The effects of matching algorithms and estimation methods using linked data

Rachel Anderson[*]

This Version: August 30, 2019

Abstract

This paper studies the effect of different matching algorithms and estimation procedures for linked data on the quality of the estimates that they produce.

1 Introduction

In applied microeconomics, identifying a common set of individuals appearing in two or more datasets is often complicated by the absence of unique identifying variables. For example, Aizer et al. (2016) link children listed on their mother's welfare program applications with their death records using individuals' names and dates of birth. However, since name and date combinations are not necessarily unique (and may be prone to typographical error), the authors identify cases where multiple death records seem to refer to the same individual. Instead of dropping these observations from their analysis, they use estimation techniques from Anderson et al. (2019) that allow for observations to have multiple linked outcomes.

[*]Mailing Address: Department of Economics, Julis Romo Rabinowitz Building, Princeton, NJ 08544. Email: rachelsa@Princeton.edu. This project received funding from the NIH (Grant 1 R01 HD077227-01).

The methods in Anderson et al. (2019) assume that each of the linked outcomes is equally likely to be the true match; however, the authors describe how to construct more efficient estimators if additional information about match quality is available. Specifically, if the researcher can estimate the probability that each individual-outcome pair is a true match, then this knowledge can be used to achieve a reduction in mean-squared error. Such probabilities are outputted by probabilistic record linkage procedures, first developed by Fellegi and Sunter (1969) in the statistics literature, but only recently applied to economics Abramitzky et al. (2018). Hence, any discussion of best practices for using linked data should address how to choose matching algorithms and estimation procedures jointly.

The goal of this paper is therefore to study the effects of different combinations of matching algorithms and estimation procedures for linked data on the quality of the estimates that they produce. First, I will compare how different matching algorithms perform in terms of the representativeness of the matched data they produce and their tolerance for type I and type II errors. Next, with multiple matched versions of the data in hand, I will compute point estimates and confidence intervals for the same parameter of interest using methods that vary by whether they allow for multiple matches, incorporate the matching probabilities, and are likelihood-based in their approach. In total, I will perform the above analysis twice – with simulated data and with real data that the simulated data are generated to imitate.

To the best of my knowledge, how data pre-processing impacts subsequent inference in economics research is not well understood. This paper adds to a recent series of papers by Bailey et al. (2017) and Abramitzky et al. (2018, 2019a), which evaluate the performance of common matching algorithms for historical data in a variety of real and simulated data settings, and offer informal discussions about the impact of matching on subsequent inference. This paper pushes these ideas further, by measuring the *joint* effects of matching and estimation; with a focus on developing estimation techniques that incorporate information from the matching process to improve efficiency and accurately reflect uncertainty.

The matching techniques that will be studied in this paper include deterministic record linkage procedures developed by Ferrie (1996) and Abramitzky et al. (2012); abe; Abramitzky et al. (2019b), and Aizer et al. (2016), and multiple implementations of probabilistic record linkage, specifically the fastLink Enamorado et al. (2019), and machine learning approaches Feigenbaum (2016). Estimation techniques will include Anderson et al. (2019), Lahiri and Larsen (2005), and a fully Bayesian approach that I will develop in this paper.

The data used in this paper consist of the unmerged files from Aizer et al. (2016), which I will pre-process using the practices developed by Abramitzky et al. (2018). The parameter of interest is the average treatment effect of receiving a cash transfer on the long-term outcomes of children in poor families.

The next section describes the general problem that this paper seeks to address. Section 3 provides an overview of the matching techniques that I will study. Section 4 describes the estimation techniques. Section 5 will have simulations and results. Section 6 will have real data and results.

2 General problem

Suppose that the researcher would like to estimate θ_0 in a parametric model of the form

$$E[m(\mathbf{z_i};\theta_0)] = 0 \tag{1}$$

where $m(\cdot)$ is a moment function and $\mathbf{z_i} = (x_i, y_i)$ is the data associated with an individual i sampled at random from the population of interest; however, instead of observing (x_i, y_i) pairs directly, the researcher observes two datasets. The first dataset contains variables x_i and identifiers w_i for individuals $i = 1, \ldots, N_0$ recorded at time 0. The second dataset contains outcomes y_j and identifiers w_j for individuals $j = 1, \ldots, N_T$, observed at time T far in the future.

To estimate (1) with standard econometric methods, the researcher must identify which of the x_i and y_j refer to the same individuals. That is, she needs to recover the matching function $\varphi : \{1, \ldots, N_0\} \to \{1, \ldots, N_T\}$, where $\varphi(i) = j$ if individual i observed at time 0 and individual j observed at time T refer to the same entity. If w_i and w_j identify individuals uniquely and do not change over time, then $\varphi(i) = j$ if and only if $w_i = w_j$. However, if the identifiers are non-unique or prone to errors introduced by the record-generating process, then φ needs to be estimated, and inference about θ needs to be adjusted accordingly.

In statistics, the task of recovering φ is called *record linkage*. A record linkage procedure consists of a set of decisions about (i) selecting and standardizing the identifiers w_i and w_j, (ii) choosing which records to consider as potential matches (especially when $N_0 \times N_T \times \dim(w_i)$ is large), (iii) defining what patterns of (w_i, w_j) constitute (partial) agreements, and (iv) designating observation pairs as matches. Each step of the record linkage process introduces the possibility that a true match is overlooked (Type II error), or that a false match is designated as correct (Type I error). [Abramitzky et al. (2019a) illustrate a tradeoff between the two...].

To fix ideas, suppose that θ_0 is the long-run impact of cash transfers on the longevity of children raised in poor families. The vector x_i includes family and child characteristics as observed in welfare program applications; and the outcomes y_j are constructed by calculating (day of death $-$ day of birth) for all of the individuals observed in a set of death records. For each observation, the identifiers w_i and w_j include the individual's first and last name, and date of birth. Additionally, some w_i and w_j contain middle initials, w_i includes i's place of birth, and w_j includes j's place of death.

In this case, an example of a (deterministic) record linkage procedure is to:

(i) match observations using phonetic spellings of first and last names, birth dates, and middle initials when available;

(ii) measure agreements between names using Jaro-Winkler string distances, and penalize disagreements in birth year more than differences in birth month;

(iii) consider as potential matches only (i, j) pairs whose (phonetically standardized) initials match, and whose birth years are within ± 2 years;

(iv) accept as matches all (i, j) pairs with scores calculated using the metrics in (ii) that exceed a pre-specified cutoff; if a record i has multiple possible matches j that exceed the cut-off, then choose the match with the highest score (or pick a random match if there is a tie).

Another record linkage procedure could be defined using the same rules for steps (i)-(iii), but replace (iv) with a probabilistic matching rule that does not enforce one-to-one matching:

(iv*) use the Expectation Maximization algorithm to estimate the probability that each record pair is a match, and then designate matchesusing the Fellegi-Sunter rule for pre-specified tolerances for Type I and Type II errors

Except in rare cases, the estimates of φ produced by the preceding procedures will be different. Whereas the first procedure associates each x_i with at most one matched y_j, the second procedure may associate the same x_i with multiple y_j (or zero!). These differences, in turn, impact the estimates of θ, and so uncertainty about θ should reflect uncertainty introduced by estimating φ.

As pointed out by Bailey et al. (2017), record linkage procedures differ according to the set of assumptions that motivate their use. However, all of the procedures discussed in this paper will be studied under a unifying set of assumptions:

1. (De-duplication) Within a given dataset, each observation refers to a distinct entity.

That is, if two observations share the same identifier, they represent two different individuals.

2. (No unobserved sample selection) The observed x_i and y_j are random samples conditional on w_i and w_j, respectively. This means that all individuals with the same identifying information have equal probability of appearing in the sample.

3. There exists a unique $\theta_0 \in \Theta$ that satisfies the relationship in (1), that can be consistently estimated using standard econometric techniques if φ_0 is known.

The next section discusses in detail the exact record linkage techniques that will be studied, and their motivating assumptions.

3 Overview of matching techniques

In statistics, techniques for recovering φ are called *record linkage* procedures; but for the purposes of this paper, I will use the term "matching procedures" interchangeably.

In the section that follows I talk about each of these.

which records to compare data pre-processing, blocking/implementation, and match assignment. Data pre-processing includes selecting and standardizing variables for matching (by applying phonetic algorithms to string variables for example), and defining a distance metric for string variables. Implementation decisions include blocking rules to limit the number of comparisons for computational feasibility and picking relative weights for disagreements in different variables. Finally, the review process involves designating record pairs as matches if a one-to-one matching is desired (or defining a cut-off for multiple matches).

Matching procedures are either deterministic or probabilistic. Deterministic methods are those like **?** and **?**, where a fixed set of rules determine which records are matching and which are not. On the other hand, probabilistic methods attempt to "let the data speak", at the cost of computational feasibility, as they involve estimating the probability that each record pair within a pre-specified block refer to a match. In many instances, it is not clear which perform better (especially when the datasets are large!)

record linkage process is x y,z,

There are 44,000 files in Dataset A. Dataset B consists of many, many million observations. The goal is to find the matches for the 44,000 individuals... this is not an insignificant task. Blocking is crucial. Although individuals with distinct names are unlikely to appear on multiple death records, individuals with common names like "John Smith" may be linked to multiple death records. Similarly, if the econometrician matches individuals using a subset of the variables in w_i (as in the case that females are matched by first name only, and place and year of birth), she is likely to find many possible links that are equally credible.

The following note formalizes the assumptions that are necessary to estimate the model (1) without dropping observations with non-unique matches.

Assumption 1. The observed (x_i, z_i, w_i) is a random sample drawn from the marginal distribution $f(x^*, z^*, w^*)$. The $\{y_{i\ell}\}_{\ell=1}^{L_i}$ is a random sample drawn from $f(y^*|w_i, L_i)$, so that $(x_i, z_i) = (x_i^*, z_i^*)$ and $y_{i\ell}$ are independent conditional on w_i for $y_{i\ell} \neq y_i^*$.

Assumption 2. There is exactly one $y_{i\ell} = y_i^*$, $\ell \in \{1, \ldots, L_i\}$ for all i. That is, we assume that one of the $\{y_{i\ell}\}_{\ell=1}^{L_i}$ is drawn from the marginal distribution $f(y_i^*|x_i^*, z_i^*) = f(y_i^*|x_i, z_i)$.

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 $(y_{i\ell}, x_i, z_i) = (y_i^*, x_i^*, z_i^*)$, and that equals 0 otherwise. Also, $\sum_{\ell=1}^{L_i} s_{i\ell} = 1$ for all i, and we can rewrite 1 as,

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

Assumption 3. The identifying variables w_i in (x_i, z_i, w_i) exactly match, or are sufficiently close to w_i in $(\{y_{i\ell}\}_{\ell=1}^{L_i}, w_i)$, such that the researcher cannot distinguish which $y_{i\ell} = y_i^*$ for $\ell \in \{1, \ldots, L_i\}$. In other words, the researcher behaves as if,

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

Assumptions 1 and 3 rule 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.

References

A. Aizer, S. Eli, J. Ferrie, and A. Lleras-Muney, "The long-run impact of cash transfers to poor families," *American Economic Review*, vol. 106, no. 4, pp. 935–71, April 2016. [Online]. Available: http://www.aeaweb.org/articles?id=10.1257/aer.20140529

R. Anderson, B. Honore, and A. Lleras-Muney, "Estimation and inference using imperfectly matched data," *Working paper*, August 2019. [Online]. Available: http://www.github.com/rachelsanderson/ImperfectMatching

I. P. Fellegi and A. B. Sunter, "A theory for record linkage," *Journal of the American Statistical Association*, vol. 64, pp. 1183–1210, 1969.

R. Abramitzky, R. Mill, and S. Perez, "Linking individuals across historical sources: a fully automated approach," National Bureau of Economic Research, Working Paper 24324, February 2018. [Online]. Available: http://www.nber.org/papers/w24324

M. Bailey, C. Cole, M. Henderson, and C. Massey, "How well do automated linking methods perform? lessons from u.s. historical data," National Bureau of Economic Research, Working Paper 24019, November 2017. [Online]. Available: http://www.nber.org/papers/w24019

R. Abramitzky, L. Boustan, K. Eriksson, J. Feigenbaum, and S. Perez, "Automated linking of historical data," *NBER Working Paper*, 2019.

J. P. Ferrie, "A new sample of males linked from the public use microdata sample of the 1850 u.s. federal census of population to the 1860 u.s. federal census manuscript schedules," *Historical Methods*, vol. 29, no. 4, pp. 141–156, 1 1996.

R. Abramitzky, L. P. Boustan, and K. Eriksson, "Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration," *American*

Economic Review, vol. 102, no. 5, pp. 1832–56, May 2012. [Online]. Available: http://www.aeaweb.org/articles?id=10.1257/aer.102.5.1832

R. Abramitzky, L. P. Boustan, and K. Eriksson, "To the new world and back again: Return migrants in the age of mass migration," *ILR Review*, vol. 72, no. 2, pp. 300–322, 2019. [Online]. Available: https://doi.org/10.1177/0019793917726981

T. Enamorado, B. Fifield, and K. Imai, "Using a probabilistic model to assist merging of large-scale administrative records," *American Political Science Review*, vol. 113, no. 2, p. 353?371, 2019.

J. J. Feigenbaum, "A machine learning approach to census record linking ?" 2016.

P. Lahiri and M. D. Larsen, "Regression analysis with linked data," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 222–230, 2005. [Online]. Available: http://www.jstor.org/stable/27590532

4 Matching Methods

For the purposes of this paper, I define a record linkage procedure as a set of rules about data pre-processing, blocking/implementation, and match assignment. Data pre-processing includes selecting and standardizing variables for matching (by applying phonetic algorithms to string variables for example), and defining a distance metric for string variables. Implementation decisions include blocking rules to limit the number of comparisons for computational feasibility and picking relative weights for disagreements in different variables. Finally, the review process involves designating record pairs as matches if a one-to-one matching is desired (or defining a cut-off for multiple matches).

Matching procedures are either deterministic or probabilistic. Deterministic methods are those like **?** and **?**, where a fixed set of rules determine which records are matching and which are not. On the other hand, probabilistic methods attempt to "let the data speak", at the cost of computational feasibility, as they involve estimating the probability that each record pair within a pre-specified block refer to a match. In many instances, it is not clear which perform better (especially when the datasets are large!)

In this paper I do two implementations of each type (not exhaustive!). X YZ already review different types.

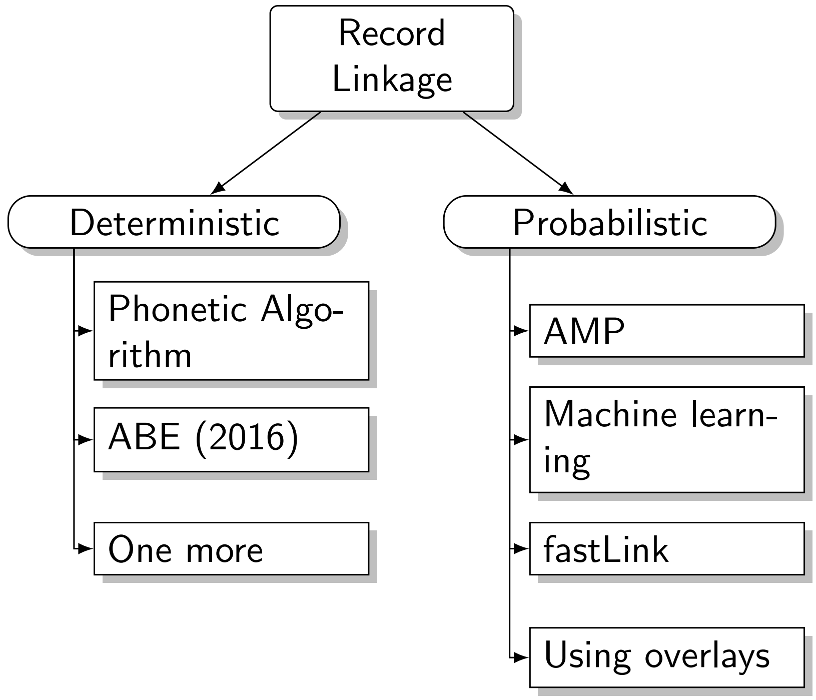
I will do 4: ABE Code, FErrie (done already), ABE R CODE, AND FASTLINK.

INSERT A GRAPHIC WITH THE METHODS I WILL TEST

TABLE WITH SURVEY OF MATCHING METHODS used in literature. categoreis are like enforces one to one matching, blocking, string comparator, phonetic algorithm (this is lit review essentially/annotated bibliography)

- Deterministic: Rely heavily on phonetic algorithms (i.e. SOUNDEX, NYSIIS, Metaphone, Spanish Metaphone) and string distance measures (SPEDIS, Jaro-Winkler).

Figure 1: Overview of matching methods



Abramitzky et al. (2019a) discuss differences in these choices and call SOUNDEX outdated.

1. Aizer et al. (2016): Described in appendix, match uses first name, middle initial, last name, day, month and years of birth. Match allows for errors in strings (using SOUNDEX and SPEDIS) and in single digits for DOB.

2. Abramtizky, Boustan and Eriksson (ABE) algorithms – discard observations that do not have unique matches; they typically perform multiple variations of the same algorithm to check robustness. Perform matching in both directions, and then take the intersection of the two matched samples. Implemented with abematch stata command.

3. ABE-JW adjustment algorithm – block by place of birth, first letter of first and last names match. Jaro-winkler can be implemented with stata jarowinkler com-

mand; R package stirngdist. They also only accept one-to-one matching.

Limitation of ABE algorithms is that "it is not clear how to appropriately weight differences in name spelling versus differences in age when comparing two records".

- Probabilistic (see Winkler 2006 for survey)
  - E-M Algorithm Abramitzky et al. (2018) still blocking by same place of birth, predicted year of birth, and first letter of first and last name. Still use decision rule to enforce one-to-one matching.
  - Training sample (Ruggles and Feigenbaum) – trained to NOT allow mutliple matches; generates a predicted probability of being a match for each pair of records in A and B
  - IPUMS linking method: trains support vector machine on training sample of manually classified records (like Feigenbaum 2016) In historical applications this is problematic due to sample attrition. The DGP changes, so a full likelihood is a good idea.

- Overview of matching methods

Important measurements: estimated type 1, type 2 errors; representativeness of sample, sample size, overlapping of samples - Comparison of matching methods from (a) theoretical perspective, (b) with simulated data, (c) with actual data

1. Estimation Methods
   - Anderson, Honore, Lleras-Muney (2019)
   - Lahiri Larsen
   - Scheuren Winkler

   - Overview of estimation methods

- Comparison of estimation methods from (a) theoretical perspective, (b) with simulated data, (c) with actual data

(3) Further investigation/follow-up simulations inspired by steps 1 and 2

Data problems in practice Bailey et al. (2017)

- "time-invariant" features not time-invariant – names misspelled, individual reported incorrectly, or individual changed name
- Age heaping – rounding ages to the nearest multiple of five
- digitization of handwritten manuscripts

String comparators: NYSIIS, Soundex, Jaro-Winkler, Levenstein, etc.

5 Annotated bibliography

- Neter, Maynes, and Ramanathan (1965): small mismatch errors in finite population sampling can lead to a substantial bias in estimating the relationship between response errors and true values
- Scheuren and Winkler (1993): propose method for adjusting for bias of mismatch error in OLS
- SW (1997, 1991): iterative procedure that modifies regression and matching results for apparent outliers
- Lahiri and Larsen (2005): provides unbiased estimator directly instead of bias correction for OLS, by applying regression to transformed model
- Abramitzky, Mill, Pérez (2019): guide for researchers in the choice of which variables to use for linking, how to estimate probabilities, and then choose which records to use

in the analysis. Created R code and stata command to implement the method

- Ferrie 1996, Abramitzky, BOustan and Eriksson (2012 2014 2017) are deterministic. Conservative methods require no other potential match with same name within a 5-year band
- Semi-automated Feigenbaum, Ruggles et al
- Abramitzky, Boustan, Eriksson, Feigenbaum, Pèrez (2019): evaluate different automated methods for record linkage, specifically deterministic (like Ferrie and ABE papers), machine learning Feigenbaum approach, and the AMP approach with the EM algorithm. Document a frontier between type I and type II errors; cost of low false positive rates comes at cost of designating relatively fewer (true) matches. Humans typically match more at a cost of more false positives. They study how different linking methods affect inference – sensitivity of regression estimates to the choice of linking algorithm. They find that the parameter estimates are stable across linking methods. Find effect of matching algorithm on inference is small.
- Bailey et al. (2017) review literature on historical record linkage in US and examines performance of automated record linkage algorithms with two high-quality historical datasets and one synthetic ground truth. They conclude that no method consistently produces representative samples; machine linking has high number of false links and may introduce bias into analyses.

  Treatment of equally likely – equal probability weighting of tied candidates (Bleakley and Ferrie 2016); weighted combo of linking features to ehlp disambiguate potential matches. Ferrie 96; Old ABE, new ABE, and Feigenbaum.
- Survey paper from handbook of econometrics
- For example, Goeken et al. (2017) document that in two enumerations of St. Louis in the 1880 Census, nearly 46 percent of first names are not exact matches, and the

Early Indicators project notes that 11.5 percent of individuals in the Oldest Old sample have a shorter first name in pension records than in the original Civil War enlistment records (Costa et al. 2017).

Overall, high variability in performance of matching methods depending on choice of variables, string comparators used.

6 Empirical Application

6.1 Data

What are the long-run effects of cash transfers to poor families? Specifically, are there lifelong benefits for children raised in poor families? Transfers may not help poor children if the amounts are insufficent, parents mis-allocate funds, or the transfers induce behavioral changes that are detrimental to the child.

Evaluating whether cash transfers improve outcomes necessitates identifying a plausible counterfactual. **?** collect administrative records from the MOthers' Pension program (1911-1935), which was the first US government-sponsored welfare program for poor mothers with dependent children. The intent of the MP program was to imrpove the conditions of "young children that have become dependent through the loss or disability of the breadwinner" (Abbott 1993, p 1). The transfers generally represented 12-25 percent of family income, and typically lasted for three years.

The authors measure the impact in terms of longetivity and other outcomes of the children whose mothers applied to the program. These data include information on thousands of accepted and rejected applciants born between 1900 and 1925, most of whom had died by 2012. The idetnifying information in the application records allows them to link children with other datasets to trace their lifetime outcomes.

Identification comes from comparing children of mothers who applied for transfers and who were initially deemed eligible, but were denied upon further investigation. This is a standard strategy in program evaluation that has been used successfully in studies of disability insurance. The validity of this strategy hinges on the assumpotion that accepted and rejected mothers and their children do not differ on unobservable characteristics. Rejected mothers were on average slightly better-off, based on observable characteristics at the time of the applciation. Authors say that the outcomes for boys of rjeected mothers provide a best-case scenario (upper bound) for what could be expected of beneficiaries in the absence of transfers.

Data collected on over 16,000 boys from 11 states who were born between 1900 and 1925, and whose mothers applied to the MP program; find that transfers increased longevity by about 1 year. Poorest families in the sample had longevity increased by 1.5 years of life. Results are robust to alternative function form specifications, counterfactual comparisons, and treatment of attrition. They interpret their results as the effect of cash transfers alone.

The authors match also a subset of the records to WWII enlistment and 1940 census records; cash transfers reduced the probability fo being underweight by half, increased educational attainment by 0.34 years, and increased income in early adulthood by 14 percent. They say that these mechanisms are responsible for 75 percent of the increase in longevitiy

6.2 Empirical Model

Basic empirical model

$$\log(\text{age at death})_{ifts} = \theta_0 + \theta_1 MP_f + \theta_2 X_{if} + \theta_3 Z_{st} + \theta_c + \theta_t + \epsilon_{if} \tag{2}$$

Model to address attrition and mutliple matches

$$P(\text{survived to age } a = 1)_{iftcs} = f(\theta_0 + \theta_1 MP_f + \theta_2 X_{if} + \theta_3 Z_{st} + \theta_c + \theta_t + \epsilon_{if}) \tag{3}$$

(i) above using unique matches – baseline , (ii) above using multiple matches – ahl baseline, (iii) above in bayesian framework – bayesian baseline, (iv) above using probabilities of matches (freq + bayesian versions)

Replicate Table 4 with different matching; different estimation techniques

Do a basic Bayesian log normal model for different states – pooled v. separate multiple matching