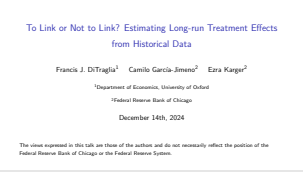
To Link or Not to Link?



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 views expressed in this talk are those of the authors and do not necessarily reflect the position of the Federal Reserve Bank of Chicago or the Federal Reserve System.

1. Thank you; it's nice to see some familiar faces and some new faces this morning.
2. This is joint work with my colleagues Camilo and Ezra from the Chicago Fed
3. Still very much work in progress. Interested to hear your thoughts.

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"

1. Many top papers estimate "long-run" policy effects: immigration quotas, education, public library access etc. Must somehow connect treatment to outcomes measured years or decades later. Typically: admin records with treatment in one dataset, outcomes in another; attempts to "link" them, often imperfectly
2. Example: Mothers' pension program was the first government-sponsored welfare program in US history. Involved making cash payments to impoverished single mothers (mainly widows). What are the long-term effects of the program on children of recipients (wage, education, longevity)? Treatment measured in admin records of program applicants (accepted / rejected); outcomes measured in 1940 census. "Linking variables" typically used, including in this example, are name, sex, race, age, and state of birth.
3. ABE algorithm: Scan through treatment and outcome data. If there is any value of the linking variables that appears *exactly once* in both datasets (e.g. Theodore Roosevelt, while, born in 1858 in NY) declare the corresponding datasets of treatments and outcomes a match, and record them in the "linked" regression dataset. End up with a dataset of "exact matches" but typically the word "exact" is a bit mis-leading here. Make multiple passes through the data and tolerate slight discrepancies in later passes, e.g. "Teddy" rather than "Theodore." Drop anyone you can't match

Methodological Challenges

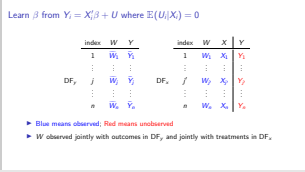
Low Match Rates
- In typical applications of ABE only ≈ 20% of observations are matched
- Women and minorities typically excluded altogether: harder to match
- Inefficient estimation; potentially unrepresentative sample

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

1. Although the ABE approach is widely-used, its shortcomings as well-known.
2. First, match rates are typically low: around 20%.
3. Some groups are especially hard to match (women and minorities) so they a typically excluded altogether.
4. Low match rates raise concerns about inefficient estimation; fact that some groups are hard to math raises concerns about sample selection, although I won't say much about that today
5. Second problem: variables typically used for linking are known to be measured with error.
6. Implicitly acknowledged in ABE algorithm when it allows "near" matches.
7. In practice: error in linking vars ignored, linked data is analyzed as though no spurious matches.
8. There is some relevant methodological literature in stats, comp sci, and a bit in Economics, but it isn't used in practice in econ history papers.
9. State of the art at the moment is to use "hand-linked" data from genealogy websites: people go find Grandma. Improves link rates to perhaps 60% but concerns about error and selection remain.

To Link or Not to Link?

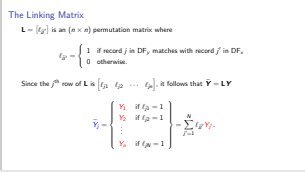
Learn β from $Y_i = X_i'\beta + U$ where $\mathbb{E}(U_i|X_i) = 0$



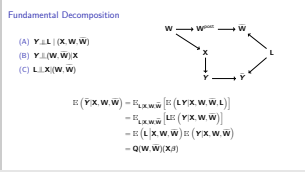
1. Today: tell you about some work we're doing on this problem to try to develop methods that we hope will be attractive to econ history researchers and improve on ABE.
2. Target parameter parameter β that *would be identified* from regression of Y on X. Problem: not observed jointly. Instead observe outcome data DF_y with (W, Y) and treatment data DF_x with (W, X) where W denotes "linking" variables, e.g. name, sex, race, age, state of birth.
3. Unlike typical "two sample" or "data fusion" problems (TS2SLS), here *same people* observed at two points in time rather than *different people* from the same population observed at the same point in time.
4. Simplifying assumption: no "extra" people in either dataset so same size (no death, birth, migration)
5. Explain notation: tildes for outcome dataset; $\widetilde{Y}$ scrambled relative to unobserved Y from DF_x.
6. Linking variables W and $\widetilde{W}$ may not agree even when they refer to the same person (measurement error)

To Link or Not to Link?

The Linking Matrix



1. This is just a slide about notation, but it's very handy notation.
2. $\mathbf{L}$ is the permutation matrix that "scrambles" $\boldsymbol{Y}$ to give $\widetilde{\boldsymbol{Y}}$.
3. Its (j, j') element is zero if record j in the outcome dataset matches record j' in the treatment dataset, zero otherwise.
4. So we can write $\widetilde{Y}_j$ as a weighted sum of $Y_{j'}$ or equivalently $\widetilde{\boldsymbol{Y}} = \mathbf{L}\boldsymbol{Y}$. Use on next slide.



1. There's a lot going on here; spend some time. LHS: three conditional independence assumptions; RHS: a DAG that implies them. Bottom, a key result that follows from the DAG / assumptions.
2. What do we need to assume about the linking variables to make progress? Several possibilities (W is a proxy for Y or W is a mediator) but only one of them really makes sense in econ history applications: *exclusion*. (B) at left; no arrow from W or $\widetilde{W}$ to Y in DAG. Allows $\mathbf{X}$ to include subset of $\mathbf{W}$.
3. Does ABE require exclusion? Implicitly yes. In applied work researchers usually exclude most of the linking variables from their final regression. Unless these are excluded $\Rightarrow$ selection-on-unobservables!
4. Subtle point: IV usually requires *relevance* (arrow from W to X) but actually this is *not* actually required. Simulation study below generates W and X independently!
5. LHS (A) and (C) are less intuitive. Roughly: X doesn't affect the measurement error process in the linking variables and the order of observations in the "tilde" datasets isn't informative.
6. Convenient decomposition: conditional mean function of $\widetilde{Y}$, the outcomes we observe, conditional on all the linking variables and all the treatments, is equal to a matrix $\mathbf{Q}$ times $\mathbf{X}\beta$. So if you knew $\mathbf{Q}$, you could simply regress $\widetilde{Y}$ on $\mathbf{QX}$!

To Link or Not to Link?

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

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 $\mathbf{Q}$

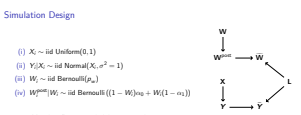
Probabilistic Linking (PL)

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

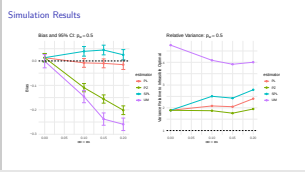
1. Previous slide: if we knew $\mathbf{Q}$, could simply regress $\widetilde{\boldsymbol{Y}}$ on $\mathbf{QX}$. Since $\mathbf{Q}$ is a matrix of probability weights (rows sum to one) $\mathbf{QX}$ "imputes" $\widetilde{\mathbf{X}}$. Don't need to "find matches." More general / fruitful to think about directly estimating β using different methods of imputing $\widetilde{\mathbf{X}}$. Several examples fit into this framework.
2. First: ABE algorithm (UM). Second: refinement due to Poirier & Ziebarth (2019). If multiple X-observations could correspond to $\widetilde{Y}_j$ for "Theodore Roosevelt", give them equal weight.
3. Error in linking variables is a problem for UM and PZ: spurious matches or groups $\Rightarrow$ poor imputations.
4. In principle: "probabilistic linking" (PL) gives a fully general solution to measurement error: specify generative model, estimate parameters, plug into Bayes' Theorem, compute $\mathbf{Q}$.
5. In practice: PL relies on strong simplifying assumptions. Only looks at agree/disagree between $\widetilde{W}$ and W, ignores the values of linking variables for all other records. Note that ABE uses this information!
6. We derived "industrial strength" version of PL: how much do the simplifying assumptions matter?
7. Remainder of talk: simulation that puts UM, PZ, simplified PL, and full-blown PL head-to-head.

To Link or Not to Link?

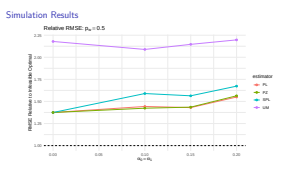
Simulation Design



1. Sim DGP is based on a DAG very similar to the one from above, except no arrow from W to X: indep!
2. Single regressor X, Uniform(0,1); normal linear model for Y; error variance equals treatment effect.
3. Single binary linking variable observed in a "pre" period (W_i) and a "post" period (W_i^{post}). Post is a noisy version of pre unless $\alpha_0 = \alpha_1 = 0$. Reduced form of latent "true" W^* with two noisy measures.
4. As in many real applications, work with a "blocking variable" i.e. a linking variable assumed to be correctly measured, so $(\widetilde{\mathbf{W}}, \widetilde{\boldsymbol{Y}})$ is a random re-ordering of the rows of $(\mathbf{W}^{\text{post}}, \boldsymbol{Y})$ within blocks.
5. 50 blocks, each with random number of individuals: add 2 to a Poisson(1) random variable.
6. How do the bias and variance of the estimators described above depend on α_0, α_1?
7. Particularly interested in the comparison of "simplified" vs "industrial strength" PL.
8. For simplicity, "infeasible" versions of PL: plug in true α_0 and α_1. In practice these would be estimated, either in a first step or using auxiliary data.



1. Selected results from our simulation study; focus on $p_w = 0.5$ and $\alpha_0 = \alpha_1$ but the pattern is general.
2. Four estimators. UM is the "unique match" estimator, i.e. ABE. PZ is Poirier & Ziebarth's refinement.
3. PL is our "full" probabilistic linking estimator that we have derived. SPL is a "simplified" probabilistic linking approach–the only one we've ever seen used in practice–that calculates the (j, j') entry of Q by examining only the linking variables for these two records. Generates a Q matrix that violates adding up constraints for probabilities, so normalize rowsums to avoid catastropically bad performance.
4. In background: infeasible "full" estimator in which all X and Y observations are correctly linked. Unbiased; lower bound for variance of others. Var and RMSE are relative to it.
5. LHS: bias of estimators and 95% CI to quantify simulation error; RHS: relative variance of estimators
6. Can show: PL, PZ and SPL *exactly coincide* when $\alpha_0 = \alpha_1 = 0$. Simulation bears this out.
7. Key takeaways. First: all estimators *except ours* are biased when there is mis-classification, and the bias of UM and PZ increases as the mis-classification worsens. Interestingly, SPL is *upward biased*. Worth trying to explore theoretically. Second: UM has high variance, and SPL is worse than PL. Interestingly PZ seems to beat us on variance.



1. Even more interesting: PZ is competitive in terms of RMSE, at least in this DGP / n.
2. But it achieves this with a considerable bias, and that bias doesn't decrease with n. (Haven't proved but have simulated).

Next Steps

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

1. As I mentioned at the beginning: this is very much work in progress. Here are our next steps
2. Head-to-head comparisons of ABE, PZ, and "simplified" PL in real census applications. Differences?
3. Key lesson from simulation is that the PL works and beats the "simplified version"– unbiased and lower variance. Problem is that it becomes very complicated in higher dimensions. Useful approximations?
4. Connections to TS2SLS: even though this isn't a "two-sample" problem, it can still be applied as long as we have exclusion, could use W as IVs. Robust to measurement error of a particular form: symmetry. But seems implausible in settings that aren't two samples from same population at the same moment in time. Unaware of work that addresses this issu more broadly in TS2SLS
5. PL can be thought of as "different first stage" that uses everyone's W, but also addresses arbitrary forms of measurement error–as long as you can implement it!
6. Point about treatment often being jointly determined by X and a linking variable. So if we think the linking variable is measured with error, so is the treatment! Extremely common in practice but ignored. Doesn't quite fit with the PL approach.
7. Haven't talked about sample selection; we're still thinking about this.