Essays on the Econometrics of Data Quality

By

Elan A. Segarra

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Economics)

at the

University of Wisconsin–Madison

2021

Date of final oral examination: 7/14/2021

The dissertation is approved by the following members of the Final Oral Committee:

Jack Porter, Professor, Economics

Felix Elwert, Professor, Sociology

Christopher Taber, Professor, Economics

Jeffrey Smith, Professor, Economics

Dedication

I dedicate this work to my parents,
who instilled in me a life long curiosity and
gave me the support to forge my own path.

Acknowledgments

Like any project worth doing, this couldn't have been completed without the support and guidance of others. Of course to my advisor, Jack Porter, I owe an enormous debt of gratitude for his mentorship over the years and boundless patience with my often nebulous questions and sluggish work pace. I thank my other committee members, Felix Elwert, Chris Taber, and Jeff Smith, for their feedback, expertise, friendliness, and generosity with their valuable time. Other faculty at UW and abroad who have aided in both the fruition of this work and my own development as a researcher include Brent Hueth, Richard Dunn, Joachim Freyberger, Xiaoxia Shi, Bruce Hansen, Mikkel Sølvsten, Alan Sorensen, Corina Mommaerts, Kenneth West, Matthew Wiswall, and Jean-Francois Houde.

Just as faculty have helped in my professional development, my family has been an important source of emotional and social support. I thank my parents, Marsha and Angel, who have been my cheerleaders throughout this process, and had faith in me even when I doubted myself. While they may not have always understood the dilemmas I faced they have always been willing to lend a sympathetic ear. I could never put into words how thankful I am to my partner in life, Ben, who has supported me with his patience and understanding during long work nights and temporary fits of insanity. When I was too distracted to feed myself he cooked me hot meals. More than anything, he has been my anchor to the real world when I've fallen too deep into the rabbit hole of research.

I must also acknowledge the numerous institutions that have contributed to

my work either through financial support or other research resources. The Wisconsin Research Data Center is where the motivation for my work in chapter one was born. While the WiscRDC provided a gateway to incredible data, it also served as a playground for engaging conversations with the other amazing researchers that use the space. I'm grateful to the Wisconsin Business Dynamics Research Consortium for allowing me to use their data as well. I thank the Institute for Research on Innovation & Science at the University of Michigan for both their financial support in my research, but also for the numerous research summits they have organized. Attending these summits over the years has exposed me to not only the brilliant scholars and administrators at IRIS, but also to the larger research community that they have built and continue to nurture. Finally, I am indebted to the tireless work of both Kim Grocholski and Becca George; without them the department would surely have fallen apart long ago.

Last but certainly not least, I must thank all the friends and peers who have accompanied me along this journey. Were it not for Joel McMurry and Nguyen Nguyen and our endless days studying for prelims, I might not have survived the first year. Gary Baker, Amrita Kulka, Moshi Ul Alam, Elise Marifian, Jonathan Becker, and Dennis McWeeny were not just fabulous office mates and friends, but they made grad school tolerable and, dare I say, even enjoyable. Our engaging discussions that ranged from mathematical quandaries to which is the best food cart on the mall will not soon be forgotten. We've served as each other's research soundboards and they have each listened patiently to my DAG evangelizing at one point or another. Additionally, I want to thank Diwakar Raisingh, Anton Babkin, Monica, Arpita Patnaik, Nicolás Badaracco, Joanna Venator, Hans Schwarz, Lois Miller, Natalie Duncombe, Sam Engle, Renata Gaineddenova, Sandra Spirovska, Andrey Zubanov and all the participants of the various seminars and working groups for their attention, brilliance, and camaraderie on this ride.

Abstract

This dissertation consists of three essays which explore scenarios where data quality issues interfere with the goals of empirical research. These situations motivate closer analysis of existing econometric methods and even the provision of new methods to account for the deficiencies present in the data. In all three cases the work presented aims to provide clarity and advice to aid researchers so they may accomplish their primary objective while simultaneously managing the shortcomings in their data.

In the first chapter I consider survival analysis when durations are subject to mismeasurement due to record linkage errors that manifest during data collection and processing. Panel data 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. When conducting survival analysis (e.g. firm death, mortality, or emigration), missed linkages induce error in the observed lifetime durations, and thus inconsistency in standard survival estimators. New methods are developed which restore consistency of the estimators of parameters 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 data 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.

The second chapter investigates the estimation of discrete choice models when market size is unobserved or mismeasured. Estimates of elasticities are a common output of interest in discrete choice models, however they can be significantly biased when the population size is misspecified. In this chapter we decompose the bias in elasticity estimates in the logit model into a direct effect and an indirect effect coming from bias in the structural parameter estimates. Since these effects can go in opposite directions addressing bias from the indirect channel, via market fixed effects, will have an indeterminate effect on the total bias in the elasticity. We provide a complete characterization of when including market fixed effects will mitigate versus exacerbate elasticity bias. Our results reveal that for own characteristic elasticities products with small shares will typically benefit most from market fixed effects while the benefit (or detriment) for cross characteristic elasticities is independent of share.

The third chapter explores instrumental variables estimation in the presence of outcome attrition and presents a novel estimator to handle this missingness. Instrumental variables (IV) methods are a ubiquitous tool for estimating causal effects. However, when data are subject to missingness the exclusion restriction can be violated leading to significant bias in IV estimators. This work proposes a

new method, termed the missingness instrumental variables (MIV) estimator, to recover causal effects in the presence of outcome attrition. The method leverages statistical independences to replace the infeasible moments of the IV estimator with moments that can be estimated using data subject to missingness. Just like IV methods with complete data, MIV is able to estimate many causal effects of interest including average treatment effects, local average treatment effects, and marginal treatment effects. The method is compared with inverse probability weighting methods and multiple imputation methods, and Monte Carlo simulations highlight how MIV fares better than alternative methods when positivity is violated or under misspecification of error distributions.

Contents

Chapter 1

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

Chapter Summary

Panel data 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 chapter 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. New methods are developed which restore consistency of the estimators of parameters 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 data 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.

1.1 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 in ways that 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 chapter I ask what can be learned about the distribution of an event of interest using panel data subject to linkage error? Moreover under what scenarios (i.e. observables

and linkage assumptions) can the object of interest be point identified? Finally, when is partial identification still informative of the objects of interest? 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 (e.g. 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 Honoré 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 chapter proceeds as follows. In section 1.2 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 1.3 presents theoretical results when point identification is achieved, while section 1.4 showcases theoretical results from a partial identification perspective. Section 1.5 displays the results of Monte Carlo simulations of the proposed estimators. Section 1.6 describes the empirical application of the methods to the estimation of firm lifetimes, and section 1.7 concludes.

1.2 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.

1.2.1 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 an integer valued variable. For example if X is an integer valued scalar random variable, then $\vec{X}$ is a binary vector of random variables with a 1 at the index matching the value of X,

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

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

operator represent the vectorization of the object in parentheses. For example, in the case where A is an $m \times n$ matrix then vec(A) will be a $mn \times 1$ column vector created by stacking the columns of A on top of each other starting with the leftmost column at the top. If A had more than two dimensions (e.g. was a tensor) then the exact transformation of A into vec(A) is left ambiguous, but, importantly, vec(A) will still be a one dimensional column vector containing all elements of A.

1.2.2 Researcher Objective

Let $(S_i^*, D_i^*) \in \mathfrak{S}^* \times \mathfrak{D}^*$ be a pair of discrete unobserved random variables with respective supports $\mathfrak{S}^*, \mathfrak{D}^* \subseteq \mathbb{N}^+$. Let D_i^* represent the time until an event of interest measured relative to the individual's 'birth' period, S_i^*. In other words, 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 distinct individuals in a sample is denoted by n^*.

The primary goal of the researcher is to learn the marginal distribution of D_i^*, denoted by f_D^*. Often it is the case that a researcher is instead focused on a moment of this distribution, such as the mean or a specific survival probability. Since the marginal distribution is sufficient for constructing these other parameters, I focus on identifying and estimating f_D^* in this chapter. Similarly, in a world with treatments or controls, if an empiricist is interested in a treatment effect it often suffices to learn the joint distribution of these variables.

Were the researcher able to observe D_i^* then identifying and estimating f_D^* would 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 the individual represented by that data point. The following section describes the process that transforms the latent variables into the observed variables.

1.2.3 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 series of events 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 vector valued random variable, with support $\mathfrak{R}^* \subseteq (\mathbb{N}^+)^L$, that represents the times between record linkage error events (RLEEs). Under this support definition there are a maximum of L breaks possible for each individual in this record linkage error model. The lth element of R_i^* represents the time until the lth RLEE measured relative to the previous RLEE, while the first element represents the time until the first RLEE relative to the individual's "birth" period. Thus a single 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 $S_i^* + R_i^* - 1$. For example, for an individual with values $(S_i^*, D_i^*, R_i^*) = (3, 9, (2, 5))$, they were born in period $t = 3$, the event of interest occurred after 9 periods, the first RLEE happened between period $t = 5$ and $t = 6$, and the second RLEE happened between period $t = 9$ and $t = 10$. A graphical representation of this example can be found in Figure 1.1a.

Depending on the timing of the RLEE the true duration may remain intact or be broken up into 2 or more constituent durations. Whether or not a duration is broken depends on the timing of the RLEEs, R_i^*, relative to the duration of interest, D_i^*. For example when $L = 1$ (i.e. there is at most one break) then the duration will be broken into two parts if the RLEE occurs before the event of interest, $R_i^* < D_i^*$. Otherwise the full duration remains intact. 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 start and duration of the first broken part (denoted by (S_{i1}^*, D_{i1}^*)) represents the time between the start and the RLEE while the second start and duration (S_{i2}^*, D_{i2}^*) is the time between the RLEE 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

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

For the general case (i.e. $L \geq 1$) it is useful to define the number of breaks that will occur given the RLEEs. Let $B_i^* \in \mathbb{N}$ be a discrete random variable that indicates how many times individual i's duration has been broken. This variable is defined as

$$B_i^* = \max\left\{ b \in \{0, 1, 2, \ldots, L\} \,\middle|\, \sum_{l=1}^{b} R_{il}^* < D_i^* \right\}.$$

If $B_i^* = 0$ then the duration has not been broken at all, if $B_i^* = 1$ then it has been broken exactly once into 2 constituent parts, if $B_i^* = 2$ then it has been broken twice into three constituent parts, and so on and so forth. In general, if $B_i^* = b$, then the duration will be broken into $b + 1$ separate constituent parts.

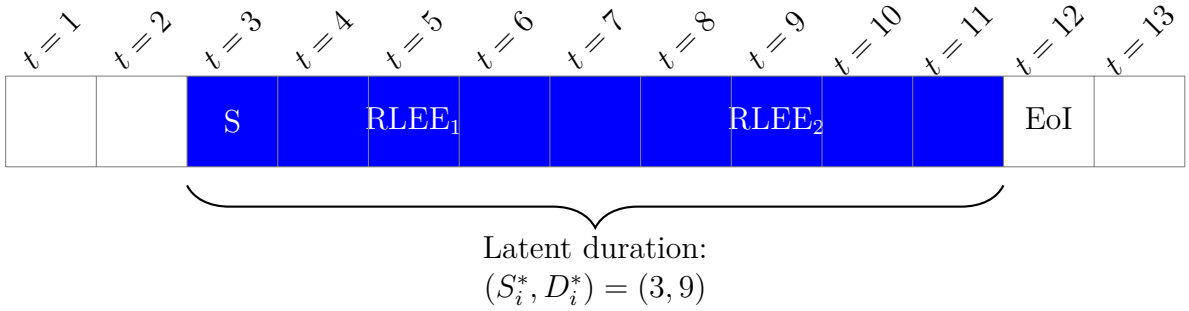
If we let the pair, (S_{ik}^*, D_{ik}^*), denote the start and duration of the kth broken part of individual i, then we have

$$(S_{ik}^*, D_{ik}^*) = \begin{cases} \left(S_i^* + \sum_{l=1}^{k-1} R_{il}^*, R_{ik}^*\right) & \text{if } k < B_i^* + 1 \\ \left(S_i^* + \sum_{l=1}^{k-1} R_{il}^*, D_i^* - \sum_{l=1}^{k-1} R_{il}^*\right) & \text{if } k = B_i^* + 1 \\ missing & \text{if } k > B_i^* + 1 \end{cases}.$$

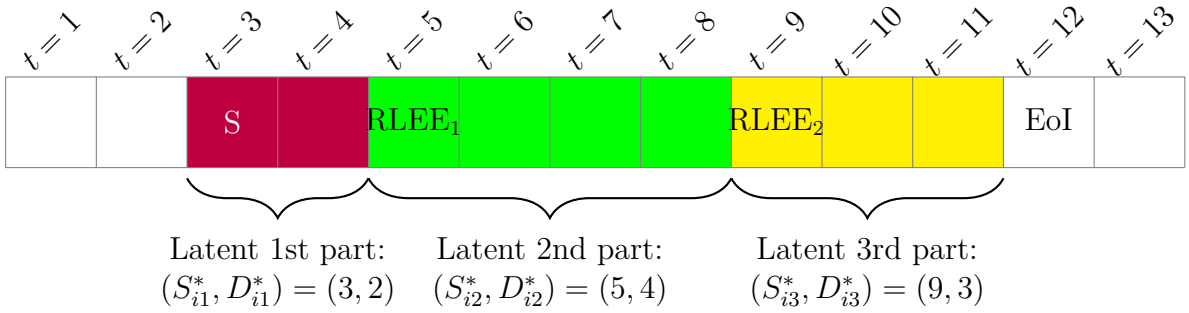
Figure 1.1 illustrates a situation where two breaks result in three latent durations from a single individual.

This process of duration breakage essentially creates an unobserved and unbalanced panel. Those with unbroken durations have one tuple consisting of a latent start and duration, while those with broken durations have two or more tuples of starts and durations (specifically $B_i^* + 1$ tuples). Table 1.1 displays an unbalanced panel that would result from a data set where individuals 1, 3, and 4 have durations broken by RLEEs. Were the panel of (S_{ik}^*, D_{ik}^*) observed it would be trivial to reconstruct the underlying values of (S_i^*, D_i^*).

(a) Latent Durations



(b) Observed Durations



Observed starts and durations:
$(S_j, D_j) = (3, 2)$ $(S_k, D_k) = (5, 4)$ $(S_m, D_m) = (9, 3)$

Figure 1.1: The above figures illustrate the latent durations (1.1a) and observed durations (1.1b) for an individual with $(S^*, D^*, R^*) = (3, 9, (2, 4))$. 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}^*)	(S_{i3}^*, D_{i3}^*)
1	(3, 6, (2,5))	1	(3, 2)	(5, 4)	-
2	(1, 5, (5,1))	0	(1, 5)	-	-
3	(2, 3, (2,4))	1	(2, 2)	(4, 1)	-
4	(4, 6, (2,3))	2	(4, 2)	(6, 3)	(9,1)
5	(2, 3, (4,1))	0	(2, 3)	-	-

Table 1.1: Example of latent unbalanced panel produced by record linkage error breaking some durations into smaller constituent parts.

The observed data on start times and durations does not include information on the unbalanced panel aspect of the latent variables. Moreover, the indices of the observed data also do not provide information on the latent unbalanced panel. To capture these features in the model, the final step flattens and randomizes the indices of the broken and unbroken durations. Let $n_b^* \equiv \sum_i \mathbb{1}\{B_i^* = b\}$ denote the number of individual with exactly b breaks. Thus there are n_0^* individuals with unbroken durations, n_1^* individuals with one break (and 2 duration parts), n_2^* individuals with two breaks (and 3 duration parts), and so on. Without loss of generality, assume the individuals are ordered by B_i^* so that $B_i^* = 0$ for all $i = 1, \ldots, n_0^*$, $B_i^* = 1$ for all $i = n_0^* + 1, \ldots, n_0^* + n_1^*$ and in general $B_i^* = b$ for all $i = \sum_{l=1}^{b-1} n_l^* + 1, \ldots, \sum_{l=1}^{b} n_l^*$. Since individual i will end up having $B_i^* + 1$ durations associated with it, there are a total of

$$n = \sum_{b=0}^{L} n_b^* (b+1) \tag{1.1}$$

separate durations in the broken sample.

To randomize the indices we first construct the set of all indices of the broken durations and then map them to the index in our final observed sample. Let the set of 2-dimensional indices over the latent broken and unbroken durations, (S_{ib}^*, D_{ib}^*), be denoted by

$$\mathcal{I} = \bigcup_{B=0}^{L} \left(\bigcup_{i:B_i=B} \{(i,b)\}_{b=1}^{B+1} \right).$$

Define the random permutation operator, $\pi : \{1, \ldots, n\} \to \mathcal{I}$, which randomly assigns the univariate indices of the observed sample to the bivariate indices of all the broken and unbroken durations. Finally, let an observed start and duration, denoted by (S_i, D_i), be mapped from one of these latent start/duration tuples:

$$(S_i, D_i) = \left(S_{\pi(i)}^*, D_{\pi(i)}^* \right).$$

Thus at the end of the record linkage process a researcher observes a sample of size n of

either durations alone or start times and durations,

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

These two cases will be treated separately since they have different ramifications for identification and estimation.

Remark 1. As modeled here, when a researcher observes a duration, D_i, they have no knowledge of whether that duration corresponds to a true unbroken duration, D_i^*, or one of the constituent parts of a broken duration. Moreover, when only durations are observed, the researcher has no information on which durations belong to the same individual. Accounting for this unobserved correlation structure is one hurdle to be addressed by the work here.

Remark 2. If there is record linkage error, i.e. $B_i^* > 0$ for some i, then the size of the observed sample, n, is strictly larger than the number of truly distinct individuals, n^*. This follows directly from equation (1.1) and the fact that $n^* = \sum_{b=0}^{L} n_b^*$.

There are two assumptions that are maintained throughout this chapter, 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 (iid) sample of size* n^*.

Assumption A2 (Bounded Support)**.** D_i^* *has bounded support on the positive integers,* $\mathfrak{D}^* = \{1, 2, \ldots, H_D^*\}$.

This first assumption is analogous to 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 potentially 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.

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

Assumption A3 (No Missing Constituents)**.** *If* $B_i^* \geq 1$ *then all parts of the broken duration,* $\left\{D_{i1}^*, D_{i2}^*, \ldots, D_{iB_i^*+1}^*\right\}$*, are found in the observed sample.*

This assumption merely guarantees that no durations are left out of the observed sample. For example, if we observe the first part of a duration which was broken into three parts, then the second and third constituents must also be in the sample somewhere.

In this section we have described a model of record linkage error using durations between error events which may seem divorced from the typical approach of having multiple data sets and modeling the matching process between records. The model described is implied in scenarios where data from several time periods is linked using a deterministic algorithm and the matching variables are sufficient necessary switchers. Refer to Appendix 1.9.2 for a full description of these types of matching variables, amenable linking algorithms, and properties which imply that the above duration model is appropriate.

1.2.4 Examples

To help further elucidate how the above model functions, I present two examples in this section: one in the context of firm dynamics and the other in the context of individual migration. In each I describe the real world counterparts to the latent and observed variables defined above, and remark on any idiosyncrasies 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 the 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, and can change over a firm's lifetime even when no substantive change in the goods, customers, or structure has occurred. Changes in EIN can be driven by incentives to manipulate unemployment insurance rates in schemes known as State Unemployment Tax Act (SUTA) dumping (see Benedetto et al. (2007) and Kearns (2006) for further discussion). Without unique IDs, other variables such as employer name and address are often used to link firms across years. However, when firms relocate, perhaps due to growing pains or business contraction, linkage error is more likely to occur. In this context, R^* represents the durations between firm relocations.

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 the fact that the model accumulates mass into the left tail due to young and old firms being more likely to relocate. Thus even a small amount of linkage error can result in much higher 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/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 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. In the absence of SSNs, the first name, last name, date of birth,

and gender become important linking variables. However, linking error can be considerable due to the large percentage of women who change their last name after getting married. In this context the time until marriage (R^*) will represent the record linkage error event.

This scenario is further complicated by the sensitive nature of releasing names to researchers, meaning even those accessing the restricted versions of these datasets often do not have access to names. In these situations, there is little hope for researchers to directly fix any linking error when they do not have access to these important matching variables.

One mitigating element is that when age or date of birth are available (which is common) durations can be more accurately measured despite 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.

1.2.5 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 $f_{RD}^*(r,d) \equiv P(R^* = r, D^* = d)$, and denote the distribution function of D_i by $f_D(k) \equiv P(D = k)$.

For pedagogical reasons it is useful to start with a simpler model where we restrict $L = 1$, meaning that there is a single RLEE per individual which may or may not result in a broken

duration. In this case, the distribution of observed durations is given by

$$f_D(k) = P(D = k) = \frac{1}{\lambda}\left[\overbrace{\sum_{r=k}^{\infty} f^*_{RD}(r,k)}^{\text{unbroken}} + \overbrace{\sum_{d=k+1}^{\infty} f^*_{RD}(k,d)}^{\text{broken 1st half}} + \overbrace{\sum_{d-r=k} f^*_{RD}(r,d)}^{\text{broken 2nd half}}\right] \tag{1.2}$$

$$\text{where} \quad \lambda = 1 + P(R^* < D^*) = 1 + \sum_{r<d} f^*_{RD}(r,d) \tag{1.3}$$

Every observed duration falls into one of three categories as represented by the separate summations. Each sum iterates over the set of latent events that would give rise to an observed duration of that type and of length k. The first sum contains latent events that result in no record linkage error because the RLEE occurs at the same time or after the event of interest. The second and third sums represent the two duration constituents that are generated when the RLEE breaks the duration. Finally, the entire object is rescaled by λ to account for the fact that the observed sample of durations has strictly more observations than the latent sample that generates it.

The general form, i.e. for any $L \in \mathbb{N}^+$, involves slightly more complex indexation for the sums, but the general structure is the same as above,

$$f_D(k) = P(D = k) = \frac{1}{\lambda}\left[\overbrace{\sum_{r\in\mathcal{I}_u(k)} f^*_{RD}(r,k)}^{\text{unbroken}} + \overbrace{\sum_{(r,d)\in\mathcal{I}_m(k)} f^*_{RD}(r,d)}^{\text{broken middle part}} + \overbrace{\sum_{(r,d)\in\mathcal{I}_e(k)} f^*_{RD}(r,d)}^{\text{broken end part}}\right] \tag{1.4}$$

$$\text{where} \quad \lambda = 1 + \sum_{k} \sum_{(r,d)\in\mathcal{I}_m(k)\cup\mathcal{I}_m(k)} f^*_{RD}(r,d) \tag{1.5}$$

$$\mathcal{I}_u(k) = \{r \in \mathfrak{R}^* | r_1 \geq k\} \tag{1.6}$$

$$\mathcal{I}_m(k) = \left\{(r,d) \in \mathfrak{R}^* \times \mathfrak{D}^* \,\middle|\, \exists l \in \mathbb{N}^+ \text{ with } r_l = k \text{ and } \sum_{j=1}^{l} r_j \leq d\right\} \tag{1.7}$$

$$\mathcal{I}_e(k) = \left\{(r,d) \in \mathfrak{R}^* \times \mathfrak{D}^* \,\middle|\, \exists l \in \mathbb{N}^+ \text{ with } d - \sum_{j=1}^{l-1} r_j = k \text{ and } \sum_{j=1}^{l} r_j > d\right\} \tag{1.8}$$

Once again every observed duration falls into one of three categories, and each sum

iterates over the set of latent events that would give rise to an observed duration of that type and of length k. The first set, $\mathcal{I}_u(k)$, contains all events that result in an unbroken duration of length k, which occurs whenever the first RLEE happens at the same time or after the event of interest. The second set, $\mathcal{I}_m(k)$, contains all events that result in broken durations with an observed duration of length k that are not the tail of the constituent durations. For example, the 1st and 2nd parts found in Figure 1.1b (colored red and green respectively) fall into this category. The third set, $\mathcal{I}_e(k)$, contains all events that result in broken durations where the tail event has length k. The 3rd duration found in Figure 1.1b (colored in yellow) is of this variety.

Remark 3. It is important to note that $\mathcal{I}_m(k)$ is a multiset, meaning that it can, and likely will, contain repetitions of the same element. This stems from the possibility that a single individual true duration (i.e. latent event) may be broken up into several constituent durations of the same length. This aspect is also what results in potential overlap between $\mathcal{I}_m(k)$ and $\mathcal{I}_e(k)$. Consider $\mathfrak{D}^* = \{1, 2\}$ and $\mathfrak{R}^* = \{1, 2\}^2$; meaning there are at most 2 linkage errors. It can be shown that

$$\begin{aligned}
\mathcal{I}_u(1) &= \{((1,1),1),((1,2),1),((2,1),1),((2,2),1)\} \\
\mathcal{I}_m(1) &= \{((1,1),2),((1,2),2)\} \\
\mathcal{I}_e(1) &= \{((1,1),2)\}.
\end{aligned}$$

The latent event $(R^*, D^*) = ((1,1),2)$ appears both in $\mathcal{I}_m(1)$ and $\mathcal{I}_e(1)$ because it results in observing 2 durations of length 1.

Next we turn to some of the properties of the observed distribution and how they relate to the latent distribution. While most properties follow along our intuition, some may seem surprising on initial inspection. A brief exploration of a few of these will hopefully convince the reader that exploring identification and estimation in this framework is nontrivial.

To start we consider how the mean of the observed durations compares to the mean of the true durations of interest (proofs of Proposition 1 and 2 are found in appendix 1.9.4.1).

Proposition 1. *Under assumptions A1-A3 we have that* $\mathbb{E}[D_i] = \frac{1}{\lambda}\mathbb{E}[D_i^*] = \frac{1}{1+\mathbb{E}[B_i^*]}\mathbb{E}[D_i^*]$.

This result aligns with our intuition in that it illustrates that observed durations are smaller on average than the truth. It goes further by displaying that the attenuation factor is driven entirely by the expected number of breaks in a given scenario. A simple corollary to this proposition guarantees that the true mean is identified from the observed mean if either the distribution of the number of breaks is known or even if only the mean of this distribution is known. In the special case where there is at most one break per individual, i.e. $L = 1$, then simply knowing the true number of individuals, n^*, is entirely sufficient to identify to identify $\mathbb{E}[B_i^*]$ and thus $\mathbb{E}[D_i^*]$ from the observed durations.[1]

This proposition additionally exhibits how partial identification approaches can be leveraged to establish sharp bounds on the mean of the latent distribution. Under assumption A2 we have that $0 \leq \mathbb{E}[B_i^*] \leq H_D^* - 1$ which implies that

$$\mathbb{E}[D_i] \leq \mathbb{E}[D_i^*] \leq H_D^* \mathbb{E}[D_i].$$

With more information on the linkage error process these bounds can be tightened further. For instance, in example 2 it might be perfectly reasonable to assume that nobody gets married and changes their name more than twice, in which case $\mathbb{E}[B_i^*] \leq 2$ and our bounds on $\mathbb{E}[D_i^*]$ become $[\mathbb{E}[D_i], 3\mathbb{E}[D_i]]$. The general idea of looking at this model and what can be learned about f_{RD}^* using partial identification approaches is explored in depth in section 1.4.

Continuing with our intuition and the result of proposition 1 it would be reasonable to surmise that the model always shifts probability mass toward the left tail, i.e. that the latent duration distribution has first order stochastic dominance over the observed duration distribution. For example, 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

[1]This follows because $(n - n^*)/n^* \to_p P(R^* < D^*) = \mathbb{E}[B_i^*]$ when $L = 1$.

proposition reveals that this is not at all guaranteed.

Proposition 2. *Let $L = 1$. If $H_D^* > 3$ then D^* and D cannot be ordered using first order stochastic dominance without further assumptions.*

This small proposition demonstrates that there exist distributions where the linkage error does not merely shift mass from right to left, but instead moves mass around in potentially non-intuitive ways. This can happen, for instance, 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. Refer to the proof in appendix 1.9.4.1 for a complete example of when this problem occurs.

Remark 4. Note that when $H_D^* \leq 3$ then D^* first order stochastically dominates D. The ability to stochastically order these distributions comes about entirely because of the limited flexibility in the set of discrete distributions with only 3 or fewer points of support.

This lack of stochastic ordering between the latent and observed duration distributions comes about from the interaction of two potentially countervailing forces. The first force is the breakdown of durations into shorter constituent parts which does indeed shift mass to the left tail. The second force comes from the generation of additional durations and the subsequent rescaling of the distribution needed to account for this. Therefore, even though there will necessarily be more observed durations of length one than true durations of length 1, the probability of observing a duration of length one can still fall relative to the latent probability if there are many record linkage breaks and thus many more durations in the observed data set relative to the latent.

1.3 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 that are considered separately. In section 1.3.1 I assume independence resulting in point identification of the distribution of event durations, while in section 1.3.1 I place no restriction on the dependence structure and explore partially identified set of distributions.

1.3.1 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 one 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 $f_D^* \equiv \left[\begin{array}{ccc} f_D^*(1) & \cdots & f_D^*(H_D^*) \end{array} \right]'$ and $f_R^* \equiv \left[\begin{array}{ccc} f_R^*(1) & \cdots & f_R^*(H_R^*) \end{array} \right]'$ 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)**.** $\max \mathfrak{D}^* \leq \max \mathfrak{R}^*$.

One implication of this assumption is that all durations have a chance of being broken by record linkage error. Together with 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*

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

The proof of Theorem 1 proceeds quite naturally after formulating the relationship between the distributions as a nearly linear system of equations (refer to appendix 1.9.4.2 for the full proof). The intuition behind this identification result can be described in two parts.

The contrapositive of the identification result is the easier of the two directions to establish. If A5 does not hold then $\max \mathfrak{D}^* > \max \mathfrak{R}^*$ which means there are durations with positive probability that are always being broken. If these durations are always being broken, then there is no observable information about the likelihood of these durations. For example, if individuals that live to 10 years (i.e. $D_i^* = 10$) always relocate before they die (i.e. $R_i^* < 10$) then we will only observe durations strictly less than 10 years and will never be able to identify the probability of truly living 10 years.

The intuition for the forward 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 permits identification of the entire distribution.

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,

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

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

$$\widehat{f_D^*} \equiv \frac{A_{R^*}^{-1} \frac{1}{n} \sum_{i=1}^{n} \vec{D}_i}{\mathbf{1}' A_{R^*} \frac{1}{n} \sum_{i=1}^{n} \vec{D}_i}, \tag{1.9}$$

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^*. Refer to section 1.9.3.1 for a detailed discussion of this matrix.

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 1.9.4.2).

Theorem 2 (Consistency)**.** *Under assumptions A1-A5 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 5. 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 (1.9), 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 if Ω is invertible then*

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

where

$$V = \lambda \left(A_{R^*}^{-1}\right)' \left(I - f_D^* \mathbf{1}'\right)' \Omega \left(I - f_D^* \mathbf{1}'\right) A_{R^*}^{-1} \qquad \textit{and} \qquad \Omega = Var\left(\sum_{l=1}^{B_j^*+1} \overrightarrow{D_{jl}^*}\right).$$

Once again care must be taken when proving Theorem 3 because the sample is not iid. Even though the dependence structure is not observed, accounting for the dependence while establishing asymptotic normality is accomplished by focusing on the underlying sample that is iid. Once the estimator is rewritten as one over the latent sample, standard asymptotics can be applied. Refer to appendix 1.9.4.2 for the full proof of Theorem 3.

At the heart of the variance of the asymptotic distribution is Ω, which at first glance may seem unwieldy. Closer inspection reveals that the random sum inside Ω merely counts the number of durations of various lengths that occur when a duration is broken into constituent parts. In other words, the variation in this estimator is primarily driven by how much the underlying data generating process, i.e. the distribution of R^*, breaks up the lifetimes of interest. At one extreme, if $L = 0$, meaning there is no linkage error occurring, then this variance simplifies to $\Omega = Var(D_i^*)$. At the other extreme, if L is large and the distribution of R_i^* is concentrated at smaller values (i.e. many breaks are occurring) then B_i^* will likely be large on average resulting in many terms in the summand of Ω and thus a large asymptotic variance V. For more discussion of these objects, and in particular how Ω is constructed from the latent distributions, f_D^* and f_R^*, refer to Appendix 1.9.3.2.

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. Normally this would be insurmountable, but there is an alternative route since the asymptotic covariance matrix is entirely determined by f_D^* and f_R^* . Consider the plugin estimator of V,

$$\widehat{V} = \widehat{\lambda}\left(A_{R^*}^{-1}\right)'\left(I - \widehat{f_D^*}\mathbf{1}'\right)'\widehat{\Omega}\left(I - \widehat{f_D^*}\mathbf{1}'\right)A_{R^*}^{-1}, \tag{1.10}$$

where the individual cells of $\widehat{\Omega}$ are estimated using

$$\left(\widehat{\Omega}\right)_{ij} = A_2(i,j)'\text{vec}\left(\widehat{f_{RD}^*}\right) - \text{vec}\left(\widehat{f_{RD}^*}\right)'A_1(i,j)'\text{vec}\left(\widehat{f_{RD}^*}\right). \tag{1.11}$$

Both $A_1(i,j)$ and $A_2(i,j)$ are integer valued matrix functions whose values are only determined by the dimension of D^* and R^*. Refer to Appendix 1.9.3.2 for a complete description of the structure and construction of these auxiliary matrices. For our purposes, all that is important is that once we have estimates of the latent distribution, $\widehat{f^*_{RD}}$, then (1.10) and (1.11) allow for the estimation of the asymptotic covariance matrix.

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

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 1.9.4.2, the procedure is a fairly straightforward repeated application of Slutsky's theorem on each of the constituent estimators found in equations (1.10) and (1.11).

1.3.2 Dependent Durations

In some scenarios it may be untenable to assume that the underlying record linkage error process is completely independent of the time until the event of interest. In both running examples this may be the case since we might expect successful growing firms, which are unlikely to die, to be more likely to change their location. Similarly, an individual's propensity to immigrate could be affected by, if not primarily driven by, changes in their marital status. To address these situations, in this section we allow for completely unrestricted dependence between R^*_i and D^*_i and we explore identification, estimation, and inference results under this regime.

In other similar frameworks, for example a competing risks or convolution framework, allowing for dependence between the two input distributions results in the loss of point identification (Tsiatis 1975), and we will find the same result here. One potential way to restore point identification would be to parameterize the dependence at the expense of the flexibility of the marginals, for example by imposing a correlated binomial distribution. Other approaches might involve including additional covariates as done in Heckman and Honoré

(1989) or Abbring and van den Berg (2003). However, we will see that the unique structure of the record linkage error model under discussion will mean that simply observing starting times will lend identification power capable of uncovering the entire joint distribution of interest.

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 1.2. 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 a true unbroken duration or individuals B and E could be the first and second half of the same individual because of their adjacency. Similarly, individuals A and C or A and D could represent pairs of broken durations. Moreover, if this figure represented the entire data set (i.e. a total of 7 observations) then the imposition of assumption A3, no missing constituents, would imply that the observations at the horizons also contains useful information. For example, individual A could not represent the second

half of a broken duration while neither individuals F nor G could represent the first halves of a broken duration.

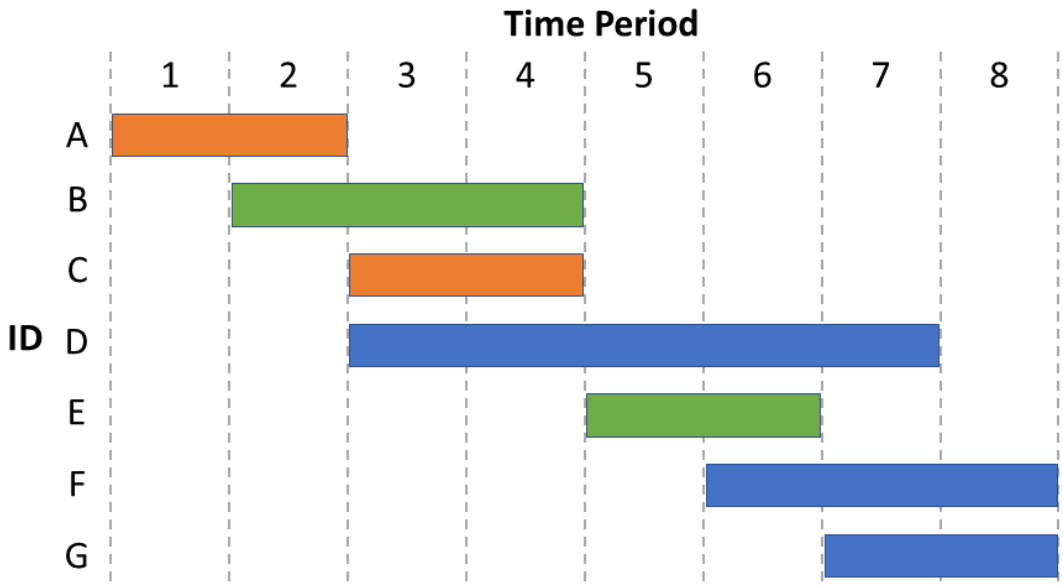


Figure 1.2: Visual example of panel data (with linkage error) illustrating potential relationships among observed durations.

Throughout this section, we will focus on a linkage error model that permits at most one linkage error per individual history, i.e. $L = 1$. This is primarily done for mathematical tractability, since even just characterizing an unrestricted joint distribution between D_i^* and a multidimensional R_i^* is already challenging. That being said, most of the results to be discussed should be readily extendable to a model of multiple linkage errors where the distribution of the time between linkage errors is both memoryless and identically distributed. The more general case, which allows for a completely unrestricted multidimensional R_i^*, remains in the space of active research.

Let the matrix form of the joint distribution of R_i^* and D_i^* be denoted by f_{RD}^*,

$$f_{RD}^* = \begin{bmatrix} f_{RD}^*(1,1) & f_{RD}^*(1,2) & \cdots & f_{RD}^*(1,H_D^*) \\ f_{RD}^*(2,1) & f_{RD}^*(2,2) & \cdots & f_{RD}^*(2,H_D^*) \\ \vdots & \vdots & \ddots & \vdots \\ f_{RD}^*(H_R^*,1) & f_{RD}^*(H_R^*,2) & \cdots & f_{RD}^*(H_R^*,H_D^*) \end{bmatrix}$$

where $f_{RD}^*(i,j) = P(R^* = i, D^* = j)$.

Though we are able to relax all restrictions on the dependence between D_i^* and R_i^*, additional assumptions are still required to achieve point identification. Consider the following two assumptions:

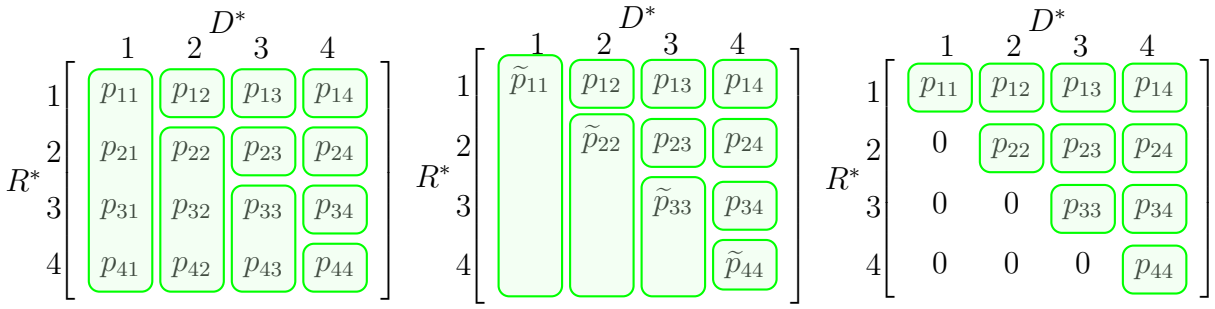
Assumption A6 (Independence of Start)**.** $S_i^* \perp R_i^*$ *and* $S_i^* \perp D_i^*$.

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

While the content of assumption A6 is straightforward its 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. Overall the assumption is innocuous enough that it is implicitly assumed throughout most theoretical work in survival analysis when modeling censoring .

Assumption A7 is useful to rectify the fact that only certain joint probabilities can be identified. In a simple model where $H_D^* = 4$ and $H_R^* = 4$ figure 1.3a 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 1.3c).[2] 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 $\widetilde{p}_{ii} \equiv P(R^* \geq i, D^* = i)$ since that is identifiable (as illustrated in figure 1.3b). This normalization is more appropriate in example 2 where it is perfectly reasonable for a person to get married (R^*) after emigrating (D^*).



(a) Identified probabilities (b) A7 as a normalization (c) A7 as an assumption

Figure 1.3: Illustrates which probabilities are separately identifiable (a), and how assumption A7 can be interpreted as either a normalization (b) or as a strict assumption (c). Circled green cells indicate which probabilities are separately identifiable, where $p_{ij} = f^*_{RD}(i,j) = P(R^* = i, D^* = j)$, and $\widetilde{p}_{ii} \equiv P(R^* \geq i, D^* = i)$.

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.

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

[2]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.

1. *the marginal distributions, f_S^* and f_D^*, are point identified, and*

2. *if A7 also holds then f_S^* and the joint distribution, f_{RD}^*, are point identified.*

The full proof of this result is found in Appendix 1.9.4.3, 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. For example, under independence of start times and the bounded support condition the distribution of observed starting times for those individuals with the maximum duration yields information on the distribution of true starting times since we know these durations are unbroken. In other words, these assumptions imply that the marginal distribution of the start times is identified from the conditional distribution of start times given maximum durations, $P\left(S_i^* = s\right) = P\left(S_i = s | D_i = H_D\right)$.

With the marginal distribution of true starting times in hand we can move on to identifying the distributions of interest in two steps. First, we can identify the joint probabilities that result in broken durations, i.e. those cells of f_{RD}^* which lie above the diagonal as illustrated in Figure 1.3. These probabilities are identified from the distribution of durations observed at the tail end of our horizon since they must all be coming from broken durations. This implication critically relies on assumption A3 since all durations observed after the final true starting period (which is identified from f_S^*) must be the second halves of broken durations. Second, with the joint probabilities that result in breaks identified, we can back out the probabilities associated with unbroken events from the rest of the observed distribution of starts and durations. As highlighted in Figure 1.3 we cannot identify the joint probabilities below the diagonal, e.g. $f_{RD}^*(r, d)$ when $r \geq d$, since these events will always be observationally equivalent. However, we can still identify the sums of columns below the diagonal of f_{RD}^*, and thus identify the sums of each column which of course correspond with the marginal probability distribution, f_D^*.

Though an estimator can be derived which follows the identification strategy just discussed, this approach would ignore all the information available to the researcher in the

observed data. In particular, the identification strategy above only requires the final column and a few of the final rows of the joint distribution, f_{SD}, however the full distribution is accessible. The additional moments could be used to test the assumptions of the model, however, here we will instead use them to improve the accuracy of our estimator since the dimension of the parameter of interest, f^*_{RD}, may be large.

To use the entire observable distribution in our estimation procedure, and to package the estimator in a familiar form to empiricists, we can use the Generalized Method of Moments (GMM) to estimate the distributions of interest. Since all distributions are discrete and finite (i.e. the parameter of interest is a finite vector) then the transformation between the latent distributions, f^*_{RD} and f^*_S, to the observed distribution, f_{SD}, can be cast as a series of moment conditions which make GMM a perfectly appropriate method. Under the assumption of independent starting times this relationship is given by

$$f_{SD} = g\left(f^*_{RD}, f^*_S\right) \equiv \frac{1}{\lambda}\left(\begin{bmatrix} f^*_S \\ 0 \end{bmatrix}\left(\mathbf{1}'\left[f^*_{RD}\right]_L + \mathbf{1}'\left[f^*_{RD}\right]'_U\right) + \sum_k f^*_S(k) L_k \left[\left[f^*_{RD}\right]_U\right]_Q\right) \tag{1.12}$$

$$\text{where } \lambda = 1 + \mathbf{1}'\left[f^*_{RD}\right]_U \mathbf{1}. \tag{1.13}$$

There are a few new pieces of notation ($[\cdot]_L, [\cdot]_U, [\cdot]_Q$ and L_k) used for the first time here. For intuition sake it is sufficient to understand that the terms found in equation (1.12) are analogous to the terms found in equation (1.4). Specifically, the term with $[\cdot]_L$ represents events with unbroken durations (i.e. the lower triangular part of f^*_{RD}), the term with $[\cdot]_U$ represents the first half of broken durations (i.e. the upper triangular part of f^*_{RD}), and the third term with $[\cdot]_Q$ represents the second half of broken durations. This third term involves a shifting matrix, L_k, since the starting times of these broken durations are being shifted forward from the true starting times. For those readers diving deeper, the exact definitions of these objects are described precisely in Appendix 1.9.1.

Let $\overrightarrow{SD_i}$ denote the $H_S \times H_D$ matrix for individual i whose cells are defined by

$$\left(\overrightarrow{SD_i}\right)_{s,d} = \mathbb{1}\{S_i = s \text{ and } D_i = d\}.$$

We can then define the following moment condition function:

$$m\left(\overrightarrow{SD_i}; f_1, f_2\right) \equiv \text{vec}\left(g\left(f_1, f_2\right) - \overrightarrow{SD_i}\right). \tag{1.14}$$

Since $g(\cdot)$ maps the latent distributions to the observed distribution then equation (1.12) can be written compactly as

$$\mathbb{E}\left[m\left(\overrightarrow{SD_i}; f^*_{RD}, f^*_S\right)\right] = 0.$$

This single equation in reality represents M moment conditions with

$$M \equiv \dim\left(m\left(\overrightarrow{SD_i}; f^*_{RD}, f_S\right)\right) = H^*_S H^*_D + \frac{1}{2} H^*_D (H^*_D - 1).$$

Given a positive definite $M \times M$ weight matrix, W, we can define the sample moment function, the GMM objective function, and the corresponding estimator as follows

$$\overline{m}\left(f_1, f_2\right) = \frac{1}{n}\sum_{i=1}^{n} m\left(\overrightarrow{SD^*_i}; f_1, f_2\right) \tag{1.15}$$

$$J_n\left(f_1, f_2\right) \equiv \overline{m}\left(f_1, f_2\right)' W \overline{m}\left(f_1, f_2\right) \tag{1.16}$$

$$\left(\widehat{f^*_{RD}}, \widehat{f^*_S}\right) = \underset{\substack{f_1 \in \Delta(\mathcal{R}\times\mathcal{D}) \\ f_2 \in \Delta(\mathcal{S})}}{\arg\min}\ J_n\left(f_1, f_2\right). \tag{1.17}$$

The statements of further assumptions and theorems are made clearer and more concise if we cast the latent distributions, f^*_{RD} and f^*_S, and the spaces where they live into a single vector of parameters:

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

$$\theta^* \equiv \begin{bmatrix} \text{vec}\left(f^*_{RD}\right) \\ f^*_S \end{bmatrix} \qquad \widehat{\theta} \equiv \begin{bmatrix} \text{vec}\left(\widehat{f^*_{RD}}\right) \\ \widehat{f^*_S} \end{bmatrix}$$

Note that we have written the estimator and its constituent parts in terms of the joint distribution, f^*_{RD}. If you are in a situation where assumption A7 holds as a normalization or does not hold at all (in which case the entire joint, f^*_{RD}, is not identified) then only a slight

modification is required for estimation of the marginal distribution, f_D^*. Proceed using the exact same estimator as described, and finish using $\widehat{f_D^*} \equiv \mathbf{1}'\widehat{f_{RD}^*}$. All of the consistency and inference results about to be discussed can be trivially extended to this estimator of f_D^*.

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 (1.12). This assumption is necessary to guarantee that the moment condition function is uniformly continuous over the parameter space. In our specific context, this assumption ensures that observing two distributions over S and D that are "close" implies that the latent distributions that generated each of these are also "close".

Assumption A8 (Regularity Conditions for Consistency)**.** *There exists a function* $h : \mathbb{R} \mapsto \mathbb{R}$ *with* $\lim_{u \to 0^+} h(u) = 0$ *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 1.9.4.3) leverages the consistency results of the clustered GMM estimators described in Hansen and Lee (2019).

Theorem 6 (Consistency)**.** *Let the researcher observe starts and durations from the sample subject to linkage error,* $\{S_i, D_i\}_{i=1}^n$*, and suppose that the assumptions of Theorem 5 hold along with A8. Then* $\widehat{\theta} \to_p \theta^*$.

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 for Asym. Normality)**.** *Let* f^*_{RD}*,* f^*_S*,* $m\left(\overrightarrow{SD_i}; f_1, f_2\right)$*, and* W *satisfy:*

1. $\theta^* \in$ *interior* (Θ)
2. *There exists a function* $h : \mathbb{R} \mapsto \mathbb{R}$ *with* $\lim_{u \to 0^+} h(u) = 0$ *such that for every pair of latent distributions,* θ_1, θ_2*, in a neighborhood of* θ^* *we have*

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

3. *The eigenvalues of* W *are bounded away from 0,* $\lambda_{\min}(W) \geq C > 0$*.*
4. *The eigenvalues of* Ω *are bounded away from 0,* $\lambda_{\min}(\Omega) \geq D > 0$*.*
5. Q *has full column rank.*

Q *and* Ω *are defined below in Theorem 7.*

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 Θ. 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

$$\sqrt{n}\left(\widehat{\theta}-\theta^{*}\right)\rightarrow_{d} N\left(0,V\right)$$

where

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

$$Q\equiv\mathbb{E}\left[\frac{\partial}{\partial\theta}g\left(\theta^{*}\right)\right]\qquad\Omega\equiv\frac{P(R^{*}<D^{*})}{\lambda}\Omega_{b}+\frac{P(R^{*}\geq D^{*})}{\lambda}\Omega_{u}\qquad\lambda=1+P\left(R_{i}^{*}<D_{i}^{*}\right).$$

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 Ω_b and Ω_u which correspond with the broken and unbroken clusters respectively. In particular these variables are defined to be

$$\Omega_{b}\equiv\mathbb{E}\left[\left(m\left(\overrightarrow{SD_{1i}};\theta^{*}\right)+m\left(\overrightarrow{SD_{2i}};\theta^{*}\right)\right)\left(m\left(\overrightarrow{SD_{1i}};\theta^{*}\right)+m\left(\overrightarrow{SD_{2i}};\theta^{*}\right)\right)'|R_{i}^{*}<D_{i}^{*}\right]$$

$$\Omega_{u}\equiv\mathbb{E}\left[m\left(\overrightarrow{SD_{i}};\theta^{*}\right)m\left(\overrightarrow{SD_{i}};\theta^{*}\right)'|R_{i}^{*}\geq D_{i}^{*}\right].$$

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.

To construct standard errors and confidence intervals associated with $\widehat{f_{RD}^{*}}$ it is necessary to have a consistent estimator of V. As before, in section 1.3.1, the typical methods of Hansen and Lee (2019) or Hwang (2021) do not apply since the clusters are unobserved, However, the constituent objects, Q,Ω_b, and Ω_u are all functions of the underlying latent distributions, so a plugin estimator can be explored.

1.4 Partial Identification and Estimation

The assumptions presented in sections 1.3.1 and 1.3.2 may prove untenable in some empirical contexts, in which case point identification may be out of reach. While point identification

may be out of reach in these situations, a natural question that arises is whether there is still information to say something about the latent variables of interest. In this section we present partial identification results which spotlight how informative bounds on the objects of interest, e.g. marginal distributions or means, can be constructed without imposing any restrictions on the relationships between S_i^*, D_i^*, or R_i^*. These methods are especially useful when an empiricist seeks to implement a sensitivity analysis to gauge how much potential record linkage error may affect the estimation of the parameter of interest.

Some additional notation relevant to this section is noted here. 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

$$\mathcal{H}(f_{RD}^*) = \left\{ f_{RD}^* \in \Delta^{H_D^2 - 1} : \mathbb{E}\left[\vec{D}_i + \left(\vec{D}_i b'_{H_D} - A_{H_D} \right) \mathrm{vec}(f_{RD}^*) \right] = 0 \right\}. \tag{1.18}$$

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 is 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, $\mathcal{H}_1\left(f_{RD}^*\right)$,

$$\begin{aligned} \mathcal{H}_1(f_{RD}^*) = \Big\{ f_{RD}^* \in \Delta^{H_D^2 - 1} : \mathbb{E}\left[\vec{D}_i + \left(\vec{D}_i b'_{H_D} - A_{H_D} \right) \mathrm{vec}(f_{RD}^*) \right] = 0 \\ \text{and } f_R^* - M_{H_D} \mathrm{vec}(f_{RD}^*) = 0 \Big\}. \end{aligned} \tag{1.19}$$

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

$$Q(f_{md*}) = \left\| \mathbb{E}\left[\vec{D}_i + \left(\vec{D}_i b'_{H_D} - A_{H_D} \right) \mathrm{vec}(f_{RD}^*) \right] \right\|^2 \tag{1.20}$$

$$Q_n(f_{md*}) = \left\| \frac{1}{n} \sum_{i=1}^{n} \left[\vec{D}_i + \left(\vec{D}_i b'_{H_D} - A_{H_D} \right) \mathrm{vec}(f_{md*}) \right] \right\|^2 . \tag{1.21}$$

These criterion functions correspond to the more general form described in CHT with the weight matrix taken to be the identity. Note that the set of minimizers of (1.20) corresponds exactly with the identified set, $\mathcal{H}\left(f_{RD}^*\right)$, which is what inspires the set estimator

$$\widehat{\mathcal{H}}(f_{RD}^*, c) = \left\{ f_{RD}^* \in \Delta^{H_D^2 - 1} : Q_n(f_{RD}^*) \leq \frac{1}{n} c \right\}, \tag{1.22}$$

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*

$$\mathcal{C}_n = \sup_{f_{RD}^* \in \mathcal{H}(f_{RD}^*)} n Q_n(f_{RD}^*),$$

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 (1.22) 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 allows 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 6. 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).

1.5 Monte Carlo Simulations

In this section we present the results from Monte Carlo simulations of the various estimators presented in the previous two sections. These exercises illustrate some of the strengths and weaknesses of the proposed estimators in an environment where the latent distributions are completely controlled.

In all the models the true starting times follow a uniform distribution, $S^* \sim Unif(1, 10)$, implying individuals are born randomly between period 1 and 10. We restrict ourselves to a model of at most one linkage error per individual, $L = 1$, so that estimators under independence and dependence can be compared side by side. The joint distribution of the record linkage error timing and duration of interest, f^*_{RD}, follows what we are calling a Truncated Discretized Bivariate Normal (TDBN) distribution. The TDBN distribution converts a bivariate normal distribution into a finite discrete random variable by truncating

at the lower and upper bounds and collapsing the probability around each integer to a point mass at that integer. For example, if $X \sim TDBN(\mu, \Sigma, l, u)$ then X has support on $\{l, \ldots, u\}$ and the probability mass function of X is given by

$$P(X = k) = \mathbb{1}\{k \in \mathbb{Z}, l \leq k \leq u\} \frac{P(k - 0.5 \leq Y < k + 0.5)}{P(l - 0.5 \leq Y < u + 0.5)}$$

where Y is a bivariate normal distribution with $Y \sim N(\mu, \Sigma)$.

There are two parameterizations of this distribution in order to cover a case of dependence and case of independence[3]. The exact parameterizations are given by

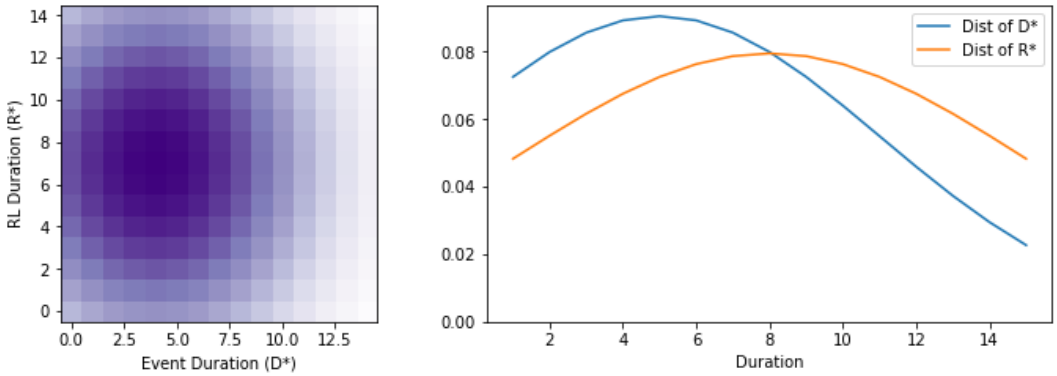
$$\textbf{Model 1 (Indep.)} : \begin{bmatrix} R^* \\ D^* \end{bmatrix} \sim TDBN \left(\mu = \begin{bmatrix} 8 \\ 5 \end{bmatrix}, \Sigma = \begin{bmatrix} 49 & 0 \\ 0 & 36 \end{bmatrix}, l = 1, u = 15 \right)$$

$$\textbf{Model 2 (Dep.)} : \begin{bmatrix} R^* \\ D^* \end{bmatrix} \sim TDBN \left(\mu = \begin{bmatrix} 8 \\ 5 \end{bmatrix}, \Sigma = \begin{bmatrix} 49 & 29.4 \\ 29.4 & 36 \end{bmatrix}, l = 1, u = 15 \right)$$

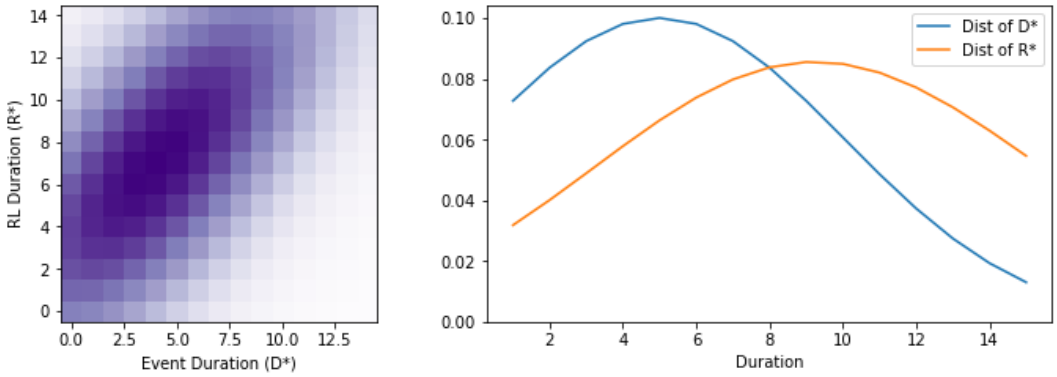
For reference, the correlation coefficient in Model 2 is $\rho = 0.7$. The primary difference between these two models is the dependence between the variables, however even just changing Σ across the models will result in them having slightly different marginal distributions. We tried to keep the rest of the distributional characteristics the same so as to maintain their comparability throughout the simulations.

Plots of the joint distribution and the marginal distributions can be found in Figure 1.4. The heat maps of the joint distributions on the left hand side are able to give some sense as to the frequency of duration breakage. If we imagine a diagonal line from the south west corner up to the north east corner, then any cells below the diagonal (i.e. $R^* < D^*$) will result in a break. With this in mind we can see that the independent model has more mass below the diagonal than the dependent model and as a result is subject to more linkage error.

[3]Note that when Σ is a diagonal matrix this does not necessarily imply that the TDBN variables are independent. To achieve a truly independent distribution we take the truncated and discretized marginal distributions of $N(\mu, \Sigma)$ and multiply them to produce a joint distribution that is close to a TDBN but which actually manifests independence between R^* and D^*.



(a) Joint Distribution and Marginal Distributions of Model 1 (Independence)



(b) Joint Distribution and Marginal Distributions of Model 2 (Dependence)

Figure 1.4: The above figures illustrate the joint distributions on the left and marginal distributions on the right of latent variables in the independent (1.4a) and dependent (1.4b) models.

For each of these two models we implement 3 different estimators of the marginal distribution, f_D^*. The first estimator, which we call the naive estimator and denote by $\widehat{\theta}_{naive}$, serves as the baseline since it represents an approach which ignores the potential linkage error in the sample. In this estimator we simply take the observed empirical distribution of D as an estimate for f_D^*, i.e. $\widehat{\theta}_{naive} \equiv \widehat{f_D}$. The second estimator, denoted $\widehat{\theta}_{indep}$, is the estimator presented in section 1.3.1 and defined in equation (1.9). As discussed earlier, this estimator assumes independence between D_i^* and R_i^*, uses only information from D_i, and requires knowledge of the marginal distribution of R^*. The third estimator, denoted $\widehat{\theta}_{dep}$, is the GMM estimator presented in section 1.3.2 and defined in equation (1.17). This estimator allows for unrestricted dependence between D^* and R^*, and uses information from both D_i and S_i.

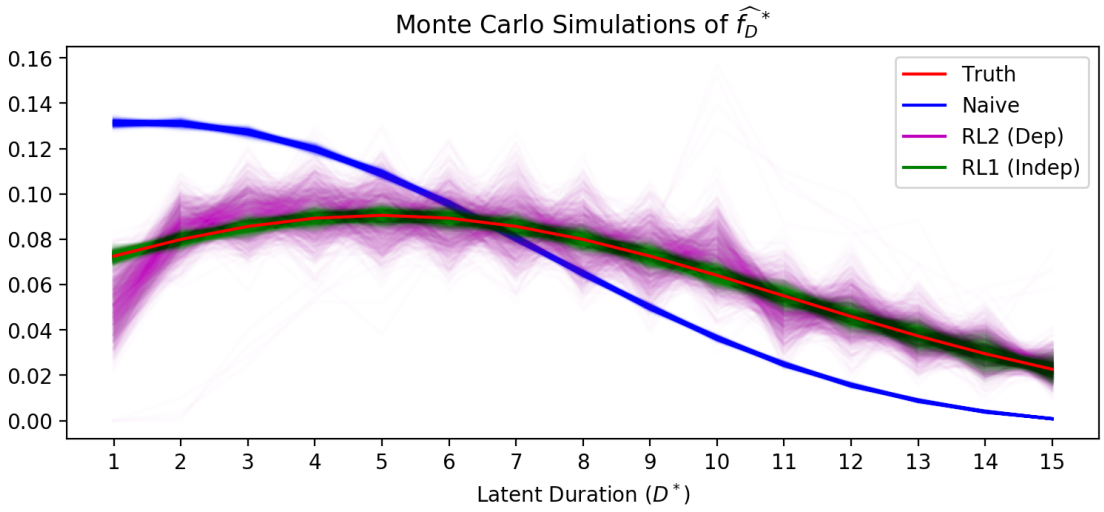
Each Monte Carlo simulation proceeds as follows. First we draw a latent sample of $n^* = 50000$ using the specified distribution of D^* and R^*. This latent sample of durations are then broken according to the model described in 1.2 to produce a sample of observed D and S which mimic the combined data set the researcher would have had they matched with linkage error (see appendix 1.9.2). Finally, we evaluate all three estimators and repeat this entire process for $s = 1000$ simulations.

The results of these Monte Carlo simulations are presented in Figure 1.5. The true marginal distribution, f_D^*, is plotted in red while the simulated estimates for $\widehat{\theta}_{naive}, \widehat{\theta}_{indep}$, and $\widehat{\theta}_{dep}$ are plotted in blue, purple, and green respectively. Focusing on panel (a), which corresponds with independent latent variables, we see can see the substantial error present in the naive estimator as it shifts mass from the right to the left. By ignoring the linkage error, the individual's would appear to have shorter lifetimes on average as compared with the truth. In contrast, both proposed estimators successfully account for the linkage error as they are centered around the truth. The second proposed estimator, $\widehat{\theta}_{dep}$, does exhibit more variance than the first, however this is to be expected since it is estimating the entire joint distribution and is thus subject to more variation with the larger dimension of the parameter space.

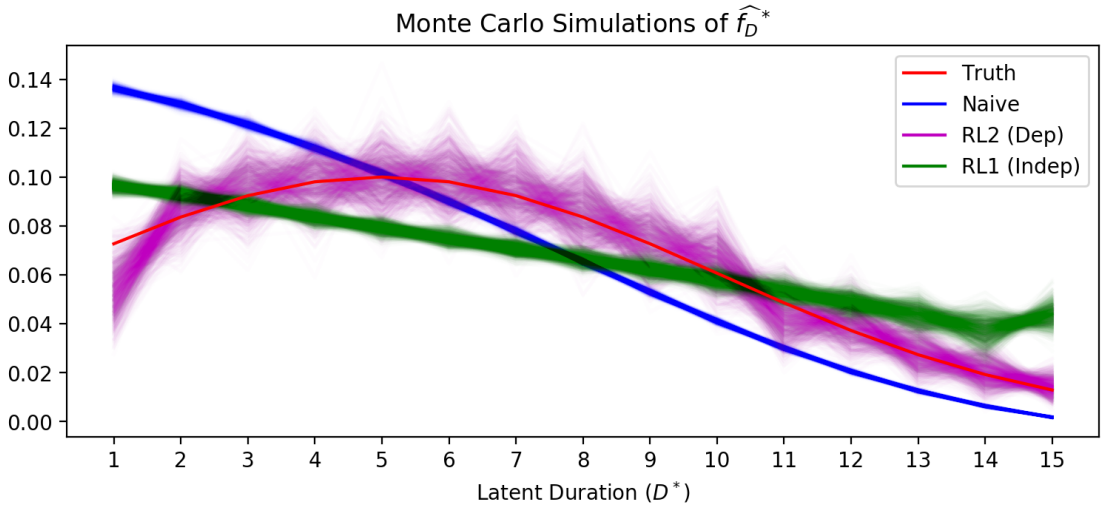
The improved accuracy of these estimators is also illustrated in the first column of Table 1.2a which presents the mean integrated squared error (MISE[4]) in the distribution estimators. There we see a substantial decrease in the MISE of the proposed estimators as compared to the naive estimator.

Turning our attention to Figure 1.5b, which presents the simulations under model 2, we see a slightly different story. Again we see that the naive estimator in blue is fairly off from the truth, but now even $\widehat{\theta}_{indep}$ departs from the true distribution. This result should not be terribly surprising given that the validity of $\widehat{\theta}_{indep}$ rests on the assumption of independence which is violated here in model 2. Despite this violation, the proposed

[4]For reference, the mean integrated squared error (MISE) is calculated as $MISE(\widehat{f_D^*}) = \sum_{k=1}^{H_D^*}(\widehat{f_D^*(k)} - f_D^*(k))^2 f_D^*(k)$



(a) Monte Carlo Simulations of Model 1 (Independence)



(b) Monte Carlo Simulations of Model 2 (Dependence)

Figure 1.5: The above figures plot the estimates of f_D^* for three different estimators () across two different models over $s = 1000$ simulations, each with a latent sample size of $n^* = 50000$. Panel (a) uses Model 1 (latent variables are independent) while panel (b) uses Model 2 (latent variables are positively correlated).

estimator still outperforms the naive estimator with respect to all statistics collected in Table 1.2b. Focusing on $\widehat{\theta}_{dep}$ (the purple estimates), we have a much more encouraging result since those distributions are centered around the truth. This should not be surprising given that this estimator does not impose any restriction on the dependence structure.

Since the marginal distribution of the lifetime durations may not be the primary object

(a) Model 1: Latent variables are independent

	f_D^*	$\mu_{D^*} = 6.762$		$P_5 = 0.418$		$P_1 = 0.073$	
Estimator	MISE $\left(\widehat{f_D^*}\right)$	$\mathbb{E}\left[\widehat{\mu}_{D^*}\right]$	MSE $\left(\widehat{\mu}_{D^*}\right)$	$\mathbb{E}\left[\widehat{P_5}\right]$	MSE $\left(\widehat{P_5}\right)$	$\mathbb{E}\left[\widehat{P_1}\right]$	MSE $\left(\widehat{P_1}\right)$
Naive ($\widehat{\theta}_{naive}$)	0.014	4.912	3.421	0.619	0.040	0.131	0.003
RL1 ($\widehat{\theta}_{indep}$)	0.000	6.763	0.001	0.418	0.000	0.073	0.000
RL2 ($\widehat{\theta}_{dep}$)	0.002	6.879	0.041	0.409	0.000	0.051	0.001

(b) Model 2: Latent variables are positively correlated

	f_D^*	$\mu_{D^*} = 6.378$		$P_5 = 0.447$		$P_1 = 0.073$	
Estimator	MISE $\left(\widehat{f_D^*}\right)$	$\mathbb{E}\left[\widehat{\mu}_{D^*}\right]$	MSE $\left(\widehat{\mu}_{D^*}\right)$	$\mathbb{E}\left[\widehat{P_5}\right]$	MSE $\left(\widehat{P_5}\right)$	$\mathbb{E}\left[\widehat{P_1}\right]$	MSE $\left(\widehat{P_1}\right)$
Naive ($\widehat{\theta}_{naive}$)	0.010	5.070	1.709	0.601	0.024	0.136	0.004
RL1 ($\widehat{\theta}_{indep}$)	0.005	6.817	0.194	0.440	0.000	0.096	0.001
RL2 ($\widehat{\theta}_{dep}$)	0.002	6.487	0.015	0.436	0.000	0.054	0.000

Table 1.2: Monte Carlo statistics from $s = 1000$ simulations with latent sample size $n^* = 50000$. Statistics include the mean integrated squared error (MISE) of the estimators of the marginal distribution, and the mean and mean squared error (MSE) of three univariate estimators: $\mu_{D^*} = \mathbb{E}\left[D_i^*\right]$, $P_5 = P\left(D_i^* \leq 5\right)$, and $P_1 = P\left(D_i^* = 1\right)$. Panel (a) corresponds with model 1 where latent variables are independent while panel (b) corresponds with model 2 where latent variables are positively correlated.

of interest, we also investigate several other statistical objects. Specifically, we estimate the mean lifetime duration, $\mu_{D^*} = \mathbb{E}\left[D_i^*\right]$, the probability of lifetimes ending in the first 5 periods, $P_5 = P\left(D_i^* \leq 5\right)$, and the probability of lifetimes of length 1, $P_1 = P\left(D_i^* = 1\right)$. For each of these estimands we use each of the estimators of f_D^* to produce estimates for each of $\mu_{D^*}^*, P_5$ and P_1.

Monte Carlo statistics related to these estimators are found in Table 1.2 in the last 6 columns. For each estimator we report the truth (in the top row), the estimated mean of the estimator, and the estimated mean squared error. Across both models and both proposed estimators we see substantial improvements with respect to our baseline naive estimators. Under independence (model 1 in Table 1.2a) the estimators for μ_{D^*} and P_5 are spot on. With respect to the estimation of P_1 the first proposed estimator, $\widehat{\theta_{indep}}$, has functionally no bias, while $\widehat{\theta_{dep}}$ does exhibit some downward bias though still less than the naive estimator. This

downward bias in the estimation of P_1 using $\widehat{\theta_{dep}}$ recurs in both models and is also evident in Figure 1.5. One possible explanation for this behavior is that it comes from the fact that identification and estimation essentially start at the right hand tail and finish at the left hand tail. Under this theory, error in the estimation of right tail probabilities can propagate into left tail probabilities, however, this would not necessarily explain the consistent downward bias. Further exploration of this issue is definitely in order.

1.6 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 (see Kearns (2006)). Due to the lack of unique identifiers, to create these panels the linkage algorithms must lean heavily on name and address matching. When

establishments relocate, this can lead to the type of linkage error discussed throughout this chapter. 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, e.g. 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. Given this shortcoming, the raw YTS data under reports relocation when it is determined by changes in zip codes across a establishment's history (since they may relocate within the same zip code). In order to more faithfully represent the rate of relocation we first flag those establishments that do not change zip codes, and generate an additional indicator meant to represent a relocation that is unseen. The distribution of these additional artificial relocation events is taken to match the distribution we observe in the zip code changes. We inject enough additional relocation events so that the aggregate number of individuals which relocate at some point during their

lifetime matches what we observe in finer data where we see relocation rates of between 15%-17%.

Some sample statistics of the raw YTS data truth are presented in the tables found in 1.3. Of particular importance in these descriptive statistics is to note that not only does the data display fairly significant association between firm lifetimes and time to relocation, there is even a slight association with the birth period. These details are pivotal since they inform which of the proposed estimators are appropriate. The other statistic of note is $P(R^* < D^*)$ which tells us that 16.4% of the firms in our sample will be subject to a record linkage error event since they relocate before the end of their lifetime. While this level of linkage error is non-trivial, it is important to remind the reader again that part of this is injected artificially, as described earlier, to match relocation rates seen in data with finer geolocation information.

$E[D^*]$	5.053
$P(D^* \leq 3)$	0.416
$P(R^* < D^*)$	0.164

Covariances

	D^*	R^*	S^*
D^*	10.187	8.695	-0.037
R^*	8.695	9.751	0.037
S^*	-0.037	0.037	2.301

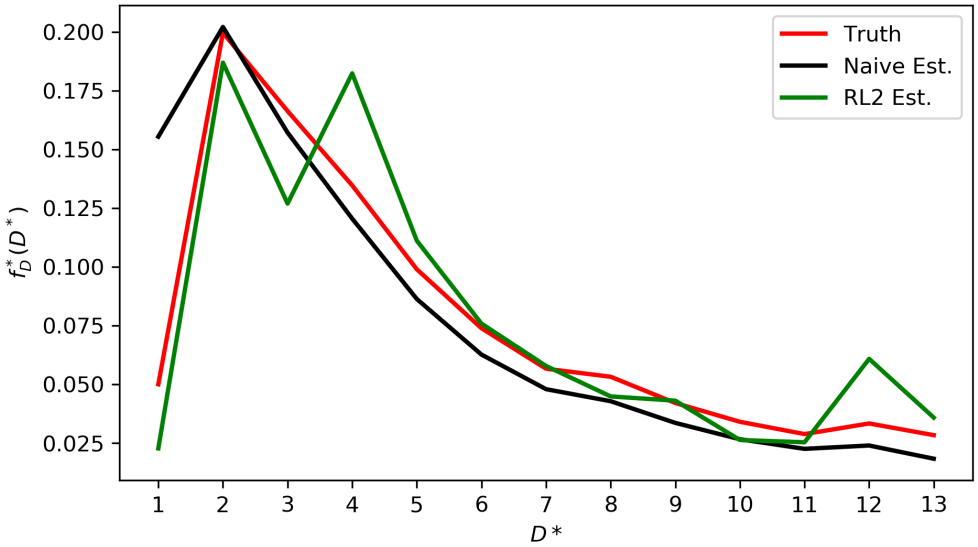
Table 1.3: Descriptive statistics of the unbroken YTS.

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 1.9.2, 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, $\widehat{f}_{Naive}$, which simply takes the observed

broken durations at face value. The second estimator, $\widehat{f}_{RL1}$, which is defined in (1.9) only uses the observed durations, marginal distribution, f_R^*, and assumes independence of the relocation timing. The third estimator, $\widehat{f}_{RL2}$, defined in (1.17) uses observed starts and durations and assumes independence of the latent start times.

(a) **Estimation of f_D^* on Broken Data**



(b) **Estimation of f_D^* on Broken and Permuted Data**

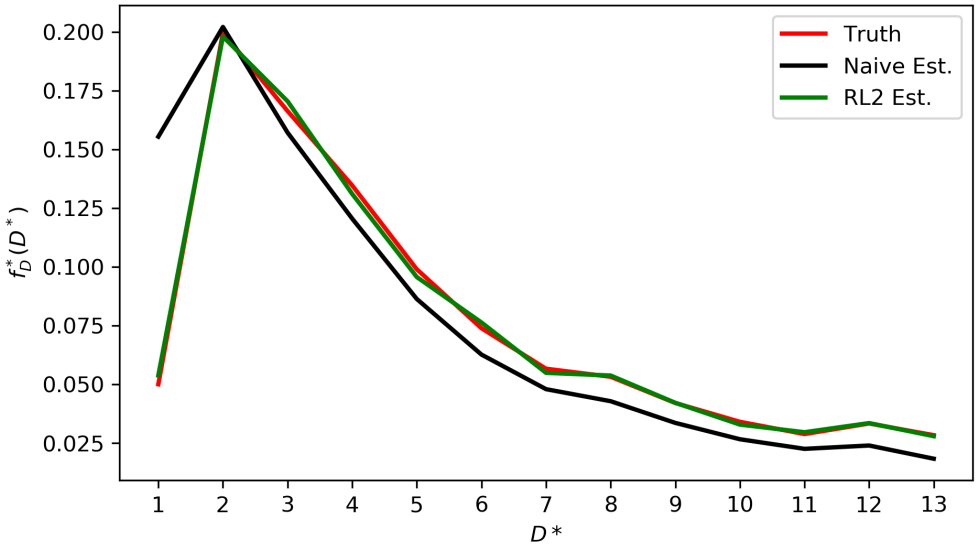


Figure 1.6: Results of estimation of the marginal distribution of firm lifetimes using synthetic empirical data that is (a) broken and (b) broken with permuted starting times.

In Figure 1.6 we display the point identified estimation results across both of the models

described above. In particular, we plot the true distribution of f_D^* along with the naive estimate of f_D^* and the second estimate, $\widehat{f}_{RL2}$. Results of the first estimate, $\widehat{f}_{RL1}$, are omitted since their underlying assumption, mainly independence of death and relocation time, is violated as illustrated by Table 1.3. Just as we saw with the Monte Carlo simulations in the previous section, these plots illustrate both the bias in the naive estimator (especially with respect to estimation of $P(D_i^* = 1)$) and how the proposed estimator yields a more accurate estimate of the likelihood of death for younger firms. We also again see how the proposed estimator can behave somewhat erratically on data where the independence of start times does not strictly hold, but behaves flawlessly in the data where we enforced this assumption by permuting the starting times.

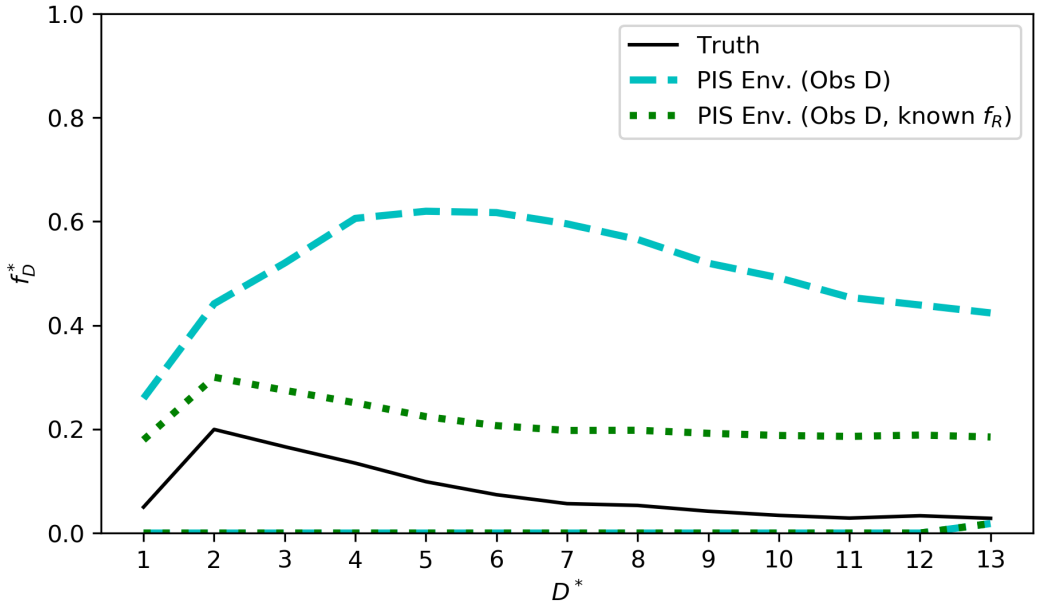
In Table 1.4 we present point estimates and statistics illustrating the fit of the marginal distribution estimates across the data sets discussed. The first column contains point estimates of the probability of death in the first 3 years, $P(D_i^* \leq 3)$, whose true value is 0.416. Both proposed estimators have smaller bias x 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.

	$P(D^* \leq 3)$		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 1.4: Estimates of Young Establishment Death Rates and Fit of $\widehat{f_D^*}$.

Finally, we present some of the estimates of various partially identified sets in Figure 1.7. We plot both the outer envelope of the partially identified set of the marginal distribution of establishment lifetimes, i.e. all the sharp bounds of probability estimates. In the second figure we plot partially identified sets two univariate statistics: the mean lifetime and the probability of death in the first 3 years. Across both plots we show how the partially

(a) **Outer Envelope of Partially Identified Set of f_D^***



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

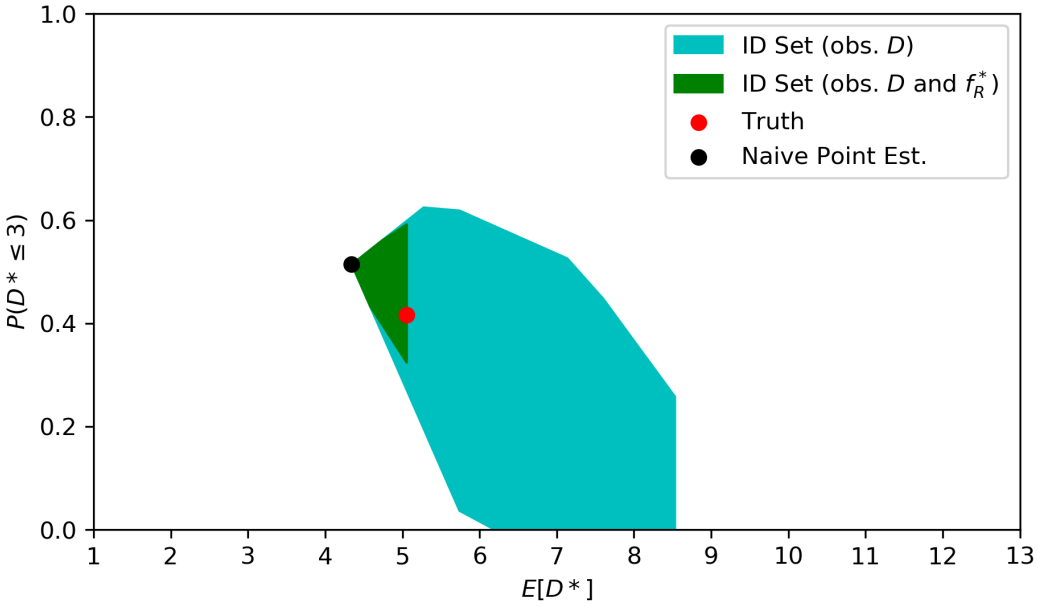


Figure 1.7: Estimates of partially identified sets using the synthetic empirical data. Figure a plots the outer envelope of the estimate of the partially identified sets of the marginal distribution of firm lifetimes and Figure b plots estimates of the partially identified set of two univariate statistics. Both plots display estimates with and without incorporating knowledge of the marginal distribution of relocation.

identified sets shrink as we incorporate more information, e.g. the marginal distribution of establishment relocation, into our estimates of the partially identified sets. Figure 1.7b is especially effective at depicting how adding extra information on the marginal distribution can greatly shrink the partially identified set from the cyan region to the much smaller green region.

1.7 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 substantially overestimated.

1.8 References

Abbring, J. H. and van den Berg, G. J. (2003). The identifiability of the mixed proportional hazards competing risks model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):701–710.

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.

Andrews, D. W. K. (2002). Generalized method of moments estimation when a parameter is on a boundary. *Journal of Business & Economic Statistics*, 20(4):530–544.

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.

Benedetto, G., Haltiwanger, J., Lane, J., and McKinney, K. (2007). Using worker flows to measure firm dynamics. *Journal of Business & Economic Statistics*, 25(3):299–313.

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 Honoré, 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.

Hwang, J. (2021). Simple and trustworthy cluster-robust gmm inference. *Journal of Econometrics*, 222(2):993–1023.

Kearns, C. C. (2006). State implementation of the SUTA dumping prevention act of 2004. *State and Local Tax Lawyer*, 11:105.

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.

1.9 Appendix

1.9.1 Notation

D^*	True duration until the event of interest
R^*	True duration until the splitting event
D	Duration observed by the econometrician
L	Maximum number of breaks permitted; $L = dim(R^*)$
B^*	Number of breaks (resulting in $B^* + 1$ constituent durations);
n^*	Number of latent (unobserved) individuals
n_b^*	Number of latent individuals with b breaks (i.e. $b + 1$ broken durations)
n	Number of observed individuals (note that $n = \sum_{b=0}^{L} n_b^*(b+1) \geq n^*$)
H_D^*	Maximum duration until event of interest; $\max supp(D^*)$
H_R^*	Maximum duration until split event; $\max supp(R^*)$
H_D	Maximum observed duration; $\max supp(D)$
$f_D^*(i)$	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; $P(D = i)$
$f_{RD}^*(i,j)$	Joint probability of splitting duration and event duration; $P(R^* = i, D^* = j)$

Table 1.5: Notation

A table that describes each of the variables and important distributional functions is presented above.

A number of additional bits of notation that appear through the text are also important to define. Let $[\cdot]_L$ and $[\cdot]_U$ denote the lower and upper triangular matrix transformations which generate lower and upper triangular matrices respectively with the same dimensions as their arguments. The lower triangular transformation zeros all elements above the diagonal while the upper triangular transformation zeros all elements on the diagonal and below. For example, if

$$X = \begin{bmatrix} a & b & c & d \\ e & f & g & h \\ i & j & k & l \end{bmatrix} \quad \text{then} \quad [X]_L = \begin{bmatrix} a & 0 & 0 & 0 \\ e & f & 0 & 0 \\ i & j & k & 0 \end{bmatrix} \quad \text{and} \quad [X]_U = \begin{bmatrix} 0 & b & c & d \\ 0 & 0 & g & h \\ 0 & 0 & 0 & l \end{bmatrix}.$$

Though triangular matrices are typically square, the above operators are still defined on

non-square matrices. To be perfectly rigorous their exact definitions are given by $\left([X]_L\right)_{ij} = \mathbb{1}\{i \geq j\}X_{ij}$ and $\left([X]_U\right)_{ij} = \mathbb{1}\{i < j\}X_{ij}$. Note that for any X we have $X = [X]_L + [X]_U$.

The function, $[\cdot]_Q$, takes a matrix as an argument and returns a matrix of the same dimensions which has shifted over the upper triangular portion of the matrix (ignoring the diagonal) and zeroed out the rest. For example, if

$$X = \begin{bmatrix} a & b & c & d \\ e & f & g & h \\ i & j & k & l \end{bmatrix} \quad \text{then} \quad [X]_Q = \begin{bmatrix} b & c & d & 0 \\ g & h & 0 & 0 \\ l & 0 & 0 & 0 \end{bmatrix}.$$

If X is a matrix with n columns then the exact definition of this function is given by $\left([X]_Q\right)_{ij} = \mathbb{1}\{i+j \leq n\}X_{i,j+i}$. Note that this function is invertible if and only if the domain is restricted to upper triangular matrices with zeros on the diagonal. This is particularly relevant for our purposes since we usually have $[[X]_U]_Q$ as in equation (1.12). Finally, because this transformation ignores the diagonal, the resulting matrix will always have zeros in the final column.

1.9.2 Linkage Error Model

The motivation for this work comes from handling record linkage error when multiple datasets need to be matched. However, there has been little discussion of combining data sets and matching variables in the main text. In this section I dive deeper into the data microfoundations, and describe a record linkage model that underlies the duration model transformation described in section 1.2. The subsequent model and assumptions discussed here are meant to help researchers decide whether the model presented in the main text, and by extension the proposed estimators, are appropriate for their empirical situation.

Let there be τ ordered data sets, indexed by $t \in \mathcal{T} = \{1, \ldots, \tau\}$, with n_t individuals in each data set, and let each data set contain variables across their n_t individuals. For example, data set t may contain the answers collected from the long-form questionnaire of the U.S. Census Bureau in time period t, where $t \in \{1970, 1980, 1990, 2000, 2010\}$. Each data set

may contain both variables that are of primary interest to the researcher, i.e. earnings and address, and variables important for matching across data set, i.e. SSNs or family names. 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, $ID_t : \mathcal{N}_t \to \mathcal{I}$,

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

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. We will on occasion need to use the inverse of this function, $ID_t^{-1} : \mathcal{I} \to \mathcal{N}_t \cup \{0\}$, which maps individual i to their index in file t and maps to 0 if individual i does not appear anywhere in file t.

The analytic goal of the researcher requires matching individuals across the data sets to create a larger more comprehensive data file, e.g. a panel with individual histories or a larger cross section with more variables. To match between two data files the researcher compares pairs of matching variables for potential individuals, i.e. comparing the SSNs between an individual in the 1990 census with one in the 2000 census. Now consider a K-dimensional vector, $X_{ti} = \left[X_{ti}^{(1)}, X_{ti}^{(2)}, \ldots, X_{ti}^{(K)}\right]$, representing K 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 s 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_U is a sufficient matcher $\Rightarrow$ M_W is a sufficient 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. Conclusion 2 in the lemma is the contrapositive of this logical statement, so it is thus proven. ■

Now we consider sets of covariates that have a very specific relationship that is pivotal to the model discussed in this work.

Definition 1. Let the pair of disjoint sets of matching variables, U_1 and U_2, be called *sufficient necessary switchers* (SNS) if

1. $M_{U_1 \cup U_2}$ is a sufficient but not necessary matcher.

2. M_{U_1} is a necessary but not sufficient matcher.

In words, the variables in U_2 help the sufficiency of a matcher, i.e. when they match they can contribute more evidence to the hypothesis that these records come from the same individual. However, these variables are not required to match, meaning that they may lead to a missed link if the records for an individual change over time. 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.

Finally, to ensure that any individual history has at most L breaks we must limit how often the matching variables can change over a given history. If $U \subseteq \{1, \ldots, K\}$ is a subset of size k_U of the matching variables, let X_{it}^U denote the $k_U \times 1$ vector of matching variables in set U for individual i at time t.

Definition 2. Let $U \subseteq \{1, \ldots, K\}$ have the *L-change* property if for any given individual i, the history of matching variables, $\{X_{it}\}_{t=1}^{\tau}$, takes at most L distinct values.

To be clear it is important to note that $X^U_{it_1}$ and $X^U_{it_2}$ are distinct if they differ in any of the k_U variables. Every period in which these matching variables differ from the previous period will trigger a record linkage error event and these changes will determine the distribution of R^*_i. For example, suppose individual i is born in period t_0 and has the following history of matching variables

$$X^U_{it_0} = \begin{bmatrix} A \\ P \\ 212 \end{bmatrix} \quad X^U_{it_0+1} = \begin{bmatrix} B \\ P \\ 212 \end{bmatrix} \quad X^U_{it_0+2} = \begin{bmatrix} B \\ P \\ 212 \end{bmatrix} \quad X^U_{it_0+3} = \begin{bmatrix} B \\ P \\ 212 \end{bmatrix} \quad X^U_{it_0+4} = \begin{bmatrix} B \\ P \\ 400 \end{bmatrix}$$

and they are used for matching across time. Since the matching variables change between period t_0 and $t_0 + 1$ as well as between $t_0 + 3$ and $t_0 + 4$ this individual will not be linked across those periods, resulting in 2 linkage errors. Moreover, this specific scenario would correspond to $R^*_i = (1, 3)$ since the first linkage error happened 1 period after birth and the second error happened 3 periods after the first error.

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 1.2 and thus be germane to the results of this chapter.

Theorem 9. *Let U_1 and U_2 be sets of covariates that are sufficient necessary switchers, and suppose U_2 has the L-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 1.2) with the distribution of R^* coming from changes in the covariates of U_2 over time.*

1.9.3 Construction of Auxiliary Objects

1.9.3.1 Construction of A_{R^*}

One vitally important object needed to identify and estimate the latent duration distribution as discussed in section 1.3.1 is A_{R^*}. In this section I precisely define this matrix and how it can be constructed from the marginal distribution of R^*, $f^*_R(r^*)$. The A_{R^*} matrix is $H_D \times H^*_D$,

and the cell in the ith row and jth column is given by

$$(A_{R^*})_{ij} \equiv \mathbb{E}\left[\omega\left(i, j, R^*\right)\right] = \sum_{r^* \in \mathfrak{R}^*} \omega\left(i, j, r^*\right) f_R^*(r^*) \tag{1.23}$$

where $\omega(\cdot)$ is a weighting function. More specifically, $\omega\left(d, d^*, r^*\right)$ is a non-negative integer valued function which counts the number of length d durations that occur when a d^* length duration is broken up by the r^* record linkage error events according to the transformation described in 1.2. Taken all together, $(A_{R^*})_{ij}$ is the expected number of durations of length i when the true underlying duration is $D^* = j$.

Examples often aid in clarity, so if $i = 3$ and $j = 10$ and $r^* = (1, 3, 2, 1, 5)$, then the observed durations would be of lengths 1, 3, 2, 1, and 3. Since the duration of length $i = 3$ occurs 2 times, then $\omega\left(3, 10, (1, 3, 2, 1, 5)\right) = 2$.

Writing out the matrix, A_{R^*}, in its entirety and in its general form is laborious and further encumbered by the fact that $\omega(\cdot)$ does not have a closed formula. However, intuition for its structure can still be developed through another concrete example. Suppose that D^* has a maximum duration of $H_D^* = 3$ and that at most two linkage error events can happen in an individual's lifetime, $R^* = (R_1^*, R_2^*)$. In this scenario a latent distribution can be broken into at most three pieces, meaning $\omega\left(d, d^*, r^*\right) \in \{0, 1, 2, 3\}$. For a given marginal distribution, $f_R^*(r_1^*, r_2^*)$, the full matrix is then given by

$$A_{R^*} = \begin{bmatrix} \sum\limits_{r_1^* \geq 1, r_2^*} f_R^*(r_1^*, r_2^*) & 2\sum\limits_{r_2^*} f_R^*(1, r_2^*) & 3f_R^*(1,1) + \sum\limits_{r_2^* \geq 2} f_R^*(1, r_2^*) + \sum\limits_{r_2^*} f_R^*(2, r_2^*) \\ 0 & \sum\limits_{r_1^* \geq 2, r_2^*} f_R^*(r_1^*, r_2^*) & \sum\limits_{r_2^* \geq 2} f_R^*(1, r_2^*) + \sum\limits_{r_2^*} f_R^*(2, r_2^*) \\ 0 & 0 & \sum\limits_{r_1^* \geq 3, r_2^*} f_R^*(r_1^*, r_2^*) \end{bmatrix}.$$

Below are formal proofs of the properties described in 1.3.1

Proof that A_{R^*} is upper triangular. Let $i > j$ and note that $\omega\left(i, j, r^*\right) = 0$ for all $r^* \in \mathfrak{R}^*$ because if $D^* = j$ then there will never be any observed durations of length $i > j$. It follows from this and equation (1.23) that $(A_{R^*})_{ij} = 0$, and thus A_{R^*} is upper triangular. ■

1.9.3.2 Construction of $\Omega, A_1,$ and A_2

This section describes Ω which is a core part of the asymptotic variance of the estimators described in section 1.3.1. For ease we recall the definition of Ω and note that it can be rewritten as a sum with deterministic, rather than random, limits.

$$\Omega \equiv Var\left(\sum_{b=1}^{B_i^*+1} \overrightarrow{D_{ib}^*}\right) = Var\left(\sum_{b=1}^{L+1} \mathbb{1}\left\{b \leq B_i^* + 1\right\} \overrightarrow{D_{ib}^*}\right).$$

As mentioned in the main text, the sum of random variables above essentially counts the number of durations of various lengths that result from a potentially broken true duration. In other words, the kth element of this random vector is equal to the number of durations of length k that result from individual i's true lifetime being broken by the record linkage process. By inspection it should be clear that the underlying joint distribution of D_i^* and R_i^* entirely determine Ω. To aid in the construction of Ω, and by extension the construction of $\widehat{\Omega}$ found in (1.11), we now make that connection concrete.

Let us focus on the kth element of the random vector inside the variance. Using the definition of $\overrightarrow{\cdot}$ we have

$$\left(\sum_{b=1}^{L+1} \mathbb{1}\left\{b \leq B_i^* + 1\right\} \overrightarrow{D_{ib}^*}\right)_k = \sum_{b=1}^{L+1} \mathbb{1}\left\{b \leq B_i^* + 1\right\} \mathbb{1}\left\{D_{ib}^* = k\right\}.$$

The Ω matrix simply consists of all the pairwise covariances of these random variables, and the fact that they are all sums of indicator functions means we can write these covariances

as weighted sums of joint probabilities. In particular consider the (k,l)th cell of Ω

$$\begin{aligned}
(\Omega)_{kl} &= Cov\left(\sum_{b=1}^{L+1}\mathbb{1}\left\{b \leq B_i^* + 1, D_{ib}^* = k\right\}, \sum_{c=1}^{L+1}\mathbb{1}\left\{c \leq B_i^* + 1, D_{ic}^* = l\right\}\right) \\
&= \sum_{b=1}^{L+1}\sum_{c=1}^{L+1} Cov\left(\mathbb{1}\left\{b \leq B_i^* + 1, D_{ib}^* = k\right\}, \mathbb{1}\left\{c \leq B_i^* + 1, D_{ic}^* = l\right\}\right) \\
&= \sum_{b=1}^{L+1}\sum_{c=1}^{L+1} \left[\mathbb{E}\left[\mathbb{1}\left\{b \leq B_i^* + 1, D_{ib}^* = k\right\}\mathbb{1}\left\{c \leq B_i^* + 1, D_{ic}^* = l\right\}\right]\right. \\
&\qquad\qquad \left. - \mathbb{E}\left[\mathbb{1}\left\{b \leq B_i^* + 1, D_{ib}^* = k\right\}\right]\mathbb{E}\left[\mathbb{1}\left\{c \leq B_i^* + 1, D_{ic}^* = l\right\}\right]\right] \\
&= \sum_{b=1}^{L+1}\sum_{c=1}^{L+1} \left[P\left(D_{ib}^* = k, D_{ic}^* = l\right) - P\left(D_{ib}^* = k\right)P\left(D_{ic}^* = l\right)\right] \\
&= \sum_{b=1}^{L+1}\sum_{c=1}^{L+1} \left[a_2(b,k,c,l)'\text{vec}(f_{RD}^*) - a_1(b,k)'\text{vec}(f_{RD}^*)a_1(c,l)'\text{vec}(f_{RD}^*)\right].
\end{aligned}$$

Since the probability of each event in the second to last line is simply the sum of the latent events that result in that event, we can replace the probabilities with a dot product between the vector form of the entire joint distribution, $\text{vec}(f_{RD}^*)$, and binary weighting vectors indexed by the events. More specifically, a_1 and a_2 are vector valued functions,

$$\begin{aligned}
a_1 &: \{1,2,\ldots,Ł+1\} \times \{1,2,\ldots,H_D\} \mapsto \{0,1\}^{H_{RD}\times 1} \\
a_2 &: \{1,2,\ldots,Ł+1\}^2 \times \{1,2,\ldots,H_D\}^2 \mapsto \{0,1\}^{H_{RD}\times 1}
\end{aligned}$$

which sum up the the relevant probabilities for a specific event. For example, if we consider $P(D_{i3}^* = 5)$, which is the probability that the third broken part of a duration has length 5, this event occurs under several latent events including the following:

$$\begin{aligned}
(D_i^* = 15, R_i^* = (4,2,5,6)) &\qquad \text{which produces breaks of length 4, 2, 5, and 4} \\
(D_i^* = 9, R_i^* = (1,3,7,5)) &\qquad \text{which produces breaks of length 1, 3, and 5} \\
(D_i^* = 12, R_i^* = (3,2,5,4)) &\qquad \text{which produces breaks of length 3, 2, 5, and 2}
\end{aligned}$$

Thus the index of $a_1(3,5)$ corresponding to each of these events in $\text{vec}(f_{RD}^*)$ will have a 1 (and so will the index for any other event that results in the the 3rd break being length 5). Since there are many ways to formulate the multidimensional joint distribution of f_{RD}^* into

the one-dimensional vector, $\text{vec}(f^*_{RD})$, it should be clear that the exact structure of a_1 and a_2 will depend on this ordering. The important point is that a_1 and a_2 are constant binary vectors that do not depend on the latent distribution outside of its dimension.

$$\begin{aligned}
(a_1(b,k))_i &= \mathbb{1}\left\{\text{vec}\left(f^*_{RD}\right)_i \in \mathcal{W}(b,k)\right\} \\
\text{where } \mathcal{W}(b,k) &= \left\{(D^*_i, R^*_i) \in \mathfrak{D}^* \times \mathfrak{R}^* : \left(D^*_i \geq \sum_{l=1}^{b} R^*_{il} \text{ and } R^*_{ib} = k\right) \text{ or } \right. \\
&\qquad \left. \left(D^*_i < \sum_{l=1}^{b} R^*_{il} \text{ and } D^*_i - \sum_{l=1}^{b-1} R^*_{il} = k\right)\right\}
\end{aligned}$$

The final expression of $(\Omega)_{kl}$ can be further manipulated and collapsed in the name of compactness to arrive at the following formulation,

$$\begin{aligned}
(\Omega)_{kl} &= \sum_{b=1}^{L+1}\sum_{c=1}^{L+1}\left[a_2(b,k,c,l)'\text{vec}(f^*_{RD}) - a_1(b,k)'\text{vec}(f^*_{RD})a_1(c,l)'\text{vec}(f^*_{RD})\right] \\
&= \left(\sum_{b=1}^{L+1}\sum_{c=1}^{L+1} a_2(b,k,c,l)'\right)\text{vec}(f^*_{RD}) - \text{vec}(f^*_{RD})'\left(\sum_{b=1}^{L+1}\sum_{c=1}^{L+1} a_1(b,k)a_1(c,l)'\right)\text{vec}(f^*_{RD}) \\
&= A_2(k,l)'\text{vec}(f^*_{RD}) - \text{vec}(f^*_{RD})'A_1(k,l)\text{vec}(f^*_{RD})
\end{aligned}$$

where $A_2(k,l)$ is a $H_{RD} \times 1$ vector and $A_1(k,l)$ is a $H_{RD} \times H_{RD}$ matrix with each defined according to the parenthesis in the preceding expression. Both of these are again constant, and do not depend on the joint distribution aside from its dimensionality.

One final note to aid any researcher who aims to construct $A_2(k,l)$ is that its constituent part, $a_2(b,k,c,l)$, can be derived from $a_1(b,k)$. In particular, it can be shown that

$$a_2(b,k,c,l) = a_1(b,k) \circ a_1(c,1)$$

where $\circ$ denotes the Hadamard product (i.e. element-wise multiplication).

1.9.4 Proofs

1.9.4.1 Properties of D and D^*

Proof of Proposition 1 (Mean Attenuation). Let all variables be as defined in Section 1.2.3. Starting with equation (1.4) the expectation of observed durations is

$$\mathbb{E}[D] = \sum_{k=1}^{H_D} k f_D(k) = \frac{1}{\lambda} \sum_{k=1}^{H_D} k \left[\sum_{r \in \mathcal{I}_u(k)} f^*_{RD}(r,k) + \sum_{(r,d) \in \mathcal{I}_m(k)} f^*_{RD}(r,d) + \sum_{(r,d) \in \mathcal{I}_e(k)} f^*_{RD}(r,d) \right].$$

Since all the inner summands are over the latent distribution, f^*_{RD}, we can combine and rewrite them as a single weighted sum,

$$\mathbb{E}[D] = \frac{1}{\lambda} \sum_{k=1}^{H_D} k \left[\sum_{(r,d) \in \mathfrak{R}^* \times \mathfrak{D}} h(k,r,d) f^*_{RD}(r,d) \right],$$

where $h(k,r,d)$ simply counts the number of observed durations of length k that result from latent event (r,d). We can swap the order of the sums and note that $\sum kh(k,r,d)$ is simply adding up all the lengths of the constituent durations resulting from latent event (r,d) which will always sum to d. This leaves us with our result

$$\mathbb{E}[D] = \frac{1}{\lambda} \sum_{(r,d) \in \mathfrak{R}^* \times \mathfrak{D}} f^*_{RD}(r,d) \sum_{k=1}^{H_D} kh(k,r,d) = \frac{1}{\lambda} \sum_{(r,d) \in \mathfrak{R}^* \times \mathfrak{D}} d f^*_{RD}(r,d) = \frac{1}{\lambda} \mathbb{E}[D^*].$$

■

Proof of Proposition 2 (Stochastic Dominance). For the first part, let $L = 1$ and $H^*_D = 3$. Using equation (1.2) and the fact that $\lambda = 1 + f^*_{RD}(1,2) + f^*_{RD}(1,3) + f^*_{RD}(2,3)$ we have

$$\begin{aligned} P(D_i = 1) - P(D^*_i = 1) &= \frac{1}{\lambda} \left[\sum_{r=1}^{\infty} f^*_{RD}(r,1) + \sum_{d=2}^{\infty} f^*_{RD}(1,d) + \sum_{d-r=1} f^*_{RD}(r,d) \right] - P(D^*_i = 1) \\ &= \frac{1}{\lambda} \left[P(D^*_i = 1) + f^*_{RD}(1,2) + \lambda - 1 \right] - P(D^*_i = 1) \\ &= \frac{1}{\lambda} \left[(\lambda - 1)\left(1 - P(D^*_i = 1)\right) + f^*_{RD}(1,2) \right] \end{aligned}$$

Since $\lambda \geq 1$ then everything on the right hand side is positive and we can conclude that

$P(D_i = 1) \geq P(D_i^* = 1)$. Similar algebraic manipulation yields

$$\begin{aligned} P(D_i \leq 2) - P(D_i^* \leq 2) &= P(D_i = 1) + P(D_i = 2) - P(D_i^* = 1) - P(D_i^* = 2) \\ &= \frac{1}{\lambda}\left[P(D_i^* = 1) + f_{RD}^*(1,2) + \lambda - 1\right] + \frac{1}{\lambda}\left[P(D_i^* = 2) - f_{RD}^*(1,2) + \right. \\ &\qquad \left. f_{RD}^*(2,3) + f_{RD}^*(1,3)\right] - P(D_i^* = 1) - P(D_i^* = 2) \\ &= \frac{\lambda - 1}{\lambda}\left[1 - P(D_i^* \leq 2)\right] + \frac{1}{\lambda}\left[f_{RD}^*(2,3) + f_{RD}^*(1,3)\right] \end{aligned}$$

Once again everything on the right hand side is positive, implying that $P(D_i \leq 2) \geq P(D_i^* \leq 2)$. Thus $P(D_i \leq d) \geq P(D_i^* \leq d)$ for all d meaning that D^* first order stochastically dominates D. An analagous approach can be used to show this also holds for $H_D^* = 2$, and the case of $H_D^* = 1$ is degenerate.

To prove the second part of proposition 2 we consider $H_D^* = 4$ and focus on the relationship between $P(D_i = 1)$ and $P(D_i^* = 1)$.

$$\begin{aligned} P(D_i = 1) - P(D_i^* = 1) &= \frac{1}{\lambda}\left[\sum_{r=1}^{\infty} f_{RD}^*(r,1) + \sum_{d=2}^{\infty} f_{RD}^*(1,d) + \sum_{d-r=1} f_{RD}^*(r,d)\right] - P(D_i^* = 1) \\ &= \frac{1}{\lambda}\left[P(D_i^* = 1) + f_{RD}^*(1,2) + \lambda - 1 - f_{RD}^*(2,4)\right] - P(D_i^* = 1) \\ &= \frac{1}{\lambda}\left[(\lambda - 1)\left(1 - P(D_i^* = 1)\right) + f_{RD}^*(1,2) - f_{RD}^*(2,4)\right] \end{aligned}$$

If $f_{RD}^*(2,4) = 0$ then the right hand side is positive, however if $f_{RD}^*(2,4)$ is large enough then the right hand side can be negative. Unlike the previous case, here we cannot order $P(D_i = 1)$ and $P(D_i^* = 1)$ without more assumptions, which implies that D and D^* cannot be ordered using first order stochastic dominance. For cases where $H_D^* > 4$ the degrees of freedom (i.e. terms with positive and negative signs) on the right hand side expand further and present the same problem. ■

1.9.4.2 Under Independent Linkage Error

Proof of Theorem 1 (Identification Under Independence). Under independence the observed duration distribution, f_D, can be written as a *nearly* linear function of the true distribution

of the duration of interest, f_D^*, as $f_D = \frac{1}{\lambda} A_{R^*} f_D^*$ where

$$f_D = \begin{bmatrix} f_D(1) \\ f_D(2) \\ f_D(3) \\ \vdots \\ f_D(H_D - 1) \\ f_D(H_D) \end{bmatrix} \qquad f_D^* = \begin{bmatrix} f_D^*(1) \\ f_D^*(2) \\ f_D^*(3) \\ \vdots \\ f_D^*(H_D - 1) \\ f_D^*(H_D) \end{bmatrix} \tag{1.24}$$

$$A_{R^*} = \begin{bmatrix} \sum_{i=1}^{H_D} f_R^*(i) & f_R^*(1) + f_R^*(1) & \cdots & f_R^*(1) + f_R^*(H_D - 2) & f_R^*(1) + f_R^*(H_D - 1) \\ 0 & \sum_{i=2}^{H_D} f_R^*(i) & \cdots & f_R^*(2) + f_R^*(H_D - 3) & f_R^*(2) + f_R^*(H_D - 2) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \sum_{i=H_D-1}^{H_D} f_R^*(i) & f_R^*(H_D - 1) + f_R^*(1) \\ 0 & 0 & \cdots & 0 & f_R^*(H_D) \end{bmatrix} \tag{1.25}$$

and $\lambda = 1 + P\left(R^* < D^*\right) = 1 + \sum_{i=1}^{H_D - 1} \sum_{j>i} f_D^*(j) f_R^*(i)$. While the above expression illustrates the form of A_{R^*} when $L = 1$, refer to section 1.9.3.1 for a description of the form of A_{R^*} when $L \geq 1$. The subsequent logic in this proof holds for any $L \geq 1$.

If the distribution of R^* is known then A_{R^*} is known. We say f_D is *nearly* a linear function of f_D^* because though it is technically nonlinear (λ is also a function of f_D^*), the scalar constant in front is easily dealt with given we are estimating a probability distribution. By definition of the support we have that $f_R^*(H_D) \neq 0$ which implies that the diagonal is non-zero, and thus that A_{R^*} is invertible. Further note that since f_D^* is a discrete probability distribution we have $\mathbf{1}' A_{R^*}^{-1} f_D = \mathbf{1}' \frac{1}{\lambda} f_D^* = \frac{1}{\lambda}$. Finally, since f_D is observed then f_D^* is identified via $f_D^* = \frac{A_{R^*}^{-1} f_D}{\mathbf{1}' A_{R^*}^{-1} f_D}$. ■

Proof of Theorem 2 (Consistency Under Independence). First we focus on the sample mean over $\overrightarrow{D}_i$ and aim to show that this is consistent for the f_D. Note that we cannot simply apply a weak law of large numbers (WLLN) since the sample is not independent, so we must

first partition this sum into a number of sums that are over iid samples. We do this first by grouping the summands by their latent individual association and then iterating over all the duration parts associated with each latent individual. Then we have

$$\frac{1}{n}\sum_{i=1}^{n}\overrightarrow{D}_i = \frac{1}{n}\sum_{j=1}^{n^*}\sum_{l=1}^{B_j^*+1}\overrightarrow{D_{jl}^*} = \frac{n^*}{n}\left(\frac{1}{n^*}\sum_{j=1}^{n^*}\sum_{l=1}^{B_j^*+1}\overrightarrow{D_{jl}^*}\right).$$

This transformation relies on having all relevant durations in the sample as provided by assumption A3. Since the latent sample is independent across individuals we can now apply the WLLN to the sample mean found within the parenthesis. Thus as $n^* \to \infty$,

$$\frac{1}{n^*}\sum_{j=1}^{n^*}\left[\sum_{l=1}^{B_j^*+1}\overrightarrow{D_{jl}^*}\right] \to_p \mathbb{E}\left[\sum_{l=1}^{B_j^*+1}\overrightarrow{D_{jl}^*}\right] = \mathbb{E}\left[\sum_{l=1}^{L}\mathbb{1}\{l \leq B_j^*+1\}\overrightarrow{D_{jl}^*}\right].$$

Turning to the ratio of latent and observed sample sizes, recall that $n = \sum_{b=0}^{L} n_b^*(b+1)$ and thus we have

$$\lim_{n^*\to\infty}\frac{n^*}{n} = \lim_{n^*\to\infty}\frac{n^*}{n^* + \sum_{b=0}^{L} n_b^* b} = \lim_{n^*\to\infty}\left[1 + \sum_{b=0}^{L}\frac{n_b^*}{n^*}b\right]^{-1} = \left[1 + \mathbb{E}\left[B_i^*\right]\right]^{-1} = \frac{1}{\lambda}.$$

Bringing these two facts together using Slutsky's theorem and focusing on the kth index we have

$$\left(\frac{1}{n}\sum_{i=1}^{n}\overrightarrow{D}_i\right)_k \to_p \left(\frac{1}{\lambda}\sum_{l=1}^{L}\mathbb{E}\left[\mathbb{1}\{l \leq B_j^*+1\}\overrightarrow{D_{jl}^*}\right]\right)_k = \frac{1}{\lambda}\sum_{l=1}^{L}P\left(B_j^* \geq l-1 \text{ and } D_{jl}^* = k\right) = f_D(k).$$

The final equality comes from close inspection of the derivation of the distribution of D as described in equation (1.4). Having shown that $\frac{1}{n}\sum\overrightarrow{D}_i \to_p f_D$ it follows from equation (1.24) that $\mathbf{1}'A_{R^*}^{-1}\frac{1}{n}\sum_{i=1}^{n}\overrightarrow{D}_i \to_p \frac{1}{\lambda}$. Finally, bringing all the pieces together we have

$$\widehat{f_D^*} = \frac{A_{R^*}^{-1}\frac{1}{n}\sum_{i=1}^{n}\overrightarrow{D}_i}{\mathbf{1}'A_{R^*}^{-1}\frac{1}{n}\sum_{i=1}^{n}\overrightarrow{D}_i} \to_p \frac{A_{R^*}^{-1}f_D}{\frac{1}{\lambda}} = \frac{\frac{1}{\lambda}f_D^*}{\frac{1}{\lambda}} = f_D^*.$$

■

Proof of Theorem 3 (Asymptotic Normality Under Independence). Focusing first on the asymptotic distribution of $\widehat{f_D}$, note that we cannot directly apply a standard central limit theorem (CLT) since the sample is not independent, as was the case when proving consistency. As

before, we rewrite the sum over the latent iid sample

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\overrightarrow{D}_i - f_D\right) = \frac{\sqrt{n^*}}{\sqrt{n}}\sqrt{n^*}\left(\frac{1}{n^*}\sum_{j=1}^{n^*}\left[\sum_{l=1}^{B_j^*+1}\overrightarrow{D_{jl}^*}\right] - \frac{n}{n^*}f_D\right).$$

Let the covariance matrix of the inner sum be denoted by Ω,

$$\Omega \equiv Var\left(\sum_{l=1}^{B_j^*+1}\overrightarrow{D_{jl}^*}\right).$$

Refer to section 1.9.3.2 for a discussion of Ω in terms of the underlying distributions. Provided Ω is invertible, then we can apply the multivariate Lindeberg-Lévy CLT to establish the asymptotic normality of the inner expression. While proving consistency we showed that $\frac{n^*}{n}$ converges in probability to $\frac{1}{\lambda}$ which, together with our asymptotic distribution, yields

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\overrightarrow{D}_i - f_D\right) \to_d N\left(0, \frac{1}{\lambda}\Omega\right).$$

Returning to the estimator of interest, $\widehat{f_D^*}$, we have that

$$\widehat{f_D^*} = g\left(\frac{1}{n}\sum_{i=1}^{n}\overrightarrow{D}_i\right) \qquad \text{where} \qquad g(x) = \frac{A_{R^*}^{-1}x}{\mathbf{1}'A_{R^*}^{-1}x}.$$

First, note that $g(x)$ is continuously differentiable in a neighborhood of $x = f_D$ because $\mathbf{1}'A_{R^*}^{-1}f_D$ is bounded away from 0 (specifically we have that $\mathbf{1}'A_{R^*}^{-1}f_D = \frac{1}{\lambda} > \frac{1}{1+L}$). Second, the Jacobian of g evaluated at f_D is given by

$$\left.\frac{\partial g}{\partial x}\right|_{x=f_D} = \frac{A_{R^*}^{-1}\left(\mathbf{1}'A_{R^*}^{-1}f_D\right) - A_{R^*}^{-1}f_D(\mathbf{1}'A_{R^*}^{-1})}{\left(\mathbf{1}'A_{R^*}^{-1}f_D\right)^2} = \lambda\left(I - f_D^*\mathbf{1}'\right)A_{R^*}^{-1}.$$

Finally, we can use the asymptotic distribution of $\widehat{f_D}$ and apply the Delta Method to derive the approximate asymptotic distribution of $\widehat{f_D^*}$,

$$\sqrt{n}\left(\widehat{f_D^*} - f_D^*\right) \to_d N\left(0, \lambda\left(A_{R^*}^{-1}\right)'\left(I - f_D^*\mathbf{1}'\right)'\Omega\left(I - f_D^*\mathbf{1}'\right)A_{R^*}^{-1}\right).$$

■

Proof of Theorem 4 (Consistency of $\widehat{V}$ Under Independence). Let $\widehat{f_D^*}$ be a consistent estimator for f_D^* and let f_R^* be known. This allows for the construction of the estimator of joint

distribution $\text{vec}\left(\widehat{f^*_{RD}}\right)$ which will be consistent for $\text{vec}\left(f^*_{RD}\right)$. Note that the exact form of $\widehat{f^*_{RD}}$ (i.e. a matrix, a vector, or a tensor) is only relevant for the form of $A_1(i,j)$ and $A_2(i,j)$. Refer to section 1.9.3.2 for a discussion of this.

This then implies the element-wise estimator of Ω is consistent,

$$\widehat{(\Omega)_{ij}} = A_2(i,j)'\text{vec}\left(\widehat{f^*_{RD}}\right) - \text{vec}\left(\widehat{f^*_{RD}}\right)' A_1(i,j)'\text{vec}\left(\widehat{f^*_{RD}}\right) \to_p (\Omega)_{ij}$$

and that the corresponding matrix estimator, $\widehat{\Omega}$, is consistent for Ω. Finally, this implies that the overall variance estimator is also consistent,

$$\widehat{V} = (A^{-1}_{R^*})'\widehat{\Omega}A^{-1}_{R^*} \to_p (A^{-1}_{R^*})'\Omega A^{-1}_{R^*} = V.$$

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.

■

1.9.4.3 Under Dependent Linkage Error

Proof of Theorem 5 (Identification). Let assumptions A1, A2, A3, A5, and A6 hold. First we identify f^*_S using the conditional distribution of observed starting times for the maximum duration individuals. If e_{H_D} is a column vector of 0s with a 1 in the last row (i.e. the H_Dth row) then the conditional distribution, $P\left(S_i = s | D_i = H_D\right)$, is given by $f_{SD}e_{H_D}\left(\mathbf{1}'f_{SD}e_{H_D}\right)^{-1}$.

Note that

$$
\begin{aligned}
f_{SD}e_{H_D} &= \frac{1}{\lambda}\left(\begin{bmatrix} f_S^* \\ 0 \end{bmatrix}\left(\mathbf{1}'\left[f_{RD}^*\right]_L e_{H_D} + \mathbf{1}'\underbrace{\left[f_{RD}^*\right]_U' e_{H_D}}_{\equiv 0}\right) + \sum_k f_S^*(k)L_k\underbrace{\left[\left[f_{RD}^*\right]_U\right]_Q e_{H_D}}_{\equiv \mathbf{0}}\right) \\
&= \frac{1}{\lambda}\begin{bmatrix} f_S^* \\ 0 \end{bmatrix}\mathbf{1}'\begin{bmatrix} 0 \\ \vdots \\ f_{RD}^*(H_R^*, H_D^*) \end{bmatrix} \\
&= \frac{f_{RD}^*(H_R^*, H_D^*)}{\lambda}\begin{bmatrix} f_S^* \\ 0 \end{bmatrix}
\end{aligned}
$$

In the first line, the first expression is 0 since $[\cdot]_U$ produces matrices with 0s in the last row, and the second expression is 0 since $[\cdot]_Q$ produces matrices with 0s in the final column. This then implies that $\mathbf{1}'f_{SD}e_{H_D} = \frac{f_{RD}^*(H_R^*, H_D^*)}{\lambda}$ and thus that $f_{SD}e_{H_D}\left(\mathbf{1}'f_{SD}e_{H_D}\right)^{-1} = \begin{bmatrix} f_S^* & 0 \end{bmatrix}'$ which identifies f_S^*.

Moving on to identifying $\left[f_{RD}^*\right]_U$ we start by considering the $(H_D - 1) \times (H_S^* + H_D - 1)$ matrix, $Z = \begin{bmatrix} \mathbf{0} & I_{H_D-1} \end{bmatrix}$, where I_{H_D-1} is the $(H_D - 1) \times (H_D - 1)$ identity matrix. Note that

$$
\begin{aligned}
Zf_{SD} &= \frac{1}{\lambda}\left(\underbrace{Z\begin{bmatrix} f_S^* \\ 0 \end{bmatrix}}_{\equiv 0}\left(\mathbf{1}'\left[f_{RD}^*\right]_L + \mathbf{1}'\left[f_{RD}^*\right]_U'\right) + Z\sum_k f_S^*(k)L_k\left[\left[f_{RD}^*\right]_U\right]_Q\right) \\
&= \frac{1}{\lambda}\underbrace{Z\sum_k f_S^*(k)L_k}_{\equiv W}\left[\left[f_{RD}^*\right]_U\right]_Q,
\end{aligned}
$$

where W is entirely known since L_k is a constant matrix and f_S^* was previously identified. When W is invertible then we have that $W^{-1}Zf_{SD} = \frac{1}{\lambda}\left[\left[f_{RD}^*\right]_U\right]_Q$. Furthermore, since $\lambda = 1 + \mathbf{1}'\left[\left[f_{RD}^*\right]_U\right]_Q\mathbf{1}$ it can be shown that

$$
\lambda = \left(1 - \mathbf{1}'W^{-1}Zf_{SD}\mathbf{1}\right))^{-1}
$$

which implies that λ is identified. Bringing this together, we have that

$$\lambda W^{-1} Z f_{SD} = [[f^*_{RD}]_U]_Q$$

which implies the right hand side is identified. Since $[\cdot]_Q$ is invertible when the domain consists of upper triangular matrices (see 1.9.1) then this implies that we can identify $[f^*_{RD}]_U$. If assumption A7 holds as a strict assumption, then identifying $[f^*_{RD}]_U$ identifies all of f^*_{RD}.

We next move on to identifying the marginal, f^*_D if either assumption A7 holds as a normalization or does not hold at all. Let e_1 denote a vector of all zeros with a one in the first cell. Then given everything that has been previously identified we have

$$\frac{1}{f^*_S(1)} e_1' \left(\lambda f_{SD} - \sum_k f^*_S(k) L_k [[f^*_{RD}]_U]_Q - \begin{bmatrix} f^*_S \\ 0 \end{bmatrix} \mathbf{1}' [f^*_{RD}]'_U \right)$$
$$= \frac{1}{f^*_S(1)} e_1' \begin{bmatrix} f^*_S \\ 0 \end{bmatrix} \mathbf{1}' [f^*_{RD}]_L = \mathbf{1}' [f^*_{RD}]_L$$

which implies that $\mathbf{1}' [f^*_{RD}]_L$ is identified. This object yields identification of $P(R^*_i = k, D^*_i \geq k)$ for all k, and with knowledge of $P(R^*_i = r, D^*_i = d)$ for all $r < d$ (which comes from identification of $[f^*_{RD}]_U$) we can construct the marginal of interest, f^*_D. ■

Proof of Theorem 6 (Consistency of GMM Estimator). The proof of consistency leverages the GMM consistency result for clustered samples found in Theorem 12 of Hansen and Lee (2019), so we proceed by verifying the conditions of their theorem. The clusters in our scenario are the potentially numerous broken durations for each individual, so in their notation the size of a cluster, n_g, is equal to the number of broken durations, $B^*_i + 1$. Since the number of broken durations is bounded by $L + 1$, then their assumption 1 is trivially satisfied.

With respect to requirements 1 and 2 of Theorem 12, our parameter space consists of probability distributions over finite discrete random variables which implies the space is compact, and the parameter is identified under the assumptions of Theorem 5. Point 3 is

satisfied because the marginal distribution of observed starts and durations is identical and the random variables are finite and discrete implying $\mathbb{E}\,\|m(X_i, \theta)\| < \infty$.

Point 4 is assumed via our assumption A8 and points 5 and 6 hold trivially since we only consider the constant weight matrix, $W_n = I$. Thus we can apply Theorem 12 and conclude that $\widehat{\theta} \to_p \theta$. ■

Proof of Theorem 7 (Asymptotic Normality of GMM Estimator). As with our consistency result we again leverage the work of Hansen and Lee (2019) via Theorem 13 of their paper. Condition 1 of their theorem is covered by our assumption A9 and conditions 2(a), 2(b), and 2(c) are trivially satisfied since all random variables are discrete and finite and the moment function is relatively well behaved. Conditions 2(d), 3, 4, and 5 are all covered via our assumption A9. With all conditions satisfied we can apply Theorem 13 of Hansen and Lee (2019) to establish asymptotic normality. Note that since the asymptotic covariance in our case converges to a constant matrix, we do not need the full generality allowed by Hansen and Lee (2019). ■

Chapter 2

Elasticity Estimation in Discrete Choice Models with Potential Demand Misspecification

(with Diwakar Raisingh)

Chapter Summary

Estimates of elasticities are a common output of interest in discrete choice models, however they can be significantly biased when the population size is misspecified. In this note we decompose elasticity estimates in the logit model into a direct effect and an indirect effect coming from bias in the structural parameter estimates. Since these effects can go in opposite directions addressing bias from the indirect channel, via market fixed effects, will have an indeterminate effect on the total bias in the elasticity. We provide a complete characterization of when including market fixed effects will mitigate versus exacerbate elasticity bias. Our results reveal that for own characteristic elasticities products with small shares will typically benefit most from market fixed effects while the benefit (or detriment) for cross characteristic elasticities is independent of share.

2.1 Introduction

Discrete choice models are an especially common workhorse in researching substitution patterns in demand. Berry et al. (1995) (henceforth BLP) logit models are utilized when researcher believes that there are unobserved (by the econometrician) shocks that are correlated with observable variables (e.g. price). This model is often estimated using conditional choice probabilities obtained with aggregate data on quantities purchased and an assumption on potential demand. Since the potential demand, the maximum quantity that could have been purchased, is assumed and typically misspecified, using it can bias the structural parameter estimates. Moreover, this misspecification biases the elasticity estimates, which capture substitution patterns, that are the primary focus in demand estimation.

In this note we study how misspecified potential demand biases the elasticity estimates. We decompose the elasticity bias into a direct effect, from the misspecified potential demand, and an indirect effect, from bias in the structural parameter estimates induced by the misspecification. Though the indirect effects can be eliminated by including market fixed effects the net effect on bias is indeterminate. We provide a complete characterization of these effects and show that correcting the bias in the structural parameter estimates can increase the bias in the elasticity estimates.

For empiricists we provide two prescriptions (Corollaries 1 and 2) dictating when to include market fixed effects. When the bias in an endogenous parameter due to market size misspecification can be estimated, we provide thresholds on market shares that delineate when market fixed effects are preferred. Alternatively if the researcher can leverage what we are calling a "targeted instrument", then we can provide conditions that guarantee that including market fixed effects will reduce bias in own price elasticity estimates.

For notation we use an open circle ($\circ$) to represent the percent deviation from the truth. For example the percent deviation of the incorrect market potential demand, $\widetilde{M}_t$, from the

true market potential demand, M_t, is

$$\mathring{M}_t = \frac{\widetilde{M}_t - M_t}{M_t}.$$

Moreover, we will use the term market size interchangeably with market potential demand.

The rest of this note is organized as follows. In Section 2.2 we detail the logit model and how misspecification results in inconsistent estimators. Section 2.3 presents the elasticity decomposition and results pertaining to when market fixed effects reduce the magnitude of bias. Section 2.4 concludes.

2.2 Model & Estimation

In a BLP logit model, each of the M_t agents in market t choose one of $j = 1, \ldots, J_t$ goods offered in the market or the outside good, $j = 0$, that gives the highest utility. Agent i's indirect utility from choosing product j is

$$u_{ijt} = \underbrace{\boldsymbol{\beta}\boldsymbol{x}_{jt} + \alpha p_{jt} + \xi_{jt}}_{\equiv \delta_{jt}} + \varepsilon_{ijt}, \tag{2.1}$$

where δ_{jt} is the mean utility of product j in market t; $\boldsymbol{x}_{jt} = (1, x_{jt}^{(1)}, \ldots, x_{jt}^{(K)})'$ and p_{jt} are the vector of observable characteristics and the price; $\boldsymbol{\beta}$ and α are the corresponding taste parameters; ξ_{jt} is unobservable (to the econometrician) characteristic correlated with p_{jt} that is common to all agents; and ε_{ijt} is the idiosyncratic shock assumed to be i.i.d. Type I Extreme Value. Utility is normalized such that choosing the outside option has mean utility zero and yields $u_{i0t} = \varepsilon_{i0t}$. M_t represents the potential market demand (also referred to as market size) because this is the maximum quantity of inside goods that can be purchased. The probability an agent chooses good j in market t:

$$\sigma_{jt} = \frac{\exp\{\delta_{jt}\}}{1 + \sum_{k=1}^{J_t} \exp\{\delta_{kt}\}}. \tag{2.2}$$

The price elasticities of good j with respect to the price change in good k is given by

$$\eta_{jkt} = \frac{p_{kt}}{\sigma_{jt}} \frac{\partial \sigma_{jt}}{\partial p_{kt}} = \begin{cases} \alpha p_{jt}(1 - \sigma_{jt}) & \text{if } j = k, \\ -\alpha p_{kt}\sigma_{kt} & \text{otherwise.} \end{cases} \tag{2.3}$$

These elasticities are the main objects of interest because they capture the substitution patterns among products.

Estimation of the parameters utilizes the well known (see Berry (1994)) inversion, $\delta_{jt} = \ln(\sigma_{jt}) - \ln(\sigma_{0t})$, to recover the mean utilities given choice probabilities, In practice when the quantities of inside goods (q_{jt}) purchased is observed and M_t is sufficiently large, the econometrician uses the market shares $s_{jt} = \frac{q_{jt}}{M_t}$ to perform the inversion because they are consistent estimators of σ_{jt}. We will refer to market shares s_{jt} for the rest of the paper and use them in lieu of the σ_{jt}'s.

In many contexts, researchers only observe the quantities of the inside goods $\{q_{1t}, \ldots, q_{J_t t}\}_{t=1}^{T}$, without knowledge of either the outside quantity, $\{q_{0t}\}_{t=1}^{T}$, nor the market size, $\{M_t\}_{t=1}^{T}$. In order to compute market shares to perform the inversion, researchers typically use an ad-hoc market size, denoted here as $\widetilde{M}_t$, allowing them to compute (possibly erroneous) shares, $\widetilde{s}_{jt} \equiv q_{jt}/\widetilde{M}_t$. Note $\widetilde{s}_{jt} = s_{jt}$ only if $\widetilde{M}_t = M_t$. In this paper we make the following assumptions with regard to estimation.

Assumption A10 (Instrument Validity)**.** $\mathbb{E}\left[\boldsymbol{x}'_{jt}\boldsymbol{z}_{jt}\right]$ *has full rank and* $\mathbb{E}\left[\boldsymbol{z}_{jt}\xi_{jt}\right] = \mathbf{0}$.

Assumption A11 (Rational Size)**.** $\widetilde{M}_t > \sum_{j=1}^{J_t} q_{jt} \quad \forall t$.

Assumption A10 is a standard assumption in BLP models and states that the researcher has valid instruments $\boldsymbol{z}_{jt}$ for price. Assumption A11 ensures that the observable data does not immediately refute the validity of the selected market size.

Using the market share inversion, α and $\boldsymbol{\beta}$ can be estimated via generalized method of moments (GMM) using the estimating equation:

$$\ln\left(\tfrac{q_{jt}}{\widetilde{M}_t}\right) - \ln\left(\tfrac{\widetilde{q}_{0t}}{\widetilde{M}_t}\right) = \boldsymbol{\beta}\boldsymbol{x}_{jt} + \alpha p_{jt} + \omega_t + \xi_{jt},$$

where

$$\omega_t = -\ln\left(\mathring{M}_t s_{0t}^{-1} + 1\right).$$

Suppose the empiricist has instruments that satisfy assumption A10 and that $\widetilde{M}_t = M_t$ (so $\mathring{M}_t = 0$ and $\omega_t = 0$) for all t. Then the GMM estimates without market fixed effects, which we denote with the subscript "$Naive$", are consistent. The price coefficient, for example, is consistent with $\text{plim}_{t\to\infty}\widehat{\alpha}_{Naive} = \alpha$ and thus the asymptotic percent deviation is zero, $\mathring{\alpha}_{Naive} = 0$.

If the empiricist did not use the true market sizes (i.e. $\mathring{M}_t \neq 0$) then $\omega_t \neq 0$. Since the true outside share, s_{0t}, is a function of observable variables ($\boldsymbol{x}_{jt}$ and p_{jt}) this implies a non-zero correlation between ω_t and all the covariates. When unaccounted for this association will violate assumption A10 and consistency will not be guaranteed, even when instruments are being employed. This means it is very likely[1] that $\text{plim}_{t\to\infty}\widehat{\alpha}_{Naive} \neq \alpha$ (equivalently $\mathring{\alpha}_{Naive} \neq 0$). This issue cannot be avoided by estimating M_t because it can be shown that market size is not estimable.[2]

Though the bias can be substantial, there is an easy solution to consistently estimating a majority of the structural parameters when the market size is misspecified. Huang and Rojas (2014) were the first to note that since ω_t is constant within a market, the inclusion of market fixed effects would enable consistent estimation of all parameters whose covariates are not constant within a market. If we define $\widehat{\alpha}_{MFE}$ to be the GMM estimator when market fixed effects are included in the estimating equation then $\text{plim}_{t\to\infty}\widehat{\alpha}_{MFE} = \alpha$ and $\mathring{\alpha}_{MFE} = 0$.

In most situations estimator bias is unobserved, however here it can be estimated by combining estimators with and without market fixed effects. If no covariates are constant within a market, $Var(\boldsymbol{x}_{jt}) > 0$ within every market t, then $(\widehat{\alpha}_{Naive} - \widehat{\alpha}_{MFE})/\widehat{\alpha}_{MFE}$ will be a

[1]There are scenarios with nontrivial market size error in which the naive estimator is still consistent for α, but these are knife edge cases.

[2]The market sizes $\{M_t\}_{t=1}^T$ are not separately identified from β_0, the coefficient on the intercept, without additional assumptions. We cannot estimate M_t by rewriting the estimating equation as $\ln(q_{jt}) = \boldsymbol{x}_{jt}\boldsymbol{\beta} + \alpha p_{jt} + \ln\left(M_t - \sum_{j=1}^{J_t} q_{jt}\right) + \xi_{jt}$ because $\sum_{j=1}^{J_t} q_{jt}$ is correlated with both M_t and q_{jt}. Thus, even if the instruments for M_t were available, assumption A10 would be violated.

consistent estimator for $\mathring{\alpha}_{Naive}$ under the model described and assumptions A10-A11. This will prove useful in the next section when trying to decide whether elasticity estimates are aided by the inclusion of market fixed effects.

Determining the bias magnitude or sign is difficult without further restrictions. Assumption A12 is one such condition which leverages a "targeted instrument": a covariate that is strongly associated with the endogenous variable and not associated with any other variables.

Assumption A12 (Targeted Instrument)**.** *The instrument, z_p, that is associated with p_{jt} has the following properties*

1. *z_p and p_{jt} are either negatively quadrant dependent (NQD) or positively quadrant dependent (PQD) conditional on other regressors.*[3]

2. *$\mathbb{E}[z_p|\chi] = 0$ where χ includes all variables aside from p_{jt} ($\boldsymbol{x}_{jt}$ and non-z_p instruments).*[4]

Using this assumption Theorem 10 signs the bias depending on the sign of the market size error.

Theorem 10. *Under assumptions A10-A12 if $\mathring{M}_t$ is the same sign for all t then $\mathring{\alpha}_{Naive}$ and $\mathring{M}$ have opposite signs.*

In other words, if the market size is overestimated in every market then $\widehat{\alpha}_{naive}$ is biased towards zero, and if the market size is underestimated in every market then $\widehat{\alpha}_{naive}$ is biased away from zero. This result will be leveraged to provide the methodological prescription described in the next section.

[3]Positive quadrant dependence holds if $F_{z,p}(z,p) \geq F_z(z)F_p(p)$ for all p, z where $F_{z,p}$ is the cdf of the joint of z_p and p_{jt} and F_z and F_p are the respective marginal cdfs. Negative quadrant dependence is defined similarly except with the inequality reversed. These measures of dependence are stronger than correlation, refer to Lehmann (1966) or Cuadras (2002) for more discussion.

[4]An example of such an instrument in the car market could be the tariff rate in a year where tariffs changed unexpectedly. Since cars have a significant design period, the unexpected tariffs would not have affected the manufacturers choice of car characteristics but would be correlated with price.

2.3 Elasticities

In the logit model the estimator of the price elasticity of good j with respect to a change in the price of good k is given by

$$\widehat{\eta}_{jkt} = \begin{cases} \widehat{\alpha} p_{jt}(1 - \widetilde{s}_{jt}) & \text{if } j = k, \\ -\widehat{\alpha} p_{kt} \widetilde{s}_{kt} & \text{otherwise.} \end{cases} \tag{2.4}$$

In both cases the estimators are functions of two estimated quantities, meaning that any error in market size will affect elasticity bias via two routes: directly through the market shares, $\widetilde{s}_{jt} = q_{jt}/\widetilde{M}_t$, and indirectly through biased estimates of the structural parameter, $\widehat{\alpha}$. In this section we investigate how including market fixed effects, which essentially shuts down the indirect pathway, compares with the standard approach which omits these fixed effects. We consider the own price and cross price elasticity separately since they yield different qualitative results.

Starting with the own price elasticity (i.e. $j = k$), we decompose the percentage deviation of the elasticity into contributions from these two pathways:

$$\mathring{\eta}_{jjt} = \underbrace{\frac{1 - s_{jt}\frac{1}{1+\mathring{M}_t}}{1 - s_{jt}} \cdot \mathring{\alpha}}_{B_\alpha} + \underbrace{\frac{s_{jt}\frac{1}{1+\mathring{M}_t}}{1 - s_{jt}} \cdot \mathring{M}_t}_{B_M}. \tag{2.5}$$

Here B_α represents the contribution of bias in the structural parameter, $\mathring{\alpha}$, while B_M represents the contribution of bias directly from mis-measuring the market size, $\mathring{M}_t$.

Inspection of this decomposition reveals that for small market shares, the coefficient in front of $\mathring{\alpha}$ is approximately 1 and the coefficient in front of the second term is approximately s_{jt}. This suggests that the contribution of bias from the structural parameter is larger than the contribution of bias directly from M_t, and thus that including fixed effects may decrease bias for small market shares. If we define $\mathring{\eta}_{jjt}^{MFE}$ and $\mathring{\eta}_{jjt}^{Naive}$ to be the percentage deviations in the elasticity estimates with and without market fixed effects respectively, then this supposition is made concrete in Theorem 11.

Theorem 11. *Under assumptions A10-A11, there exist sets* $\mathcal{S}_t \subseteq [0,1]$ *such that*

$$s_{jt} \in \mathcal{S}_t \quad \Longleftrightarrow \quad \left|\mathring{\eta}_{jjt}^{MFE}\right| > \left|\mathring{\eta}_{jjt}^{Naive}\right|.$$

The set $\mathcal{S}_t$ *is determined by* $\mathring{\alpha}_{Naive}$, $\mathring{M}_t$ *and*

$$c_{1t} = 1 + \mathring{M}_t \qquad\qquad c_{2t} = \frac{\left(1 + \mathring{M}_t\right)\mathring{\alpha}_{Naive}}{\mathring{\alpha}_{Naive} - 2\mathring{M}_t}$$

and is fully enumerated in Figure 2.1.

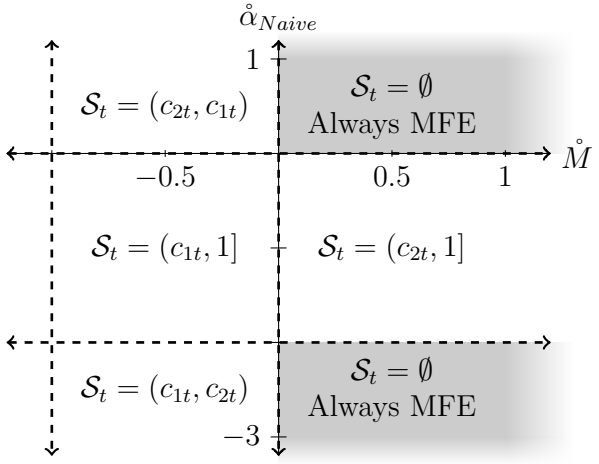


Figure 2.1: This figure accompanies Theorem 11 and illustrates sets of shares, $\mathcal{S}$, for which $\left|\mathring{\eta}_{jjt}^{MFE}\right| > \left|\mathring{\eta}_{jjt}^{Naive}\right|$. Regions shaded gray correspond to situations where the magnitude of the elasticity bias will always be reduced by including market fixed effects.

Three implications of Theorem 11 are worth noting. First it illustrates that while market fixed effects restore the consistency of $\widehat{\alpha}$, they do not necessarily translate to a reduction in bias in the own price elasticities (i.e. $\mathcal{S}$ is not always empty). Second since every $\mathcal{S}$ does not contain a neighborhood of zero, this theorem guarantees that market fixed effects will reduce bias in own price elasticities for products whose shares are close to zero. Finally when the market size is overestimated and $\mathring{\alpha} > 0$ or $\mathring{\alpha} < -2$ (regions highlighted gray in Figure 2.1) then market fixed effects provide bias improvements regardless of product share (i.e. $\mathcal{S} = \emptyset$).

The preceding theorem can be leveraged to provide two methodological prescriptions for empiricists. These prescriptions, codified in Corollaries 1 and 2, provide sufficient conditions

that imply the bias in the own price elasticity will be smaller when including market fixed effects.

Corollary 1. *Under assumptions A10 and A11, if* $\mathring{M}_t > 0$ *for all* t*, and either* $\mathring{\alpha}_{Naive} > 0$ *or* $\mathring{\alpha}_{Naive} < -2$ *then* $\left|\mathring{\eta}^{MFE}_{jjt}\right| < \left|\mathring{\eta}^{Naive}_{jjt}\right|$ *for all* j *and* t*.*

Corollary 2. *Under assumptions A10, A11, and A12, if* $\mathbb{E}\left[\boldsymbol{x}_{jt}p_{jt}\right] = 0$*, and* $\mathring{M}_t > 0$ *for all* t*, then* $\left|\mathring{\eta}^{MFE}_{jjt}\right| < \left|\mathring{\eta}^{Naive}_{jjt}\right|$ *for all* j *and* t*.*

Corollary 1, which follows directly from Theorem 11, is appropriate in situations where $\mathring{\alpha}_{Naive}$ can be estimated as discussed in the previous section. If $\widehat{\mathring{\alpha}}_{Naive} > 0$ or $\widehat{\mathring{\alpha}}_{Naive} < -2$ then overestimating the market size suggests that including market fixed effects is preferred for reducing bias in own price elasticity estimates.

If $\mathring{\alpha}$ cannot be estimated, for example when market level controls are present, then Corollary 2 provides an alternative. Provided the researcher can find a 'targeted instrument,' as defined in assumption A12 then they can once again be assured that including market fixed effects is advisable to mitigate bias in own price elasticities when market size is over estimated.

To illustrate this process in action Table 2.1a displays the results of several Monte Carlo simulations. Each row is generated by a different market size error model and presents the average percent deviation in the own price elasticity (across products, markets, and simulations) and average decomposition constituents ($\mathbb{E}\left[B_\alpha\right]$ and $\mathbb{E}\left[B_M\right]$) with and without market fixed effects. As predicted by the theory market fixed effects virtually eliminate the indirect contribution which leads to either exacerbating or mitigating elasticity bias.

	Naive Estimator			Market FE Estimator		
Error Model	$\mathbb{E}[\mathring{\eta}_{jt}]$	$\mathbb{E}[B_\alpha]$	$\mathbb{E}[B_M]$	$\mathbb{E}[\mathring{\eta}_{jt}]$	$\mathbb{E}[B_\alpha]$	$\mathbb{E}[B_M]$
$\mathring{M}_t = -0.1$	0.0729	0.0828	−0.0099	−0.0097	0.0002	−0.0099
$\mathring{M}_t = 0.1$	−0.0177	−0.0258	0.0081	0.0083	0.0002	0.0081
$\mathring{M}_t \sim N(0, 0.06^2)$	0.0131	0.0134	−0.0003	−0.0001	0.0002	−0.0003
$\hookrightarrow \{j,t : s_{jt} \notin \mathcal{S}_t\}$	0.0182	0.0171	0.0011	0.0034	0.0023	0.0011
$\hookrightarrow \{j,t : s_{jt} \in \mathcal{S}_t\}$	−0.0041	0.0010	−0.0051	−0.0120	−0.0069	−0.0051

(a) Own-price elasticity.

	Naive Estimator			Market FE Estimator		
Error Model	$\mathbb{E}[\mathring{\eta}_{0kt}]$	$\mathbb{E}[B_\alpha]$	$\mathbb{E}[B_M]$	$\mathbb{E}[\mathring{\eta}_{0kt}]$	$\mathbb{E}[B_\alpha]$	$\mathbb{E}[B_M]$
$\mathring{M}_t = -0.1$	0.2040	0.0929	0.1111	0.1113	0.0002	0.1111
$\mathring{M}_t = 0.1$	−0.1142	−0.0232	−0.0909	−0.0907	0.0002	−0.0909
$\mathring{M}_t \sim N(0, 0.06^2)$	0.0171	0.0135	0.0036	0.0038	0.0002	0.0036
$\hookrightarrow \{t : \lvert\mathring{M}_t\rvert < \lvert\mathring{\alpha}^{\text{Naive}} - \mathring{M}_t\rvert\}$	0.0466	0.0149	0.0317	0.0325	0.0008	0.0317
$\hookrightarrow \{t : \lvert\mathring{M}_t\rvert \geq \lvert\mathring{\alpha}^{\text{Naive}} - \mathring{M}_t\rvert\}$	−0.0192	0.0117	−0.0309	−0.0315	−0.0006	−0.0309

(b) Cross-price elasticity for the outside option.

Table 2.1: Results are generated from 1,000 simulations of data sets that have 10,000 markets with 10 products in each market. Agents have indirect utility $u_{ijt} = 1 + x_{jt} - p_{jt} + \xi_{jt} + \varepsilon_{ijt}$, where $p_{jt} = 1.4 + x_{jt} + w_{jt} + 0.25\xi_{jt} + \omega_{jt}$. The variables have distributions $x_{jt}, w_{jt} \sim U(0.5, 1)$, $\xi_{jt}, \omega_{jt} \sim N(0, 0.5^2)$, and $\varepsilon_{ijt} \sim$ T1EV. The true outside share in a market is about 0.21. In each row the smaller magnitude of $\mathbb{E}[\mathring{\eta}_{jt}]$ is boxed to indicate whether the MFE or naive estimator is preferred. When computing $\mathbb{E}[\mathring{\eta}_{jt}]$ for the final two rows in each table, in each simulation where $\widetilde{M}_t = \zeta_t M_t$ we compute $\mathring{\eta}_{jt}$ for each product, tracking if the listed criteria is satisfied. Then we average across all simulations, weighting each simulation by the number of products that satisfied the criteria.

Turning to cross price elasticities, η_{jkt}, the decomposition of the bias is given by

$$\mathring{\eta}_{jkt} = \underbrace{\frac{1}{1+\mathring{M}_t} \cdot \mathring{\alpha}}_{B_\alpha} - \underbrace{\frac{1}{1+\mathring{M}_t} \cdot \mathring{M}_t}_{B_M} \tag{2.6}$$

In contrast to decomposition (2.5), the indirect and direct contributions here have the same weight, and both are entirely independent of the product share. Thus whether market fixed effects reduce or inflate elasticity bias, it will do so in the same direction across all market shares, and this is made concrete in Theorem 12.

Theorem 12. *Under assumptions A10-A11, for cross price elasticities (i.e. $j \neq k$) we have*

$$\left|\mathring{M}_t\right| < \left|\mathring{\alpha} - \mathring{M}_t\right| \quad \Longleftrightarrow \quad \left|\mathring{\eta}_{jkt}^{MFE}\right| < \left|\mathring{\eta}_{jkt}^{Naive}\right|.$$

Table 2.1b displays the same Monte Carlo simulations but focuses on the cross-price elasticities. Once again the results coincide with the theoretical predictions.

2.4 Conclusion

When estimating a logit model with misspecified market sizes, the bias in structural parameters can be corrected with market fixed effects. This note shows that correcting the structural parameter estimates can both decrease and increase the bias in elasticity estimates relative to the naive estimates. We derive conditions under which own-characteristic elasticity estimates under the fixed effects estimator have smaller bias than under the naive estimator.

This small note represents just the beginning of a larger investigation into market misspecification. Ongoing work includes extending the above results to the random coefficients logit model as well as extending the theory to heterogeneous market size error.

2.5 References

Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, 25(2):242–262.

Berry, S. T., Levinsohn, J., and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, 63(4):841–890.

Cuadras, C. (2002). On the covariance between functions. *Journal of Multivariate Analysis*, 81(1):19–27.

Huang, D. and Rojas, C. (2014). Eliminating the outside good bias in logit models of demand with aggregate data. *Review of Marketing Science*, 12(1):1–36.

Lehmann, E. L. (1966). Some concepts of dependence. *The Annals of Mathematical Statistics*, 37(5):1137–1153.

Nevo, A. (2000a). Mergers with differentiated products : The case of the ready-to-eat cereal industry. *The RAND Journal of Economics*, 31(3):395–421.

Nevo, A. (2000b). A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of Economics & Management Strategy*, 9(4):513–548.

Nevo, A. (2001). Measuring market power in the ready-to-eat cereal industry. *Econometrica*, 69(2):307–342.

2.6 Appendix

The proof of Theorem 10 leverages the following lemma, which follows from direct application of a generalization of Hoeffding's covariance identity as found in Theorem 1 of Cuadras (2002).

Lemma 2. *Given random variables X and Y and function $h : \mathbb{R} \to \mathbb{R}$, if h is differentiable, monotonic, has bounded variation, and $\mathbb{E}[|X|], \mathbb{E}[|h(Y)|], \mathbb{E}[|Xh(Y)|]$ exist, then*

$$\begin{aligned} X \text{ and } Y \text{ are PQD} \quad &\Rightarrow \quad Cov(X, h(Y))\, h'(\cdot) > 0 \\ X \text{ and } Y \text{ are NQD} \quad &\Rightarrow \quad Cov(X, h(Y))\, h'(\cdot) < 0 \end{aligned}$$

Proof of Theorem 10. Without loss of generality we can demean all variables with the only consequence being that $\omega_t = -\ln\left(\mathring{M}_t s_{0t}^{-1} + 1\right) - \mathbb{E}\left[-\ln\left(\mathring{M}_t s_{0t}^{-1} + 1\right)\right]$. Denote the matrix and vector equivalents of $\widetilde{\delta}_{jt}, \boldsymbol{x}_{jt}, \boldsymbol{z}_{jt}, p_{jt}, z_{jt}^{(p)}, \omega_t, \xi_{jt}$ by $\widetilde{\boldsymbol{\delta}}, \boldsymbol{X}_1, \boldsymbol{Z}_1, \boldsymbol{p}, \boldsymbol{z}_p, \boldsymbol{\omega}$, and $\boldsymbol{\xi}$ respectively (all of which have dimension $n \times 1$ except for $\boldsymbol{X}_1$ and $\boldsymbol{Z}_1$ which are $n \times k$ where $n = \sum_t J_t$). Further define the matrices

$$\gamma = \begin{bmatrix} \boldsymbol{\beta} \\ \alpha \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{p} \end{bmatrix}, \quad \boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_1 & \boldsymbol{z}_p \end{bmatrix},$$

which allows for the compact reduced form representation $\widetilde{\boldsymbol{\delta}} = \boldsymbol{X}\gamma + \boldsymbol{\omega} + \boldsymbol{\xi}$. The just identified instrumental variable estimator is then given by

$$\widehat{\gamma} = (\boldsymbol{Z}'\boldsymbol{X})^{-1} \boldsymbol{Z}'\widetilde{\boldsymbol{\delta}} = \gamma + (\boldsymbol{Z}'\boldsymbol{X})^{-1} \boldsymbol{Z}'\boldsymbol{\omega} + (\boldsymbol{Z}'\boldsymbol{X})^{-1} \boldsymbol{Z}'\boldsymbol{\xi}.$$

After leveraging the identification condition in assumption A10 the probability limit of this estimator is given by

$$\widehat{\gamma} \to_p \gamma + \mathbb{E}[\boldsymbol{Z}'\boldsymbol{X}]^{-1}\, \mathbb{E}[\boldsymbol{Z}'\boldsymbol{\omega}]$$

illustrating the bias resulting from the mismeasured market size. Expanding the matrices and leveraging $\mathbb{E}\left[\boldsymbol{z}_p'\boldsymbol{X}_1\right] = 0$ (which comes from the second characteristic of assumption

A12) reveals that

$$\widehat{\alpha} - \alpha = \mathbb{E}\left[\boldsymbol{z}_p'\boldsymbol{p}\right]^{-1}\mathbb{E}\left[\boldsymbol{z}_p'\boldsymbol{\omega}\right].$$

The requirement that the targeted instrument be negative quadrant dependent with p_{jt} implies that $\mathbb{E}\left[\boldsymbol{z}_p'\boldsymbol{p}\right]^{-1} < 0$, and thus the sign of the bias will match the sign of $-\mathbb{E}\left[\boldsymbol{z}_p'\boldsymbol{\omega}\right]$.

It will be useful to write $\boldsymbol{\omega}$ as a function: $\boldsymbol{\omega}\left(p_{jt}, \chi\right)$ where $\chi = \left(\boldsymbol{x}_{jt}, \xi_{jt}, \mathring{M}_t\right)$. Using the law of total covariance together with $\mathbb{E}\left[z_p|\chi\right] = 0$ yields

$$\mathbb{E}\left[z_p\boldsymbol{\omega}\left(p_{jt}, \chi\right)\right] = \mathbb{E}\left[\mathrm{Cov}\left(z_p, \boldsymbol{\omega}\left(p_{jt}, \chi\right)|\chi\right)\right].$$

Taking the partial derivative of $\boldsymbol{\omega}\left(p_{jt}, \chi\right)$ with respect to p_{jt} yields

$$\frac{\partial\omega(p_{jt}, \chi)}{\partial p_{jt}} = -\alpha\mathring{M}_t\frac{\exp\{\delta_{jt}\}s_{0t}}{\mathring{M}_t + s_{0t}}.$$

Since assumption A11 guarantees that $\mathring{M}_t + s_{0t} > 0$ then the sign of the derivative depends entirely on $-\alpha\mathring{M}_t$. Since $\boldsymbol{\omega}\left(p_{jt}, \chi\right)$ is monotonic this implies it has bounded variation on the support of p_{jt} and we can invoke Lemma 2 to conclude that the sign of $\mathbb{E}\left[z_p, \boldsymbol{\omega}\left(p_{jt}, \chi\right)\right]$ is the same as that of $-\alpha\mathring{M}_t$, and thus opposite from $\widehat{\alpha} - \alpha$. Thus the signs of $\mathring{\alpha} \equiv \frac{\widehat{\alpha}-\alpha}{\alpha}$ and $\mathring{M}_t$ are the opposite. ■

Proof of Theorem 11. First we define $\mathring{\nu}_{jt} = (1 - s_{jt})\left(1 + \mathring{M}_t\right)\mathring{\eta}_{jt}$ and note that assumption A11 and $s_{jt} \in (0, 1)$ imply $\mathring{\nu}_{jt} > 0$ if and only if $\mathring{\eta}_{jt} > 0$. Given this we can focus on comparing the significantly simpler forms of

$$\left|\mathring{\nu}_{jt}^{Naive}\right| = \left|(1 + \mathring{M}_t)\mathring{\alpha} + (\mathring{M}_t - \mathring{\alpha})s_{jt}\right|$$
$$\left|\mathring{\nu}_{jt}^{MFE}\right| = \left|\mathring{M}_t s_{jt}\right|.$$

Since these are piecewise linear functions of s_{jt} simple algebra reveals that the only points of intersection are

$$s_1 = 1 + \mathring{M}_t \qquad s_2 = \left(1 + \mathring{M}_t\right)\frac{\mathring{\alpha}}{\mathring{\alpha} - 2\mathring{M}_t},$$

and thus these points must delineate the intervals where $\left|\mathring{\nu}_{jt}^{Naive}\right| > \left|\mathring{\nu}_{jt}^{MFE}\right|$. Since $\left|\mathring{\nu}_{jt}^{Naive}\right| >$

$\left|\mathring{\nu}_{jt}^{MFE}\right|$ at $s_{jt} = 0$ and $s_1 > 0$ then if $s_2 > 0$ we have that $\left|\mathring{\nu}_{jt}^{Naive}\right| > \left|\mathring{\nu}_{jt}^{MFE}\right|$ for $s \in (0, \min\{s_1, s_2\}) \cup (\max\{s_1, s_2\}, \infty)$. Similarly, if $s_2 < 0$ then we have $\left|\mathring{\nu}_{jt}^{Naive}\right| > \left|\mathring{\nu}_{jt}^{MFE}\right|$ for $s \in (0, s_1)$. From here we can proceed by various cases to determine these intervals exactly.

Case 1 ($\mathring{\alpha} > 0, \mathring{M}_t > 0$): Here $s_1 > 1$ and when s_2 is positive it follows that $s_2 > s_1 > 1$. Therefore $(s_1^*, s_2^*) = (1, 1)$ for this case (with the same result when $s_2 < 0$).

Case 2 ($\mathring{\alpha} < 0, \mathring{M}_t < 0$): Here $s_1 < 1$ and when s_2 is positive it follows again that $s_2 > s_1$. Therefore $(s_1^*, s_2^*) = (s_1, s_2)$ for this case (with the same result when $s_2 < 0$).

Case 3 ($\mathring{\alpha} > 0, \mathring{M}_t < 0$): Here $s_1 < 1$, $s_2 > 0$, and $s_2 < s_1$. Therefore $(s_1^*, s_2^*) = (s_2, s_1)$ for this case.

Case 4 ($\mathring{\alpha} < 0, \mathring{M}_t > 0$): Here $s_1 > 1$, $s_2 > 0$, and $s_2 < s_1$. Therefore $(s_1^*, s_2^*) = (s_2, 1)$ for this case.

Collecting these cases and their results yields the theorem's statement. ■

Proof of Theorem 12. The decomposition implies that

$$\left|\mathring{\eta}_{jkt}^{MFE}\right| - \left|\mathring{\eta}_{jkt}^{Naive}\right| = \frac{1}{1 + \mathring{M}_t}\left(\left|\mathring{M}_t\right| - \left|\mathring{\alpha} - \mathring{M}_t\right|\right).$$

and the result immediately follows.

■

Chapter 3

Instrumental Variables Estimation in the Presence of Outcome Attrition

(with Felix Elwert)

Chapter Summary

Instrumental variables (IV) methods are a ubiquitous tool for estimating causal effects. However, when data are subject to missingness the exclusion restriction can be violated leading to significant bias in IV estimators. This work proposes a new method, termed the missingness instrumental variables (MIV) estimator, to recover causal effects in the presence of outcome attrition. The method leverages statistical independences to replace the infeasible moments of the IV estimator with moments that can be estimated using data subject to missingness. Just like IV methods with complete data, MIV is able to estimate many causal effects of interest including average treatment effects, local average treatment effects, and marginal treatment effects. The method is compared with inverse probability weighting methods and multiple imputation methods, and Monte Carlo simulations highlight how MIV fares better than alternative methods when positivity is violated or under misspecification of error distributions.

3.1 Introduction

Causal inference is essential for decision making in politics, business, and health care, and instrumental variables analysis is a primary tool for estimating causal effects in empirical research. An instrumental variable isolates an element of random variation in treatment receipt, which is then used to construct comparable treated and untreated groups. In observational studies, instrumental variables can remove the hidden bias that results from non-random assignment of treatment ("unmeasured confounding"). In randomized experiments, they can remove the bias that results from the breakdown of randomization when individuals refuse their originally assigned treatment ("non-compliance").

Instrumental variables estimation, however, is highly vulnerable to bias from missing data. This bias can be especially hard to quantify when some values of the outcome are missing as a consequence of the treatment ("treatment-induced attrition," or loss to follow up) (Elwert and Segarra 2020). Missing data and attrition are a common problem in observational studies (Groves 2006; Curtin et al. 2005; de Leeuw and de Heer 2002) and randomized experiments (Hewitt et al. 2010). While the amount of missingness naturally varies across data sets, even the use of administrative data that are supposed to capture entire populations is no panacea.

For illustration, consider three well-known instrumental-variables applications, for which, to our knowledge, the problem of missing data has not previously been examined analytically.

- Draft Lottery and Wages: Angrist (1990) analyzed the causal effect of Vietnam-War ear military service (treatment) on later civilian earnings (outcome), using random variation in service induced by a draft-lottery as an instrumental variable. Missing data exist because earnings are only measured for men who survive the war. Hence, outcomes are likely missing as a function of treatment (treatment-induced attrition).

- Judges and Recidivism: Aizer and Doyle (2015) analyze the causal effect of youthful incarceration (treatment) on adult (re-)incarceration (outcome), using plausibly-random

assignment of juvenile offenders in Chicago to judges with differing sentencing propensities (instrument). Treatment-induced attrition is likely, since youthful incarceration likely affects geographic mobility, and the outcome is only ascertained within the state of Illinois (25% missingness).

- Schooling and Smoking: Hughes et al. (2019) study the effect of an additional year of schooling (treatment) on smoking behavior (outcome) using a policy change in the mandatory schooling age as an instrument. The authors utilize the UK Biobank Survey data which has a response rate of only 5.5%, meaning 94.5% missingness in the outcome. They further find that educational attainment significantly predicts survey response.

Much recent methodological work has explored the vulnerability of instrumental variables analysis to various forms of missing data. Most of this work merely notes the existence of bias under various missingness scenarios (Swanson et al. 2015; Ertefaie et al. 2016), or explores the size of the bias using stylized simulations (Canan et al. 2017; Gkatzionis and Burgess 2019; Hughes et al. 2019). Empirical work typically excludes all cases with any missing data ("casewise deletion," "sample selection," or "sample truncation") and proceeds as if analyzing complete data. Elwert and Segarra (2020) derive the first analytic bias expressions for instrumental variables estimation with truncated samples and show that, even in relatively simple scenarios, sign and size of instrumental-variable bias induced by missingness can be large and hard to predict.

Prior work on analytic solutions for missing data in instrumental variables research is relatively limited. Existing work mostly considers inverse probability weighting and multiple imputation. While valuable, these approaches have limitations. Inverse weighting requires ancillary statistical models for the missingness process; and popular multiple imputation approaches rely on strong distributional assumptions. Violation of these assumptions can lead to new biases. Neither approach appears to be especially widely used in practice.

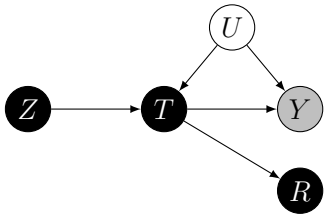
The work undertaken here seeks to introduce a new estimators for handling missing outcome data in instrumental variables analysis. We call this new estimator the *Missingness Instrumental Variables* (MIV) estimator. Intuitively, the new MIV estimator reformulates the conventional instrumental variables estimator by exploiting statistical independencies in the missing-data process. Mechanically, the new estimator replaces the moment conditions on which the conventional instrumental variables estimator relies in the absence of missing data with adjusted moment conditions that are valid in the presence of missing data.

The remainder of this chapter proceeds as follows. Section 3.2 presents a simple model to illustrate the new estimator in a familiar empirical context. Section 3.3 presents the full model of outcome attrition and how instrumental variables can be invalidated by the missing data process. Section 3.4 introduces recovery covariates, establishes identification, and defines the full MIV estimator. Section 3.5 discusses graphical causal models and their utility in identifying recovery covariates. Section 3.6 considers alternative estimators and compares their assumptions while section 3.7 illustrates how all of the potential estimators fare in Monte Carlo simulations. Finally, section 3.8 concludes.

3.2 Illustrative Example

To motivate and illustrate the new method proposed in this paper, we start with a simple example. Consider the investigation of the effects of veteran status on long run earnings conducted in Angrist (1990). In that paper, the author sought to estimate the effect of veteran status, a binary variable we denote by T_i, on later life earnings, a continuous variable we denote by Y_i. The treatment effect, denoted τ, can be defined using either the potential outcomes framework (see Rubin (2005)) or directed acyclic graphs (see Pearl (1995) and Pearl (2009)), both of which are discussed in more detail in section . For pedagogical purposes, let us assume that the treatment effect is homogeneous across individuals, and the case of heterogeneous treatment effects will be considered in the full model discussed in section 3.3.

Identifying the causal effect of T_i on Y_i using purely observational data can be a challenge given the potential presence of unobserved confounders. For example, if individuals who enlist in the army are more likely to be self driven, and this drive also makes them more likely to be higher earners later in life, then statistical associations between T_i and Y_i include not only the direct causal effect, but also the effect of this confounder, self-drive. To tease out the causal effect, the author leverages an *instrumental variable*, denoted by Z_i, which essentially provides a source of exogenous variation in the treatment to mimic random assignment. One of the instruments used in Angrist's work is the binary variable which indicates whether an individual's draft lottery number was called at some point during the draft. The random nature of the lottery...



(a)

$$
\begin{aligned}
Z &= \varepsilon_Z \\
U &= \varepsilon_U \\
T &= \mathbb{1}\{\pi Z + U + \varepsilon_T > 0\} \\
R &= \mathbb{1}\{\gamma T + \varepsilon_R > 0\} \\
Y &= \tau T + U + \varepsilon_Y = \tau T + \omega_Y
\end{aligned}
$$

(b)

Figure 3.1: Illustrative model of an instrumental variables (IV) framework equivalently displayed as a causal graph (a) and as a linear structural equations model (b). In the context of Angrist (1990), Y is later life earnings, T is veteran status, Z is draft lottery assignment, U represent unobservables (e.g. self-drive), and R_y indicates survival (i.e. whether the researcher is able to observe earnings). All idiosyncratic errors, $\varepsilon_Z, \varepsilon_U, \varepsilon_T, \varepsilon_R$, and ε_Y, are exogenous and pairwise independent. Note that the unobserved error in Y_i is defined as $\omega_Y = U + \varepsilon_Y$.

Figure 3.1 presents one simple characterization of the model described here. While figure 3.1b defines the model as a system of linear structural equations, 3.1a presents the model as a directed acyclic graph (DAG). For the purposes of this toy model the graphical causal diagram is not strictly necessary, however those familiar with DAGs will find them incredibly useful tools for quickly determining conditional independencies, which will prove indispensable in later discussions. For readers unfamiliar with graphical causal diagrams, a primer is

presented in section 3.5.1.

In order for Z_i to be a valid instrument, it must satisfy two conditions termed relevance and exclusion. Instrument relevance requires that $\text{Cov}(Z_i, T_i) \neq 0$, which is clearly satisfied upon inspection of either representation in Figure 3.1. The instrument exclusion restriction requires that $\text{Cov}(Z_i, \omega_Y) = 0$, which also holds in the above model, though perhaps takes a mite more work to show. Both of these conditions are discussed in more detail in section 3.4 from both the econometric perspective and the graphical causal model perspective. With both conditions satisfied, we can identify the causal effect using the fact that

$$\tau = \frac{\text{Cov}(Z_i, Y_i)}{\text{Cov}(Z_i, T_i)} = \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[T_i|Z_i = 1] - \mathbb{E}[T_i|Z_i = 0]}. \tag{3.1}$$

The sample analog of either expression[1] is termed the instrumental variables (IV) estimator, $\widehat{\tau}_{IV}$, for the causal effect of T_i on Y_i. With a random sample of observations, $\widehat{\tau}_{IV}$ is a consistent estimator for τ, the causal effect can be easily estimated, and all the standard IV theory comes into play.

The potential problem in this model, and the motivational crux for the work undertaken in this paper, is that the researcher may not observe all variables for every individual in the sample. In particular, since earnings are measured later in life after the conclusion of the Vietnam War it may not be the case that all individuals in our sample were alive (or present in the U.S.) for their earnings to be measured. Let R_i be 1 if the earnings of individual i are observed in the sample and 0 otherwise. Given the potentially dangerous nature of deployment it is reasonable that the likelihood of survival is affected by enlistment which is represented by T_i having a direct effect on R_i (i.e. an arrow from T_i to R_i).

Using the complete case data, i.e. the subset of observations for which we observe all variables, results in issues for the standard IV estimator. These issues stem from the fact that the instrument is no longer valid in the conditional subsample where $R_i = 1$. While

[1]The first representation (the ratio of covariances) is the usual form of the IV estimator for the simple univariate case, while the second representation is used by Angrist (1990) and follows due to the binary nature of our instrument.

the relevance of Z_i still holds, $\text{Cov}\left(Z_i, T_i | R_i = 1\right) \neq 0$, the exogeneity condition is violated, $\text{Cov}\left(Z_i, \omega_Y | R_i = 1\right) \neq 0$. This violation can be seen algebraically,

$$\begin{aligned}\text{Cov}\left(Z_i, \omega_Y | R_i = 1\right) &= \text{Cov}\left(Z_i, U_i + \varepsilon_Y | \pi Z_i + U_i + \varepsilon_T > 0\right) \\ &= \text{Cov}\left(Z_i, U_i | \pi Z_i + U_i + \varepsilon_T > 0\right) \neq 0,\end{aligned}$$

or in the DAG since conditioning on descendants of colliders opens previously closed non-causal pathways (this type of graphical analysis is discussed in section 3.5.1). This violation thus implies that $\widehat{\beta_{IV}}$ on the observed subsample will not be a consistent for the causal effect, τ.

Though the standard IV estimator is inconsistent for the causal effect in this conditional sample, we can still recover the moments necessary to identify the causal effect. Referring back to equation (3.1), if we were able to recover all of these moments using data from the subsample, then we could construct a new estimator which was consistent for τ. Inspection of the DAG or structural equations show that R_i is independent of Y_i after conditioning on Z_i and T_i[2]. This conditional independence implies that

$$\mathbb{E}\left[Y_i | T_i, Z_i, R_i = 1\right] = \mathbb{E}\left[Y_i | T_i, Z_i\right]$$

which can be coupled with the law of iterated expectation to show that

$$\begin{aligned}\mathbb{E}\left[Y_i | Z_i = z\right] =& P(T_i = 1 | Z_i = z)\mathbb{E}\left[Y_i | T_i = 1, Z_i = z, R_i = 1\right] \\ &+ P(T_i = 0 | Z_i = z)\mathbb{E}\left[Y_i | T_i = 0, Z_i = z, R_i = 1\right]. \qquad (3.2)\end{aligned}$$

This relationship is key since we have transformed the moment on the right, which we do not observe directly but need for equation (3.1), into a function of moments on the left hand side that are all observed in the data.

The relationship derived in (3.2) forms the basis of the new estimator we are proposing in this work. For notational brevity denote the conditional probability as $p_{t|z} = P(T_i =$

[2]Algebraically, we have $\text{Cov}\left(R_i, Y_i | Z_i = z, T_i = t\right) = \text{Cov}\left(\mathbb{1}\{\gamma t + \varepsilon_R > 0\}, \tau t + \varepsilon_U + \varepsilon_Y | Z_i = z, T_i = t\right) = 0$ since $\varepsilon_R, \varepsilon_U$, and ε_Y are all pairwise independent.

$t|Z_i = z)$ and the conditional mean as $y_{tzr} = \mathbb{E}\left[Y_i|T_i = t, Z_i = z, R_i = r\right]$. Combining (3.1) with (3.2) yields

$$\tau = \frac{\mathbb{E}\left[Y_i|Z_i = 1\right] - \mathbb{E}\left[Y_i|Z_i = 0\right]}{\mathbb{E}\left[T_i|Z_i = 1\right] - \mathbb{E}\left[T_i|Z_i = 0\right]} = \frac{p_{1|1}y_{111} + p_{0|1}y_{011} - p_{1|0}y_{101} - p_{0|0}y_{001}}{p_{1|1} - p_{1|0}}$$

which shows that we can still identify τ despite only observing Y_i for those with $R_i = 1$, i.e. observing earnings only for those individuals that survive to have earnings. Plugging in the sample analogs of each of these objects yields the *missingness instrumental variables* (MIV) estimator defined here as

$$\widehat{\tau}_{MIV} = \frac{\widehat{p}_{1|1}\widehat{y}_{111} + \widehat{p}_{0|1}\widehat{y}_{011} - \widehat{p}_{1|0}\widehat{y}_{101} - \widehat{p}_{0|0}\widehat{y}_{001}}{\widehat{p}_{1|1} - \widehat{p}_{1|0}}.$$

For clarity's sake, in our given context $\widehat{p}_{0|1}$ would be the proportion of individuals who did not enlist ($T_i = 0$) among those whose draft lottery number was called ($Z_i = 1$), while $\widehat{y}_{101}$ is the average earnings of individuals who enlisted ($T_i = 1$), whose draft lottery numbers were not called ($Z_i = 0$), and who survived to have their earnings measured later in life ($R_i = 1$). Since the population moments of $\widehat{\tau}_{MIV}$ recover the moments found in equation (3.1) using the subsample with missing earnings data, we have thus constructed an estimator that is consistent for the causal effect, τ, and robust to missing outcome data provided it follows the model defined in Figure 3.1.

The model and estimator presented in this section may be restrictive, however, the general approach is amenable to a wider array of situations. The remainder of the paper will present a model and associated MIV estimator which permits continuous or binary variables and additional covariates. Furthermore, the methods discussed will allow for outcome attrition that is not solely driven by treatment (as in the model presented here) provided the proper conditional associations exist, which is discussed at length in sections 3.4 and 3.5. Finally, while we limited ourselves to homogeneous treatment effects, the full model can encompass any causal effects of interest that are identified by traditional IV analysis which include heterogeneous treatment effects.

3.3 Model

Let Y_i and T_i be the outcome and treatment respectively for individual i, both of which may be discrete or continuous random variables. Let X_i be a k dimensional vector of covariates which includes T_i as an element, and let Z_i be an r dimensional vector of instruments (with $r \geq k$), to be defined shortly. Given a random sample of n individuals, let the $n \times k$ and $n \times r$ matrices, $\mathbf{X}$ and $\mathbf{Z}$, be formed from stacking the individual X_i's and Z_i's respectively. While X_i and Z_i are observable for all n individuals, the outcome is only observable for a subset, $n^* \leq n$, of the full sample. Let $R_i = 1$ indicate that the outcome of individual i is observed while $R_i = 0$ indicates the outcome is unobserved. Similarly let the $n \times 1$ matrix, $\mathbf{Y}$, be formed from stacking the observed Y_i. Let an asterisk denote the analogous matrix of variables that includes only those observations for which $R_i = 1$. Thus $\mathbf{X}^*, \mathbf{Z}^*$, and $\mathbf{Y}^*$ are $n^* \times k, n^* \times r$ and $n^* \times 1$ matrices respectively.

The end goal of the researcher is to identify and estimate the causal effect of the treatment on the outcome. We follow the potential outcomes framework of causality (Rubin 1974), where $Y_i(t)$ denotes the potential outcome defined to be the outcome that would be observed if individual i were exposed to treatment $T = t$. If the treatment is binary then the relationship between the observed outcome and potential outcomes in this framework is given by $Y_i = T_i Y_i(1) + (1 - T_i)Y_i(0)$. The causal effect, denoted τ_i, of a binary treatment on the outcome for individual i is given by the difference in potential outcomes: $\tau_i = Y_i(1) - Y_i(0)$. Causal effects of multi-valued and continuous treatments can be defined in analogous manners, however these effects may not be captured by a scalar as they are with a binary instrument. Unfortunately, individual-level causal effects cannot be observed directly, because one never observes more than one of any individual's potential outcomes. This hurdle is commonly referred to as the "fundamental problem of causal inference" (Holland 1986) and constitutes the primary difficulty that modern causal inference attempts to overcome.

Treatment effect heterogeneity is a common characteristic of many models and empiri-

cal endeavors in the social sciences. Therefore, investigators often pursue aggregate causal effects that are averaged across the entire population or across specific sub-populations. For example, the average treatment effect (ATE) averages across the entire population and is given by $\tau_{ATE} = \mathbb{E}[\tau_i]$. In other settings, analysts are interested in the average treatment effect for members of some group, defined by observed or unobserved variables $V = v$, $\tau(v) = \mathbb{E}[\tau_i|V_i = v]$. The local average treatment effect (LATE), which averages τ_i over those individuals whose treatment uptake is affected by their instrument exposure, is one such effect of interest (see Imbens and Angrist (1994) and Imbens (2010)). The effect of the treatment on the treated (ETT), defined as $\mathbb{E}[\tau_i|T_i = 1]$, is yet another aggregate effect that is especially relevant in the medical and policy literature.

Many of the most common treatment effects, including all those mentioned here, can be identified using instrumental variables methods under suitable assumptions. The flexibility of the identifying power of this estimator, which will be defined fully in the next section, is one reason we focus on it in our current endeavors. For the same reason, we leave the exact treatment effect of interest ambiguous so as to encompass any of the situations that fall under the purview of the instrumental variables estimator. In particular, let β represent the $k-$dimensional vector of causal effects of X_i on Y_i and let τ represent the specific causal effect of T_i on Y_i (i.e. if T_i is the mth element of X_i then $\tau = e_m'\beta$).

3.4 Identification and MIV

In the model described in the previous section there are two hurdles that must be overcome in order to identify the causal effect of interest. First there is the generic problem of identifying a causal effect in the presence of unobserved confounding, which is often surmounted using a method such as the instrumental variables estimator. The second issue concerns the potentially non-random missingness in the outcome variable which we will resolve through appropriate conditioning variables that allow us to swap out unobserved moments for ob-

servable moments. This section describes these two aspects and how they can be combined to identify causal effects in the presence of outcome attrition.

3.4.1 Instrumental Variables

In order to overcome the fundamental problem of causal inference, investigators typically attempt to compare the outcomes of treated and untreated individuals that are comparable in all respects except the treatment they have received (Holland 1986, Imbens and Rubin 2015). *Ideal* randomized experiments produce comparability by randomizing treatment receipt, so that potential outcomes are statistically independent of treatment, $Y_i(t) \perp T_i$, known as "ignorability". In observational studies, where treatment receipt is not randomized, investigators often assert as-if random assignment conditional on a set of observed control variables, X_i, that are assumed to govern treatment receipt, so that $Y_i(t) \perp T_i|X_i$ ("conditional ignorability"), which creates comparable groups within strata defined by $X = x$.

In many settings, however, neither ignorability nor conditional ignorability are fully credible. Even in real randomized experiments, treated and untreated groups may be incomparable because individuals refuse to take their assigned treatment or otherwise break randomization ("non-compliance"). In observational studies, analysts may lack data to fully adjust for the necessary covariates X_i such as the unobserved confounder of self-drive discussed in section 3.2. In these scenarios, popular regression, matching, and weighting estimators are liable to exhibit bias.

Instrumental variables (IV) analysis is one of the primary tools for estimating causal effects when neither ignorability nor conditional ignorability hold (e.g., Angrist et al. 1996; Wooldridge 2011; Imbens and Rubin 2015). This method works by exploiting an element of random assignment in treatment receipt in order to construct comparable treated and untreated groups. The validity of IV methods rests on two main assumptions that are often termed instrument relevance and instrument exogeneity which are defined below. Though this definition is the multivariate extension of the assumptions that were discussed in section

3.2, the intuition behind each assumption is the same. Relevance (E1) ensures that the instrument has some association with the treatment and is equivalent to the requirement that $\text{Cov}(Z_i, T_i) \neq 0$ in the univariate case (i.e. when $r = k = 2$ and one of those variables is a constant). The exclusion restriction (E2) requires that our instrument is uncorrelated with the structural error of the outcome variable. An alternative interpretation of this assumption is that it requires that the only association between Z_i and Y_i must must be mediated by the treatment T_i.

Definition 3. An $r-$dimensional variable, Z_i, is an *instrumental variable* for the causal effect of the $k-$dimensional variable X_i on Y_i if,

E1 (Relevance): $\mathbb{E}[Z_i' X_i]$ has full column rank,

E2 (Exclusion): $\mathbb{E}[Z_i' \omega_Y] = 0$.

With a valid instrument in hand we can define the two stage least squares estimator which is so often utilized to estimate causal effects.

$$\widehat{\beta}_{2SLS} = \left(\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{Y},$$

If T_i is the mth element of X_i then the estimate of the causal effect of interest is given by $\widehat{\tau}_{2SLS} \equiv e_m' \widehat{\beta}_{2SLS}$. Furthermore, when the model is just identified ($k = r$) the estimator reduces to the familiar instrumental variables estimator as a special case,

$$\widehat{\beta}_{IV} = \left(\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{Z}'\mathbf{Y},$$

as seen in our earlier motivating example. Let the probability limit of this estimator, denoted β, be the causal effect of interest

$$\beta \equiv \text{plim}\,\widehat{\beta}_{2SLS} = \left(\mathbb{E}[X_i' Z_i]\,\mathbb{E}[Z_i' Z_i]^{-1}\,\mathbb{E}[Z_i' X_i]\right)^{-1}\mathbb{E}[X_i' Z_i]\,\mathbb{E}[Z_i' Z_i]^{-1}\,\mathbb{E}[Z_i' Y_i]. \tag{3.3}$$

As discussed in the previous section, β can represent several different types of causal effects depending on the specific context and additional assumptions.

3.4.2 Recovery from Missingness

Even though the instrumental variables estimator is consistent for the causal effect of interest, this estimator is infeasible in the model described here. The estimator, and it's subsequent consistency, rely on observing all variables for the entire sample, while our context has the researcher only observing Y_i for those with $R_i = 1$. From an identification standpoint we observe all the constituent moments found in equation (3.3) except for $\mathbb{E}[Z_i'Y_i]$. Unless $\mathbb{E}[Z_i'Y_i|R_i] = \mathbb{E}[Z_i'Y_i]$, which is generally false without further assumptions, then the 2SLS estimator will not be consistent for β.

Following our approach in section 3.2, our goal is to find a way to replace the infeasible moments found in equation (3.3) with feasible moments. If the missingness, R_i, were completely independent of all observed and unobserved variables (i.e. missing completely at random, MCAR) then $\mathbb{E}[Z_i'Y_i|R_i = 1] = \mathbb{E}[Z_i'Y_i]$ and then $\widehat{\beta}_{2SLS}$ would still be consistent even when using the selected sub-sample. Assuming that the missingness follows an MCAR mechanism is quite strict, so instead we opt to find sets of covariates which, when conditioned on, make the missingness independent of the outcome.

Definition 4. A set of observable variables, V_R, are called *recovery covariates* for the instrument Z_i if $R_i \perp Y_i|Z_i, V_R$.

Finding recovery covariates can be just as tricky as finding suitable instruments, so section 3.5 provides some exploration of the multiple features to consider. Since conditional independences are at the heart of these recovery covariates, we also provide a primer on graphical causal models (i.e. DAGs) as they are especially well-suited for quick determination of conditional independences in a given model.

One important aspect to note regarding the definition of recovery covariates is that they depend not only on the underlying missingness model, i.e. the relationship between R_i and Y_i, but also on the specific instruments being leveraged. Therefore, a set of recovery covariates that is valid for one instrument may not be valid for another instrument, even

within the same model. This facet is discussed further in section 3.5.

One final caveat concerns the untestable nature of recovery covariates. Just like the exclusion restriction of an instrumental variable, the very nature of the missingness implies that we cannot test whether $R_i \perp Y_i | Z_i, V_R$ holds for a given set of covariates. Rather, these assumptions must be argued and debated based on the researcher's knowledge of the empirical context. Given the difficulty of conceptualizing conditional independence in complex systems, this factor lends all the more justification for leveraging DAGs to aid in the deliberation of a given set of recovery covariates.

Given a set of recovery covariates we are now in a position to handle the missingness and identify the causal effect of interest. As alluded to earlier, we are able to identify β, even in the sample with outcome attrition, precisely because the recovery covariates allow us to substitute the infeasible moments, i.e. $\mathbb{E}[Z_i' Y_i]$, with functions of moments that we can identify. The proof of this identification result is quick and fundamentally only relies on the law of iterated expectations.

Theorem 13 (Identification)**.** *Let Z_i be a valid instrument for identifying the causal effect, β, of X_i on Y_i in a fully observed sample, and let V_R be a set of observed recovery covariates with respect to Z_i and the missingness process of Y_i. Then β is identified in a sample where Y_i is subject to missingness.*

Proof. Using the law of iterated expectations and the definition of recovery covariates we have

$$\mathbb{E}[Z_i' Y_i] = \mathbb{E}[\mathbb{E}[Z_i' Y_i | Z_i, V_R]] = \mathbb{E}[Z_i' \mathbb{E}[Y_i | Z_i, V_R]] = \mathbb{E}[Z_i' \mathbb{E}[Y_i | Z_i, V_R, R_i = 1]].$$

The result immediately follows from substitution into equation (3.3),

$$\begin{aligned}\beta &= \left(\mathbb{E}[X_i' Z_i]\,\mathbb{E}[Z_i' Z_i]^{-1}\,\mathbb{E}[Z_i' X_i]\right)^{-1} \mathbb{E}[X_i' Z_i]\,\mathbb{E}[Z_i' Z_i]^{-1}\,\mathbb{E}[Z_i' Y_i] \\ &= \left(\mathbb{E}[X_i' Z_i]\,\mathbb{E}[Z_i' Z_i]^{-1}\,\mathbb{E}[Z_i' X_i]\right)^{-1} \mathbb{E}[X_i' Z_i]\,\mathbb{E}[Z_i' Z_i]^{-1}\,\mathbb{E}[Z_i' \mathbb{E}[Y_i | Z_i, V_R, R_i = 1]]\end{aligned}$$

since all expectations in the final expression are identified in the sample subject to missingness. ■

3.4.3 Estimation

With identification of the causal effect nailed down we can turn to the complete definition of the new estimator which is one of the main contributions of this work. Estimation will prove fairly straightforward and will only involve one extra step from the standard methods used when implementing the 2SLS estimator. At the heart of the new estimator will be a conditional mean function which we briefly discuss followed by a description of the estimator implementation.

Our work will be made more succinct if we first define the conditional expectation function, $g_y(\cdot)$, as follows

$$g_y(z, v) = \mathbb{E}\left[Y_i | Z_i = z, V_R = v\right],$$

and let G_i denote the value of this conditional mean for individual i, $G_i \equiv g_y(Z_i, V_i)$. As with our previous notation, let $\mathbf{G}$ denote the $n \times 1$ vector which stacks the individual values of G_i.

The full estimator, which we call the missingness instrumental variables (MIV) estimator is given by

$$\widehat{\beta}_{MIV} = \left(\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{G}. \tag{3.4}$$

Computation of this estimator can essentially be broken up into two easy steps; first the researcher generates $\mathbf{G}$ and second they compute the 2SLS estimator using $\mathbf{G}$ instead of $\mathbf{Y}$.

The first step, generation of $\mathbf{G}$, bears further discussion given its importance. In the rare event that the function g_Y is known then the generation of G_i is a simple application the function to the observed data, Z_i and V_i, in the full sample.

More often than not this function is not known beforehand, however it can be estimated.

When V_R is a valid set of observed recovery covariates, then g_y is nonparametrically identified because

$$g_y(z, v) = \mathbb{E}\left[Y_i | Z_i = z, V_R = v, R_i = 1\right]$$

and this conditional mean is nonparametrically identified from the selected sample. Therefore, this function can be estimated on the sub-sample of observed outcomes, where $R_i = 1$, but then applied to the complete sample since Z_i and V_R are observed for all individuals. Researchers can use the Nadaraya-Watson kernel regression estimator or any another nonparametric estimator to estimate g_y (see Henderson and Parmeter 2015, Li and Racine 2006, or Härdle and Linton 1994). For example, the multivariate Nadaraya-Watson estimator, $\widehat{g_y}$, of g_y is given by

$$\widehat{g_y}(z, v) \equiv \frac{1}{n^* h_n^{r+m} \widehat{f}_{Z,V}(z, v)} \sum_{i=1}^{n^*} Y_i \left[\prod_{d=1}^{r} K\left(\frac{z_d - Z_{di}}{h_n}\right) \prod_{d=1}^{m} K\left(\frac{v_d - V_{di}}{h_n}\right)\right]$$

where Z_i has dimension r, V_R has dimension m, h_n is the sample size specific bandwidth, $K(\cdot)$ is a kernel function[3], and $\widehat{f}_{Z,V}(z, v)$ is the kernel density estimator of the joint distribution of Z_i and V_R.

After estimating this function, a researcher can apply it to the observed complete data and implement the full MIV estimator. If $\widehat{g_y}$ denotes the estimator of g_y then the full MIV estimator becomes

$$\widehat{\beta}_{MIV} = \left(\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\widehat{\mathbf{G}}, \tag{3.5}$$

where $\widehat{\mathbf{G}}$ is the $n \times 1$ vector of estimates of G_i, $\widehat{G_i} = \widehat{g_y}(Z_i, V_i)$, over the whole sample.

Consistency of $\widehat{\beta}_{MIV}$ follows from the consistency of $\widehat{g_y}$ for g_y, and standard assumptions associated with the estimation of nonparametric conditional mean functions. In particular, for the Nadaraya-Watson estimator typical assumptions include the following:

Assumption A13 (Sufficient conditions for consistency of $\widehat{g_y}$)**.**

[3]Any valid probability density function (pdf) can serve as a kernel function, however they are typically chosen so that they are the pdfs of distributions which are mean zero, symmetric, and bounded.

1. $\int |K(u)|du < \infty$ *and* $\lim_{|u|\to\infty} uK(u) = 0$.

2. $g_y, f_{Z,V}$, *and* $\sigma(z,v) = Var(Y - g_y(z,v)|Z = z.V = v)$ *are continuous functions.*

3. $f_{Z,V}(z,v) > 0$.

4. $\lim_{n\to\infty} h_n = 0$ *and* $\lim_{n\to\infty} nh_n = \infty$.

Refer to Härdle and Linton (1994) and Schuster (1972) for further discussion of pointwise consistency and other asymptotic properties of $\widehat{g_y}$.

Altogether, the previous assumptions and above assumptions imply that the MIV etimator is consistent for the causal effect, β. The theorem statement and straightforward proof follow below.

Theorem 14 (Consistency)**.** *Let Z_i be a valid instrument for identifying the causal effect, β, of X_i on Y_i in a fully observed sample, and let V_R be a set of observed recovery covariates with respect to Z_i and the missingness process of Y_i. Under the assumptions of A13 we have that $\widehat{\beta}_{MIV} \to_p \beta$.*

Proof. Leveraging the weak law of large numbers and the consistency of $\widehat{g_y}$ (from Theorem 1 in Härdle and Linton (1994)) we have that

$$\begin{aligned}
\widehat{\beta}_{MIV} &= \left(\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\widehat{\mathbf{G}} \\
&\to_p \left(\mathbb{E}\left[X_i'Z_i\right]\mathbb{E}\left[Z_i'Z_i\right]^{-1}\mathbb{E}\left[Z_i'X_i\right]\right)^{-1}\mathbb{E}\left[X_i'Z_i\right]\mathbb{E}\left[Z_i'Z_i\right]^{-1}\mathbb{E}\left[Z_i'G_i\right] \\
&= \left(\mathbb{E}\left[X_i'Z_i\right]\mathbb{E}\left[Z_i'Z_i\right]^{-1}\mathbb{E}\left[Z_i'X_i\right]\right)^{-1}\mathbb{E}\left[X_i'Z_i\right]\mathbb{E}\left[Z_i'Z_i\right]^{-1}\mathbb{E}\left[Z_i'Y_i\right] = \beta.
\end{aligned}$$

The last line follows from the properties of the recovery covariates which imply that $\mathbb{E}\left[Z_i'G_i\right] = \mathbb{E}\left[Z_i'Y_i\right]$. ■

3.5 Choosing Valid Recovery Covariates

Finding sets of variables that form valid recovery covariates is at the core of what makes the MIV estimator function correctly. Just like the search for a valid instrument, this task can be challenging. Therefore, in this section we provide tools and examples to aid researchers in this endeavor. First, we provide a primer on graphical causal models because they are acutely valuable for assessing the implied conditional independences which define recovery covariates. Then, we follow with multiple examples of recovery covariates across a range of models to illustrate the various factors researchers should keep in mind.

3.5.1 Primer on Graphical Causal Models

We integrate our econometric presentation with graphical causal models (Pearl 2009 and Maathuis et al. 2018, Shpitser 2018) in order to transparently display the causal assumptions and derive implied statistical independences on which the proposed work relies (Brito 2010), and to draw out new links with relevant recent work in biostatistics and computer science (e.g., Mohan and Pearl 2014, 2018).

Let $\mathcal{G}(\mathbf{V}, E)$ be a directed acyclic graph (DAG, Pearl 2009), where $\mathbf{V}$ is a set of nodes, and E is a set of directed edges. When $\mathbf{V}$ represents the variables and E represents the causal effects in the DGP, then the DAG $\mathcal{G}$ is called a *causal graph*. The causal graph can be interpreted as a *non-parametric structural equation model* (NPSEM), in which each variable $V \in \mathbf{V}$ is generated by a structural equation $V := f_V(pa(V), \varepsilon_V)$, where f_V is a possibly unknown function that takes as arguments the *parent set*, $pa(V)$, of V that contains all variables in $\mathbf{V}$ with a directed edge into V and a random (exogenous) shock, ε_V, representing all influences on V not captured in the graph. By convention, causal graphs do not explicitly display the idiosyncratic shocks ("error terms"), ε_V, that affect individual variables V; "correlated errors" between variables are notated as shared dependence on an unmeasured variable, e.g., $V_1 \leftarrow U \rightarrow V_2$. The NPSEM is non-parametric in the sense that

it imposes no restrictions on the distribution of the variables $\mathbf{V}$ or the functional form of the causal effects E.

The causal graph links to potential outcomes via the NPSEM. The potential outcomes for any variable V with respect to a hypothetical intervention on any parent(s) of V are generated by inserting specific value(s) for $pa(V)$ into $f_V(.)$. Potential outcomes of V with respect to hypothetical intervention(s) on non-parents of V are generated through recursive substitution (Shpitser 2018).

Under mild conditions, the graphical rules of *d-separation* determine the statistical independences between variables in data generated according to $\mathcal{G}$ (Pearl 1988, Verma and Pearl 1988, Geiger et al. 1990). The notions of paths, collider variables, and descendants play a central role in these rules. A *path* is an acyclic sequence of adjacent edges between two variables, regardless of the direction of the arrows. In a *causal path* from treatment to outcome, all edges point toward the outcome. In a *non-causal*, or *spurious*, path between treatment and outcome, at least one edge points away from the outcome. A variable is called a *collider* with respect to a specific path if it receives two inbound edges on the path. For example, T is a collider on the path $Z \rightarrow T \leftarrow U \rightarrow Y$. The *descendant set* of a variable V, $de(V)$, contains all variables directly and indirectly caused by it.

Definition 5. Variables $X_1, X_2 \in \mathbf{V}$ are *d-separated* conditional on a set of variables $W \subseteq \mathbf{V}$ if all paths between X_1 and X_2 are closed. A path is *closed* iff either (a) it contains a collider and neither the collider nor any of its descendants are in W, or (b) it contains a non-collider and that variable is in W. Conversely, two variables are *d-connected* if there is at least one open path between them. A path is *open* iff it is not closed.

The critical connection is the link between d-separation and conditional independence, as made concrete in the lemma below (see Geiger et al. 1990 for further discussion). Though conditional independencies are at the heart of many econometric methods, as they are here with respect to recovery covariates, they can be difficult to interrogate directly, even with a fully specified structural equations models. However, the causal connections between nodes in

a DAG align with the natural way researchers tend to conceptualize the relationships between real world variables. Checking for d-separation in a given causal DAG is straightforward, and since this implies conditional independence, these tools can aid researchers in their quest to determine suitable variables that have required statistical properties.

Lemma 3. *Under mild regularity conditions, if variables X_1 and X_2 are d-separated conditional on $W \subseteq \boldsymbol{V}$, then $X_1 \perp X_2 | W$.*

Importantly, when a path contains only one collider, then conditioning on this collider or any of its descendants, while not conditioning on any non-colliders on the path, opens the path. From these rules, it follows that each variable is statistically independent of its non-descendants conditional on its parent set, so that the joint distribution of variables in a causal graph factors as $p(\mathbf{V}) = \prod_{V \in \mathbf{V}} p(V|pa(V))$.

From the graphical causal model perspective, the definition of recovery covariates can also be stated using d-separation. The set of observable covariates $V_R \subseteq \mathbf{V}$ is a valid set of recovery covariates if R and Y are d-separated conditional on Z and V_R.

Mohan et al. (2013) introduced *m-graphs* to facilitate the graphical analysis of missing data problems. M-graphs are DAGs like those discussed above, except that they additionally notate the presence of, and reasons for, missingness. M-DAGs consist of four disjunct sets of variables, $\mathbf{V} = V_o \cup V_m \cup U \cup R$, where V_o denotes fully observed variables, V_m are partially missing variables, U are fully unobserved variables, and R are a set of missingness indicators, respectively. We display fully observed, partially observed, and fully unobserved variables with black, grey, and white nodes, respectively. We write $R_{Vi} = 1$ if the value of variable V is missing for individual i, and $R_{Vi} = 0$ if the value is observed. A directed edge from some variable V_S into R_{V_T} indicates that V_S caused missingness in V_T. Variables can cause their own missingness; for example, income causes its own missingness when high-earners are less likely to report their income than low earners.

To illustrate, in Figure 3.1a, variables Z_i and T_i are fully observed, whereas U_i is fully unobserved. By contrast, Y is only partially observed, and missingness in Y, R_Y is caused

by the fully observed variable T, possibly in addition to missingness caused by some random shock ε_{R_Y}, which, by convention, is not displayed in the graph. Typically in m-graphs there is more than a single variable that is partially observed, however since this paper only concerns models with outcome attrition we suppress the subscript on R_y and it is generally understood that R refers to missingness in Y_i.

3.5.2 Examples of Recovery Covariates

Having defined causal DAGs and how to use them to determine the statistical independences that they imply, we now turn to several example models to illustrate ways in which recovery covariates may or may not manifest.

Figure 3.2 presents three models which show that there may be one, multiple, or no sets of recovery covariates in a given scenario. In figure 3.2a, the same simple model considered in section 3.2, we have the simplest situation which has exactly one recovery covariate set which can be used, $V_R = \{T_i\}$. At the other extreme, the model in figure 3.2b reveals that there may not be any recovery covariates that can be leveraged. The problem exhibited here, mainly that there is an unobserved confounder, U_i, of Y_i and its missingness R_i, is a generic issue, and will always result in the inability to ever close off the open path $R \leftarrow U \rightarrow Y$. In these situations, it is not just the case that MIV will not be able to account for the missingness, but rather it is unlikely any other estimators will be able to help without further restrictions being placed on the model.

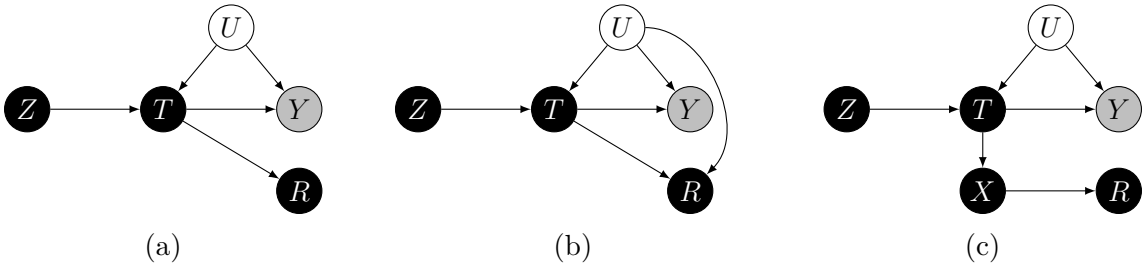


Figure 3.2: Examples of causal DAGs of various IV models subject to outcome attrition. These illustrate situations where there exist (a) one set of recovery covariates, (b) no sets of recovery covariates, or (c) multiple possible sets.

Moving our attention to Figure 3.2c we can consider a scenario where multiple sets of recovery covariates are available. If we were to think back to the draft lottery context in our motivating example, X could be a health measure that is tied more closely with veteran status (e.g. an indicator for post traumatic stress disorder). Since X_i mediates the effect of treatment on missingness, this admits three distinct recovery covariate sets: $V_R = \{T_i\}, V_R = \{X_i\}$, and $V_R = \{T_i, X_i\}$. Each of these V_R can be used to construct a distinct MIV estimator which is consistent for the treatment effect of interest.

The end choice among these estimators could be driven by variable types, efficiency motives, or robustness characteristics. For example, if variable X_i were continuous while T_i were binary, then accurate estimation of g_y could be more easily achieved when conditioning on the discrete variable T_i rather than worrying about the difficulties inherent in estimating a continuous conditional mean function. If efficiency concerns were more prevalent, an analysis similar to Rotnitzky and Smucler (2020) could be carried out to determine the minimum variance estimator among the choices. Finally, the existence of multiple estimators hints at the possibility of constructing a single estimator that leverages all the potential V_R and provides a robustness to misspecification. For example, estimators can be constructed so that only one of the g_y functions needs to be specified correctly in order to maintain the consistency of the estimator.

When defining recovery covariates it was alluded to that these sets are intimately tied to the missingness model and the specific instrument being leveraged in the infeasible estimator. Figure 3.3 provides two models that showcase this connection. In the first case, figure 3.3a, the model reveals why having a covariate which d-separates R_i and Y_i is insufficient on its own to function as a recovery covariate. While it is true that $R_i \perp Y_i | T_i$, additionally conditioning on the instrument, which is an important facet of the definition, breaks this independence: $R_i \not\perp Y_i | Z_i, T_i$. When conditioning on Z_i and T_i we in fact open up the path $R \leftarrow X_2 \rightarrow Z \leftarrow X_1 \rightarrow T \leftarrow U \rightarrow Y$. There are two valid sets of recovery covariates in this model, $V_R = \{T_i, X_1\}$ and $V_R = \{T_i, X_2\}$, and both close all pathways between R_i and Y_i

when conditioning on Z_i and V_R.

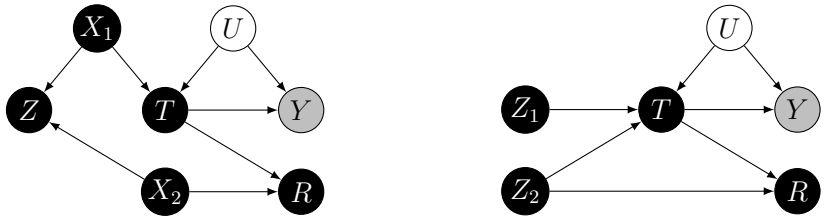


Figure 3.3: More examples of causal DAGs which illustrate how any recovery covariates are intimately tied to the instrument that is being leveraged.

This model is also noteworthy as it presents an IV context where Z_i is a valid instrument despite not having a direct effect on treatment. Moreover, a researcher who blindly includes all observed covariates as controls would find that Z_i becomes invalid conditional on X_1 since instrument relevance would no longer be satisfied.

Finally, Figure 3.3b shows how a single model with multiple possible valid instruments may have different recovery covariates associated with each instrument. Both Z_1 and Z_2 are valid instruments, each satisfying relevance and exclusion on their own. When exploiting Z_1 as the instrument there is one set of recovery covariates, $V_R = \{T_i, Z_{2i}\}$. In contrast, when leveraging Z_2 as the instrument there is a single different set of recovery covariates, $V_R = \{T_i\}$. These recovery sets lead to different g_y and thus distinct MIV estimators depending on which instrument you choose. Obviously this is a case where the model is overidentified, so the third option would be to include both instruments and take care with respect to the definition of the resulting MIV estimator[4].

3.6 Comparison with Alternative Estimators

Despite the ubiquity of IV methods and large literature on missing data, relatively little work connects these subjects beyond noting the potential biases described above. One explanation

[4]The current form of the MIV estimator in equation (3.5) requires V_R to be valid for every instrument in the vector Z_i, however it is possible to break this equation up and allow for distinct V_R for each instrument.

for lack the of work at this intersection is that it is difficult for a single method to handle two types of endogeneity simultaneously[5].

In this section we explore two sets of methods that can applied to instrumental variables with missing data: inverse probability weighting and multiple imputation. This analysis will proceed in the same spirit as Kennedy et al. (2019), but will provide guidance on when certain estimators will be preferable over others in the context of instrumental variables and outcome attrition. For each method we describe the approach and compare and contrast their core assumptions with that of the MIV estimator proposed in this work. Finally, we finish with Monte Carlo simulations to highlight how each of these three methods fare. For expository purposes, the following sections will all focus on the simple model of treatment-induced attrition represented by Figure 3.1 without any extra covariates.

3.6.1 Inverse Probability Weighting Estimators

Inverse probability weighting (IPW) methods comprise an enormous literature and enjoy wide application both in estimating causal effects (Angrist et al. (1996); Imbens and Rubin 2015) as well as accounting for missing data (Robins et al. 1994; Seaman and White 2011; Li et al. 2011). When being leveraged for causal identification IPW works by reweighting observations to account for non-random treatment assignment. For missing data IPW reweights observations to recover the moments of interest in the fully observed data from the sample of non-missing data. Since IV methods can be motivated from a simple moment condition, we will be using IPW to handle the missing data aspect of our problem.

In practice this is done by weighting the complete-case observations, i.e. individuals for which we observe all variables, by the inverse of the probability of being observed. For example, suppose the function $M(T_i, Y_i; \beta)$ identifies the parameter of interest, β^*, by satisfying

[5]Searching for methods that act at this intersection is made all the harder by the mistaken conflation of the two problems. Causal inference is sometimes framed as a missing data problem, because you don't observe the counterfactual outcome. Similarly, missing data and sample selection can be viewed as an endogeneity problem (see Heckman 1979). However the problems are distinct enough that addressing them simultaneously is nontrivial.

the moment restriction $\mathbb{E}[M(T_i, Y_i; \beta^*)] = 0$. Furthermore, suppose that T_i is always observed while Y_i is only partially observed (indicated by R_i). If we define $h(T_i) = P(R_i = 1|T_i)$ then using the law of iterated expectation we have that

$$\mathbb{E}\left[\frac{\mathbb{1}\{R_i = 1\}}{h(T_i)} M(T_i, Y_i; \beta)\right] = \mathbb{E}\left[\frac{\mathbb{E}[\mathbb{1}\{R_i = 1\}|T_i, Y_i]}{h(T_i)} M(T_i, Y_i; \beta)\right] = \mathbb{E}\left[M(T_i, Y_i; \beta)\right]. \quad (3.6)$$

This motivates the weighting estimator, $\widehat{\beta}_W$, that solves

$$\sum_i \left[\frac{\mathbb{1}\{R_i = 1\}}{h(T_i)} M(T_i, Y_i; \widehat{\beta}_W)\right] = 0 \quad (3.7)$$

which will consistently estimate β^* provided $h(T_i)$ is correctly specified.

Since $\widehat{\beta}_W$ throws out information by only using complete cases (i.e. individuals with $R_i = 1$), Li et al. (2011) show that this estimator can be augmented to leverage all available data to improve efficiency. In particular, if data are MAR then they define the efficient weighting estimator, $\widehat{\beta}_E W$, that solves

$$\sum_i \left[\frac{\mathbb{1}\{R_i = 1\}}{h(T_i)} M(T_i, Y_i; \widehat{\beta}_{EW}) + \left(1 - \frac{\mathbb{1}\{R_i = 1\}}{h(T_i)}\right) \mathbb{E}[M(T_i, Y_i; \widehat{\beta}_{EW})|T_i]\right] = 0. \quad (3.8)$$

This estimator also enjoys a double-robustness property in the sense that it is consistent for β^* if either $h(T)$ or $\mathbb{E}[M(T_i, Y_i; \widehat{\beta}_{EW})|T_i]$ (but not necessarily both) is correctly specified.

The IV estimator can be derived from the moment conditions defined the exclusion restriction: $\mathbb{E}\left[Z_i \omega_Y\right] = 0$. Applying the above definitions to this moment condition, we can derive the IPW estimator which we call the weighted IV (WIV) estimator. This estimator takes the form

$$\widehat{\beta}_{WIV} = \left[\sum_i \frac{\mathbb{1}\{R_{Yi} = 1\}}{h(T_i)} Z_i T_i\right]^{-1} \left[\sum_i \frac{\mathbb{1}\{R_{Yi} = 1\}}{h(T_i)} Z_i Y_i\right]. \quad (3.9)$$

where the function $h(T) = P(R = 1|T)$ models the selection mechanism which is typically estimated from the observed data.

As mentioned above, the WIV throws out information by only utilizing the complete cases. If we apply the relatively efficient weighted estimator, proposed by Li et al. (2011),

to the IV framework we get the efficient weighted IV (EWIV) estimator:

$$\widehat{\beta}_{EWIV} = \left[\sum_i Z_i T_i\right]^{-1} \left(\left[\sum_i \frac{\mathbb{1}\{R_{Yi}=1\}}{h(T_i)} Z_i Y_i\right] + \left[\sum_i \left(1 - \frac{\mathbb{1}\{R_{Yi}=1\}}{h(T_i)}\right) Z_i g(Z_i, T_i)\right]\right).$$

where $g(Z,T) = \mathbb{E}[Y|Z,T,R_Y=1]$. Careful readers will notice that EWIV contains terms that are similar to MIV, specifically the g_y function described in section 3.4 is also found in the second term of the EWIV estimator. Though MIV appears as a constituent part of EWIV, Li et al. (2011) fail to acknowledge that this portion of their estimator can stand on it's own as a consistent estimator of β.

One pivotal assumption underlying both WIV and EWIV is that of positivity, which requires that the probability of missingness be bounded away from 0, i.e. there exists $\epsilon > 0$ such that $h(T_i) > \epsilon$ for all T_i. When this assumption is violated, the weights can be extremely volatile, leading to a few observations having undue influence on the final estimate (see Seaman and White 2011; Crump et al. 2009; Yoshida et al. 2018).

While the MIV estimator does not explicitly have the same weakness to violations of positivity, it will still be implicitly affected by such issues. For example, if $P(R_i = 1|Z_i = z)$ is close to 0 for certain values of z this will cause difficulties in estimating $g_y(z,v) = \mathbb{E}\left[Y_i|Z_i = z, V_i = v, R_i = 1\right]$ because relatively few observations will be present in those regions. However, WIV and EWIV are much more prone to violations of positivity simply because of the algebraic manner in which $h_(T_i)$ enters the estimators. Furthermore, EWIV is double susceptible since it is dividing by $h(T_i)$ and trying to estimate g_y.

3.6.2 Multiple Imputation

Multiple imputation methods (first proposed by Rubin 1978) approach the missing data problem by attempting to recreate complete data sets through repeated generation of missing values.

Consider the same setup for Y_i and T_i as in the previous sections, and let $f_Y(Y_i|T_i)$ denote the conditional distribution of Y_i given T_i. These methods start by estimating this

conditional distribution, usually using one of two methods. Simple MI methods assume that missing variables follow a multivariate normal distribution conditional on non-missing variables, and more sophisticated methods like multiple imputation with chained equations (van Buuren et al. 1999) lean on assuming that estimated coefficients in the intermediary regressions are normally distributed.

After estimating this conditional distribution the next step is to generate new data from this estimated object. For every missing value of Y_i, a random value is drawn from $f_Y(y|Z_i, T_i)$ until a complete set of data is drawn. The standard IV estimator is applied on this complete data set to yield estimate $\widehat{\beta}_{IV}^{(1)}$. This process is repeated with a new sample of imputed values to yield another estimate, $\widehat{\beta}_{IV}^{(2)}$, and this is repeated again until M estimates are obtained. Finally these estimates are pooled, usually by simple averaging to arrive at the multiply imputed IV estimate

$$\widehat{\beta}_{MI} = \frac{1}{M} \sum_{m=1}^{M} \widehat{\beta}_{IV}^{(m)}. \tag{3.10}$$

To the best of our our knowledge no papers have explored how well multiple imputation (MI) methods fare in an instrumental variables framework subject to missing data. The typical Achilles heel of MI is their reliance on distributional assumptions. Both the simple and more complex methods, assume normality at some point and in either case, violations of these assumptions can result in small-sample bias and invalid inference (Murray 2018). MIV is partially protected from this bias, in that it only needs the first moment of these distributions, e.g. $g(Z, T) = \mathbb{E}[Y|Z, T, R_Y = 1]$, rather than the entire distribution.

3.7 Monte Carlo Simulations

To compare how WIV, EWIV, and M-IV fare under violations of positivity, we ran preparatory Monte Carlo simulations under a simple parameterization with linear effects ($\beta = 3$) and normal errors. The full data generating process is given by

$$
\begin{aligned}
Z &= \varepsilon_Z \\
U &= \varepsilon_U \\
T &= \mathbb{1}\{2Z + 0.5U + \varepsilon_T > 0\} \\
R &= \mathbb{1}\{-1.5T + \varepsilon_R > 0\} \\
Y &= 3T + U + \varepsilon_Y
\end{aligned}
\qquad
\begin{bmatrix} \varepsilon_Z \\ \varepsilon_U \\ \varepsilon_T \\ \varepsilon_R \\ \varepsilon_Y \end{bmatrix} \sim N\left(\begin{bmatrix} 0.5 \\ 0 \\ 2 \\ 5 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & Var(\varepsilon_R) & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \right)
$$

Two models are considered that vary *only* in $Var(\varepsilon_R)$:

- Model 1 (Positivity Holds): $Var(\varepsilon_R) = 7$ leads to substantial variation in R_Y due to unobserved variables. This results in $P(R = 1|T = t) \in (0, 1)$ for most observable t.
- Model 2 (Positivity Violated): $Var(\varepsilon_R) = 1$ leads to little variation in R_Y due to unobserved variables. This results in $P(R = 1|T = t) \approx 0$ or 1 for many values of t.

Figure 3.4 illustrates the violation of positivity for Model 2.

(a) Positivity Holds (Model 1)

(b) Positivity Fails (Model 2)

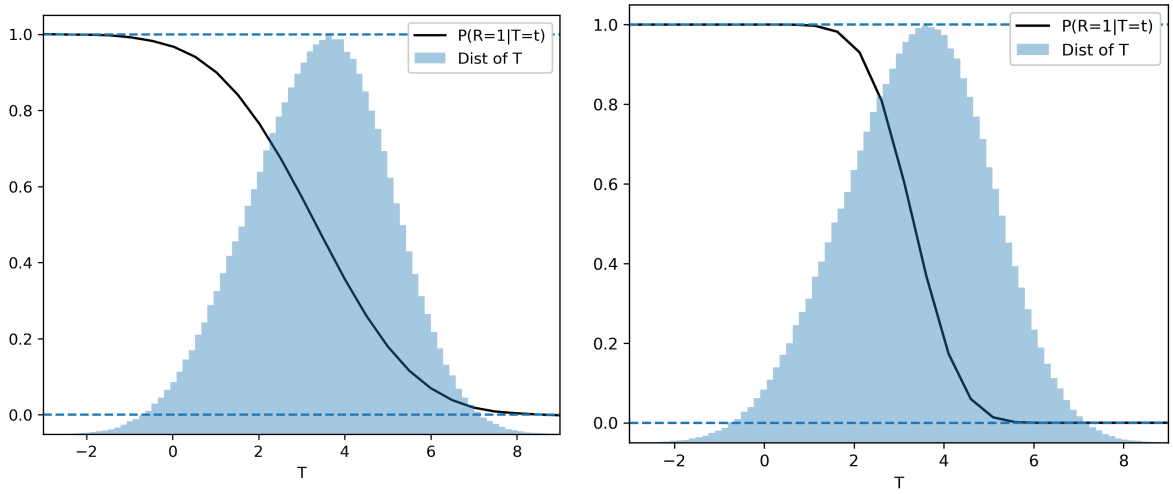


Figure 3.4: Plots display the conditional probability of missingness and the distribution of T_i to illustrate that positivity holds in Model 1 (a and is violated in Model 2 (b).

A total of 2,000 simulations were run, each with a sample size of $n = 10,000$ and 5 estimates were computed: (1) The infeasible conventional IV on the entire sample, (2)

(a) Distribution of $\widehat{\beta}$ under Model 1

(b) Distribution of $\widehat{\beta}$ under Model 2

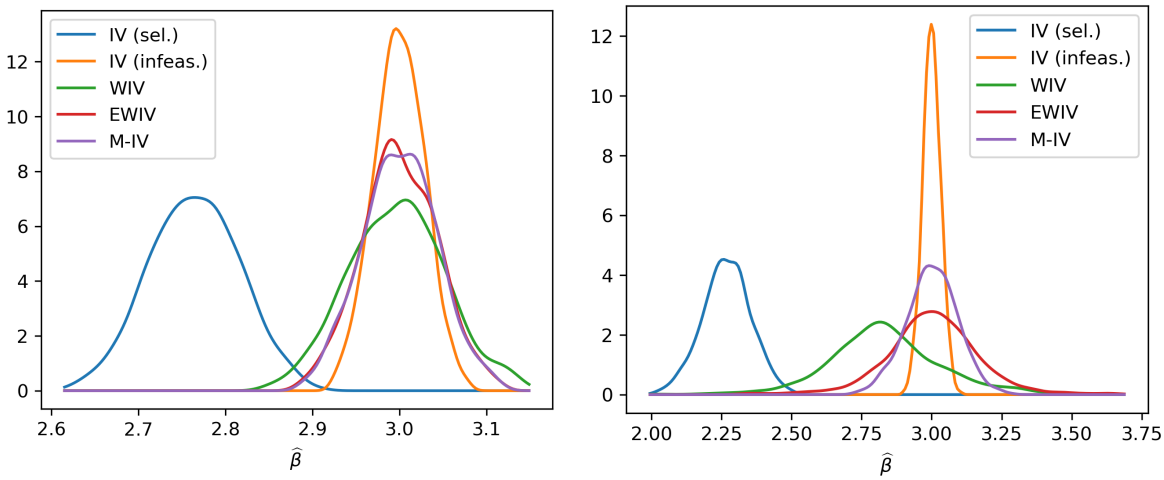


Figure 3.5: MC simulations of the distributions of treatment effects for various estimators when positivity holds in Model 1 (a) and when positivity fails in Model 2 (b). The true effect is $\beta = 3.0$. The proposed M-IV estimator is least biased and efficient.

	Model 1 (Positivity Holds)			Model 2 (Positivity Viol.)		
	Bias	Emp SE	MSE	Bias	Emp SE	MSE
IV (infeas.)	0.0001	0.0288	0.0008	0.0001	0.0319	0.0010
IV (selected)	-0.2357	0.0521	0.0583	-0.7320	0.0864	0.5432
WIV	-0.0030	0.0555	0.0031	-0.1623	0.2065	0.0690
EWIV	0.0002	0.0433	0.0019	0.0031	0.1688	0.0285
M-IV	0.0008	0.0429	0.0018	0.0012	0.0915	0.0084

Table 3.1: Monte Carlo simulation statistics illustrating effects of positivity assumption on estimators.

conventional IV on the selected sample (i.e. $R_Y = 1$), (3) WIV using estimated weights from a probit (4) EWIV using weights from a probit (5) our newly proposed M-IV as described in Section 3.6.2. Statistics associated with the simulations are presented in Table 3.1 and the distribution of estimates for Model 2 is plotted in Figure 3.5b.

When positivity holds (Figure 3.5a), all estimators (besides the strongly biased naive IV on the selected sample) perform well, although the efficiency gains of EWIV over WIV are even present there. When positivity is violated (Figure 3.5b) WIV exhibits small sample bias (even at 10,000 observations) while EWIV and M-IV maintain very little bias. However, the

volatility in the weights significantly increases the standard errors of EWIV to nearly double that of M-IV, and the MSE of M-IV remains very small as compared to all the estimators here.

This robustness to the positivity assumption is one important difference between the weighting estimators and M-IV that will be explored further. We will also investigate how misspecification in either the probability of missingness, $h(T) = P(R = 1|T)$, or the conditional mean function, $g(Z, T) = \mathbb{E}[Y|Z, T, R_Y = 1]$, affect the relative merits of the three estimators.

3.8 Discussion

In this paper we investigated the identification and estimation of causal effects in the presence of outcome attrition. When the missingness of the outcome is affected by the treatment received, then the standard exclusion restriction will be violated in the sub-sample of complete cases, even when the instrument is perfectly exogenous and exclusively affects treatment. To solve this problem we proposed a new estimator, called the missingness instrumental variables (MIV) estimator. This new estimator transforms the infeasible IV estimator by leveraging recovery covariates to replace the problematic moments with those that can be consistently estimated using the sub-sample of complete cases. Since MIV identifies the same statistical objects that the standard IV estimator would if it were not subject to missing data, this estimator is applicable in the estimation of a wide variety of causal objects including average treatment effects and local average treatment effects.

A vital aspect of this new estimator is the necessary provision of a set of recovery covariates which allow the moment transformation to occur. Since this set of covariates is defined by a conditional independence assumption, we discussed graphical causal models given their usefulness for exploring implied statistical associations. We also provided several examples of recovery covariates across a range of models to help develop intuition for researchers looking

to implement our method.

Finally, we compared and contrasted our method with a couple alternative methods that have yet to be fully explored in our specific context. While both inverse probability weighting and multiple imputation are applicable, they are sensitive to positivity assumptions and distributional assumptions respectively, while MIV is more robust to these issues. Monte Carlo simulations were carried out to further compare all three methods and illustrate where MIV excels over the other two alternatives.

Though this paper focused on treatment induced outcome attrition, there are many aspects of our approach that may be amenable to more general scenarios. In fact, most of the results discussed here can be trivially extended to situations where the outcome attrition is driven by variables other than treatment. Moreover, the general approach of transforming infeasible moments into feasible moments on the sub-sample of complete cases can be exported to situations where variables other than the outcome are subject to missingness, and even cases where multiple variables are missing.

3.9 References

Aizer, A. and Doyle, J. J. (2015). Juvenile incarceration, human capital, and future crime: Evidence from randomly assigned judges. *The Quarterly Journal of Economics*, 130(2):759–803.

Angrist, J. D. (1990). Lifetime earnings and the vietnam era draft lottery: Evidence from social security records. *American Economic Review*, 80(3):313–336.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.

Brito, C. (2010). Instrumental sets. In Dechter, R., Geffner, H., and Halpern, J., editors, *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, chapter 17, pages 295–307. College Publications.

Canan, C., Lesko, C., and Lau, B. (2017). Instrumental variable analyses and selection bias. *Epidemiology*, 28(3):396–398.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199.

Curtin, R., Presser, S., and Singer, E. (2005). Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Quarterly*, 69(1):87–98.

de Leeuw, E. and de Heer, W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. In Groves, R. M. and Dillman, D., editors, *Survey Nonresponse*, chapter 3, pages 41–54. New York: Wiley.

Elwert, F. and Segarra, E. (2020). *Instrumental Variables with Treatment-Induced Selection: Exact Bias Results*, pages 1–17. ACM Books.

Ertefaie, A., Small, D., Flory, J., and Hennessy, S. (2016). Selection bias when using instrumental variable methods to compare two treatments but more than two treatments are available. *The International Journal of Biostatistics*, 12(1):219–232.

Geiger, D., Verma, T., and Pearl, J. (1990). Identifying independence in bayesian networks. *Networks*, 20(5):507–534.

Gkatzionis, A. and Burgess, S. (2019). Contextualizing selection bias in mendelian randomization: how bad is it likely to be? *International Journal of Epidemiology*.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70(5):646–675.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1):153–161.

Henderson, D. J. and Parmeter, C. F. (2015). *Applied Nonparametric Econometrics*. Cambridge University Press.

Hewitt, C. E., Kumaravel, B., Dumville, J. C., and Torgerson, D. J. (2010). Assessing the impact of attrition in randomized controlled trials. *Journal of Clinical Epidemiology*, 63(11):1264–1270.

Holland, P. W. (1986). Statistics and causal inference: Rejoinder. *Journal of the American Statistical Association*, 81(396):968–970.

Härdle, W. and Linton, O. B. (1994). *Chapter 38 Applied nonparametric methods*, pages 2295–2339. Elsevier.

Hughes, R. A., Davies, N. M., Davey Smith, G., and Tilling, K. (2019). Selection bias when estimating average treatment effects using one-sample instrumental variable analysis. *Epidemiology*, 30(3):350–357.

Imbens, G. W. (2010). Better late than nothing: Some comments on deaton (2009) and heckman and urzua (2009). *Journal of Economic Literature*, 48(2):399–423.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 1st edition.

Kennedy, E. H., Mauro, J. A., Daniels, M. J., Burns, N., and Small, D. S. (2019). Handling missing data in instrumental variable methods for causal inference. *Annual Review of Statistics and Its Application*, 6(1):125–148.

Li, L., Shen, C., Li, X., and Robins, J. M. (2011). On weighting approaches for missing data. *Statistical Methods in Medical Research*, 22(1):14–30.

Li, Q. and Racine, J. S. (2006). *Nonparametric Econometrics Theory and Practice*. Princeton University Press.

Maathuis, M. H., Drton, M., Lauritzen, S., and Wainwright, M., editors (2018). *Handbook of Graphical Models*. CRC Press, 1st edition.

Mohan, K., Pearl, J., and Tian, J. (2013). Graphical models for inference with missing data. In *Advances in neural information processing systems*, Advances in neural information processing systems.

Murray, J. S. (2018). Multiple imputation: A review of practical and theoretical findings. *Statistical Science*, 33(2):142–159.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–710.

Pearl, J. (2009). *Causality*. Cambridge University Press, second edition.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Rotnitzky, A. and Smucler, E. (2020). Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *Journal of Machine Learning Research*, 21:1–86.

Rubin, D. B. (1974). Estimating causal effects of treatments inrandomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rubin, D. B. (1978). Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, volume 1.

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331.

Schuster, E. F. (1972). Joint asymptotic distribution of the estimated regression function at a finite number of distinct points. *The Annals of Mathematical Statistics*, 43(1):84–88.

Seaman, S. R. and White, I. R. (2011). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22(3):278–295.

Shpitser, I. (2018). *Identification in Graphical Causal Models*, pages 381–404. CRC Press, 1 edition.

Swanson, S. A., Robins, J. M., Miller, M., and Hernán, M. A. (2015). Selecting on treatment : A pervasive form of bias in instrumental variable analyses. *American Journal of Epidemiology*, 284:1–7.

van Buuren, S., Boshuizen, H. C., and Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18(6):681–694.

Verma, T. and Pearl, J. (1988). Causal networks: Semantics and expressiveness. In *Proceedings of the Fourth Annual Conference on Uncertainty in Artificial Intelligence*, pages 69–78. North-Holland Publishing Co.

Wooldridge, J. M. (2011). *Econometric Analysis of Cross Section and Panel Data*. MIT press.

Yoshida, K., Solomon, D. H., Haneuse, S., Kim, S. C., Patorno, E., Tedeschi, S. K., Lyu, H., Franklin, J. M., Stürmer, T., Hernández-Díaz, S., and Glynn, R. J. (2018). Multinomial extension of propensity score trimming methods: A simulation study. *American Journal of Epidemiology*, 188(3):609–616.