Birth, Death, and Record Linkage: Conducting Survival Analysis in the Presence of Linkage Error

Elan Segarra[*]

April 2020

Abstract

Longitudinal panels have a long history of use across the social sciences; however, they can be imperfect representations of reality when fuzzy linkage methods are employed during their creation. In this paper I study survival estimation (e.g. firm death, mortality, or emigration) when missed linkages induce error in the observed lifetime durations, and thus inconsistency in standard survival estimators. Importantly, the error introduced does not take the form of a standard competing risks model, and the methods developed here illustrate that consistency of the parameters of interest can be restored without correcting the linkages. This work makes three distinct theoretical contributions: First, under a known independent linkage error process I show that the marginal distribution of time to death is non-parametrically identified from linkage error induced durations, and I provide consistent estimators. Second, I provide sharp informative bounds on the marginal distribution of death when independence is relaxed. Third, when start and end periods are also observed, I show the marginal distribution of death can be point identified without imposing any dependence structure on the linkage error. The methods are then applied to longitudinal business data (where linkage error occurs due to firm relocation) to show that traditional estimates of survival rates of new firms are significantly overestimated. Finally, I discuss additional applications to the estimation of household migration and mortality where linkage error is induced by family name changes at marriage.

[*] University of Wisconsin–Madison, Department of Economics, United States

1 Introduction

Combining distinct data sets has not only been a common practice throughout social science research, but it has also been a pivotal step for cleverly answering some of the most important questions we have entertained. When data combination is accomplished without error, such as when unique individual identifiers can be leveraged, this step of the research process is trivial and often disregarded. In the absence of unique identifiers the possibility of linkage error implies that the produced data sets may be imperfect representations of reality which affect the validity of downstream analysis. This problem becomes particularly acute with survival analysis using longitudinal data, where error in linking individuals across time will directly affects the observed durations. Investigating the ramifications of record linkage error in this context, and providing novel estimators that account for it is the subject of this work.

One concrete example that will be examined thoroughly concerns the estimation of firm lifetimes and exit patterns. Declining firm dynamism has been observed across multiple sectors, and while recent research has considered economic explanations (Decker et al. 2016 and Akcigit and Ates 2019) few have considered data construction artifacts. Given the surprising lack of unique firm identifiers, the panel data utilized to study firm dynamics may be subject to linkage error which can bias standard survival analysis estimators. A particular problem stems from firm relocation which can induce linkage error and is prominent among young firms that are experiencing rapid growth. Since correcting the linkages can be extremely costly, and sometimes impossible given the observable data, empiricists need to account for the linkage error in downstream estimates. This work illustrates that under various assumptions about the linkage error process, the true distributions of firm lifetimes can be recovered without directly correcting the erroneous linkages in the panel data.

In this paper I ask what can be learned about the distribution of an event of interest using panel data subject to linkage error? Moreover under what situations (i.e. observables and linkage assumptions) can the object of interest be point identified, and when only partial identification is available, when will results still be informative? I make three distinct theoretical contributions. First, under a known independent linkage error process I show that the marginal distribution of time to the event of interest is non-parametrically identified from linkage error induced durations. In this scenario I provide consistent estimators and tools for inference. Second, I characterize the partially identified set and provide sharp informative bounds on the distribution of interest when the dependence between the event and linkage error is completely unrestricted. Third, when start and end periods are also observed, I show point identification of the distribution of interest can be restored without imposing any dependence structure. Finally an empirical application of the methods developed demonstrates that the true distribution of firm lifetimes can be recovered from panel data subject to record linkage error, and that traditional estimates of young firm exit are overestimated.

The work undertaken here crosses three very different strands of research: record linkage, survival analysis, and partial identification. Here I briefly discuss the previous related literature in each of these subfields as well as the contributions made by this project.

While there has been substantial work on the general theory of record linkage (Fellegi and Sunter 1969,Winkler 1999, Ridder and Moffitt 2007, Sadinle and Fienberg 2013, and Ruggles et al. 2018) there has been much less attention paid to addressing the implications of record linkage on downstream analysis. Nonetheless there have been recent acknowledgments that linkage procedures are imperfect and can have substantial effects on our analyses (Bailey et al. 2017). Most of the previous work on correcting this error has looked at linear regressions when the outcome and treatment reside in different files that must first be matched (Neter et al. 1965, Scheuren and Winkler 1993, Lahiri and Larsen 2005, and Hirukawa and Prokhorov 2018). Hof et al. (2017) is the most germane to the project at hand since it also tackles survival analysis in the presence of record linkage. However, their context concerns a situation with only two data files to be linked: one that contains durations and the other that contains covariates. In my work I consider the much more difficult, and pervasive, problem that occurs when multiple periods (data sets) are imperfectly linked together to form the panel data from which the durations are constructed. Rather than record linkage error simply involving the substitution of one individual's outcome or treatment with that of another's, I consider a scenario where the linkage error actually alters observed distributions.

Survival analysis has an extensive history that is summed up well in van den Berg (2001). As will become apparent when the model is described the work here is similar to the competing risks frameworks (Tsiatis 1975 and Heckman and Honor 1989) since the linkage error truncates durations before the event of interest. What makes the model at hand different and more complex is that linkage error will not only truncate durations, but also produce additional spurious observations representing the time between the linkage error event and the event of interest. Complicating the problem further is the notion that the event that occurs (i.e. linkage error or the event of interest) is unobserved, representing a major departure form the traditional competing risks framework. Most closely related in spirit is the work of Peterson (1976) which derived bounds on the latent distributions of interest.

Finally the partial identification literature (started by Manski 1989 and wonderfully surveyed by Molinari 2019) has received considerable attention in recent years. Of particular import for this project is work on partial identification in moment equality models (Chernozhukov et al. 2007 and Chernozhukov et al. 2013). Though I do not necessarily contribute to the expansion of the *theoretical* partial identification literature, this project does represent the first application of partial identification to the record linkage problem. While all previous work on record linkage error has focused on restoring point identification with strong assumptions or impractical requirements, I explore what can still be learned about the distribution of interest if we leave the record linkage error relatively unrestricted and approach this from a partial identification perspective.

The remainder of this paper proceeds as follows. In section 2 I describe the model that represents how record linkage error transforms the latent unobserved durations of interest into the start times and durations observed using imperfect panel data. Section 3 present theoretical results in scenarios where the researcher observes only durations, while section 4 showcases theoretical results when the start times of the durations are additionally observed. Section 6 describes the empirical application of the methods to the estimation of firm lifetimes, and section 7 concludes.

2 Model

In this section I describe the model which transforms the latent variables, which are the primary interest of the researcher, into the observed variables. More specifically the subsequent setup is meant to model the type of record linkage error that can occur as well as how it interferes with the true durations and the results of standard survival analysis estimators. Additionally I give examples mapping the theoretical objects to real world instances to give context to the type of situations this model is appropriate for.

A full table of notation can be found in the appendix, but a few general points are worth mentioning here. Throughout asterisks will indicate latent unobserved variables, and $\mathbb{1}$ represents the indicator function which is 1 when the argument is true and 0 otherwise. Arrows $(\vec{\cdot})$ over variables represent the vectorized dummy version of the discrete variable. More specifically $\vec{X}$ is a vector of 0s with a 1 in the index matching the value of X:

$$\vec{X} = \begin{bmatrix} \mathbb{1}\{X = 1\} \\ \mathbb{1}\{X = 2\} \\ \vdots \end{bmatrix}.$$

For clarity note that if the marginal distribution of discrete random variable X is f_x (structured in vector form so that $(f_x)_k = P(X = k)$), then we have $f_x = \mathbb{E}\left[\vec{X}\right]$.

2.1 title

2.2 Model of Latent and Observed Data

Let $(S_i^*, D_i^*, R_i^*) \in \mathbb{Z} \times \mathbb{N}^+ \times \mathbb{N}^+$ be a tuple of discrete unobserved random variables representing individual i's starting period, duration to an event of interest, and time to a record linkage error event respectively. More specifically D_i^* is the time until an event of interest measured relative to the individual's 'birth' period, S_i^*. In other words if it could be observed, an individual would start being tracked in period $t = S_i^*$ and the event of interest would occur in period $t = S_i^* + D_i^* + 1$ so that the individual had a lifetime duration of D_i^*. The primary goal of the researcher is to learn information about D_i^* such as its mean or distribution. The number of truly distinct individuals in a sample is denoted by n^*.

In a survival analysis context the main problem resulting from record linkage error is that it breaks true durations into smaller constituent parts. To stay within a survival analysis framework I model record linkage error as another event which prevents the record at that time from being linked with the record in the previous time period. Let R_i^* be the time until a record linkage error event (RLEE) occurs relative to S_i^*. Thus a RLEE occurring in period $S_i^* + R_i^*$ indicates that the record in period $S_i^* + R_i^*$ was not successfully linked with the appropriate record in period $S_i^* + R_i^* - 1$. A full accounting of relevant record linkage mechanisms and how their properties translate into the distribution of R_i^* is presented in appendix A.1.

Depending on the timing of the RLEE the true duration may end up broken or remain intact. Let r_i^* be an indicator of whether a record linkage error breaks the true duration for individual i, where $r_i^* = \mathbb{1}\{R_i^* < D_i^*\}$. When this happens the true duration gets split into two smaller durations:

$$(S_{i1}^*, D_{i1}^*) = (S_i^*, R_i^*) \qquad \text{and} \qquad (S_{i2}^*, D_{i2}^*) = (S_i^* + R_i^*, D_i^* - R_i^*)\,.$$

The first start and duration (denoted by (S_{i1}^*, D_{i1}^*)) represents the time between the start and the linkage error while the second start and duration ((S_{i2}^*, D_{i2}^*)) is the time between the linkage error and the event of interest. If the RLEE happens at the time of or after the event of interest, $R_i^* \geq D_i^*$, then no breakage occurs, and there is only a single start and duration matching the truth

$$(S_{i1}^*, D_{i1}^*) = (S_i^*, D_i^*)\,.$$

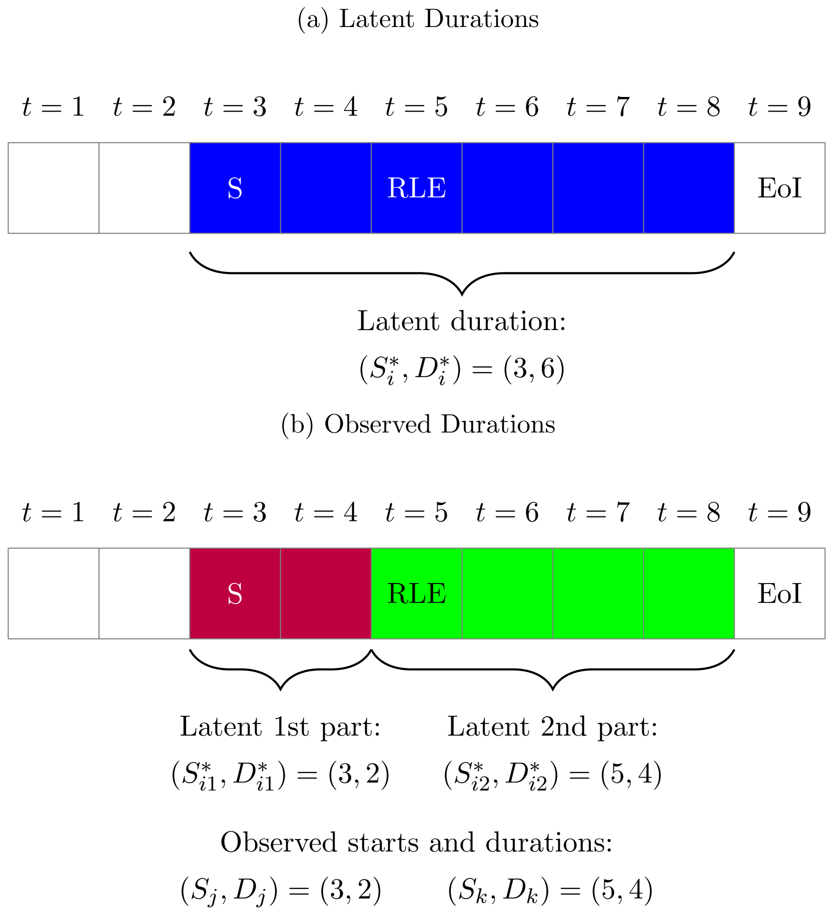
Refer to figure 1 for an example of this duration breakage for a specific individual.

The process of duration breakage can be described by the creation of an unbalanced panel which is then subsequently flattened and permuted. Those with unbroken durations have one tuple consisting of a latent start and duration, while those with broken durations have two tuples of starts and durations. Table 1 displays an unbalanced panel that would result from a fictitious data set where individuals 1, 3, and 5 have durations broken by the RLEE.

The final step takes all of the broken and unbroken durations and both flattens them and randomizes the indices. Letting $n_b^* = \sum r_i^*$ and $n_u^* = n^* - n_b^*$ denote the number of individuals with broken and unbroken durations respectively, we then have a total of $n = n_u^* + 2n_b^*$ tuples of starts and durations. Without loss of generality assume the individuals are ordered so that $r_i^* = 0$ for all $i = 1, \ldots, n_u^*$ and $r_i^* = 1$ for all $i = n_u^* + 1, \ldots, n^*$. Define the random permutation operator

$$\pi : \{1, \ldots, n\} \to \{(i,1)\}_{i=1}^{n_u^*} \cup \{(i,1), (i,2)\}_{i=n_u^*+1}^{n^*}$$

which randomly assigns the univariate indices of the to be observed sample, to the bivariate indices of all the broken and unbroken durations. Finally let an observed start and duration, (S_i, D_i), be


(a) Latent Durations

(b) Observed Durations

Figure 1: The above figures illustrate the latent durations (1a) and observed durations (1b) for an individual with $(S^*, D^*, R^*) = (3, 6, 2)$. The various events of start (S), the event of interest (EoI), and a record linkage error event (RLEE) are indicated in the appropriate time periods.

i	(S_i^*, D_i^*, R_i^*)	r_i^*	(S_{i1}^*, D_{i1}^*)	(S_{i2}^*, D_{i2}^*)
1	(3, 6, 2)	1	(3, 2)	(5, 4)
2	(1, 5, 5)	0	(1, 5)	-
3	(2, 3, 2)	1	(2, 2)	(4, 1)
4	(2, 3, 4)	0	(2, 3)	-
5	(4, 5, 2)	1	(4, 2)	(6, 3)
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 1: Example of latent unbalanced panel produced by record linkage error breaking some durations into smaller constituent parts.

mapped from one of these latent start duration tuples:

$$(S_i, D_i) = \left(S^*_{\pi_1(i)\pi_2(i)}, D^*_{\pi_1(i)\pi_2(i)}\right).$$

Thus at the end of the record linkage process (with linkage error) a researcher observes a sample of size n of either durations alone or start times and durations,

$$\{D_i\}_{i=1}^n \qquad \text{or} \qquad \{(S_i, D_i)\}_{i=1}^n$$

where these two cases will be treated separately since they have different ramifications on identification and estimation.

Remark 1. When a researcher observes a duration, D_i, they have no knowledge of whether that duration corresponds to a true unbroken duration, D_i^*, the first half of a broken duration, R_i^*, or the second half of a broken duration, $D_i^* - R_i^*$, without further information.

Remark 2. If there is record linkage error, then the size of the observed sample, n, is strictly larger than the number of truly distinct individuals (i.e. the size of the latent sample, n^*). Specifically we have $n = n_u^* + 2n_b^* = n^* + n_b^*$.

There are two assumptions that are maintained throughout this paper and they are codified here.

Assumption A1 (Latent IID Sample). $\{(S_i^*, D_i^*, R_i^*)\}_{i=1}^{n^*}$ *forms an independent and identically distributed sample of size* n^*.

Assumption A2 (Finite Support). D_i^* *and* R_i^* *have finite support with maximums given by* $H^{d*} = \max support D_i^*$ *and* $H^{r*} = \max support R_i^*$.

This first assumption is the analog of the standard iid assumption, with the subtle difference being that it is made on the latent sample. As will be discussed in subsequent analysis, the observed sample is in fact not *iid* since some pairs of observations come from the same individual and are thus correlated.

The second assumption is also fairly innocuous since distributions with infinite support will never by fully identified with finite data. Alternatively this can be viewed as a slight transformation of a variable with infinite support where all the mass in the tail past a certain threshold is aggregated in that threshold. A natural cut off point would be the largest duration observed in the sample since only the probabilities of durations less than that will have any hope of being estimated.

2.3 Examples

To help further elucidate how the above model functions I present two examples here: one in the context of firm dynamics and another in the context of individual migration. In each I describe the real world counterparts to theoretical objects defined above and remark on any idiosyncracies related to the context.

Example 1 (Firm Dynamism)**.** Consider a researcher interested in firm dynamics with a specific focus on survival rates of young firms. In other words your event of interest is the death or exit of a firm (D^*) relative to the birth of the firm (S^*). To investigate this, panel data is created by linking yearly censuses of firms where existence in each census is an indicator of that firm being active that period. It may come as a surprise, but finding unique firm identifiers to link perfectly across years is no easy task. Even objects such at Employer Identification Numbers (EINs) are not necessarily unique to a firm given the incentive to change them during the life of a firm to reap lower unemployment insurance rates (). This means that employer name and address are often used to link firms across years. Unfortunately when firms relocate, perhaps because of growing pains or business contraction, this can easily result in linkage error. In this context, R^* represents the time from birth until a firm changes location.

Given that broken durations are necessarily smaller than the true time to death, this linkage error will, in general, give the impression that firms are living shorter lives on average. Perhaps more problematic is that the model moves much of the mass of durations onto the left tail, meaning that even a small amount of linkage error can result in much high estimated rates of death among young firms than actually occurs.

Example 2 (Individual Migration)**.** Consider a researcher investigating migration flows in and out of the United States and when they occur during the lifetime of an individual. The event of interest here would be the time between birth or immigration (S^*) and emigration out of the country (D^*). In the United States, Social Security Numbers (SSNs) theoretically function as unique identifiers and could be used to link perfectly to create an accurate panel of people. Unfortunately most surveys or censuses that are linked across time do not contain this information. Those datasets that do potentially have access to this microdata, such as the Longitudinal Employer-Household Dynamics (LEHD) file are highly restricted and thus not available to the majority of researchers. When it comes to creating these panels the first name, last name, date of birth, and gender become important linking variables. Unfortunately a large percentage of women change their last name after getting married, which creates problems for this linking strategy. In this context the time until marriage (R^*) will represent the record linkage error event.

This scenario is further complicated by the fact that names and birth dates are often deemed too personal to release to researchers, even those accessing the restricted versions of these datasets. Therefore when these linking problems exist, the first best solution of fixing the linkages is not even available to researchers.

One mitigating element to this context is that when age or date of birth are available (which is common) durations can be more accurately measured even in the presence of linkage error. For example if S_i^* is the year they were born, then even the second half of a broken duration can be measured correctly ($D_{i2}^* = D_i^*$ instead of $D_{i2}^* = D_i^* - R_i^*$). However problems still persist since the first half of the broken link, $D_{i1}^* = R_i^*$ will remain and affect downstream survival analysis.

2.4 Properties of Observed Durations

In this section I describe various properties of the observed distribution that will be useful for discussions of identification and estimation in subsequent sections. Specifically I describe the distribution of the observed durations as it relates to the latent distributions, and discuss how the means and probability of small durations relate among latent and observed distributions.

Constructing the probability mass function of the observed durations, D_i, is fairly straightforward provided care is taken regarding the increase in sample size from the latent sample to the observed sample. Denote the joint distribution function of D^* and R^* by $f_{RD}^*(i,j) \equiv P\left(R^* = i, D^* = j\right)$, and denote the distribution function of D_i by $f_d(k) \equiv P\left(D = k\right)$. The distribution of observed durations is then given by

$$f_d(k) = P\left(D = k\right) = \frac{1}{\chi}\left[\underbrace{\sum_{r=k}^{\infty} f_{RD}^*(r,k)}_{\text{unbroken}} + \underbrace{\sum_{d=k+1}^{\infty} f_{RD}^*(k,d)}_{\text{broken 1st half}} + \underbrace{\sum_{d-r=k} f_{RD}^*(r,d)}_{\text{broken 2nd half}}\right] \tag{1}$$

$$\text{where} \quad \chi = 1 + P\left(R^* < D^*\right) = 1 + \sum_{r<d} f_{RD}^*(r,d) \tag{2}$$

This distribution exhibits numerous characteristics, some of which are intuitive and others completely unintuitive. A brief exploration of a few of these will hopefully convince the reader that exploring identification and estimation in this framework is nontrivial and worthwhile.

To start we consider how the mean of the observed durations compares to the mean of the true durations of interest (proofs of Theorem 1 and 2 are found in appendix A.2).

Proposition 1. $\mathbb{E}\left[D_i\right] = \frac{1}{\chi}\mathbb{E}\left[D_i^*\right]$ *where* $\chi = 1 + P\left(R^* < D^*\right)$.

This result has a nice elegance because it confirms what our intuition suggests, mainly that the observed durations are smaller on average than the truth. It goes further by displaying that the attenuation is only a function of the probability of a linkage error. A simple corollary to this theorem guarantees that the true mean is identified from the observed mean if either the probability of an error is known or equivalently if the true number of individuals in the latent sample is known (because $(n - n^*)/n^* \to_p P\left(R^* < D^*\right)$).

Continuing with our intuition and the previous result it would be reasonable to surmise that we always have a shift in the probability mass toward the left tail. More specifically since durations can only be broken and the smallest duration possible is of length 1 we might expect the probability of observing a duration of 1 to be weakly larger than the probability of the latent duration being 1. However the following theorem reveals that this is not at all guaranteed.

Proposition 2. *If* $\max support(D_i^*) \leq 3$ *then* $P(D_i = 1) \leq P(D_i^* = 1)$ *with strict inequality when* $P(R^* < D^*) > 0$*. If* $\max support(D_i^*) > 3$ *then the sign of* $P(D_i = 1) - P(D_i^* = 1)$ *is ambiguous without further information.*

This result demonstrates that there exist distributions where the linkage error does not merely shift mass to smaller observed durations, but instead move mass around in nonintuitive ways. This can happen when there is a large chance of a broken duration, but little relative chance that either of the broken durations have length 1. For example a joint distribution with more mass on outcomes of the form $R_i^* \approx D_i^*/2$ (i.e. durations are often split in half) can lead to this result.

Given that the transformation of D_i^* to D_i is not so well-behaved that identification or bounding results are obvious, this further motivates the investigation into what can be learned about D_i^* under various assumptions about observables and the linkage model.

3 Point Identification and Estimation

- Independence and Observed Durations

In this section I present scenarios and assumptions that allow for point identification of the objects of interest. In addition to identification, estimation and inference results are also discussed. Point identification is obtained under two different scenarios and are considered separately. In section REF SECTION the researcher observes a sample of durations, $\{S_i\}$, of size n, but critically does not observe the starting periods of each of these durations. The subsequent results are split into two sections depending on relationship between the durations of interest and the record linkage error events. In section 3.1 we assume independence resulting in point identification of the distribution of event durations, while in section 3.1 we place no restriction on the dependence structure and explore partially identified set of distributions.

3.1 Independent Durations

A natural place to begin our investigation is under the particularly strong assumption of independence between the duration of interest, D_i^*, and the time until a RLEE, R_i^*. While this assumption is most likely a difficult to maintain in many scenarios, it is still instructive and provides a foundation for more general cases to be discussed in subsequent sections. **Discuss analogs with competing risks and convolution**

Assumption A3 (Independence)**.** $R^* \perp D^*$

Let the marginal distributions of the durations of interest and the duration until RLEE be denoted by the vectors $f_R^* \equiv \begin{bmatrix} f_R^*(1) & \cdots & f_R^*(H) \end{bmatrix}'$ and $f_D^* \equiv \begin{bmatrix} f_D^*(1) & \cdots & f_D^*(H) \end{bmatrix}'$ respectively.

If f_R^* is either known (or estimated from secondary data) then we can achieve point identification under the assumption of independence if an additional support condition holds.

Assumption A4 (Support Condition)**.** $\max \mathit{support}\,(D_i^*) \leq \max \mathit{support}\,(R_i^*)$.

One implication of this assumption is that all durations have a chance of being broken by record linkage error. Under our previous assumptions assumption A4 becomes a necessary and sufficient condition for point identification of the distribution of D_i^*.

Theorem 1 (Identification)**.** *Under assumptions A1-A3 if f_R^* is identified and $\{D_i\}_{i=1}^n$ is observed then*

$$f_D^* \text{ is point identified} \qquad \Leftrightarrow \qquad \text{A4 holds.}$$

The proof of Theorem 1 proceeds quite naturally after formulating the relationship between the distributions as a linear system of equations (refer to the full proof in appendix A.2.2). The intuition for this identification result comes from the idea that we can essentially 'unzip' the distribution of D_i^* from the right tail of the distribution of D_i.

First note that $D_i = t$ implies that $D_i^* \geq t$ which suggests that observing the likelihood of $D_i = t$ tells us about the likelihood of $D_i^* = t$ and $D_i^* = t + 1$ and $D_i^* = t + 2$ etc. It can also be shown that if the support condition holds then the observed durations will have the same support as the underlying event of interest. Therefore the probability of the longest observed duration, $D_i = H$, will be proportional to the probability of the longest latent duration, $D_i^* = H$, and thus that probability is identified (up to scale). Similarly the second longest observed duration, $D_i = H - 1$ is related only to the longest latent duration, $D_i^* = H - 1$, the second longest $D_i^* = H - 2$, and the distribution of R_i^*. Since we know about all of these objects except the likelihood of $D_i^* = H - 2$, that identifies the likelihood of $D_i^* = H - 2$. Continuing in this fashion allows identification of the entire distribution.

If the support condition doesn't hold, meaning $\max \text{support}\,(D_i^*) > \max \text{support}\,(R_i^*)$, then this strategy fails right at the beginning. Since the maximum value of D_i is one less than that of D_i^* the probability of $D_i = H$ depends on both the probability of $D_i^* = H$ and $D_i^* = H + 1$. Without further information these probabilities can never be disentangled and the entire distribution remains unidentified.

The system of linear equations that relates the distribution of observed durations to latent durations suggests a natural estimator for f_D^*. Let $\vec{D}_i$ be a $H \times 1$ vector of dummy variables representing the outcome of observed individual i,

$$\vec{D}_i = \begin{bmatrix} \mathbb{1}\{D_i = 1\} & \mathbb{1}\{D_i = 2\} & \cdots & \mathbb{1}\{D_i = H\} \end{bmatrix}'.$$

Following in the spirit of the proof for identification consider the estimator, $\widehat{f_D^*}$, of f_D^* defined as

$$\widehat{f_D^*} \equiv \frac{A_{r*}^{-1}\frac{1}{n}\sum_{i=1}^{n}\vec{D}_i}{\|A_{r*}^{-1}\frac{1}{n}\sum_{i=1}^{n}\vec{D}_i\|_1}, \tag{3}$$

where A_{r*} is an $H \times H$ matrix which is upper diagonal and only a function of the distribution of R_i^*.

Though the estimator is a simple linear transformation of a standard mean estimator, consistency does not immediately follow because the observed sample is not *iid*. For individuals whose true duration was broken the two observed durations are correlated, meaning the standard weak law of large numbers does not immediately apply. Nonetheless this estimator is consistent for f_D^* as established by the following theorem (proved in appendix A.2.2).

Theorem 2 (Consistency)**.** *Under assumptions A1-A3 if f_R^* is known then $\widehat{f_D^*} \to_p f_D^*$.*

Thinking of the correlated observations as belonging to clusters leads to a simple proof of consistency that does not require any additional assumptions beyond those leveraged for identification.

Remark 3. The distribution of R_i^* can be estimated from a second independent sample and does not need to be known. As long as the estimator of R_i^* is also consistent then it can be plugged into (3), and that estimator will still be consistent for f_D^*.

The structure of the distribution of D_i is also rich enough to imply that the estimator is asymptotically normal without further assumptions.

Theorem 3 (Asymptotic Normality)**.** *Under assumptions A1-A4*

$$\sqrt{n}\left(\widehat{f_D^*} - f_D^*\right) \to_d N\left(0, \chi^2 (A_{r*}^{-1})'\Omega A_{r*}^{-1}\right)$$

where

$$\Omega = \frac{P(R^* < D^*)}{\chi}\Omega_b + \frac{P(R^* \geq D^*)}{\chi}\Omega_u, \qquad \chi = 1 + P\left(R^* < D^*\right),$$

and Ω_b and Ω_u are conditional variances that are completely determined by f_D^* and f_R^*.

Once again care must be taken when proving Theorem 3 because the sample is not *iid* (full proof found in appendix A.2.2). While the asymptotic theory for clustered samples developed in Hansen and Lee (2019) cannot be used directly (because their theorems require nonsingular covariance matrices) the general approach can be adapted to the scenario at hand. The fingerprints of the clustering are evident in the asymptotic covariance matrix, Ω, where the two parts, Ω_b and Ω_u, correspond with the clusters of broken durations and clusters of unbroken durations respectively. More specifically

$$\Omega_u = Var\left(\vec{D}_{1,i}^u\right) \qquad \Omega_b = Var\left(\vec{D}_{1,i}^b + \vec{D}_{2,i}^b\right)$$

where $D^u_{1,i}$ are the durations corresponding to individuals with $r^*_i = 0$ and $\left(D^b_{1,i}, D^b_{2,i}\right)$ are the duration of individuals with $r^*_i = 1$.

In standard clustered sample scenarios one could use the consistent covariance estimators proposed by Hansen and Lee (2019) however we do not observe the clusters as required by their method. While this would normally be insurmountable, since the asymptotic covariance matrix, Ω, is entirely determined by f^*_D and f^*_R there is an alternative route to it's estimation. The plugin estimator of Ω, using $\widehat{f^*_D}$, is likely to be consistent given the simple nature of Ω, but this still needs to be rigorously proven.

3.2 Dependent Durations

As mentioned earlier the assumption of independence is strong and unlikely to be tenable in many situations. In this section we relax this assumption and allow the distribution between R^*_i and D^*_i to be completely unrestricted. In related situations, e.g. a competing risks or convolution framework, allowing dependence between the two input distributions results in the loss of point identification and we will find the same result here. However there is still useful information when considered from a partial identification perspective.

All previous results discussed here pertain to an environment where the researcher only observes durations, however it is very common for the researcher to have access to the panel from which these durations were constructed. In standard survival analysis frameworks having the panel provides little extra benefit, but I will illustrate that there is ample extra identification power to be leveraged in the presence of record linkage error. Throughout this section I assume that the researcher observes a sample of start times and durations, $\{(S_i, D_i)\}_{i=1}^n$, opposed to just durations.

The intuition behind the extra identification power is best illustrated in the visual example of a panel data set found in figure 2. Every row represents a different observation while each column is a different time period, and a cell is filled in if that 'individual' was observed in that time period. Keep in mind that this is the panel data observed after linking across time (possibly with error) so that individual A's duration of length 2 could be the true time to the event of interest or the first half of a broken duration (where linkage error prevented linking them between period 2 and 3). Having the start times available means there is extra information in the observed adjacencies between individuals. For example individual B could be an unbroken duration or individuals B and E could be the same individual that experienced a linkage error event in period 5 (thus breaking a duration of 5 into 3 and 2).

I will discuss two methods for exploiting this extra information: a distributional approach and a reconstruction approach. In the distribution approach we leverage the fact that the distribution over latent start time, event durations, and linkage error events implies a distribution over observed start times and durations. In the reconstruction approach use the adjacencies to construct potential

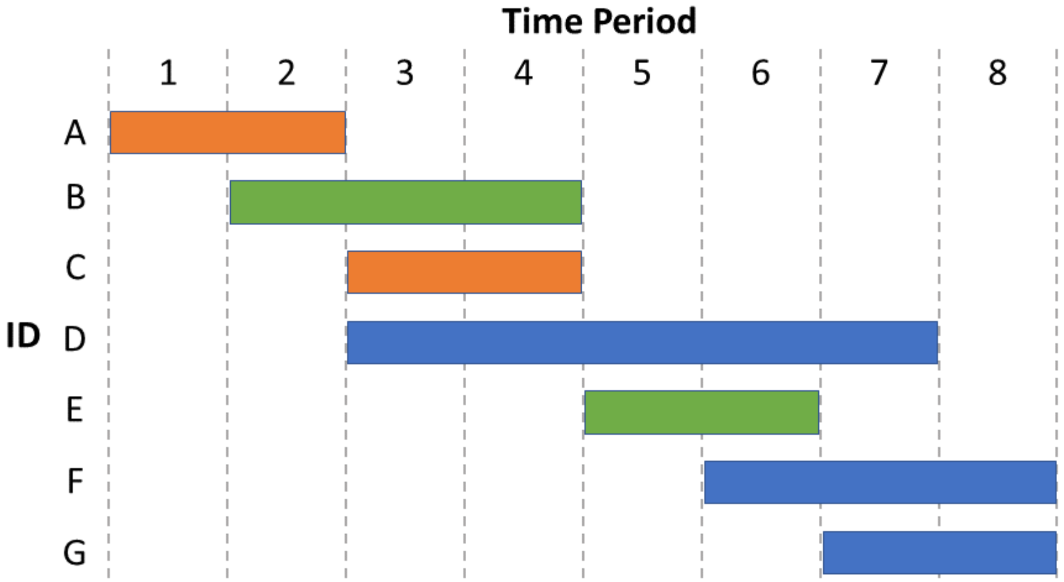


Figure 2: Visual example of panel data (with linkage error) illustrating potential relationships among observed durations.

unobserved panel on which standard survival analysis methods can be applied.

3.3 Distributional Approach

Throughout the previous section there was an implicit assumption regarding a lack of censoring, but making this explicit here will be important.

Assumption A5 (Missing Constituents)**.** *If* $r_i^* = 1$ *then both parts of the broken duration,* R_i^* *and* $D_i^* - R_i^*$ *are in the observed sample.*

This assumption merely guarantees that no durations are left out of the observed sample, meaning if we observe the first half of a broken duration, then the second half must also be in the sample somewhere. Additional assumptions to obtain identification results are also given here.

Assumption A6 (Independence of Start)**.** $S_i^* \perp R_i^*$ *and* $S_i^* \perp D_i^*$*.*

Assumption A7 (Terminal Event)**.** $P\left(R_i^* > D_i^*\right) = 0$*.*

Note that assumption A7 is not requiring that every duration be broken by a record linkage event because it still permits $R_i^* = D_i^*$ which does not result in breakage. What it is saying is that the event of interest is terminal (i.e. like a a death event) so that the record linkage error event cannot occur afterward. Essentially all this is imply that all cell below the diagonal of f_{RD}^* are 0. Together these assumptions yield the following identification result

Theorem 4 (Identification). *Under assumptions A1, A2, A4, A5, and A6 the marginal distribution of D_i^*, f_D^* is point identified.*

If additionally A7 holds then the entire joint distribution f_{RD}^* is point identified.

Remark 4. The assumptions above are all sufficient for point identification, but it is unlikely that they are necessary.

4 Partial Identification and Estimation

/

Some additional notation relevant to this section is noted here. Let the matrix form of the joint distribution of R_i^* and D_i^* be denoted by f_{RD}^*,

$$f_{RD}^* = \begin{bmatrix} f_{RD}^*(1,1) & f_{RD}^*(1,2) & \cdots & f_{RD}^*(1,H) \\ f_{RD}^*(2,1) & f_{RD}^*(2,2) & \cdots & f_{RD}^*(2,H) \\ \vdots & \vdots & \ddots & \vdots \\ f_{RD}^*(H,1) & f_{RD}^*(H,2) & \cdots & f_{RD}^*(H,H) \end{bmatrix} \quad \text{where } f_{RD}^*(i,j) = P(R^* = i, D^* = j)$$

and let $\text{vec}\left(f_{RD}^*\right)$ denote the vectorization of that matrix (i.e. stacking the columns into a single column vector). Let Δ^K denote the probability $k-$simplex representing the set of discrete distributions over $\{1, 2, \ldots, K+1\}$

$$\Delta^K = \left\{ p \in [0,1]^{K+1} : \sum_{k=1}^{K+1} p_k = 1 \right\}.$$

If we denote the partially identified set of joint distributions by $\mathcal{H}\left(f_{RD}^*\right)$ then the set can be defined by the following moment equality

$$\mathcal{H}(f_{RD}^*) = \left\{ f_{RD}^* \in \Delta^{H^2-1} : \mathbb{E}\left[\vec{D}_i + \left(\vec{D}_i b_H' - A_H \right) \text{vec}(f_{RD}^*) \right] = 0 \right\}. \tag{4}$$

In the above definition A_H is an $H \times H$ matrix, b_H is an $H \times 1$ vector, and both are constant, known, and only depend on H. **Put link to examples of these matrices** This moment equality essentially defines the transformation of the joint distribution of R_i^* and D_i^* into the distribution of D_i. Therefore any characteristics of the identified set (such as bounds on expectations or marginal distributions) will necessarily be sharp as they include all information available about the latent joint distribution. This set almost surely not a a singleton because there are H moment equations but H^2 unknown parameters.

If the marginal distribution of R^*, f_R^*, is also known then we can define a further restricted identified set, $\mathcal{H}_1\left(f_{RD}^*\right)$,

$$\mathcal{H}_1(f_{RD}^*) = \left\{ f_{RD}^* \in \Delta^{H^2-1} : \mathbb{E}\left[\vec{D}_i + \left(\vec{D}_i b_H' - A_H \right) \text{vec}(f_{RD}^*) \right] = 0 \text{ and } f_R^* - M_H \text{vec}(f_{RD}^*) = 0 \right\}. \tag{5}$$

The matrix M_H is simply the linear transformation from the joint distribution to the marginal distribution of R_i^* (and thus is constant, known, and only depends on H).

Since these partially identified sets are characterized by moment equalities we can apply the tools developed in Chernozhukov et al. (2007) (henceforth CHT) to produce both consistent estimators and confidence regions for $\mathcal{H}(f^*_{RD})$ and $\mathcal{H}_1(f^*_{RD})$. Moving forward all results in this section will be with respect to estimating $\mathcal{H}(f^*_{RD})$, however they trivially extend to estimation of $\mathcal{H}_1(f^*_{RD})$.

We start by defining the population criterion function and sample criterion functions

$$Q(f_{md*}) = \left\| \mathbb{E}\left[\vec{D}_i + \left(\vec{D}_i b'_H - A_H\right) \operatorname{vec}(f^*_{RD})\right] \right\|^2 \tag{6}$$

$$Q_n(f_{md*}) = \left\| \frac{1}{n}\sum_{i=1}^{n} \left[\vec{D}_i + \left(\vec{D}_i b'_H - A_H\right) \operatorname{vec}(f_{md*})\right] \right\|^2 . \tag{7}$$

These criterion functions correspond to the more general form described in CHT with the weight matrix taken to be the identity. Note that set of minimizers of (6) correspond exactly with the identified set, $\mathcal{H}\left(f^*_{RD}\right)$, which is what inspires the set estimator

$$\widehat{\mathcal{H}}(f^*_{RD}, c) = \left\{ f^*_{RD} \in \Delta^{H^2-1} : Q_n(f^*_{RD}) \leq \frac{1}{n} c \right\}, \tag{8}$$

where c is a constant that parameterizes the contour set of the criterion function used in the estimator. This estimator, $\widehat{\mathcal{H}}(f^*_{RD}, c)$, will serve as both the set estimator and the confidence region of $\mathcal{H}(f^*_{RD})$.

Our goal here is to have a consistent set estimator, where a set estimator is *consistent* provided the distance between the estimator and identified set converges in probability to 0 (in the Hausdorff metric). Applying the work of CHT to our scenario provides the following consistency result.

Theorem 5 (Consistency)**.** *Under assumptions A1 and A2 if $c \geq \mathcal{C}_n$ where*

$$\mathcal{C}_n = \sup_{f^*_{RD} \in \mathcal{H}(f^*_{RD})} nQ_n(f^*_{RD}),$$

then $d_{haus}\left(\widehat{\mathcal{H}}(f^*_{RD}), \mathcal{H}(f^*_{RD})\right) = o_p(1)$ and $\mathcal{H}(f^*_{RD}) \subseteq \widehat{\mathcal{H}}(f^*_{RD})$ w.p. approaching 1.

The above result gives conditions on the contour threshold, c, that will imply our set estimator in (8) is consistent for the true partially identified set. While intuition would suggest $c = 0$ as a natural threshold, CHT show that problems can arise if the threshold converges to 0 faster than the rate at which the sample criterion function converges to the population criterion function.

The work of CHT also allow us to choose an alternative threshold so that our estimator has a confidence region property. The estimator, $\mathcal{H}(f^*_{RD})$, is a $1-\alpha$ confidence region if $P\left(\mathcal{H}\left(f^*_{RD}\right) \subseteq \widehat{\mathcal{H}}\left(f^*_{RD}\right)\right)$ goes to α as n goes to infinity.

Remark 5. One idiosyncracy of using the set estimators proposed by CHT concerns how the set estimate compares to a confidence region. Due to the nature of the estimator definition it is possible for the confidence region to be a proper subset of the estimate of the identified set. Given that this is a rather unintuitive property future work will investigate and apply the half-median unbiased estimators proposed in Chernozhukov et al. (2013).

5 Monte Carlo Simulations

In this section I present Monte Carlo simulation results of the estimators presented in the previous two sections.

5.1 Observed Durations

5.2 Observed Start and Stop Times

6 Empirical Application

In this section I discuss the empirical application of the methods described earlier to the case of firm dynamics. As introduced in example 1 the estimation of firm dynamics, and especially the contribution and dynamism of young firms, is of substantial interest to economists. Unfortunately given the lack of unique identification numbers many firm longitudinal data sets were created using less ideal linking variables including firm name and address. Since firm relocation can then result in record linkage error we take R_i^* to be the time until firm i changes address.

The data set first utilized in this application was the Longitudinal Business Database (LBD) as described in Jarmin and Miranda (2002). This annual panel contains information on all firms and establishments in the United States that have at least one individual on their payroll in a given year. Though the full data set covers 1975-2016 however I focus only on data from 1990-2016 due to data quality issues associated with the early years. The methods described in section 3 under the assumption of independence between firm relocation and time to firm death were applied to the Agriculture Services Sector of the economy (NAICS 1151XX). Initial results showed that record linkage error lead to significant overestimation of the probability of firm death in the first 5 years of operation. Unfortunately the Covid-19 pandemic necessitated the temporary closure of all Federal Research Data Centers which house the LBD, so further investigation using this data has been suspended.

Subsequent application of the methods developed here will instead utilize the Your-economy Time Series (YTS) data provided through the Wisconsin Business Dynamics Research Consortium. This is also a panel of U.S firms and establishments created from the Infogroup Business Historical Databases and covers 1997-2019. One potential hurdle with this panel is that it does not include establishment addresses, so business relocation is difficult to perfectly identify. The smallest geographic information included is the zip code, so subsequent analyses using this data will use zip code changes over time to indicate firm relocation. Application of the methods described above are currently ongoing.

7 Conclusion

In this paper I have explored the estimation of duration models in the presence of record linkage error during data construction. Since even minor record linkage error can cause fairly substantial error in standard analysis, the issue should be addressed in the estimation if the linkages themselves cannot be improved. This problem can be accounted for by either imposing extra structure to point identify the distribution of interest or by using partial identification methods to analyze the set of estimates that are rationalized by the observed data. In the former situation I have shown that either independence of the record linkage process or observation of start times is sufficient for point identification of the marginal distribution of interest. Additionally I have provided estimators and inference methods in these situations. In the latter scenario I have adapted standard partial identification methods to both estimate the partially identified set and provide confidence regions. All available information is leveraged in these estimators so further statistics derived from the set estimator, such as bounds on survival probabilities, will be sharp. Finally I have begun to apply the methods developed to longitudinal business data where firm relocation is a major cause of record linkage error. Initial results show that failing to account for the linkage error can lead to survival rates of young firms being significantly overestimated.

References

Akcigit, U. and Ates, S. (2019). Ten facts on declining business dynamism and lessons from endogenous growth theory.

Bailey, M., Cole, C., Henderson, M., and Massey, C. G. (2017). How well do automated linking methods perform in historical samples? evidence from new ground truth.

Chernozhukov, V., Hong, H., and Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models. *Econometrica*, 75(5):1243–1284.

Chernozhukov, V., Lee, S., and Rosen, A. M. (2013). Intersection bounds: Estimation and inference. *Econometrica*, 81(2):667–737.

Decker, R. A., Haltiwanger, J. C., Jarmin, R. S., and Miranda, J. (2016). Where has all the skewness gone? the decline in high-growth (young) firms in the u.s. *European Economic Review*, 86:4–23.

Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210.

Hansen, B. E. and Lee, S. (2019). Asymptotic theory for clustered samples. *Journal of Econometrics*, 210(2):268–290.

Heckman, J. J. and Honor, B. E. (1989). The identifiability of the competing risks model. *Biometrika*, 76(2):325.

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

Hof, M. H., Ravelli, A. C., and Zwinderman, A. H. (2017). A probabilistic record linkage model for survival data. *Journal of the American Statistical Association*, 112(520):1504–1515.

Jarmin, R. S. and Miranda, J. (2002). The longitudinal business database. *SSRN Electronic Journal*.

Lahiri, P. and Larsen, M. D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, 100(469):222–230.

Manski, C. F. (1989). Anatomy of the selection problem. *The Journal of Human Resources*, 24(3):343.

Molinari, F. (2019). Econometrics with partial identification.

Neter, J., Maynes, E. S., and Ramanathan, R. (1965). The effect of mismatching on the measurement of response error. *Journal of the American Statistical Association*, 60(312):1005–1027.

Peterson, A. V. (1976). Bounds for a joint distribution function with fixed sub-distribution functions: Application to competing risks. *Proceedings of the National Academy of Sciences*, 73(1):11–13.

Ridder, G. and Moffitt, R. (2007). *The Econometrics of Data Combination*, volume 58, chapter Ch. 75, pages 5469–5547.

Ruggles, S., Fitch, C. A., and Roberts, E. (2018). Historical census record linkage. *Annual Review of Sociology*, 44(1):19–37.

Sadinle, M. and Fienberg, S. E. (2013). A generalized fellegi-sunter framework for multiple record linkage with applications to homicide record systems. *Journal of the American Statistical Association*, 108(502):385–397.

Scheuren, F. and Winkler, W. E. (1993). Regression analysis of data files that are computer matched - part i. *Survey Methodology*, 19(1):39–58.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, 72(1):20–22.

van den Berg, G. J. (2001). *Duration Models: Specification, Identification and Multiple Durations*, chapter 55, pages 3381–3460. Elsevier.

Winkler, W. E. (1999). The state of record linkage and current research problems.

A Appendix

D^*	True duration until the event of interest
R^*	True duration until the splitting event
D	Duration observed by the econometrician
r_i^*	Indicates whether individual i's duration is broken; $r_i^* \equiv \mathbb{1}\{R_i^* < D_i^*\}$
n^*	Number of latent (unobserved) individuals
n_u^*	Number of latent individuals with broken durations; $n_u^* \equiv \sum r_i^*$
n_u^*	Number of latent individuals with unbroken durations; $n_u^* \equiv n^* - n_b^*$
n	Number of observed individuals (note $n \geq n^*$)
H^{d*}	Maximum duration until event of interest; $\max supp(D^*)$
H^{r*}	Maximum duration until split event; $\max supp(R^*)$
H	Maximum observed duration; $\max supp(D)$
$f_D^*(i)$	Marginal probability mass function of main event duration; $P\left(D^* = i\right)$
$f_R^*(i)$	Marginal probability mass function of splitting event duration; $P\left(R^* = i\right)$
$f_d(i)$	Marginal probability mass function of observed duration; $P\left(T = i\right)$
$f_{RD}^*(i,j)$	Joint probability of splitting duration and event duration; $P\left(R^* = i, D^* = j\right)$

Table 2: Notation

A.1 Linkage Error Model

In this section I dive deeper into the data microfoundations, and describe a record linkage model that would result in the duration model transformation described in section 2.

Consider τ ordered data files, indexed by $t \in \mathcal{T} = \{1, \ldots, \tau\}$, with n_t individuals in each file. Individuals within file t are indexed by $i \in \mathcal{N}_t = \{1, \ldots, n_t\}$, but note that individual i in file t and individual i in file t' need not be the same individual (the indices are only labels within a file). Let $\mathcal{I} = \{1, \ldots, n\}$ be the set of all distinct individuals across all files. To compare individuals across files we define the identity function, $ID_t : \mathcal{N}_t \to \mathcal{I}$,

$$ID_t(i) = \text{Identity of individual } i \text{ in population } t,$$

and note that $ID_t(i) = ID_{t'}(j)$ means that individual i in population t and individual j in population t' are the same individual.

Now consider a K-dimensional vector, $X_{ti} = \left[X_{ti}^{(1)}, X_{ti}^{(2)}, \ldots, X_{ti}^{(K)}\right]$, representing potential matching variables associated with individual i in file t. A *deterministic matching algorithm*, $M_U : \mathcal{N}_t \times \mathcal{N}_s \to \{0, 1\}$, between file t and t' indicates whether a pair of observations across the data files should be linked by comparing all covariates in $U \subseteq \{1, \ldots, K\}$. Mathematically this is

$$M_U(i,j) = \prod_{k \in U} \mathbb{1}\{X_{ti}^{(k)} = X_{sj}^{(k)}\}$$

We can now define certain properties of a given matching algorithm. Let the matching algorithm M_U be a *sufficient matcher* if

$$M_U(i,j) = 1 \quad \Rightarrow \quad ID_t(i) = ID_{t'}(j).$$

In other words matching in the subset of covariates is sufficient to be a true match. Analagously M_U is a *necessary matcher* if

$$ID_t(i) = ID_{t'}(j) \quad \Rightarrow \quad M_U(i,j) = 1,$$

meaning true matches will always have matching covariates in U. Note that if a matching algorithm is necessary and sufficient then the matching algorithm will always match correctly, and there are never any linking errors.

Very specific types of matching errors can occur if a given algorithm lacks one or both of the above characteristics. A matching algorithm that is sufficient but not necessary will occasionally miss matches because they did not match on the covariates but corresponded to the same individual nonetheless. For example if $U = \{\text{First Name}, \text{Last Name}\}$, then in a small community first name and last name may uniquely identify individuals across time, but this matching strategy may miss linking individuals if they change their last name. Similarly an algorithm that is necessary but not sufficient could match records that are not the same individual. For example if $U = \{\text{First Name}\}$, then the same individual will likely have the same first name throughout the data, but if several different individual's share the same first name, matching on U could lead to linking different people. If the algorithm has neither property than both matching errors can occur.

When comparing matching algorithms that use overlapping sets of characteristics we can deduce how the above properties transmit via the following lemma.

Lemma 1. *Consider matching algorithms M_U and M_W. If $U \subseteq W$ then all of the following hold:*

- *M_W is a necessary matcher $\Rightarrow$ M_U is a necessary matcher.*
- *M_W is not a sufficient matcher $\Rightarrow$ M_U is not a sufficient matcher .*
- *M_U is a sufficient matcher $\Rightarrow$ M_W is a sufficient matcher.*
- *M_U is not a necessary matcher $\Rightarrow$ M_W is not a necessary matcher.*

Proof. Regarding the first implication let $U \subseteq W$ and M_W be a necessary matcher. Consider individual i and j and suppose $ID_t(i) = ID_{t'}(j)$. Because it is a necessary matcher this implies

$M_W(i,j) = 1$, and that for all $k \in W$ we have $X_{ti}^{(k)} = X_{sj}^{(k)}$. However since $U \subseteq W$ this implies $M_U(i,j) = 1$ and thus that M_U is a necessary matcher.

Now suppose instead that M_W is not a sufficient matcher. Then there exist individuals i and j such that $M_W(i,j) = 1$ but $ID_t(i) \neq ID_{t'}(j)$. Since $U \subseteq W$ it follows that $M_U(i,j) = 1$, and thus M_U is not a sufficient matcher.

Proving implications 3 and 4 proceeds in an analogous fashion. ■

Now we consider sets of covariates that have a very specific relationship. Let $\{U_1, U_2\}$ be called if

1. $U_1 \cap U_2 = \emptyset$
2. $M_{U_1 \cup U_2}$ is a sufficient but not necessary matcher.
3. M_{U_1} is a necessary but not sufficient matcher.

In words this means that the variables in U_2 help the sufficiency of a matcher, but true links may still not agree on these variables. Furthermore, when these variables are omitted we introduce potential mismatches because U_1 is not enough on its own to identify individuals. Variables in U_2 will come to form the basis of the record linkage error events that are referenced in the main model.

One additional assumption is required to ensure that record linkage errors occur at most once in an individual's history. Let $U \subseteq \{1, \ldots, K\}$ have the *single change* property if for all $i \in \mathcal{I}$ and $i_t \equiv ID_t^{-1}(i)$ there exists $t_0 \geq 1$ such that

$$\begin{aligned} X_{i_t t}^{(k)} &= X_{i_0 0}^{(k)} \qquad \forall t < t_0, k \in U \\ X_{i_t t}^{(k)} &= X_{i_\tau \tau}^{(k)} \qquad \forall t \geq t_0, k \in U \end{aligned}$$

This property says that for any given individual, all of the characteristics in U associated with that individual change at most one time (jointly) across observations of that individual over all data files. If the characteristics never change this would be satisfied by $t_0 = \tau + 1$. If the characteristics do change then t_0 indicates the period when this change occurs and will be a period where a linkage error occurs if matching off covariates in U_1.

These definitions can now be aggregated to form a sufficient set of conditions on a matching algorithm which will exhibit the linkage error characterized by the model in section 2 and thus be germane to the results of this paper.

Theorem 6. *Let U_1 and U_2 be sets of covariates such that $\{U_1, U_2\}$ is and U_2 has the single change property. Then panel data created under the matching algorithm $M_{U_1 \cup U_2}$ will result in durations following the distribution of D (in section 2) with the distribution of R^* coming from changes in the covariates of U_2 over time.*

Once the model of section 2 is relaxed to allow for multiple linkage errors in a single individual's history, the above theorem can be relaxed by omitting the single change covariates property.

A.2 Proofs

A.2.1 Properties of D

Proof of Theorem 1. content... ■

Proof of Theorem 2. content... ■

A.2.2 Estimation under Independence

Proof of Theorem 1 (Identification). Under independence the observed duration distribution, P_d, can be written as a linear function of the true distribution of the duration of interest, P_{d*},

$$\underbrace{\begin{bmatrix} f_d(1) \\ f_d(2) \\ f_d(3) \\ \vdots \\ f_d(H-1) \\ f_d(H) \end{bmatrix}}_{=P_d} = \frac{1}{\chi} \underbrace{\left[\begin{matrix} \sum_{i=1}^{H} f_R^*(i) & f_R^*(1)+f_R^*(1) & f_R^*(1)+f_R^*(2) & \cdots & f_R^*(1)+f_R^*(H-2) & f_R^*(1)+f_R^*(H-1) \\ 0 & \sum_{i=2}^{H} f_R^*(i) & f_R^*(2)+f_R^*(1) & \cdots & f_R^*(2)+f_R^*(H-3) & f_R^*(2)+f_R^*(H-2) \\ 0 & 0 & \sum_{i=3}^{H} f_R^*(i) & \cdots & f_R^*(3)+f_R^*(H-4) & f_R^*(3)+f_R^*(H-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \sum_{i=H-1}^{H} f_R^*(i) & f_R^*(H-1)+f_R^*(1) \\ 0 & 0 & 0 & \cdots & 0 & f_R^*(H) \end{matrix}\right.}_{=A_{r*}}$$

$$P_d = \frac{1}{\chi} A_{r*} P_{d*}$$

where $\chi = 1 + P\left(R^* < D^*\right) = 1 + \sum_{i=1}^{H-1} \sum_{j>i} f_D^*(j) f_R^*(i)$. If the distribution of R^* is known then A_{r*} is known. By definition of the support we have that $f_R^*(H) \neq 0$ which implies A_{r*} is invertible. Further note that since P_{d*} is a discrete probability distribution we have $\|A_{r*}^{-1} P_d\|_1 = \|\frac{1}{\chi} P_{d*}\|_1 = \frac{1}{\chi}$. Finally since P_d is observed then P_{d*} is identified via $P_{d*} = \frac{A_{r*}^{-1} P_d}{\|A_{r*}^{-1} P_d\|_1}$. ■

Proof of Theorem 2 (Consistency). First we focus on the sample mean over $\vec{D}_i$ and note that this sample consists of three different types of observations: unbroken durations, $\vec{D}_i^u$, the first half of a broken duration, $\vec{D}_{i1}^b$, and the second half of a broken duration, $\vec{D}_{i2}^b$. Breaking up this sum we have

$$\frac{1}{n} \sum_{i=1}^{n} \vec{D}_i = \frac{1}{n} \sum \vec{D}_i^u + \frac{1}{n} \sum \left(\vec{D}_{i1}^b + \vec{D}_{i2}^b \right) = \frac{n_u^*}{n} \left(\frac{1}{n_u^*} \sum \vec{D}_i^u \right) + \frac{n_b^*}{n} \left(\frac{1}{n_b^*} \sum \left(\vec{D}_{i1}^b + \vec{D}_{i2}^b \right) \right)$$

where n_u^* and n_b^* represent the number of latent individuals whose durations were unbroken and broken respectively. Since the latent sample is independent across individuals we can now apply the WLLN to each of these sums as $n^* \to \infty$:

$$\frac{1}{n_u^*} \sum \vec{D}_i^u \to_p \mathbb{E}\left[\vec{D}_i | i \text{ was unbroken} \right] \qquad \frac{1}{n_b^*} \sum \left(\vec{D}_{i1}^b + \vec{D}_{i2}^b \right) \to_p \mathbb{E}\left[\vec{D}_{i1} + \vec{D}_{i2} | i \text{ was broken} \right]$$

Since $n = 2n_b^* + n_u^*$ we can also calculate the limit of the weights in front of each sample mean

$$\lim_{n^* \to \infty} \frac{n_u^*}{2n_b^* + n_u^*} = \lim_{n^* \to \infty} \frac{n_u^*/n^*}{2n_b^*/n^* + n_u^*/n^*} = \frac{P\left(\text{unbroken}\right)}{2P\left(\text{broken}\right) + P\left(\text{unbroken}\right)} = \frac{1}{\chi} P\left(D^* \leq R^*\right)$$

and similarly find that $\lim_{n^* \to \infty} \frac{n_b^*}{n} = \frac{1}{\chi} P\left(D^* > R^*\right)$. Bringing it altogether and focusing on the kth index we have

$$\begin{aligned}\left(\frac{1}{n}\sum_{i=1}^{n} \vec{D}_i\right)_k \to_p & \left(\frac{1}{\chi}\left[P\left(D^* \leq R^*\right) \mathbb{E}\left[\vec{D}_i | D_i^* \leq R_i^*\right] + P\left(D^* > R^*\right) \mathbb{E}\left[\vec{D}_{i1} + \vec{D}_{i2} | D_i^* > R_i^*\right]\right]\right)_k \\ =& \frac{1}{\chi}\left[P\left(D^* \leq R^*\right) P\left(D_i = k | D_i^* \leq R_i^*\right) + P\left(D^* > R^*\right)\left[P\left(D_{i1} = k | D_i^* > R_i^*\right) + P\left(D_{i2} = k | D_i^* > R\right.\right.\right. \\ =& \frac{1}{\chi}\left[P\left(D_i^* = k, D_i^* \leq R_i^*\right) + P\left(R_i^* = k, D_i^* > R_i^*\right) + P\left(D_i^* - R_i^* = k, D_i^* > R_i^*\right)\right] \\ =& f_d(k)\end{aligned}$$

Where the final line comes from our derivation of the distribution of D in equation (1). Having shown that $\frac{1}{n}\sum \vec{D}_i \to_p P_d$ it follows that $\|A_{r*}^{-1}\frac{1}{n}\sum_{i=1}^{n} \vec{D}_i\|_1 \to_p \frac{1}{\chi}$. Finally bringing all the pieces together we have

$$\widehat{P}_{d*} = \frac{A_{r*}^{-1}\frac{1}{n}\sum_{i=1}^{n} \vec{D}_i}{\|A_{r*}^{-1}\frac{1}{n}\sum_{i=1}^{n} \vec{D}_i\|_1} \to_p \frac{A_{r*}^{-1} P_d}{\frac{1}{\chi}} = \frac{\frac{1}{\chi} P_{d*}}{\frac{1}{\chi}} = P_{d*}.$$

■

Proof of Theorem 3 (Asymptotic Normality). First note that the dependence between observations can be seen as dependence within clusters where each cluster is made up of all durations from the same individual (as done in the proof for consistency above). Thus we have two types of clusters: one type for individuals who's durations are unbroken and one type for individuals who's durations are broken. There will be n_u^* clusters of the first type, each with exactly one observation (the true duration) and there will be n_b^* clusters of the second type, each with exactly two observations (the first half and the second half). The proof here nearly follows the same procedure as the proof for the central limit theorem in (Theorem 2) of Hansen and Lee (2019), with the adjustment that will not require the covariance matrix to be nonsingular (since it is here). In the notation of Hansen and Lee (2019) we have

$$\begin{aligned} n_1 &= n_u^* & \widetilde{X}_1 &= \vec{D^*}_{i1} \\ n_2 &= n_b^* & \widetilde{X}_2 &= \vec{D^*}_{i1} + \vec{D^*}_{i2}\end{aligned}$$

and thus the finite sample covariance matrix is

$$\Omega_n = \frac{1}{n^*}\sum_{g=1}^{2} Var\left(\widetilde{X}_g\right)$$

Since the covariance matrices of the clusters here have a well defined limit, I do not need the level of generality that necessitates sample specific cluster covariance matrices. ■