distribution of interest with probability π_L/L (these are mutually exclusive). Let

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n}(X_{i1} + ... + X_{iL})$$

the

$$E\left[\overline{X}\right] = \pi_L\mu + (L-1)\,\kappa + (1-\pi_L)\,\kappa \tag{19}$$

or

$$\mu = \frac{E\left[\overline{X}\right] - (L-1)\kappa - (1-\pi_L)\kappa}{\pi_L}$$

This is monotone in π_L so if we know that $\pi_L \in [\pi_L^o, 1]$ then μ must be between $E\left[\overline{X}\right]$ and $\frac{E\left[\overline{X}\right]-(L-1)\kappa-\left(1-\pi_L^o\right)\kappa}{\pi_L^o}$.

If we know how π_L varies with L then identification may be possible. But not always. Suppose, for example that π_L is constant, then (19) with $L = 2$ and 3 yields

$$\begin{aligned} E_{L=2}\left[\overline{X}\right] &= \pi\mu + \kappa + (1-\pi)\kappa \\ E_{L=3}\left[\overline{X}\right] &= \pi\mu + 2\kappa + (1-\pi)\kappa \end{aligned}$$

Thought of as functions of $(\pi\mu)$ and π these are two linear equations in two unknowns... But they are collinear...

5.2 Linear IV

We now return to the setup in section 3.1. Suppose that there are L_i matched y's for x_i. There is probability π_i that one of them is teh correct match. The others are drawn from the distribution of y connditional on w_i

Suppose we know π_i. We can then write

$$\sum_\ell y_{i\ell} = (L_i - \pi_i)\, g(w_i) + \pi_i x_i'\beta + u_i \tag{20}$$

where u_i is potentially correleted with x_i but uncorreleated with the instrument z_i. We can rewrite to get

$$\sum_\ell y_{i\ell} - L_i g(w_i) = -\pi_i g(w_i) + \pi_i x_i'\beta + u_i \tag{21}$$

or

$$\gamma_i \sum_\ell y_{i\ell} - \gamma_i L_i g(w_i) + g(w_i) = x_i'\beta + \varepsilon_i$$

where $\gamma_i = 1/\pi_i$. 2sls applied to thsi yields

$$\left(\left(\sum x_i z_i'\right)\left(\sum z_i z_i'\right)^{-1}\left(\sum x_i z_i'\right)'\right)^{-1}\left(\sum x_i z_i'\right)\left(\sum z_i z_i'\right)^{-1} \left(\sum z_i\left(\gamma_i\left(\sum_\ell y_{i\ell} - L_i g\left(w_i\right)\right) + g\left(w_i\right)\right)\right)$$

Let $M = \left(\left(\sum x_i z_i'\right)\left(\sum z_i z_i'\right)^{-1}\left(\sum x_i z_i'\right)'\right)^{-1}\left(\sum x_i z_i'\right)\left(\sum z_i z_i'\right)^{-1}$ and $m_i = Mz_i$ (a column vector) the the 2SLS estimator is

$$\sum_i \gamma_i m_i\left(\sum_\ell y_{i\ell} - L_i g\left(w_i\right)\right) + \sum_i m_i g\left(w_i\right)$$

Suppose we are interested in the first first element of $\widehat{\beta}$. It can be written as

$$\widehat{\beta}_1 = \sum_i \gamma_i m_{1i}\left(\sum_\ell y_{i\ell} - L_i g\left(w_i\right)\right) + \sum_i m_{1i} g\left(w_i\right)$$

where m_{1i} is the first element of. Or

$$\widehat{\beta}_1 = \sum_i \gamma_i a_{1i} + a_1$$

where $a_{1i} = m_{1i}\left(\sum_\ell y_{i\ell} - L_i g\left(w_i\right)\right)$ and $a_1 = \sum_i m_{1i} g\left(w_i\right)$

Of course, in practive we dont actually know π (or $\gamma = \pi^{-1}$) except that we know that $\underline{\pi} \leq \pi \leq 1$ and hence $1 \leq \gamma \leq \overline{\gamma} = \underline{\pi}^{-1}$. Then the lower and upper bounds for $\widehat{\beta}_1$ are obtained by

$$\left[\sum_i 1\left\{a_{1i} > 0\right\} a_{1i} + \sum_i \overline{\gamma} 1\left\{a_{1i} < 0\right\} a_{1i} + a_1, \sum_i 1\left\{a_{1i} < 0\right\} a_{1i} + \sum_i \overline{\gamma} 1\left\{a_{1i} > 0\right\} a_{1i} + a_1\right]$$

This is the sample analog of the set of 2SLS estimands generated by all possible assignments π_i

5.3 Tradeoff between interval size and estimation precision when choosing to discard low-probability matches

5.4 Maximum Likelihood

6 Application

In this section we use data from **?** to illustrate. [2]

7 Conclusion

8 Appendix

[2]The data, mp_data.dta is posted as part of the publication: https://www.aeaweb.org/articles?id=10.1257%2Faer.20140529, and it provides multiple matches for a subset of the observations.