To Link or Not to Link? Estimating Long-run Treatment Effects from Historical Data

Francis J. DiTraglia[1] Camilo García-Jimeno[2] Ezra Karger[2]

[1]Department of Economics, University of Oxford

[2]Federal Reserve Bank of Chicago

December 14th, 2024

The Linking Problem

Aizer et al (2016; AER)

- Long-term effects of "Mothers' Pension Program" on adult outcomes
- Outcomes from 1940 U.S. Census; Treatment from 1911-1935 Admin. Records

Abramitzky, Boustan & Eriksson (ABE) Algorithm

- Form linked dataset of treatments/outcomes for a subset of individuals
- Exact or near agreement of linking variables: name, sex, race, age, state of birth
- Abramitzky et al. (2021; JEL) "Automated Linking of Historical Data"

Methodological Challenges

Low Match Rates

- In typical applications of ABE only $\approx$ 20% of observations are matched
- Women and minorities typically excluded altogether: harder to match
- Inefficient estimation; potentially unrepresentative sample

Noisy Linking Variables

- Variables typically used for linking are known to be measured with error
- Implicitly acknowledged in ABE algorithm: permits "near" matches
- Ignored in practice: linked dataset analyzed as though it has no spurious matches

Learn β from $Y_i = X_i'\beta + U$ where $\mathbb{E}(U_i|X_i) = 0$

	index	W	Y
	1	$\widetilde{W}_1$	$\widetilde{Y}_1$
	$\vdots$	$\vdots$	$\vdots$
DF_y	j	$\widetilde{W}_j$	$\widetilde{Y}_j$
	$\vdots$	$\vdots$	$\vdots$
	n	$\widetilde{W}_n$	$\widetilde{Y}_n$

	index	W	X	Y
	1	W_1	X_1	Y_1
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
DF_x	j'	$W_{j'}$	$X_{j'}$	$Y_{j'}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	n	W_n	X_n	Y_n

- Blue means observed; Red means unobserved
- W observed jointly with outcomes in DF_y and jointly with treatments in DF_x

The Linking Matrix

$\mathbf{L} = [\ell_{jj'}]$ is an $(n \times n)$ permutation matrix where

$$\ell_{jj'} = \begin{cases} 1 & \text{if record } j \text{ in } \text{DF}_y \text{ matches with record } j' \text{ in } \text{DF}_x \\ 0 & \text{otherwise.} \end{cases}$$

Since the j^{th} row of $\mathbf{L}$ is $\begin{bmatrix} \ell_{j1} & \ell_{j2} & \cdots & \ell_{jn} \end{bmatrix}$, it follows that $\widetilde{\boldsymbol{Y}} = \mathbf{L}\boldsymbol{Y}$

$$\widetilde{Y}_j = \left\{ \begin{array}{ll} Y_1 & \text{if } \ell_{j1} = 1 \\ Y_2 & \text{if } \ell_{j2} = 1 \\ \vdots & \\ Y_n & \text{if } \ell_{jN} = 1 \end{array} \right\} = \sum_{j'=1}^{N} \ell_{jj'} Y_{j'}.$$

Fundamental Decomposition

$$
\begin{tikzcd}
\mathbf{W} \arrow[r] \arrow[dr] & \mathbf{W}^{\text{post}} \arrow[r] & \widetilde{\mathbf{W}} & \\
& \mathbf{X} \arrow[d] & & \mathbf{L} \arrow[ul] \arrow[dl] \\
& \boldsymbol{Y} \arrow[r] & \widetilde{\boldsymbol{Y}} &
\end{tikzcd}
$$

(A) $\boldsymbol{Y} \perp\!\!\!\perp \mathbf{L} \mid (\mathbf{X}, \mathbf{W}, \widetilde{\mathbf{W}})$

(B) $\boldsymbol{Y} \perp\!\!\!\perp (\mathbf{W}, \widetilde{\mathbf{W}}) | \mathbf{X}$

(C) $\mathbf{L} \perp\!\!\!\perp \mathbf{X} | (\mathbf{W}, \widetilde{\mathbf{W}})$

$$
\begin{aligned}
\mathbb{E}\left(\widetilde{\boldsymbol{Y}} | \mathbf{X}, \mathbf{W}, \widetilde{\mathbf{W}}\right) &= \mathbb{E}_{\mathbf{L}|\mathbf{X},\mathbf{W},\widetilde{\mathbf{W}}}\left[\mathbb{E}\left(\mathbf{L}\boldsymbol{Y} | \mathbf{X}, \mathbf{W}, \widetilde{\mathbf{W}}, \mathbf{L}\right)\right] \\
&= \mathbb{E}_{\mathbf{L}|\mathbf{X},\mathbf{W},\widetilde{\mathbf{W}}}\left[\mathbf{L}\mathbb{E}\left(\boldsymbol{Y} | \mathbf{X}, \mathbf{W}, \widetilde{\mathbf{W}}\right)\right] \\
&= \mathbb{E}\left(\mathbf{L} \,\middle|\, \mathbf{X}, \mathbf{W}, \widetilde{\mathbf{W}}\right) \mathbb{E}\left(\boldsymbol{Y} | \mathbf{X}, \mathbf{W}, \widetilde{\mathbf{W}}\right) \\
&= \mathbf{Q}(\mathbf{W}, \widetilde{\mathbf{W}})(\mathbf{X}\beta)
\end{aligned}
$$

Estimators that Regress $\widetilde{\boldsymbol{Y}}$ on $\mathbf{QX} \equiv$ Imputed $\widetilde{\mathbf{X}}$

Unique Match (UM)

Run ABE; fill $\mathbf{Q}$ with ones and zeros to indicate unique matches; drop many $\widetilde{Y}_j$

Poirier & Ziebarth (PZ)

Run ABE; form groups of "potential matches"; give equal weight to each in $\mathbf{Q}$

Probabilistic Linking (PL)

- Compute $\mathbf{Q}(\mathbf{W}, \widetilde{\mathbf{W}}) \equiv \mathbb{E}(\mathbf{L}|\mathbf{W}, \widetilde{\mathbf{W}})$ using Bayes' Theorem
- Tempting simplification: $\mathbb{P}(\ell_{jj'} = 1 | \mathbb{1}\{\widetilde{W}_j = W_{j'}\})$

Simulation Design

(i) $X_i \sim$ iid Uniform$(0,1)$

(ii) $Y_i | X_i \sim$ iid Normal$(X_i, \sigma^2 = 1)$

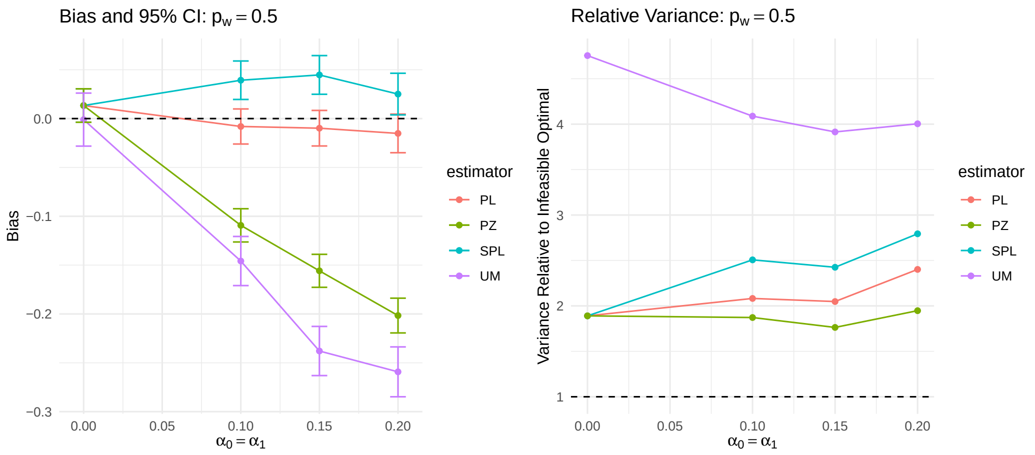
(iii) $W_i \sim$ iid Bernoulli(p_w)

(iv) $W_i^{\text{post}} | W_i \sim$ iid Bernoulli$\left((1 - W_i)\alpha_0 + W_i(1-\alpha_1)\right)$

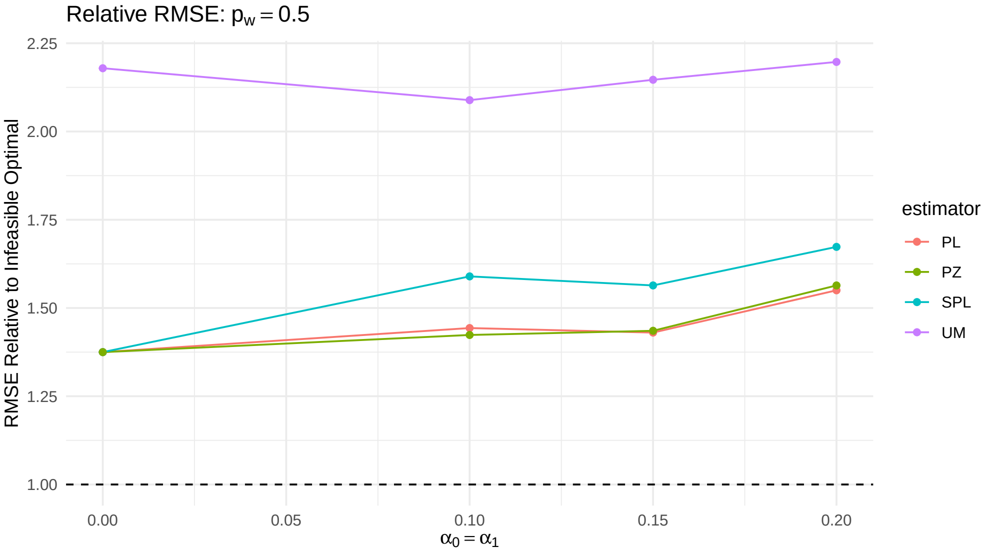
$$\begin{tikzcd}
\mathbf{W} \arrow[d] & & \\
\mathbf{W}^{\text{post}} \arrow[r] & \widetilde{\mathbf{W}} & \\
\mathbf{X} \arrow[d] & & \mathbf{L} \arrow[ul] \arrow[dl] \\
\boldsymbol{Y} \arrow[r] & \widetilde{\boldsymbol{Y}} &
\end{tikzcd}$$

- Mis-classification probabilities α_0 and α_1
- $(\widetilde{\mathbf{W}}, \widetilde{\boldsymbol{Y}})$ random re-ordering of the rows of $(\mathbf{W}^{\text{post}}, \boldsymbol{Y})$ within "blocks"
- 50 blocks ("states") each containing $[2 + \text{Poisson}(1)]$ individuals ($n \approx 150$)

Simulation Results



Simulation Results



Next Steps

- Head-to-head comparisons using real data: how different are the results?
- Our "full" PL estimator works well, but is hard to scale up. Approximations?
- Connection with TS2SLS using mis-classified instruments.
- Additional complication: "treatment" is often a function of X and W
- Start thinking more carefully about sample selection / heterogeneity