

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

Elan Segarra\*

December 1st, 2020

[Link to most recent version](#)

## Abstract

Longitudinal panels have a long history of use across the social sciences; however, they can be imperfect representations of reality when record linkage methods are employed during their creation. In this paper I study survival analysis (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 is not accommodated by a standard competing risks model. Hence new methods are developed which restore consistency of the parameters of interest without correcting the linkages. This work makes three distinct theoretical contributions under increasingly relaxed assumptions. First, under the strong assumption of a known independent linkage error process I show that the marginal distribution of time to death is nonparametrically identified from linkage error induced durations. Second, when observations on start and end dates are introduced, I show that nonparametric point identification of the joint distribution of lifetimes and linkage error is typically achieved. Third, when no restriction is placed on the dependence structure, I apply partial identification methods to derive sharp informative bounds on the marginal distribution of lifetimes. New estimators and inference methods are introduced across all scenarios and their validity is established formally. The methods are applied to longitudinal business data (where linkage error occurs due to establishment relocation), and show that establishment death rates in the first 3 years can be overestimated by as much as 10 percentage points with naive methods, while those proposed here are able to recover true rates of survival from mis-linked data.

**JEL Codes:** C1, C23, C41, M13, R3

**Keywords:** record linkage error, survival analysis, duration analysis, competing risks, panel data quality, firm dynamics

---

\*University of Wisconsin–Madison, Department of Economics. Email: segarra@wisc.edu

<sup>0</sup>I am grateful to my advisors, Jack Porter, Chris Taber, and Felix Elwert. Additionally I must thank Jeff Smith, Bruce Hansen, Xiaoxia Shi, and Mikkel Solvsten for their feedback during the lifetime of this project.

# 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 affect 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. In particular, since linking algorithms often leverage addresses, firm relocation can induce linkage error which will be especially prominent among young firms that are experiencing rapid growth. Correcting the linkages can be extremely costly, and sometimes impossible given the observable data, so empiricists need to account for the linkage error in downstream estimation. 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 reliably estimated 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 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 from 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). This project represents 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 when using panel data subject to linkage error. Section 3 presents theoretical results in scenarios where the researcher observes only durations, while section 4 shows cases theoretical results when the start times of the durations are additionally observed. Section 5 describes the empirical application of the methods to the estimation of firm lifetimes, and section

6 concludes.

## 2 Model

In this section I describe the model which transforms the latent variables, whose distribution is 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.

### 2.1 Notation

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, i.e.  $\vec{X}$  is a binary vector with a 1 in the element 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  $f_x = \mathbb{E}[\vec{X}]$ .

### 2.2 Researcher Objective

Let  $(S_i^*, D_i^*) \in \mathfrak{S}^* \times \mathfrak{D}^*$  be a tuple of discrete unobserved random variables representing individual  $i$ 's starting period and the duration to an event of interest with respective supports  $\mathfrak{S}^*, \mathfrak{D}^* \in \mathbb{N}^+$ . 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 number of truly distinct individuals in a sample is denoted by  $n^*$ .

I take the primary goal of the researcher to be the simple task of learning information about  $D_i^*$ , such as its marginal distribution, denoted  $f_D^*$ . Often it is the case that a researcher is focused on a moment of this distribution, such as the mean or a specific survival probability. However, since the marginal distribution is sufficient for these other parameters, I focus on identifying and estimating  $f_D^*$  in this paper. Similarly, in a world with treatments or controls, if an empiricist is interested in a treatment effect it suffices to learn the joint distribution of these variables.

Were the researcher able to observe  $D_i^*$  then identifying and estimating  $f_D^*$  would indeed be a trivial task. However, since observed durations often come from panel data which has been created after linking multiple periods together, the job may not be straightforward. If there is any error in the record linkage process, then researchers will instead observe  $D_i$  which may be very different from the true underlying duration of that individual. The following section describes the process that transforms the latent variables into the observed variables.

### 2.3 Model of Latent and Observed Data

In a survival analysis context the main consequence of 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  $R_i^* \in \mathfrak{R}^*$  be a discrete random variable representing 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  $b_i^* b^*$  be an indicator of whether a record linkage error breaks the true duration for individual  $i$ , where  $b_i^* = \mathbb{1}\{R_i^* < D_i^*\}$ . When this happens the true duration gets split into two smaller durations:

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

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

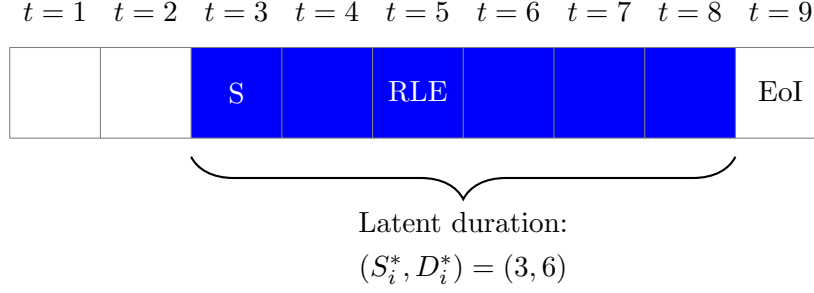
$$(S_{i1}^*, D_{i1}^*) = (S_i^*, D_i^*).$$

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  $n_b^* = \sum b_i^*$  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  $b_i^* = 0$

(a) Latent Durations



(b) Observed Durations

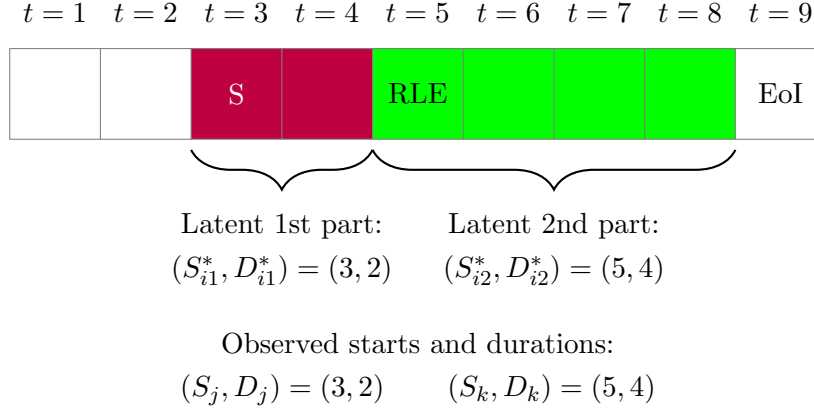


Figure 1: The above figures illustrate the latent durations (1a) and observed durations (1b) for an individual with  $(S^*, D^*, R^*) = (3, 6, 2)$ . The various events of start (S), the event of interest (EoI), and a record linkage error event (RLEE) are indicated in the appropriate time periods.

$i$	$(S_i^*, D_i^*, R_i^*)$	$b_i^*$	$(S_{i1}^*, D_{i1}^*)$	$(S_{i2}^*, D_{i2}^*)$
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.

for all  $i = 1, \dots, n_u^*$  and  $b_i^* = 1$  for all  $i = n_u^* + 1, \dots, n^*$ . Define the random permutation operator

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

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 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,

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

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).  $D_i^*$  has finite support with maximum given by  $H_D^* = \max \mathcal{D}^*$ .

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 be 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.

A third assumption is implicit in the above model, however making it explicit is useful both because of it's importance for identification, and because extensions discussed in later sections will allow for it's relaxation.

**Assumption A3** (No Missing Constituents). *If  $b_i^* = 1$  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, and vice versa.

## 2.4 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. In practice finding unique firm identifiers to link perfectly across years is nontrivial. Even objects such as 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 as done in [Akee and Jones \(2019\)](#). 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.

## 2.5 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 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  $f_{RD}^*(i, j) \equiv P(R^* = i, D^* = j)$ , and denote the distribution function of  $D_i$  by  $f_D(k) \equiv P(D = k)$ . The distribution of observed durations is then given by

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

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

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

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  $(n - n^*)/n^* \rightarrow_p P(R^* < D^*)$ ).

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  $\max \text{support}(D_i^*) \leq 3$  then  $P(D_i = 1) \leq P(D_i^* = 1)$  with strict inequality when  $P(R^* < D^*) > 0$ . If  $\max \text{support}(D_i^*) > 3$  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 from right to left, 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

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 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.

**Assumption A4** (Independence).  $R^* \perp D^*$

Let the marginal distributions of the durations of interest and the duration until RLEE be denoted by the vectors  $f_R^* \equiv [f_R^*(1) \ \cdots \ f_R^*(H_D)]'$  and  $f_D^* \equiv [f_D^*(1) \ \cdots \ f_D^*(H_D)]'$  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 A5** (Support Condition).  $\max \text{support}(D_i^*) \leq \max \text{support}(R_i^*)$ .

One implication of this assumption is that all durations have a chance of being broken by record linkage error. Under our previous assumptions assumption A5 becomes a necessary and sufficient condition for point identification of the distribution of  $D_i^*$ .

**Theorem 1** (Identification). *Let the researcher observe broken durations,  $\{D_i\}_{i=1}^n$ , know  $f_R^*$ , and suppose that assumptions A1-A4 hold. Then*

$$f_D^* \text{ is point identified} \quad \Leftrightarrow \quad \text{A5 holds.}$$

Let the researcher observe broken starts and durations,  $\{S_i, D_i\}_{i=1}^n$ , and suppose that assumptions

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,  $D_i = H_D$ , will be proportional to the probability of the longest latent duration,  $D_i^* = H_D$ , and thus that probability is identified (up to scale). Similarly the second longest observed duration,  $D_i = H_D - 1$  is related only to the longest latent duration,  $D_i^* = H_D - 1$ , the second longest  $D_i^* = H_D - 2$ , and the distribution of  $R_i^*$ . Since we know about all of these objects except the likelihood of  $D_i^* = H_D - 2$ , that identifies the likelihood of  $D_i^* = H_D - 2$ . 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  $D_i = H_D$  depends on both the probability of  $D_i^* = H_D$  and  $D_i^* = H_D + 1$ . 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  $H_D \times 1$  vector of dummy variables representing the outcome of observed individual  $i$ ,

$$\vec{D}_i = \left[ \mathbb{1}\{D_i = 1\} \quad \mathbb{1}\{D_i = 2\} \quad \cdots \quad \mathbb{1}\{D_i = H_D\} \right]'$$

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

$$\widehat{f}_D^* \equiv \frac{A_{r^*}^{-1} \frac{1}{n} \sum_{i=1}^n \vec{D}_i}{\|A_{r^*}^{-1} \frac{1}{n} \sum_{i=1}^n \vec{D}_i\|_1}, \quad (3)$$

where  $A_{r^*}$  is an  $H_D \times H_D$  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-A5 if  $f_R^*$  is known then  $\widehat{f}_D^* \rightarrow_p f_D^*$ .*

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-A5*

$$\sqrt{n} \left( \widehat{f}_D^* - f_D^* \right) \rightarrow_d N(0, V)$$

where

$$V = \lambda^2 (A_{r^*}^{-1})' \Omega A_{r^*}^{-1}, \quad \Omega = \frac{P(R^* < D^*)}{\lambda} \Omega_b + \frac{P(R^* \geq D^*)}{\lambda} \Omega_u, \quad \lambda = 1 + P(R^* < D^*),$$

and  $\Omega_b$  and  $\Omega_u$  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,  $\Omega$ , where the two parts,  $\Omega_u$  and  $\Omega_b$ , correspond with clusters of unbroken durations and clusters of broken durations respectively. More specifically

$$\Omega_u = \text{Var} \left( \overrightarrow{D}_i^* | D_i^* \leq R_i^* \right) \quad \Omega_b = \text{Var} \left( \overrightarrow{R}_i^* + \overrightarrow{D}_i^* - \overrightarrow{R}_i^* | D_i^* > R_i^* \right).$$

Recall that  $\vec{\cdot}$  transforms the discrete variable into the indicator vector, which is why  $R_i^*$  does not cancel in the second expression.

To perform proper inference we need to estimate the asymptotic covariance matrix,  $V$ . 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 is entirely

determined by  $f_D^*$  and  $f_R^*$  there is an alternative route to it's estimation. Consider the plugin estimator of  $V$ ,

$$\widehat{V} = \widehat{\lambda}^2 (A_{r^*}^{-1})' \widehat{\Omega} A_{r^*}^{-1}, \quad (4)$$

with constituent plugin estimators,

$$\widehat{\Omega} = \frac{\widehat{\lambda} - 1}{\widehat{\lambda}} \widehat{\Omega}_b + \frac{2 - \widehat{\lambda}}{\widehat{\lambda}} \widehat{\Omega}_u \quad (5)$$

$$\widehat{\lambda} = 1 + \sum_{r < d} \widehat{f}_D^*(d) f_R^*(r) \quad (6)$$

$$\widehat{\Omega}_u = \text{diag} \left( \widehat{f}_{1|u} \right) - \widehat{f}_{1|u} \widehat{f}_{1|u}' \quad (7)$$

$$\widehat{\Omega}_b = \text{diag} \left( \widehat{f}'_{2|b} \mathbf{1} + \widehat{f}_{2|b} \mathbf{1} \right) - \left( \widehat{f}'_{2|b} \mathbf{1} + \widehat{f}_{2|b} \mathbf{1} \right) \left( \widehat{f}'_{2|b} \mathbf{1} + \widehat{f}_{2|b} \mathbf{1} \right)' + \widehat{f}_{2|b} + \widehat{f}'_{2|b} \quad (8)$$

$$\left( \widehat{f}_{1|u} \right)_d = \frac{1}{2 - \widehat{\lambda}} \sum_{d \leq r} \widehat{f}_D^*(d) f_R^*(r) \quad (9)$$

$$\left( \widehat{f}_{2|b} \right)_{rw} = \frac{1}{\widehat{\lambda} - 1} \widehat{f}_D^*(r + w) f_R^*(r) \quad (10)$$

For clarification  $f_{1|u}$  is a  $X \times 1$  column vector of the probability mass function of  $\overrightarrow{D^*}$  conditional on  $D^* \leq R^*$ , i.e. the distribution of observed durations for those records that are unbroken (unaltered by linkage error). Similarly  $f_{2|b}$  is an  $X \times X$  matrix representing the joint probability mass function of  $\overrightarrow{R^*}$  and  $\overrightarrow{D^* - R^*}$  conditional on  $D^* > R^*$ , i.e. the distribution of observed durations for those records that are broken (altered by linkage error).

**Theorem 4** (Consistency of  $\widehat{V}$ ). *If  $\widehat{f}_D^* \rightarrow_p f_D^*$  and  $f_R^*$  is known then  $\widehat{V} \rightarrow_p V$ .*

Consistency of  $\widehat{V}$  follows from the consistency of the estimator for  $f_D^*$ , as made concrete in Theorem 4. While the formal proof can be found in Appendix A.2.2, the procedure is a fairly straightforward repeated application of Slutsky's theorem on each of the constituent estimators found in equations (4)-(10).

### 3.2 Dependent Durations

As mentioned earlier the assumption of independence is strong and likely untenable in many situations. In this section we relax this assumption, and allow the dependence 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 pertain to an environment where the researcher only observes durations, however it is very common to also 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 start times and durations,  $\{(S_i, D_i)\}_{i=1}^n$ , as opposed to just durations,  $\{(D_i)\}_{i=1}^n$ .

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. Note that this is the data observed after linking across time (possibly with error) so that individual B's duration of length 3 could be the true time to the event of interest, the first half of a broken duration (with linkage error occurring between period 1 and 2), or the second half of a broken duration (with linkage error occurring between period 4 and 5).

Despite the persistence of this ambiguity, observing start times means there is extra information in the observed adjacencies between individuals and at the tails of the data set. For example individual B could be an unbroken duration or individuals B and E could be the first and second half of the same individual because of their adjacency. Similarly individual A and C or A and D could represent pairs of broken duration. Moreover if this figure represents the entire data set

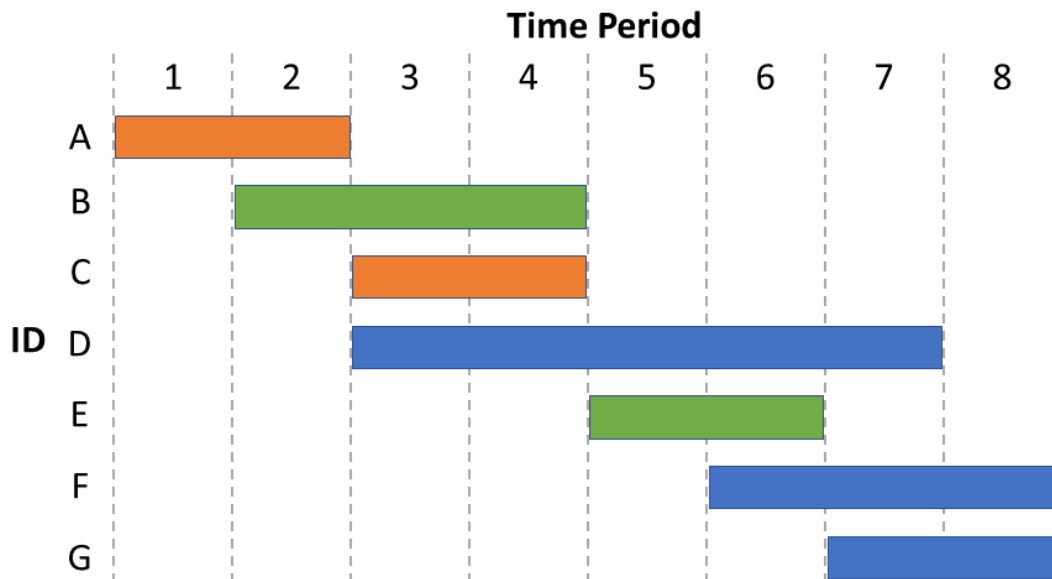


Figure 2: Visual example of panel data (with linkage error) illustrating potential relationships among observed durations.

Though we are able to relax any assumptions surrounding the dependence between  $D_i^*$  and  $R_i^*$ , additional assumptions are still required to achieve point identification. Consider the following two assumptions.

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

**Assumption A7** (Terminality of Event).  $P(R_i^* > D_i^*) = 0$ .

While the content of assumption A6 is straightforward it's plausibility will entirely depend on

the empirical context and what  $S_i^*$ ,  $D_i^*$ , and  $R_i^*$  represent. In example 2 where units are people and  $S_i^*$  is the birth year it could be reasonable to assume that a person's birth year is independent of both time until marriage and time until death/emmigration provided the horizon is relatively short (e.g. only those born between 1960-1970). However, if the sample's horizon contains both people born in the 1920s and 1990s then this assumption is more suspect given changes in marriage norms over the decades.

Assumption A7 is useful to rectify the fact that only certain joint probabilities can be identified. In a simple model whose latent variables have support up to 4, figure 3a illustrates the finest groups of joint probabilities that are separately identifiable as indicated by the green outlines. For example while  $p_{12} = P(R_i^* = 1, D_i^* = 2)$  is identified, we will never be able to separately identify  $p_{22}$  from  $p_{32}$  or from  $p_{42}$  because they are all observationally equivalent, i.e. all three events result in observing a single duration  $D_i = 2$ . Assumption A7 can be viewed either as a strict assumption or as a normalization depending on the empiricist's preferences. If the event of interest is truly terminal, in the sense that the event that drives  $R^*$  cannot occur afterward, then A7 can be viewed as a strict assumption (represented in figure 3c)<sup>1</sup>. This is likely appropriate in example 1 where establishments cannot relocate ( $R^*$ ) after they have died ( $D^*$ ). Alternatively we can normalize the tail probabilities by redefining  $\tilde{p}_{ii} \equiv P(R^* \geq i, D^* = i)$  since that is identifiable (as illustrated in figure 3b). This is more applicable in example 2 where it is perfectly reasonable for a person to get married ( $R^*$ ) after emmigrating ( $D^*$ ).

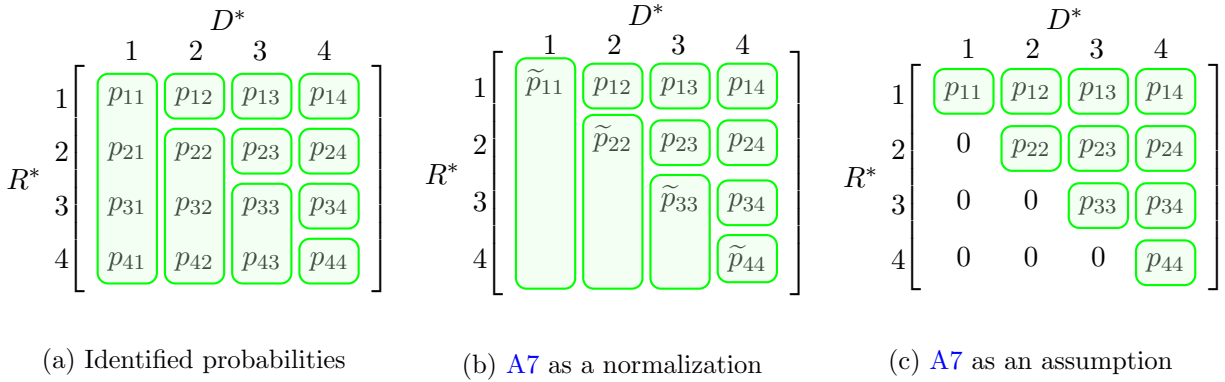


Figure 3: Illustrates which probabilities are separately identifiable (a), and how assumption ?? can be interpreted as either a normalization (b) or as a strict assumption (3c). Circled green cells indicate which probabilities are separately identifiable, where  $p_{ij} = P(R^* = i, D^* = j)$ , and  $\tilde{p}_{ii} \equiv P(R^* \geq i, D^* = i)$ .

Together these assumptions yield the identification results found in Theorem 5. The first result shows that identification of the marginal distribution of  $D^*$  is achieved when start times are independent of the other events. The second result, which imposes the terminality condition (A7), illustrates that the entire joint distribution can also be identified.

<sup>1</sup>Note that assumption A7 does not require that every duration be broken by a record linkage error event because it still allows for positive probability on  $R_i^* = D_i^*$  which does not result in breakage.

**Theorem 5** (Identification). *Let the researcher observe broken starts and durations,  $\{S_i, D_i\}_{i=1}^n$ , and suppose that assumptions A1, A2, A3, A5, and A6 hold. Then*

1. *the marginal distributions,  $f_S^*$  and  $f_D^*$ , are point identified.*
2. *if A7 holds the entire joint distribution,  $f_{SRD}^*$ , is point identified.*

The full proof of this result is found in Appendix A.2.3, however the intuition borrows some of the same identification strategies leveraged in the previous section. In particular identification once again proceeds from the edges of the observed distribution, this time from both the  $D$  dimension and  $S$  dimension.

Since observable distributions identify latent distributions and all distributions are over finite discrete events (i.e. they are represented by a finite number of parameters) we can use the Generalized Method of Moments (GMM) for estimation. Recall equation that related the the latent distributions,  $f_{SRD}^*$ , to the observed distribution of starts and durations,  $f_{SD}$ . Under the assumption of independent starting times equation becomes

$$f_{SD} = g(f_{RD}^*, f_S^*) \equiv \frac{1}{\lambda} \left( \begin{bmatrix} f_S^* \\ 0 \end{bmatrix} (\mathbf{1}' [f_{RD}^*]_L + \mathbf{1}' [f_{RD}^*]'_U) + \sum_k f_S^*(k) L_k h([f_{RD}^*]_U) \right)$$

where  $\lambda = 1 + \mathbf{1}' [f_{RD}^*]_U \mathbf{1}$ .

Let  $\overrightarrow{SD}_i$  denote the  $(H_S^* + H_D^* - 1) \times H_D^*$  binary matrix whose cells are defined by

$$\left( \overrightarrow{SD}_i^* \right)_{s,d} = \mathbb{1}\{S_i = s \text{ and } D_i = d\},$$

so we can define the following moment condition function:

$$m \left( \overrightarrow{SD}_i^*; f_1, f_2 \right) \equiv \text{vec} \left( g(f_1, f_2) - \overrightarrow{SD}_i^* \right). \quad (11)$$

Since  $g(\cdot)$  maps the latent distributions to the observed distribution we have that

$$\mathbb{E} \left[ m \left( \overrightarrow{SD}_i^*; f_{RD}^*, f_S^* \right) \right] = 0$$

resulting in  $l$  moment conditions with

$$l \equiv \dim \left( m \left( \overrightarrow{SD}_i^*; f_{RD}^*, f_S^* \right) \right) = H_S^* H_D^* + \frac{1}{2} H_D^* (H_D^* - 1).$$

Given a positive definite  $l \times l$  weight matrix,  $W$ , the sample moment functions, GMM objective function, and corresponding estimator are then given by

$$\overline{m}(f_1, f_2) = \frac{1}{n} \sum_{i=1}^n m \left( \overrightarrow{SD}_i^*; f_1, f_2 \right) \quad (12)$$

$$J_n(f_1, f_2) \equiv \overline{m}(f_1, f_2)' W \overline{m}(f_1, f_2) \quad (13)$$

$$\left( \widehat{f}_{RD}^*, \widehat{f}_S^* \right) = \arg \min_{\substack{f_1 \in \Delta(\mathcal{R} \times \mathcal{D}) \\ f_2 \in \Delta(\mathcal{S})}} J_n(f_1, f_2). \quad (14)$$



To establish consistency of this estimator we require one more regularity condition. Assumption A8 is fairly innocuous given the functional form of the moment condition function (11). This assumption is necessary to guarantee that the moment condition function is uniformly continuous over the parameter space. The statements of further assumptions and theorems are made clearer if we cast the latent distributions,  $f_{RD}^*$  and  $f_S^*$ , and the spaces where they live into a single vector of parameters:

$$\Theta \equiv \left\{ \begin{bmatrix} \text{vec}(f_1) \\ f_2 \end{bmatrix} : f_1 \in \Delta(\mathcal{R} \times \mathcal{D}), \quad f_2 \in \Delta(\mathcal{S}) \right\}$$

$$\theta^* \equiv \begin{bmatrix} \text{vec}(f_{RD}^*) \\ f_S^* \end{bmatrix} \quad \hat{\theta} \equiv \begin{bmatrix} \text{vec}(\widehat{f_{RD}^*}) \\ \widehat{f_S^*} \end{bmatrix}$$

**Assumption A8** (Regularity Conditions 1). *There exists a function  $h : \mathbb{R} \mapsto \mathbb{R}$  with  $\lim_{u \rightarrow 0^+} h(u) = 0$  such that for every pair of latent distributions  $\theta_1, \theta_2 \in \Theta$  we have*

$$\|g(\theta_1) - g(\theta_2)\| \leq h(\|\theta_1 - \theta_2\|).$$

With this extra assumption in hand we can establish the consistency of  $\widehat{f_{RD}^*}$  and  $\widehat{f_S^*}$  for  $f_{RD}^*$  and  $f_S^*$  respectively in Theorem 6. The full proof of this result (details found in Appendix A.2.3) leverages the consistency results of the clustered GMM estimators described in Hansen and Lee (2019).

**Theorem 6** (Consistency). *Let the researcher observe broken starts and durations,  $\{S_i, D_i\}_{i=1}^n$ , and suppose that the assumptions of theorem 5 hold along with A8. Then  $\hat{\theta} \rightarrow_p \theta^*$ .*

As in our previous section we now establish asymptotic normality to provide empiricists with proper inference. Despite the utilization of a fairly well defined method like GMM, recall that the observed sample is not *iid* since it includes correlation within unobserved clusters which makes asymptotics nonstandard. Nonetheless another set of regularity conditions yields asymptotic normality of the estimators.

**Assumption A9** (Regularity Conditions 2). *Let  $f_{RD}^*$ ,  $f_S^*$ , and  $m(\overrightarrow{SD}_i; f_1, f_2)$  satisfy:*

1.  $\theta^* \in \text{interior}(\Theta)$
2. *There exists a function  $h : \mathbb{R} \mapsto \mathbb{R}$  with  $\lim_{u \rightarrow 0^+} h(u) = 0$  such that for every pair of latent distributions,  $\theta_1, \theta_2$ , in a neighborhood of  $\theta^*$  we have*

$$\left\| \frac{\partial}{\partial \theta} g(\theta_1) - \frac{\partial}{\partial \theta} g(\theta_2) \right\| \leq h(\|\theta_1 - \theta_2\|).$$

The regularity conditions described in A9 are fairly standard in the GMM literature. In particular requiring the true parameter value to lie in the interior of the parameter space (A9.1) is typical since it allows for the optimum to be determined by a well defined first order condition (see Andrews (2002) for results that relax this assumption). In our context A9.1 requires that the

latent distributions of  $D^*$ ,  $R^*$ , and  $S^*$  have full support. If the empiricist expects certain outcomes to have zero probability, then the parameter space can be redefined to omit these events from the support so that this assumption is still satisfied (recall that all random variables are discrete).

Assumption A9.2 is also standard, and functions in the same manner as A8 by ensuring that the Jacobian of the moment function is uniformly continuous over  $\Theta$ . () With these regularity assumptions along with previous assumptions we have that the asymptotic distribution of  $\widehat{f}_{RD}^*$  and  $\widehat{f}_S^*$  is normally distributed as described in Theorem 7.

**Theorem 7** (Asymptotic Normality). *Let the researcher observe broken starts and durations,  $\{S_i, D_i\}_{i=1}^n$ , let  $W$  be a positive definite matrix, and suppose that the assumptions of Theorem 6 and A9 hold, then*

$$\sqrt{n}(\widehat{\theta} - \theta^*) \rightarrow_d N(0, V)$$

where

$$V = (Q'W^{-1}Q)^{-1}Q'W^{-1}\Omega W^{-1}Q(Q'W^{-1}Q)^{-1}$$

$$Q \equiv \mathbb{E} \left[ \frac{\partial}{\partial \theta} g(\theta^*) \right] \quad \Omega \equiv \frac{P(R^* < D^*)}{\lambda} \Omega_b + \frac{P(R^* \geq D^*)}{\lambda} \Omega_u \quad \lambda = 1 + P(R_i^* < D_i^*).$$

The variance of the asymptotic distribution,  $V$ , takes on the familiar sandwich form that is typical in GMM estimators. Moreover we can see evidence of the different types of clusters in  $\Omega_b$  and  $\Omega_u$  which correspond with the broken and unbroken clusters respectively. In particular these variables are defined to be

$$\Omega_b \equiv \mathbb{E} \left[ \left( m(\overrightarrow{SD}_{1i}; \theta^*) + m(\overrightarrow{SD}_{2i}; \theta^*) \right) \left( m(\overrightarrow{SD}_{1i}; \theta^*) + m(\overrightarrow{SD}_{2i}; \theta^*) \right)' \mid R_i^* < D_i^* \right]$$

$$\Omega_u \equiv \mathbb{E} \left[ m(\overrightarrow{SD}_i; \theta^*) m(\overrightarrow{SD}_i; \theta^*)' \mid R_i^* \geq D_i^* \right].$$

Notice that the first expression, which corresponds to broken individuals, sums the moment condition twice, once for terms related to the first half of a broken lifetime and another for terms related to the second half of a broken lifetime.

## 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  $f_{RD}^*$ ,

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

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

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

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

$$\mathcal{H}(f_{RD}^*) = \left\{ f_{RD}^* \in \Delta^{H_D^2-1} : \mathbb{E} \left[ \vec{D}_i + \left( \vec{D}_i b'_{H_D} - A_{H_D} \right) \text{vec}(f_{RD}^*) \right] = 0 \right\}. \quad (15)$$

In the above definition  $A_{H_D}$  is an  $H_D \times H_D$  matrix,  $b_{H_D}$  is an  $H_D \times 1$  vector, and both are constant, known, and only depend on  $H_D$ . 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 singleton because there are  $H_D$  moment equations but  $H_D^2$  unknown parameters.

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

$$\mathcal{H}_1(f_{RD}^*) = \left\{ f_{RD}^* \in \Delta^{H_D^2-1} : \mathbb{E} \left[ \vec{D}_i + \left( \vec{D}_i b'_{H_D} - A_{H_D} \right) \text{vec}(f_{RD}^*) \right] = 0 \text{ and } f_R^* - M_{H_D} \text{vec}(f_{RD}^*) = 0 \right\}. \quad (16)$$

The matrix  $M_{H_D}$  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  $H_D$ ).

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

$$Q(f_{md*}) = \left\| \mathbb{E} \left[ \vec{D}_i + \left( \vec{D}_i b'_{H_D} - A_{H_D} \right) \text{vec}(f_{RD}^*) \right] \right\|^2 \quad (17)$$

$$Q_n(f_{md*}) = \left\| \frac{1}{n} \sum_{i=1}^n \left[ \vec{D}_i + \left( \vec{D}_i b'_{H_D} - A_{H_D} \right) \text{vec}(f_{md*}) \right] \right\|^2. \quad (18)$$

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 (17) correspond exactly with the identified set,  $\mathcal{H}(f_{RD}^*)$ , which is what inspires the set estimator

$$\hat{\mathcal{H}}(f_{RD}^*, c) = \left\{ f_{RD}^* \in \Delta^{H_D^2-1} : Q_n(f_{RD}^*) \leq \frac{1}{n} c \right\}, \quad (19)$$

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 8** (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  $d_{haus}(\widehat{\mathcal{H}}(f_{RD}^*), \mathcal{H}(f_{RD}^*)) = o_p(1)$  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 (19) 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  $P(\mathcal{H}(f_{RD}^*) \subseteq \widehat{\mathcal{H}}(f_{RD}^*))$  goes to  $\alpha$  as  $n$  goes to infinity.

*Remark 4.* One idiosyncrasy 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 Empirical Application

To illustrate these methods in action this section describes an application of the proposed methods to the estimation of firm dynamics as described in example 1. Understanding the life cycle of firms and establishments is a core question in macroeconomics (). In labor economics it has been repeatedly illustrated that young establishments are significant drivers of job creation (Haltiwanger (2012), Haltiwanger et al. (2013)), however we have also seen a decrease in firm dynamism over the past two decades (Decker et al. (2016), Akcigit and Ates (2019)). In order to further study these phenomena it is of the utmost importance that we are properly measuring the volatility and death rates of establishments With this in mind, the goal of this application is to estimate the distribution of establishment lifetimes as well as the death rates of young establishments (i.e. the probability of death in the first few years).

As alluded to earlier, the creation of panel data on establishments in the United States typically comes from linking administrative data sets over time. Though there appear to be unique identifiers for this process, such as the employer identification number (EIN), these IDs are not entirely

trustworthy. For example, in many states some businesses are incentivized to change or reset their EIN because this will reset their employee insurance contribution rates . Due to the lack of unique identifiers, to create these panels the linkage algorithms must lean heavily on name and address matching. Therefore when establishment relocate, this can lead to the type of linkage error discussed throughout this paper. Therefore in this application  $D_i^*$  is the time until an establishment dies,  $S_i^*$  is the time period an establishment starts, and  $R_i^*$  is the time until an establishment relocates, and the object of interest is  $f_D^*$ .

The data used for this application is the Your-economy Time Series (YTS) provided by the Wisconsin Business Dynamics Research Consortium (WBDRC). This is a panel of all U.S. establishments including non-profit, for-profit, and public entities from 1997 to 2019. Since YTS is a cleaned derivative of the InfoGroup establishment data I will take YTS to be the truth, i.e. I will approach the analysis assuming that there is minimal record linkage error in the construction of the YTS. Alternatively the methods could be applied to the raw YTS however the ground truth would be unknown and it would be difficult to assess how the new estimates from the proposed methods compare with naive estimators. When YTS is used as the ground truth and estimators are run on broken versions of the YTS, eg. when records are broken at relocations simulating record linkage error, then comparisons can be drawn. This approach is sometimes referred to as an Empirical Monte Carlo Simulation but it also has overlap with synthetic data sets .

One shortcoming of the YTS data is that the finest available annual geodata is the zip code of the establishment. This is in comparison to either InfoGroup or the Longitudinal Business Database in the Wisconsin RDC which both contain either annual addresses of establishments or even annual latitude and longitude.

Some sample statistics of the raw YTS data truth are presented in the following tables.

		Covariances		
		$D^*$	$R^*$	$S^*$
$E[D^*]$	5.053			
$P(D^* \leq 3)$	0.416	$D^*$	10.187	8.695
$P(R^* < D^*)$	0.164	$R^*$	8.695	9.751
		$S^*$	-0.037	0.037
			2.301	

Table 2: Descriptive statistics of the unbroken YTS.

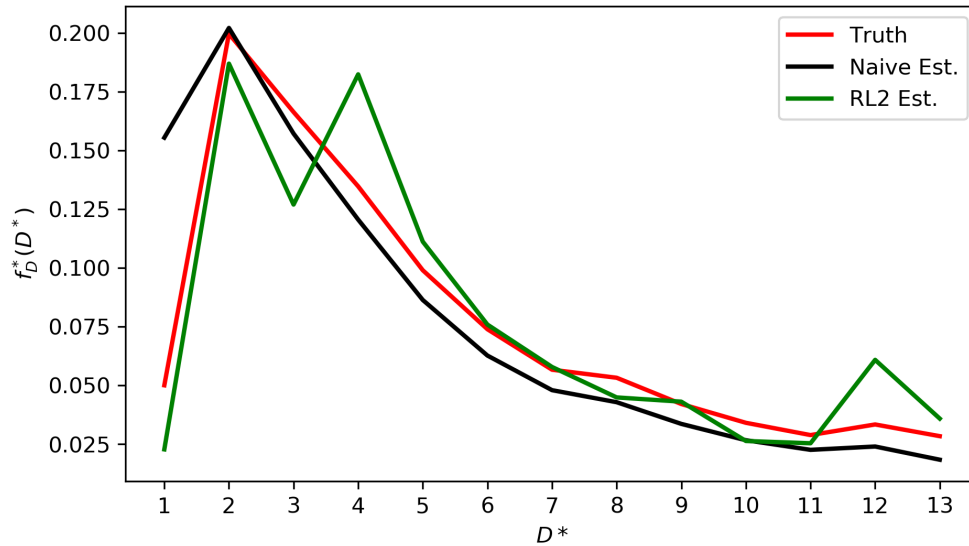
There are two transformations of the raw YTS that will be considered for this exercise. The first takes establishments and breaks their histories at the identified relocation time periods. This process mimics the panel data that would be obtained had the data been subject to a deterministic linking algorithm, as described in A.1, which included address/location as a matching variable. The second data first permutes all of the starting times of the establishments in the YTS data and then breaks them in the same manner as the first data set. This permutation of true latent starting times is meant to create an analogous data set where assumption A6 is known to hold exactly.

In each of these datasets we apply three different estimators for the distribution of establishment lifetimes. The first is the naive estimator,  $\hat{f}_{Naive}$ , which simply takes the observed broken durations

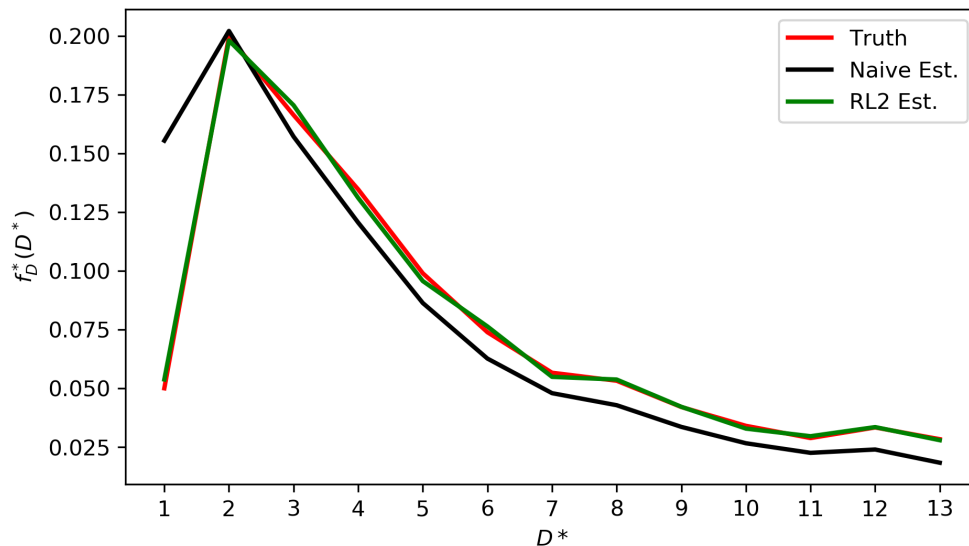
at face value. The second estimator,  $\hat{f}_{RL1}$ , defined in (3) which only uses the observed durations, marginal distribution,  $f_R^*$ , and assumes independence of the relocation timing. The third estimator,  $\hat{f}_{RL2}$ , as defined in (14) which uses observed starts and durations, and assumes independence of the latent start times.

The two following figures plot the three estimates of  $f_D^*$  across both of the broken data sets.

**Estimation of  $f_D^*$  on Broken Data**



**Estimation of  $f_D^*$  on Broken and Permuted Data**



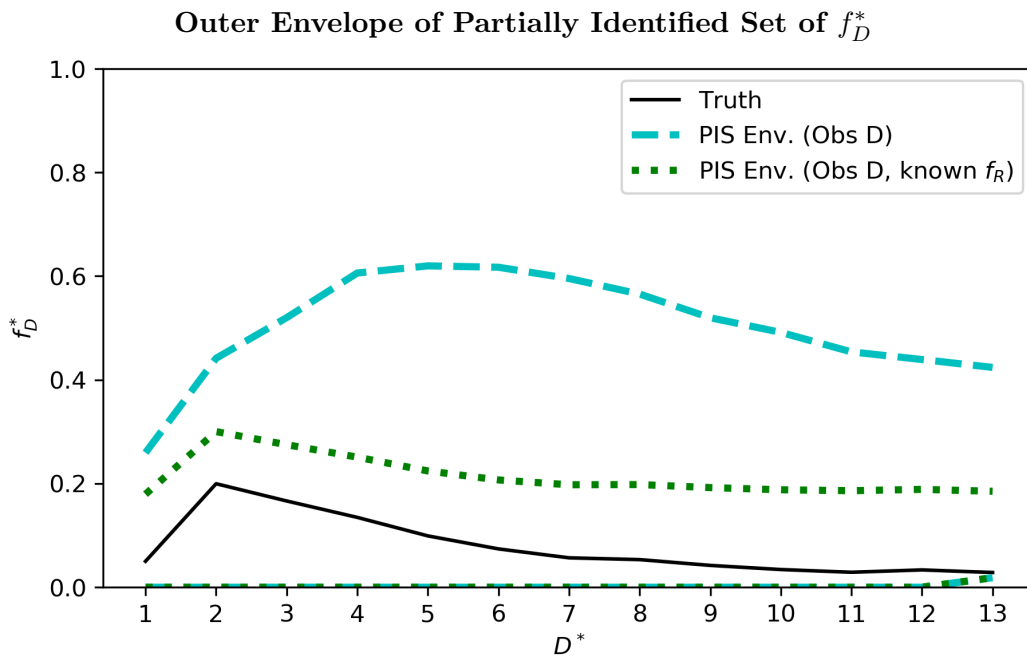
The next table present both estimates of the probability of establishment death in the first 3 years and univariate statistics on the overall error of the estimates of the marginal distributions.

Finally we present the estimates of the partially identified set of both the marginal distribution

	$P(D^* \leq 3)$		$ISE$	
	Broken	Broken/Perm.	Broken	Broken/Perm.
Naive Est.	0.515	0.515	0.01222	0.01222
RL1 Est.	0.460	0.460	0.01987	0.01987
RL2 Est.	0.336	0.422	0.00591	0.00004

Table 3: Estimates of Young Establishment Death Rates and Fit of  $\widehat{f}_D^*$ .

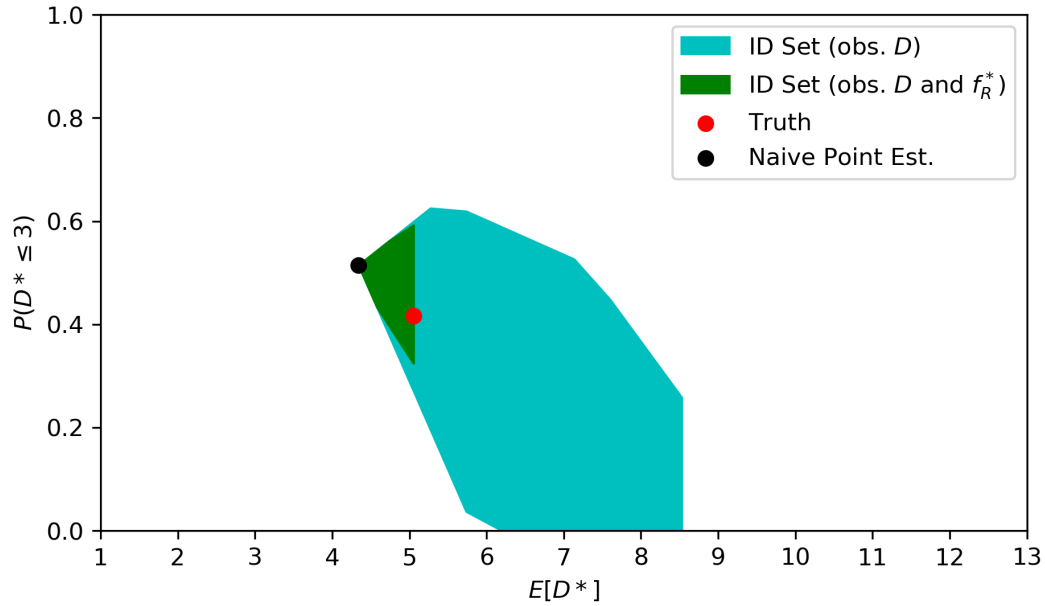
and a few univariate statistics of interest.



## 6 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

Partially Identified Set of  $E[D^*]$  and  $P(D^* \leq 3)$



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.
- Akee, R. and Jones, M. (2019). Immigrants earnings growth and return migration from the u.s.: Examining their determinants using linked survey and administrative data.
- 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.
- Haltiwanger, J. (2012). Job creation and firm dynamics in the united states. *Innovation Policy and the Economy*, 12(1):17–38.
- Haltiwanger, J. C., Jarmin, R. S., and Miranda, J. (2013). Who creates jobs? small vs. large vs. young. *The Review of Economics and Statistics*, 95(2):347–361.
- 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.
- 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; $n_u^* \equiv \sum r_i^*$
$n_b^*$	Number of latent individuals with unbroken durations; $n_b^* \equiv n^* - n_u^*$
$n$	Number of observed individuals (note $n \geq n^*$ )
$H_D^*$	Maximum duration until event of interest; $\max \text{supp}(D^*)$
$H_R^*$	Maximum duration until split event; $\max \text{supp}(R^*)$
$H_D$	Maximum observed duration; $\max \text{supp}(D)$
$f_D^*(i)$	Marginal probability mass function of main event duration; $P(D^* = i)$
$f_R^*(i)$	Marginal probability mass function of splitting event duration; $P(R^* = i)$
$f_D(i)$	Marginal probability mass function of observed duration; $P(T = i)$
$f_{RD}^*(i, j)$	Joint probability of splitting duration and event duration; $P(R^* = i, D^* = j)$

Table 4: 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  $\tau$  ordered data files, indexed by  $t \in \mathcal{T} = \{1, \dots, \tau\}$ , with  $n_t$  individuals in each file. Individuals within file  $t$  are indexed by  $i \in \mathcal{N}_t = \{1, \dots, n_t\}$ , 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  $\mathcal{I} = \{1, \dots, n\}$  be the set of all distinct individuals across all files. To compare individuals across files we define the identity function,  $ID_t : \mathcal{N}_t \rightarrow \mathcal{I}$ ,

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

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

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

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. Analogously  $M_U$  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  $U = \{\text{First Name, Last Name}\}$ , 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  $U \subseteq \{1, \dots, K\}$  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

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

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 9.** *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. To be transcribed.* ■

*Proof of Theorem 2. To be transcribed.* ■

## 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,  $P_{d^*}$ ,

$$\underbrace{\begin{bmatrix} f_D(1) \\ f_D(2) \\ f_D(3) \\ \vdots \\ f_D(H_D-1) \\ f_D(H_D) \end{bmatrix}}_{=P_d} = \frac{1}{\lambda} \underbrace{\begin{bmatrix} \sum_{i=1}^{H_D} f_R^*(i) & f_R^*(1) + f_R^*(1) & \cdots & f_R^*(1) + f_R^*(H_D-2) & f_R^*(1) + f_R^*(H_D-1) \\ 0 & \sum_{i=2}^{H_D} f_R^*(i) & \cdots & f_R^*(2) + f_R^*(H_D-3) & f_R^*(2) + f_R^*(H_D-2) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \sum_{i=H_D-1}^{H_D} f_R^*(i) & f_R^*(H_D-1) + f_R^*(1) \\ 0 & 0 & \cdots & 0 & f_R^*(H_D) \end{bmatrix}}_{=A_{r^*}} \underbrace{\begin{bmatrix} f_D^*(1) \\ f_D^*(2) \\ f_D^*(3) \\ \vdots \\ f_D^*(H_D-1) \\ f_D^*(H_D) \end{bmatrix}}_{=P_{d^*}}$$

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

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

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

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  $n^* \rightarrow \infty$ :

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

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

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

and similarly find that  $\lim_{n^* \rightarrow \infty} \frac{n_b^*}{n} = \frac{1}{\lambda} P(D^* > R^*)$ . Bringing it altogether and focusing on the  $k$ th index we have

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

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

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

*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

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

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.  $\blacksquare$

*Proof of Theorem 4 (Consistency of  $\widehat{V}$ ).* Let  $\widehat{f}_D^*$  be a consistent estimator for  $f_D^*$  and let  $f_R^*$  be known. By Slutsky's theorem we have

$$\widehat{\lambda} \rightarrow_p 1 + \sum_{r < d} f_D^*(d) f_R^*(r) = 1 + P(R_i^* < D_i^*) \equiv \lambda$$

Next we consider three cases: either  $P(R_i^* < D_i^*) \in (0, 1)$ ,  $P(R_i^* < D_i^*) = 0$  or  $P(R_i^* < D_i^*) = 1$ . These cases correspond with some duration breakage, no duration breakage, and always duration breakage respectively.

First assume that  $P(R_i^* < D_i^*) \in (0, 1)$ , which implies that  $\lambda \in (1, 2)$ . By Slutsky's theorem we have

$$\begin{aligned} (\widehat{f}_{1|u})_d &\rightarrow_p \frac{1}{2 - \lambda} \sum_{d \leq r} f_D^*(d) f_R^*(r) = P(D_i^* = d | D_i^* \leq R_i^*) \\ (\widehat{f}_{2|b})_{rw} &\rightarrow_p \frac{1}{\lambda - 1} f_D^*(r + w) f_R^*(r) = P(R_i^* = r, D_i^* - R_i^* = w | D_i^* > R_i^*) \end{aligned}$$

Thus  $\widehat{f}_{1|u}$  is consistent for the vector  $f_{1|u}$ , and  $\widehat{f}_{2|b}$  is consistent for the matrix  $f_{2|b}$ . Since  $\lambda \neq 0$ , Slutsky's theorem then also implies that  $\widehat{\Omega}_u, \widehat{\Omega}_b$ , and  $\widehat{\Omega}$  are consistent for  $\Omega_u, \Omega_b$ , and  $\Omega$  respectively. Finally one last application of Slutsky yields that  $\widehat{V}$  converges in probability to  $V$ .

In the second and third cases we must be careful with the denominators of  $\widehat{f}_{1|u}$  and  $\widehat{f}_{2|b}$ . When  $P(R_i^* < D_i^*) = 0$  (i.e.  $\lambda = 1$ )... ■

### A.2.3 Estimation under Dependence

*Proof of Theorem 5 (Identification). To be transcribed.* ■

*Proof of Theorem 6 (Consistency of GMM Estimator). To be transcribed.* ■

*Proof of Theorem 7 (Asymptotic Normality of GMM Estimator). To be transcribed.* ■