Dynamic confounding and Long term treatment effect estimation by data combination: point and partial identification

Yechan Park

October 2022

目次

概要

We combine experimental and observational data that have complementary role in estimating long term treatment effects. The underlying important objective is to highlight the decade old idea of the inherently complementary role that the experiments and economic theory should play in program evaluation in a newly emerging setting of long term treatment effect by data combination. Economic theory provides way to differentiate and choose between non-nested competing identification strategies, and can posit a reliable model for the dynamics of labor outcomes even under selection. On the other hand, the exogenous variation of the experimental data can provide ways to choosing between different

functional and distributional assumptions that economic models place, and also provides great credibility for tightening bounds even under no point identification.

1 Introduction

We are interested in assessing the long term effect of a job training program. We have a result from a randomized control trial which measured the change in unemployment rate a year after the program participation, but no long term follow up has been, possibly due to cost of follow up. On the other hand, observational data documenting the results up to 3 years is available, but probably contaminated by self-selection of workers into the program. Is there anyway to get the selection corrected long term effect of the job training program by combining the two data? This is a question that has received considerable interest recently, reflected in the increasing number of papers that have documented conditions under which the long term average treatment effect can be nonparametrically point identified. One of the major and leading works by (Athey et al., 2020) provides causal assumptions of internal and external validity of experimental data, and a form of conditional internal validity called latent unconfounding assumptions. While illuminating, there are assumptions like external validity or latent unconfounding assumption that have not been explicitly placed in the traditional program evaluation, and we would like to look deeper into the nature of these assumptions, and what to do when the assumptions seems to be difficult to hold in the application of interest.

Specifically, we first look into the external validity assumption, and (a) show that their external validity of long term potential outcome was actually unnecessary for average treatment effect estimation (b) under the relaxed assumption, provide a formal test based on partial identification literature, providing *uniformly sharp identification bounds valid even in the high dimensional regime*

Next, we will investigate the latent unconfounding assumption. We illustrate how it may be difficult to hold for time varying unobserved confounders or unobserved confounding that simultaneously affect the treatment, short, and long term outcome. To highlight this point, we turn from identification of average treatment effect to a average treatment effect on the treated survivors (ATETS), a modified form of average treatment effect that highlights the difficulty of dealing with dynamic confounding, and is of interest in its own right when unemployment rate is concerned.

We will approach identification of this challenging in three ways. The first two aims for point identification. The first approach aims for assumptions in the form of potential outcomes that is standard in the current literature of program evaluation. We show that the previously maintained assumptions by (Athey et al., 2020)is insufficient for point identification, due to the counterfactual quantity in the conditioning set, and provide additional assumption that suffice for point identification, though may be hard to hold in many settings. The second approach turns to more traditional economic modeling, by constructing a dynamic analogue of the generalized toy model(Heckman and Navarro, 2007) (Roy, 1951)that can flexibly embed assumptions motivated by economic theory on the time varying confounding, The final approach gives up point identification, and provides *sharp identification bounds* on the treatment effect quantity with economically motivated shape restrictions in increasing strength, elucidating the role each identifying assumption plays in shaping inference.

The rest of the paper is organized as follows: Section 2 provides a short introduction to the identification approach of our baseline paper (Athey et al., 2020), Section 3 discusses the external validity assumption, Section 4 discusses the nature of latent unconfounding assumptions. Section 5 motivates the modified average treatment effect estimates and the experimental approach of identification. Section 6 illustrates the economic modeling approach. Section 7 discusses partial identification strategy. Section 8 discusses the estimation strategy, and Section 9 shows the empirical application. Section 10 concludes.

2 Setup

First we introduce the general notation that will be used throughout our paper. Given a probability space $(\Omega, \mathcal{F}, \mathcal{P})$, todo

A researcher conducts a randomized experiment aimed at assessing the effects of a policy or intervention. For each individual in the experiment, they measure a q-vector X_i of pre-treatment covariants, a binary variable W_i denoting assignment to treatment, and a d-vector Y_{1i} of of short term post-treatment outcomes. The researcher is interested in the effect of the treatment on a scalar long term post treatment outcome Y_{2i} that is not measured in the experimental data. They are able to

obtain an auxiliary, observational data set containing measurements, for a separate population of individuals, which consist of the same covariates X_i, treatment, W_i, short term outcome Y_{1i} and also the long term outcome Y_{2i}. For the introduction, the data available for our case is presented in Figure 2.

As aforementioned, we have two data sets.

We first introduce the identification assumption of (Athey et al., 2020) to identify the long-term ATE. They employ the following four assumptions. (Athey et al., 2020)

仮定 2.1 (Experimental internal validity). $W \amalg Y_2(1), Y_2(0), Y_1(1), Y_1(0)|X, G = 0$.

This will be satisfied by construction in cases where

仮定 2.2 (Exernal validity of experiment). $G \amalg Y_2(1), Y_2(0), Y_1(0), Y_1(0)|X$.

仮定 2.3 (strict overlap). The probability of being assigned to treatment or of being measured in the observational data set is strictly bounded away from zero and one, i.e., for each w and g in {0,1}, the conditional probabilities

$$P(W = w|Y_1(1), X, G = g) \quad \text{and} \quad P(G = g|Y_1(1), X, W = w) \tag{1}$$

are bounded between ϵ and $1 - \epsilon$, λ- almost surely, for some fixed constant $0 < \epsilon < 1/2$.

仮定 2.4 (latent unconfounding). $W \amalg Y_2(w) \mid Y_1(w), X, G = 1 \quad (w = 0, 1)$.

In Theorem 1 of (Athey et al., 2020), they prove the following identification results.

定理 2.5 ((Athey et al., 2020) Theorem 1). $E[Y_2(1)|G = 1]$ is nonparametrically identified.

[証明]. This itself is very insightful, one of the early works that proposed an effective usage of experimental and observational data. However, at the same time, we notice that assumptions not commonly made in the conventional settings are necessary for identification. In particular, a cross locational independence assumption (Assumption 2.2), and a form of unconfoudning conditioned on a *latent variable* (Assumption 2.4) appears to be quite new, and may or may not hold depending on certain contexts. We examine both of them in the subsequent sections, how we may be able to relax, interpret, or test these assumptions. ■

3 External validity

In this section, we will focus on the external validity assumption (Assumption 2.2). We specifically do two things. We first show that the conditional independence assumption between G and $(Y_2(1), Y_1(1)$ can be relaxed to only that of G and $Y_1(1)$. Second, based on those relaxations, we consider various partial identifying assumptions specialized for this setting accompanied with sharp identification bounds for them.

We will first show that Assumption 2.2 can be relaxed of only the short term outcome, formalized in the following assumption and proposition. The key intuition for why this holds is to notice that the external validity assumptions is used in the identification proof of (Athey et al., 2020) two times, and the aggregate bias that appears without the external validity with the long term potential outcome can be shown to be of a simple form, which exactly becomes zero under the conditional independence of the short-term outcome.

仮定 3.1 (short term external validity). $G \amalg Y_1(1), Y_1(0) \mid X$.

命題 3.2 . *ii* The long term average treatment effect in the observational population $E[Y_2(1) - Y_2(0)|G = 1]$ is nonparametrically identified replacing Assumption 2.2 with Assumption 3.1.

[証明]. It suffices to show that $E[Y_2(1)|X, G = 1]$ is nonparametrically identified under the relaxed external validity assumptions. We will show that

$$\begin{aligned} &E[Y_2(1)|G = 1] \\ =&E[E[Y_2(1)|X, G = 1]|G = 1] \end{aligned}$$

$$
\begin{aligned}
&= E[E[Y_2(1)|X, G=0]|G=1]] + A \qquad (A := E[E[Y_2(1)|X, G=1] - E[Y_2(1)|X, G=0]|G=1]) \\
&= E[E[E[Y_2(1)|Y_1(1), X, G=0]|X, G=0]|G=1] + A \\
&= E[E[E[Y_2(1)|Y_1(1), X, G=1]|X, G=0]|G=1] + B + A \\
&\qquad\qquad (B := E[E[E[Y_2(1)|Y_1(1), X, G=0]|X, G=0]|G=1] - \\
&\qquad\qquad\qquad E[E[E[Y_2(1)|Y_1(1), X, G=1]|X, G=0]|G=1]) \\
&= E[E[E[Y_2(1)|Y_1(1), W=1, X, G=1]|X, G=0]|G=1] + B + A \qquad (\because \text{Assumption 2.4}) \\
&= E[E[E[Y_2(1)|Y_1(1), W=1, X, G=1]|W=1, X, G=0]|G=1] + B + A \qquad (\because \text{Assumption 2.1}) \\
&= E[E[E[Y_2|Y_1, W=1, X, G=1]|W=1, X, G=0]|G=1] + B + A \qquad (\because \text{consistency})
\end{aligned}
$$

Because $E[E[E[Y_2|Y_1, W=1, X, G=1]|W=1, X, G=0]|G=1]$ is identified from the data, if we show that B + A =0 under the relaxed external validity assumption, our goal is achieved.

Note that

$$
\begin{aligned}
&A + B \\
&= E[E[Y_2(1)|X, G=1] - E[Y_2(1)|X, G=0]|G=1] \\
&\quad - E[E[Y_2(1)|X, G=0]|G=1] - E[E[E[Y_2(1)|Y_1(1), X, G=1]|X, G=0]|G=1] \\
&= E[E[Y_2(1)|X, G=1]|G=1] - E[E[E[Y_2(1)|Y_1(1), X, G=1]|X, G=0]|G=1] \\
&= E[Y_2(1)|G=1] - E\Bigg[\iint y_2(1) p(y_2(1)|y_1(1), X, G=1) \\
&\qquad\qquad p(y_1(1)|X, G=0) dy_2(1) dy_1(1) \Big| G=1\Bigg] \\
&= E[Y_2(1)|G=1] - E\Bigg[\iint y_2(1) \frac{p(y_1(1)|X, G=0)}{p(y_1(1)|X, G=1)} \\
&\qquad\qquad p(y_2(1)|y_1(1), X, G=1) p(y_1(1)|X, G=1) dy_2(1) dy_1(1) \Big| G=1\Bigg] \\
&= E[Y_2(1)|G=1] - E\left[y_2(1) \frac{p(y_1(1)|X, G=0)}{p(y_1(1)|X, G=1)} \Big| G=1\right] \\
&= E\left[Y_2(1)\left(1 - \frac{p(y_1(1)|X, G=0)}{p(y_1(1)}|X, G=1)\right) \Big| G=1\right] = 0 \qquad (\because \text{Assumption 3.1})
\end{aligned}
$$

■

Under this relaxation, we show how to test using partial identification similar in spirit to (Blundell et al., 2007), which used it to test the differential wage gap in the UK among gender.

3.1 bounds for testing external validity

An implication of (the relaxed) external validity is that $F(Y_1(1)|X, G=1) = F(Y_1(1)|X, G=0)$. Rewriting from what we know, the right hand side is nonparametrically identified because

$$
\begin{aligned}
F(Y_1(1)|X, G=0) &= F(Y_1(1)|W=1, X, G=0) \qquad (\because \text{Ass.2.1}) &(2)\\
&= F(Y_1|W=1, X, G=0) &(3)
\end{aligned}
$$

For the left hand side, by the law of iterated expectation and chain rule, and consistency

$$
\begin{aligned}
&F(Y_1|W=1, X, G=1) P(W=1|X, G=1) &(4)\\
&\quad + F(Y_1(1)|W=0, X, G=1) P(W=0|X, G=1) &(5)
\end{aligned}
$$

and hence the counterfactual quantity is only $F(Y_1(1)|W=0, X, G=1)$. A pointwise sharp bound under no additional assumption can be easily seen to be

$$
\begin{aligned}
&F(Y_1|W=1, X, G=1) P(W=1|X, G=1) &(6)\\
&\leqslant F(Y_1|W=1, X, G=0) &(7)\\
&\leqslant F(Y_1|W=1, X, G=1) P(W=1|X, G=1) + P(W=0|X, G=1) &(8)
\end{aligned}
$$

and was indeed implemented in works like (Blundell et al., 2007),(Manski, 2009).

The bound above under no additional assumptions is pointwise sharp for each t. However in practice, this will of ten be quite wide. On the other hand, researchers may have prior beliefs motivated by the economics of the problem, and they are interested in how the assumptions may translate into narrower bounds and the results of the test under such assumptions. The partial identification framework we provide below allows for such procedures in a systematic manner.

We briefly mention the related literature. There is an increasing amount of research in the sensitivity analysis literature that assess the generalizability, or transportability of a result in one location or one research design, e.g., imai egami, etc. However, while the partial identification on the distribution function under various assumption have been widely used (Blundell et al., 2007), manski2003partial, manski horo, there are few papers that consider such framework for assessing external validity. manski 2013 provides a framework of 'external assessment' using bounds, it takes a more social decision theory, minimax perspective, which is a different approach from ours.

The novelty of the assumptions that we consider can be rephrased as the following. consider the diagram below. While past literature have focused on the assumption that assess the assumption pertaining to selection (across treatment groups), the random variable W, (Blundell et al., 2007), here we also consider the cross location/design assumptions , and moreover on the intersection of those two, delivering sharp identification bounds for each case. As we will see, this is particularly useful, because to gain informative bounds only using cross treatment (W) variation, strong assumptions such as stochastic dominance may be necessary. However, with the aid of additional reasonable cross locational variation assumptions, by taking the intersection, we could get informative bounds even under no clearly implausible assumptions.

We will propose the assumptions here and provide the sharp bounds in the subsequent sections.

(1) cross treatment (W) assumptions

stochastic dominance

$$\forall_{t\in\mathcal{T}}\ \forall_{x\in\mathcal{X}}\ F(Y_1(1)\leqslant t|X=xW=1,G=1)\leqslant F(Y_1(1)\leqslant t|X=x,W=0,G=1)$$

this is a form of negative selection into treatment, which in the case of job training program, past research like ashenfelter 1985, card, heckman 1999 have showed how people with lower potential wage or employment probability tend to apply for job training, for all covariates, and quantiles of the potential outcome.

if one thinks that this result holding uniformly might be too stringent, but is convinced that at least for the lower 25 percent quantile of the treated, such negative selection could be true, then we could consider the following.

lower quantile dominance

$$\forall_{t\leqslant t_{0.25}}\ \forall_{x\in\mathcal{X}}\ F(Y_1(1)\leqslant t|X=xW=1,G=1)\leqslant F(Y_1(1)\leqslant t|X=x,W=0,G=1)$$

where $t_{0.25}$ is the lower 25 percent quantile of $F(Y_1(1)\leqslant t|X=xW=1,G=1)$, which is observable.

Moreover, one could consider the monotone iv assumption, which states that for a particular obserbale covariate the value of which is indicative of the unobserved potential outcomes distribution, formalized below.

monotone iv

$$\begin{aligned}\forall_{v_1,v_2\in\mathcal{V}}\ v_1\leqslant v_2\implies F(Y_1(1)\leqslant t|V=v_1X=x,W=0,G=1)\\ \leqslant F(Y_1(1)\leqslant t|V=v_2X=x,W=0,G=1)\end{aligned}$$

in our problem the V may be a baseline covariate like test scores in the past, with lower values in such test scores may indicate lower values of the unobserved employment rate.

(2) cross location/design (G) assumptions

heterogeneity

$$\begin{aligned}\exists_{t_0}\ \forall_{x\in\mathcal{X}}[\forall_{t\leqslant t_0}\ F(Y_1(1)\leqslant t|V=X=x,W=0,G=1)\\ \leqslant F(Y_1(1)\leqslant t|V=X=x,W=0,G=0)\\ \wedge\ \forall_{t\geqslant t_0}\ F(Y_1(1)\leqslant t|V=X=x,W=0,G=0)\\ \leqslant F(Y_1(1)\leqslant t|V=X=x,W=0,G=1)]\end{aligned}$$

this to our knowledge is new, and is captured in the illustration below. It is often the case that the observational data have people with covariates with wider range, dispersed and heterogeneous, while the experimental data based

on eligibility rules and the design of experiments are restricted to a more concentrated covariate value. In that case, the distribution of the observational quantile would be much wider than the experimental and before and after some threshold t_0, the cumulative probability may be reverse.

3.2 point-wise sharp bounds

3.3 uniformly sharp bounds

In the previous section, we have provided point wise sharp bounds However, if we see the bound as a bound on a function(in this case the distribution function), the point-wise bound contains is not sharp in that it can contain distribution functions that is not compatible with the data. For example, it does not impose the restriction that for any $t_0 \leqslant t_1$,

$$P(t_0 \leqslant Y_1(1) \leqslant t_1|\mathbf{x}) \geqslant P(t_0 \leqslant Y_1(1) \leqslant t_1|\mathbf{x} = x, W = 1)P(W = 1|\mathbf{x} = x) \tag{9}$$

This is easiest to see in a simple counterexample as in the following: omit $\mathbf{x}$, and $G = 1$ for simplicity, and let $P(\mathbf{W} = 1) = \frac{2}{3}$ and let

$$P(Y_1(1) \leqslant t|W = 1) = \begin{cases} 0 & \text{if} \quad t < 0 \\ \frac{1}{3}t & \text{if} \quad 0 < t < 3 \\ 1 & \text{if} \quad t \geqslant 3 \end{cases} \tag{10}$$

Consider the distribution function

$$F(t) = \begin{cases} 0 & \text{if} \quad t < 0 \\ \frac{5}{9}t & \text{if} \quad 0 \leqslant t < 1 \\ \frac{1}{9}t + \frac{4}{9} & \text{if} \quad 1 \leqslant t \leqslant 2 \\ \frac{1}{3}t & \text{if} \quad 2 \leqslant t < 3 \\ 1 & \text{if} \quad t \geqslant 3 \end{cases} \tag{11}$$

For each $t \in \mathbb{R}$, $F(t)$ lies in the tube defined by the point-wise sharp bound above. However it cannot be the CDF of Y_1, because F(2) - F(1) $= \frac{1}{9} < P(1 \leqslant Y_1(1) \leqslant 2|W = 1)P(W = 1)$, directly contradicting (9).

What is then the uniformly sharp identification bound in this case? The following proposition provides an answer.

命題 3.3 . The uniformly sharp identification bound for $F(Y_1(1)|X, G = 1)$ under the maintained assumptions in (Athey et al., 2020)

$$\forall_{K \subset \mathcal{Y}} \ \mathcal{H}_P[Q(y| = \tau(x) \in \mathcal{T} : \tau_{K(x)} \geqslant \mathrm{P}(Y_1(1) \in K|\mathbf{x} = x, W = 1, G = 1)\mathrm{P}(W = 1|\mathbf{x} = x, G = 1)$$

where $\mathcal{Y} = \{\{0, 1\}\}$.

[証明]. ■

This shows how intuition do not necessarily yield the desirable uniformly sharp bounds. I therefore rely mainly on a newly emerging approach from random set theory (Molchanov, 2005) to mechanically derive sharp bounds.

I provide the fundamental definitions and theorems that I will use below, and refer the reader to (Molchanov, 2005; Molchanov and Molinari, 2018) for the proofs and more in depth explanations.

定義 3.4 (Random closed set). A map $\mathbf{X}$ from a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ to the family $\mathbf{F}$ of closed subsets of $\mathbb{R}^d$ is called a random closed set if

$$\mathbf{X}^-(K) := \{\omega \in \Omega : \mathbf{X}(\omega) \cup K \neq \varnothing\} \tag{12}$$

belongs to the σ-algebra $\mathfrak{F}$ on Ω for each compact set K in $\mathbb{R}^d$

定義 3.5 (Capacity functional and containment functional).

1. A functional $T_{\mathbf{X}}(K) : \mathcal{K} \to [0,1]$ given by

$$\mathrm{T}_{\mathbf{X}}(K) = \mathbb{P}\{\mathbf{X} \cup K \neq \varnothing\}, \qquad (K \in \mathcal{K}) \tag{13}$$

is called **capacity (or hitting) functional** of $\mathbf{X}$.

2. A functional $\mathrm{C}_{\mathbf{X}}(F) : \mathcal{F} \to [0,1]$ given by

$$\mathrm{C}_{\mathbf{X}}(F) = \mathbb{P}\{\mathbf{X} \in F\}, \qquad (F \in \mathcal{F}) \tag{14}$$

is called the **containment functional** of $\mathbf{X}$.

定義 3.6 (Measurable selection). For any random set $\mathbf{X}$, a (measurable) selection of $\mathbf{X}$ is a random element $\mathbf{x}$ with values in $\mathbb{R}^d$ such that $\mathbf{x}(\omega) \in \mathbf{X}(\omega)$ almost surely. I denote by $\mathrm{Sel}(\mathbf{X})$ the set of all selections from $\mathbf{X}$.

below we have the fundamental theorem we heavily use to characterize sharp identification bounds.

定理 3.7 (Artstein inequality). A probability distribution μ on $\mathbb{R}^d$ is the distribution of a selection of a random closed set $\mathbf{X}$ in $\mathbb{R}^d$ if and only if

$$\mu(K) \leqslant \mathrm{T}(K) = \mathbb{P}\{\mathbf{X} \cup K \neq \varnothing\} \tag{15}$$

for all compact sets $K \subseteq \mathbb{R}^d$. Equivalently, if and only if

$$\mu(F) \geqslant \mathrm{C}(F) = \mathbb{P}\{\mathbf{X} \subset F\} \tag{16}$$

for all closed sets $F \subset \mathbb{R}^d$. If $\mathbf{X}$ is a compact random closed set, it suffices to check 16 for compact sets F only.

4 Latent unconfounding assumption

Different from the previous section, we will focus on the internal validity of the observational data, i.e., under what conditions does W be independent of $Y_t(w)$? As hinted in the introduction, a problem of nonnested approaches under the same availability of data may be the first issue that practitioners have to deal with when having the dataset and deciding ways for establishing identification. Specifically, the alternative approach for identification of the long term ATE was recently proposed by ghassami et al 2022 , which posited an *equi-confounding bias assumption*, which is a form of parallel trends assumption applied to the data-combination setting,formally presented below.

仮定 4.1 (Equiconfounding bias assumption)**.**

(i) $E[Y_2(0) - Y_1(0)|G = 1] = E[Y_2(0) - Y_1(0)|G = 0]$.
(ii) $E[Y_2(1) - Y_1(1)|G = 1] = E[Y_2(1) - Y_1(1)|G = 0]$.

They proved the following theorem in their paper.

定理 4.2 .

(1) Replacing latent unconfounding assumption with 4.1 (i) and maintaining all the other assumptions, the long term ATT is identified.
(2) Replacing latent unconfounding assumption with 4.1 (i) (ii), and maintaining all the other assumptions, the long term ATE is identified.

The equiconfounding assumption can be intuitively explained that the potential growth between the short term and long term outcome is the same among the treated and the untreated. Usually only the the untreated potential outcome (i) is maintained to identify the ATT in the standard did setup. Additionally assuming the equivalent growth among the treated potential outcomes identify the ATE , the proof of which is in their paper ghassami 202.

While this is interesting, it leads to an additional problem coming from the nonnested nature of equi-confounding and the latent unconfounding assumption. The latent unconfounding has us to imagine what happens under condition on an *latent potential outcome*, while the equiconfounding assumption has us to think about the dynamics between two post-treatments, arguably more complicated than the canonical parallel trends in the usual did setup. How should a policy maker be able to

map these various assumptions onto their current problem? One solution that I posit in the following is that, taking into account the selection mechanism provides insight into when exactly these assumptions hold or fail to hold.

I illustrate this idea in two canonical selection mechanisms that are commonly used in empirical research, namely that of Card, and the Roy model (Roy, 1951). I provide koko(necessary and sufficient?) conditions under which each assumption holds.

koko(covariates)

Formally, for the main section, we model the potential outcome as

$$\forall_{i\in[0,N]}\ \forall_{t=1,2}\ Y_{it}(0) = \alpha_i + \lambda_t + \alpha_i\lambda_t + \epsilon_{it},\ E[\epsilon_{it}] = 0.$$

$Y_{it}(1) = Y_{it}(0) + \delta_{it}$, which generalizes the classical two-way-fixed effect model by (i) allowing for interactive fixed effect as in bai 2009, abadie 2021(ii) allowing for arbitrary treatment effect heterogeneity while the two-way fixed effect model usually assumes constant treatment effect. We first consider the selection mechanism proposed in (Ashenfelter and Card, 1985).

In (Ashenfelter and Card, 1985) page koko, the selection mechanism was posited as

$$W_i = 1\{Y_{i1}(0) + \beta Y_{i2}(0) \leqslant c\} = 1\{(1+\beta)\alpha_i + \epsilon_{i1} + \beta\epsilon_{i2} \leqslant \tilde{c}\}$$

where $\beta \in [0,1]$ is a discount factor and $\tilde{c} = c - \lambda_1 - \beta\lambda_2$.

The following theorem can be shown.

定理 4.3 . Consider the setting specified above. Then Latent unconfounding holds iff $\beta = 0$. Equi-confounding assumption does not hold for any

[証明]. Note that

$$W_i = 1\{Y_{i1} + \beta Y_{i2} > 0\} = 1\{\alpha_i(1+\lambda_1+\beta(1+\lambda_2)) + \lambda_1 + \beta\lambda_2 + \epsilon_{i1} + \beta\epsilon_{i2} > 0\}$$

When $\beta = 0$, we have that $W_i =$.

Next, we consider the Roy selection mechanism (Roy, 1951), with its slight extension by e.g., (Heckman and Singer, 1984). The essence of the Roy selection model is that, it is based on *treatment effects*, opposed to the untreated potential outcome in card ashenfelter. ■

I fix the outcome model to be a canonical non-separable panel data model

$$Y_{it}(0) = \alpha_i + \lambda_t + \alpha_i\lambda_t + \epsilon_{it}$$

where α_i is the individual fixed effect, λ_t is the time fixed effect, which we regard as non-stochastic (by conditioning on its realizations) as commonly assumed, in e.g., (Ghanem et al., 2022; Arellano and Bonhomme, 2011). The

To do this, I embed the two different approaches for ATE estimation in a first separable, then nonseparable panel data model, explicitly modeling the constant and time varying observables and unobservables and when the latent unconfounding and the equiconfounding hold or not hold in the respective settings.

4.1 comparison of latent unconfounding and equiconfounding in nonseparable panel data setting

following the convention of did in assuming only the parallel trend on the untreated (i), and focusing on ATT, although the ATE will also follow with a similar setup.

we formalize the problem as the following. we first assume a model for potential outcomes that is separable in the time-invariant and time-varying unobsrevables

仮定 4.4.

$$\forall_{i\in[0,N]}\ \forall_{t=1,2}\ Y_{it}(0) = \alpha_i + \lambda_t + \alpha_i\lambda_t + \epsilon_{it}, E[\epsilon_{it}] = 0$$

in the above assumption, α_i is the time invariant unobservable, λ_t is the (nonstochastic, without l.o.g) time fixed effect, and ϵ_{it} is the tie varying individual specific unobservable. This is close to the commonly assumed two way fixed effect model, only that treatment effect heterogeneity is allowed

next we model the selection mechanism W, as

$$W_i = w(\alpha_i, \epsilon_{i1}, \epsilon_{i2}, \nu_i, \eta_{i1}, \eta_{i2})$$

where the selection into treatment may in general depend on the unobservable determinants of the untreated potential outcomes

$$(\alpha_i, \epsilon_{i1}, \epsilon_{i2})$$

as well as additional time-invariant and time-varying vectors of random variables $(\nu_i, \eta_{i1}, \eta_{i2})$.

This general selection mechanism accommodats many different types of selection, including , random assignement, selection based on past outcomes, roystyle selections basd on treatment effects and othe selection mechanisms based on economc decision problems (heckman and robb).

In the case of the equiconfouding assumption (i), a necessary sufficient condition to hold is provided in ghanem et al as in the following.

定理 4.5 . Suppose that the separable outcome and the general selection mechansim holds. Suppose further that $P(G_i = 1) \in (0,1), \nu_i^1 \amalg (\alpha_i, \epsilon_{i1}, \epsilon_{i2}) P(\nu_i^1 > c) \in (0,1)$ for some $c \in \mathbb{R}$, and $P(\epsilon_{i2} \geqslant \epsilon_{i1}) > 0$. Then equiconfounding assumption (i) holds if and only if $\epsilon_{i1} = \epsilon_{i2}$ a.s.

4.1.1 example: selection mechanism of Ashenfelter and Card (1985)

In ashenfelter and card 19 85, the selection mechanism was posited as

$$W_i = 1\{Y_{i1}(0) + \beta Y_{i2}(0) \leqslant c\} = 1\{(1+\beta)\alpha_i + \epsilon_{i1} + \beta\epsilon_{i2} \leqslant \tilde{c}\}$$

where $\beta \in [0,1]$ is a discount factor and $\tilde{c} = c - \lambda_1 - \beta\lambda_2$.

In this case, selection depends only on the unobservable determinants of the untreated potential outcomes $(\alpha_i, \epsilon_{i1}, \epsilon_{i2})$.

When $\beta = 0$, so that selection depends only on $Y_{i1}(0)$,

$$W_i = 1\{Y_{i1}0) \leqslant c\} = 1\{\alpha_i + \epsilon_{i1} \leqslant \tilde{c}\} \tag{17}$$

In this case, the latent unconfounding assumptions is satisfied without any further assumptions.

On the other hand, the equiconfounding assumption requires the following additional assumptions.

(1) $(\nu_i, \eta_{i1}, \eta_{i2})|\alpha_i, \epsilon_{i1}, , \epsilon_{i2} =^d (\nu_i, \eta_{i1}, \eta_{i2})|\alpha_i$
(2) $E[\epsilon_{i2}|\alpha_i, \epsilon_{i1}] = \epsilon_{i1}$

5 estimation of confidence bands

5.1 uniform inference on the confidence band

In this section, we provide inference on the external validity bounds provided in Section3.1 that allows for possible high dimensional covariates or with many moment inequalities.

In this section, we will be looking at the estimation for the sharp identification bounds. We will first consider the usual case, with two extensions, (i) when the covariates are high-dimensional and non-Donsker (ii) when there are many moment inequalities.

Most of the bounds can be calculated readily by similar methods

5.2 the case when the covariates are high dimensional

For instance, in the case of a binary outcome like unemployment, the uniformly sharp bounds for $F(Y_1|W=1,X,G=0)$ coincide with the point-wise sharp bounds since there are only two compact sets to apply the Artstein Inequalities ($K = \{0\},\{1\}$), so we have to construct a confidence band to test whether the relationship

$$F(Y_1|W=1,X,G=1)P(W=1|X,G=1) \tag{18}$$

$$\leqslant F(Y_1|W=1,X,G=0) \tag{19}$$
$$\leqslant F(Y_1|W=1,X,G=1)P(W=1|X,G=1)+P(W=0|X,G=1) \tag{20}$$

hold for all $x \in \mathcal{X}$.

In the case where the covariates X may be high dimensional, which is often the case for observational data, the so-called Smirinov-Bickel-Rosenblatt(SBR) condition may not hold.

A major breakthrough by (Chernozhukov et al., 2014a) justified a *generalized SBR condition* using anticoncentration and gaussian approximation for suprema of empirical processes. In fact, we can show that the gaussian approximation rate can be improved using the symmetrization trick of the Stein-exchangable pair approach to Gaussian approximation. Recall the main theorem in (Chernozhukov et al., 2014b) Theorem 2.1.

定理 5.1 (Original Gaussian approximation to suprema of empirical processes (Chernozhukov et al., 2014b)2.1). Assume the following conditions:

(A1) point wise measurability of class $\mathcal{F}$.
(A2) the integrability of the envelope F; $\exists_{q \geqslant 3} : [F \in \mathcal{L}^q(P)]$
(A3) class $\mathcal{F}$ is pre-Gaussian

Let $Z = \sup_{f \in \mathcal{F}} \mathbb{G}_n f$. Let $\kappa > 0$ be any positive constant such that $\kappa^3 \geqslant E[||E_n[|f(X_i)|^3]||_{\mathcal{F}}]$Then for every $\epsilon \in (0,1]$ and $\gamma \in (0,1]$, there exists a random variable $\tilde{Z} =^d \sup_{f \in \mathcal{F}} G_P f$ such that

$$P\{|Z-\tilde{Z}| > K(q)\Delta_n(\epsilon,\gamma)\} \leqslant \gamma(1+\delta_n(\epsilon,\gamma)) + \frac{C\log n}{n} \tag{21}$$

where $K(q) > 0$ is a constant that depends only on q, and

$$\begin{aligned}\Delta_n(\epsilon,\gamma) &:= \phi_n(\epsilon) + \gamma^{-1/q}\epsilon||F_{P,2} \\ &\quad + n^{-1/2}\gamma^{-1/q}||M||_q + n^{-1/2}\gamma^{-2/q}||M||_2 \\ &\quad + n^{-1/4}\gamma^{-1/2}(E[||\mathbb{G}_n||_{\mathcal{F}.\mathcal{F}}])^{1/2}H_n^{1/2}(\epsilon) \\ &\quad + n^{-1/6}\gamma^{-1/3}\kappa H_n^{2/3}(\epsilon) \\ \delta_n(\epsilon,\gamma) &:= \frac{1}{4}P\{(F/\kappa)^3 1(F/\kappa > c\gamma^{-1/3}n^{1/3}H_n(\epsilon)^{-1/3})\}\end{aligned}$$

Note the worst rate in $\Delta_n(\epsilon,\gamma)$ is the last term with the order $n^{-1/6}$, which I show can be refined to be $n^{-1/4}$, which seems to be a large gain in the high dimensional regime. Formally, I prove the following refined Gaussian approximation to suprema of empirical processes.

定理 5.2 . Consider the same setting in addition of the finiteness of fourth moment kokoThen the

$$P\{|Z-\tilde{Z}| > K(q)\Delta_n(\epsilon,\gamma)\} \leqslant \gamma(1+\delta_n(\epsilon,\gamma)) + \frac{C\log n}{n}$$

where

$$\begin{aligned}\Delta_n(\epsilon,\gamma) &:= \phi_n(\epsilon) + \gamma^{-1/q}\epsilon\|F_{P,2}\| + n^{-1/2}\gamma^{-1/q}\|M\|_q + n^{-1/2}\gamma^{-2/q}\|M\|_2 \\ &\quad + n^{-1/4}(\gamma^{-1/2}(E[\|\mathbb{G}_n\|_{\mathcal{F}.\mathcal{F}}])^{1/2}H_n^{1/2}(\epsilon) + \gamma^{-1/3}\kappa H_n^{2/3}(\epsilon)) \\ \delta_n(\epsilon,\gamma) &:= \frac{1}{4}P\{(F/\kappa)^3 1(F/\kappa > c\gamma^{-1/3}n^{1/3}H_n(\epsilon)^{-1/3})\}\end{aligned}$$

補題 5.3 (Refinement of coupling inequality for maxima of sum of random vectors of Theorem 4.1 in (Chernozhukov et al., 2014b)). Let $X_1,...X_n$ be independent random vectors in $\mathbb{R}^p$ with mean zero and finite absolute third moments, that is $E[X_{ij}] = 0$ and $E[|X_{ij}|^3] < \infty$ for all $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant p$. Consider the statistic $Z = \max_{1 \leqslant j \leqslant p} \sum_{i=1}^n X_{ij}$ Let $Y_1,...Y_n$ be independent random vectors in $\mathbb{R}^p$ with $Y_i \sim N(0, E[X_i X_i^T])$ $(1 \leqslant i \leqslant n)$. Then, for every $\beta > 0$, and $\delta > 1/\beta$, there exists a random variable $\tilde{Z} =^d \max_{1 \leqslant j \leqslant p} \sum_{i=1}^n Y_{ij}$such that

$$P(|Z-\tilde{Z}| > 2\beta^{-1}\log p + 3\delta) \leqslant \frac{\epsilon + C\beta\delta^{-1}\{B_1 + \beta\mathbf{n}(B_2+B_3)\}}{1-\epsilon} \tag{22}$$

where $\epsilon = \epsilon_{\beta,\delta}$ is given by

$$\epsilon = \sqrt{e^{-\alpha}(1+\alpha)} < 1, \alpha = \beta^2\delta^2 - 1 > 0 \tag{23}$$

$$B_1 = E[\max_{1\leqslant j,k\leqslant p} |\sum_{i=1}^{n}(X_{ij}X_{ik} - E[X_{ij}X_{ik}])], \tag{24}$$

$$B_2 = E[\max_{1\leqslant j\leqslant p}\sum_{i=1}^{n}|X_{ij}|^4] \tag{25}$$

$$B_3 = \sum_{i=1}^{n} E[\max_{1\leqslant j\leqslant p}|X_{ij}^4]\cdot 1(\max_{1\leqslant j\leqslant p}|X_{ij}| > \beta^{-1}/2) \tag{26}$$

[証明]. We will mostly omit the parts that are common with the original.

$$|E[nR]| = \left|E\left[\frac{1}{2}\sum_{i=1}^{n}\sum_{j,k,l,m=1}^{p}\Delta_{ij}\Delta_{ik}\Delta_{il}\Delta_{im}(1-2\theta)\partial_{jklm}h(S_n' + \theta D + \theta'(1-2\theta)\Delta_i)\right]\right| \tag{27}$$

$$\leqslant \frac{1}{2}E\left[\sum_{i=1}^{n}\sum_{j,k,l,m=1}^{p}|\Delta_{ij}\Delta_{ik}\Delta_{il}\Delta_{im}|\cdot\left|\partial_{jklm}h(S_n' + \theta D + \theta'(1-2\theta)\Delta_i)\right|\right] \tag{28}$$

Let $\chi_i = 1(\max_{1\leqslant j\leqslant p}|\Delta_{ij}| \leqslant \beta^{-1})$ and $\chi_i^c := 1 - \chi_i$. Then

$$\begin{aligned}(27) &= \frac{1}{2}E[\sum_{i=1}^{n}\chi_i\cdot\sum_{j,k,l,m=1}^{p}|\Delta_{ij}\Delta_{ik}\Delta_{il}\Delta_{im}|\cdot|\partial_{jklm}h(S_n' + \theta D + \theta'(1-2\theta)\Delta_i)|]] \\ &\quad + \frac{1}{2}E[\sum_{i=1}^{n}\chi_i^c\cdot\sum_{j,k,l,m=1}^{p}|\Delta_{ij}\Delta_{ik}\Delta_{il}\Delta_{im}|\cdot|\partial_{jklm}h(S_n' + \theta D + \theta'(1-2\theta)\Delta_i)|] \\ &=: \frac{1}{2}[(A) + (B)]\end{aligned}$$

Observe that

$$\begin{aligned}(A) &\leqslant E\left[E\left[\sum_{j,k,l,m=1}^{p}\max_{1\leqslant i\leqslant n}(\chi_i\left|\partial_{jklm}h(S_n' + \theta D + \theta'(1-2\theta)\Delta_i)\right|)\cdot\max_{1\leqslant j,k,l,m\leqslant p}\sum_{i=1}^{n}|\Delta_{ij}\Delta_{ik}\Delta_{il}\Delta_{im}|\right]\right] \\ &\leqslant C\beta^3\delta^{-1}E\left[\max_{1\leqslant j,k,l,m\leqslant p}\sum_{i=1}^{n}|\Delta_{ij}\Delta_{ik}\Delta_{il}\Delta_{im}|\right] \qquad \text{(by)} \\ &\leqslant C\beta^3\delta^{-1}E\left[\max_{1\leqslant j\leqslant p}\sum_{i=1}^{n}|\Delta_{ij}|^4\right] \\ &\leqslant C\beta^3\delta^{-1}E\left[\max_{1\leqslant j\leqslant p}\sum_{i=1}^{n}|X_{ij}|^4\right] = C\beta^3\delta^{-1}B_2\end{aligned}$$

and

$$\begin{aligned}(B) &\leqslant C\beta^3\delta^{-1}\sum_{i=1}^{n}E\left[\chi^c\max_{1\leqslant j\leqslant p}|\Delta_{ij}|^4\right] \\ &\leqslant C\beta^3\delta^{-1}\sum_{i=1}^{n}E\left[\chi_i^c\max_{1\leqslant j\leqslant p}|X_{ij}|^4\right] \qquad \text{(by symmetry)}\end{aligned}$$

By the same procedure in (Chernozhukov et al., 2014b) partitioning χ_i^c and using the Chebyshev's association inequalities, we get that

$$(B) \leqslant C\beta^3\delta^{-1}B_3.$$

Therefore, we can conclude that

$$|E[f(S_n)] - E[f(T_n)]| \leqslant C\beta\delta^{-1}\left(B_1 + \frac{\beta}{n}(B_2 + B_3)\right).$$

Combining the above steps, one has

$$P(Z\in A) \leqslant (1-\epsilon)^{-1}E[g\circ F_\beta(T_n)] + \frac{C\beta\delta^{-1}\{B_1 + \beta/n(B_2+B_3)\}}{1-\epsilon}$$

$$\leqslant P(F_\beta(T_n) \in A^{e_\beta+3\delta}) + \frac{\epsilon + C\beta\delta^{-1}\{B_1 + \beta/n(B_2+B_3)\}}{1-\epsilon} \qquad \text{(by construction of } g\text{)}$$
$$\leqslant P(\tilde{Z}^* \in A^{2e_\beta+3\delta}) + \frac{\epsilon + C\beta\delta^{-1}\{B_1 + \beta/n(B_2+B_3)\}}{1-\epsilon} \qquad \text{(by)}$$

This completes the proof ■

Taking $\beta = 2\delta^{-1}\log(p\vee n)$ in Lemma 5.3, we get

$$\alpha = \beta^2\delta^2 - 1 = 4\log^2(p\vee n) - 1 \geqslant 2\log(p\vee n) \qquad \text{(recall } n \geqslant 3 > e\text{)}$$

so that

$$\epsilon \leqslant 2\log(p\vee n)/(p\vee n) \leqslant 2n^{-1}\log n.$$

Then we get from Lemma 5.3 that for every $\delta > 0$,

$$P(|Z - \tilde{Z}| > 16\delta) \leqslant \delta^{-2}\{B_1 + \delta^{-2}(B_2+B_3)\log(p\vee n)^2\}\log(p\vee n) + \frac{\log n}{n}$$

We finally apply this to the discretized empirical process. Construct a tight Gaussian random variable G_P in $\ell^\infty(\mathcal{F})$ given in assumption (A3), independent of $X_1, ...X_n$. We note that one can extend G_P to the linear hull of $\mathcal{F}$ in such a way that G_P has linear sample paths. Let $\{f_1, ...f_N\}$ be a minimal $\epsilon||F||_{P,2}$ -net of $(\mathcal{F}, e_P)$ with $N = N(\mathcal{F}, e_P, \epsilon\|F\|_{P,2})$. Then for every $f \in \mathcal{F}$, there exists a function f_j $(1 \leqslant j \leqslant N)$ such that $e_P(f, f_j) < \epsilon\|F\|_{P,2}$. Recall that

$$F_\epsilon = \{f - g : f, g \in \mathcal{F}, e_P(f,g) < \epsilon||F||_{P,2}\}$$

and define

$$Z^\epsilon := \max_{1\leqslant j\leqslant N} \mathcal{G}_n f_j, \qquad \tilde{Z}^\epsilon := \sup G_P f, \tilde{Z^{*\epsilon}} = \max_{1\leqslant j\leqslant N} G_P f_j.$$

Observe that $|Z - Z^\epsilon| \leqslant \|\mathcal{G}_n\|_{\mathcal{F}_\epsilon}$ and $|\tilde{Z}^{*\epsilon} - \tilde{Z}^*| \leqslant ||G_P||_{\mathcal{F}_\epsilon}$ Under the setup of (Chernozhukov et al., 2014b), $\log(N\vee n) = H_n(\epsilon)$.Then for every Borel subset A of $\mathbb{R}$ and $\delta > 0$,

$$P(Z^\epsilon \in A) - P(\tilde{Z}^{*\epsilon} \in A^{16\delta}) \lesssim \delta^{-2}\{B_1 + \delta^{-1}(B_2+B_4)H_n(\epsilon) + n^{-1}\log n$$

Clearly, $B_1 \leqslant n^{-1/2}E[\|\mathbb{G}_n\|_{\mathcal{F}.\mathcal{F}}], B_2 \leqslant n^{-3/2}\kappa^4$, and $B_4 \leqslant n^{-3/2}P[F^3 1(F > \delta\sqrt{n}H_n(\epsilon)^{-1})]$. Hence, choosing $\delta > 0$ in such a way that

$$C\delta^{-2}n^{-1/2}E[||\mathbb{G}_n||_{\mathcal{F}.\mathcal{F}}]H_n(\epsilon) \leqslant \frac{\gamma}{4}$$
$$C\delta{-}3n^{-3/2}\kappa^4 H_n^{7/2}\epsilon \leqslant \frac{\gamma}{4}$$

that is,

$$\delta \geqslant C\max\{\gamma^{-1/2}n^{-1/4}(E[||\mathbb{G}_n||_{\mathcal{F}.\mathcal{F}}])^{1/2}H_n^{1/2}(\epsilon), \gamma^{-1/3}n^{-1/2}\kappa^{4/3}H_n^{7/6}(\epsilon)\}$$

we have

$$P(Z^\epsilon \in A) \leqslant P(\tilde{Z}^{*\epsilon} \in A^{16\delta}) + \frac{\gamma}{2} + \frac{\gamma}{4}\kappa^{-4}P[F^4 1(F > \delta\sqrt{n}H_n(\epsilon)^{-1})] + \frac{C\log n}{n}$$

Note that $\delta \geqslant c\gamma^{-1/2}n^{1/4}(E[\|\mathbb{G}_n\|_{\mathcal{F}.\mathcal{F}}])^{1/2}H_n^{-1/2}(\epsilon)$, so that

$$P[F^4 1(F > \delta\sqrt{n}H_n(\epsilon)^{-1})] \leqslant P[F^4 1(F > c\gamma^{-1/2}n^{1/4}(E[||\mathbb{G}_n||_{\mathcal{F}.\mathcal{F}}])^{1/2}H_n^{-1/3}(\epsilon)]$$

Hence,

$$P(Z^\epsilon \in A) \leqslant P(\tilde{Z}^{*\epsilon} \in A^{16\delta}) + \frac{\gamma}{2} + \frac{\gamma}{4}\kappa^{-4}P\left(F^4 1(F > c\gamma^{-1/2}n^{1/4}(E[||\mathbb{G}_n||_{\mathcal{F}.\mathcal{F}}])^{1/2}H_n^{-1/3}(\epsilon)\right) + \frac{C\log n}{n}$$
$$=: P(\tilde{Z}^{*\epsilon} in A^{16\delta}) + \frac{\gamma}{2} + (\text{error})$$

By bounding $\|\mathbb{G}_n\|_{\mathcal{F}_\epsilon}$ and $\|G_P\|_{\mathcal{F}_\epsilon}$ by Theorem 5.1 of (Chernozhukov et al., 2014b) and Borell-Sudakov-Tsirel'son inequality of (Van den Berg and Ridder, 1998) respectively as in the original proof we have

$$P(Z \in A) \leqslant P(\tilde{Z}^* \in A^{a+b+16\delta}) + \gamma + (\text{error})$$

where a, b is

$$a := K(q)\{\phi_n(\epsilon) + (\epsilon||F||_{P,2} + n^{-1/2}||M||_q)\gamma^{-1/q} + n^{-1/2}||M||_2\gamma^{-2/q}\}$$
$$b := \phi_n(\epsilon) + \epsilon||F||_{P,2}\sqrt{2\log(4/\gamma)}$$

where $K(q)$ is a constant that depends only on q. Then, the conclusion follows from Strassen's theorem (Lemma 4.1 of (Chernozhukov et al., 2014b)).

注 5.4 (Connection to (Chernozhukov et al., 2014a))**.** How does this refined Gaussian approximation result affect the theoretical results in (Chernozhukov et al., 2014a)? In (Chernozhukov et al., 2014a), in order to construct honest confidence bands, they place high level assumptions on the non-parametric estimation procedure (H1) - (H6), where (H1) corresponds to Gaussian approximation. To verify that (H1) is satisfied for the procedure proposed in the paper for VC-type classes, they invoke a corollary of the Gaussian approximation presented above (Theorem 2.1 in the original paper) specialized to VC-type classes (Corollary 2.1 in the original paper) to guarantee the existence of sequences ϵ_{1n} and δ_{1n} bounded from above by Cn^{-c} for some constant. Thanks to the refinement, we see that ϵ_{1n} can be taken to have a tighter upper bound from $c = -1/6$to $c = n^{-1/4}$. However, since the c is common for all conditions (H1) -(H6), it is questionable how tangible the difference of improving one of the rate guarantees is. Nevertheless, we think this type of incremental improvements is also important, and leave developments on the other methods for future work.

6 ATE on the treated survivors :motivation

定義 6.1 (Average treatment effect on the treated survivors (ATETS)).

$$E[Y_2(1) = 1|Y_1(1) = 1, G = 1] - E[Y_2(0) = 1|Y_1(1) = 1, G = 1]$$

We note that if this causal estimand may be of quite interest. It provides us with what will be the causal effect on the transition probability from unemployment to employment for the todo people who would have remain unemployed in the short term. Policymakers must take into consideration such quantities, since the policy that seemed to have no effect in the short term, may turn out to be very effective later on, and vice versa.

However, it can be easily seen that there is a fundamental challenge for identification. The crucial point here is the $E[Y_2(0) = 1|Y_1(1) = 1]$ part, where while the $Y_1(1)$ is the potential outcome for the treated case,whereas $Y_2(0)$ is the untreated . In other words, there are two counterfactual worlds coinciding in this case, which is at the heart of the fundamental problem of casual inference (Imbens and Rubin, 2015)

It should be quite clear that only under the maintained assumptions in the previous section, this quantity is not nonparametrically identified under the observaional population G=1 because in any case the $Y_1(1)$ cannot be reduced to a factual quantity when $W \amalg Y_2(1), Y_2(0), Y_1(1), Y_1(0)|G = 1.$ does not hold. Is there any way to nonparametrically identify this quantity under additional assumptions?

One possibility may be the no state dependence assumption that has be introduced and elaborated in papers like (Heckman, 1981; Heckman and Singer, 1984; Torgovitsky, 2019). The issue in the literature is whether the commonly observed serial correlation between sequential outcomes come from unobserved heterogeneity , that affects both treatment and all the subsequent outcomes, or that there is state dependence, that the fact that one is placed in a state of unemployment in the previous stage itself has an effect on the possibility of unemployment this term(e.g. , in the form of less opportunity for gaining social skills in unemployment). In the latter case, the past potential outcome can be seen as a randomized treatment for the current potential outcome.

If we posit , therefore, that there is no state dependence in our context,i.e. there is no direct (treatment) effect of the past outcome on the current outcome, we could argue that

$$(Y_2(1), Y_2(0)) \amalg (Y_1(1), Y_1(0))|G = 1 \tag{29}$$

holds, [†1]

[†1] there are few papers that use potential outcome notations to formalize no state dependence, and arguably, the appropriate no state dependence

In this case it is easy to see that under the previously maintained assumptions and no state dependence, the ATETS reduces to the usual LTATE (long term average treatment effect) because $E[Y_2(1)|Y_1(1), G = 1] = E[Y_2(1)|G = 1]$, and $E[Y_2(0)|Y_1(0), G = 1] = E[Y_2(0)|G = 1]$, and the identification strategy of (Athey et al., 2020) can be used.

Nevertheless, many empirical papers heckman1981heterogeneity,heckman1984method, torgovitsky2019nonparametric have argued from economic theory and empirical evidence how , while the observed serial correlation between outcomes are often cases not solely by unobserved heterogeneity, it has been argued quite strongly that in the unemployment context, state dependence is also a crucial factor.

Therefore, in the next section , we will build economic models that incorporate the rich accumulation of empirical research on unemployment to gain the joint potential outcomes across counterfactual states, to identify the proposed quantity above.

koko

注 6.2. kokoSince transition probability can be interpreted as one form of hazard parameter, naturally is of major concern. First in terms of right censoring, The way we calculate the transition probability , we do not have to worry about the censoring,

定理 6.3 . dd

7 Economic modeling

8 point identification by structural approach: experimental data as a tool for model-validiation

In this section, we will consider how short-term experimental and observational data can be combined to get a more reliable causal effect under initial selection and dynamic treatment effects. As illustrated above, in order to nonparmetrically point identify the ATETS, strong assumptions like no serial dependence of potential outcomes had to be assumed. This difficultly may motivate movements to come up with alternative procedures for identification and estimation, like building economic models based on the economics of the problem. Since ATETS has an interpretation as a hazard ratio, we explore duration models that enable such derivations. We first briefly introduce the commonly used Mixed-Proportional-Hazard (MPH) model which is a generalization of the proportional hazard as in (Cox, 1972), but show how the multiplicative strucutre has little justification in economic terms. We then turn to the main piece for this section; we introduce an on-the-job search model with endogenous effort, and discuss how the short term experimental data can be used as a reliable source of model specification, similar in spirit to e.g., (Todd and Wolpin, 2020).

We will consider how short-term experimental and observational data can be combined to get a more reliable causal effect under initial selection and dynamic treatment effects.

8.1 mixed proportional hazard models

We will focus on the arguably most popular duration model in social sciences, the mixed proportional hazard model (MPH), and vary various functional forms on it and see how the experimental data can aid in selecting the appropriate one.

While there may be various reasons for the popularity of Mixed proportional hazard models in economics, according to (Heckman and Singer, 1984), van2001duration, abbring2003nonparametric, it is the most parsimonious model that contains the three indispensible elements; baseline hazard, observed covariates, and unobserved heterogeneity in its model.

We basically follow the definition in (Van den Berg, 2001) The MPH model has as its component,

Let $\psi(t)$ be the 'baseline hazard', a possibly time varying function that is common across all individuals, $\theta_0(x)$ be the individual specific component that affects the individual hazard(often specified as $\theta_0(x) = exp(x'\beta)$), and v be the unobserved heterogeneity term.

should not be the joint independence across counterfactuals but for each treatment potential outcome, i.e. $Y_2(w) \amalg Y_2(w)\, for w = 0, 1$. This is similar to the strong ignorability (Rosenbaum and Rubin, 1983) and the weak ignorability (Imbens and Rubin, 2015) difference. For our purposes, we will stick with the strong no state dependence assumption in this paper

定義 8.1 . (Standard MPH model) Let the hazard function of the random variable T evaluated at the duration t be denoted by $\theta(t|x,v)$.
Then MPH specification of a hazards is one sucht that there are functions ψ and θ_0 such that for every t and every x and v, the relationship

$$\theta(t|x,v) = \psi(t) \cdot \theta_0(x) \cdot v \tag{30}$$

holds.

Note that the above definition requires the duration(t), and explanatory variables(x), and the unobserved hetereogeneity(v) to be multiplicatitvely separable. While this simple, reduced form model is widely used, the interpretation of the coefficient may be difficult due to the multiplicative structure. It is easiest to see by a simple example of a job search model, introduced by (Mortensen, 1986; Mortensen and Pissarides, 1999; Van den Berg, 2001) Define the following notations

- $\lambda(t,x)$: parameter of a possion distribution(possibly time and individual specific) which describes the random intervals at which job offers arrive
- F(w) : the distribution function of the wage offer distribution
- b(t,x): unemployment benefit received in the unemployment spell
- $\rho(t,x)$: discount rate $\in [0,1]$
- R(t,x): expected present value of search when the individual follows the optimal strategy
- $\phi(t,x)$: reservation wage, i.e., a job offer is accepted iff the offer exceeds $\phi(t,x)$

Under regularity conditions , the Bellman equation for R(t,x) satisfies the Bellman equation

$$\rho(t,x)R(t,x) = b(t,x) + \lambda(t,x)E_w \max\{0, \frac{w}{\rho(t,x)} - R(t,x)\} \tag{31}$$

, where the expectation is with respect to the wage distribution.

When an offer of w arrives at time t, then the individual either rejects or accepts the offer. They accept the offer if and only if $\frac{w}{\rho(t,x)} - R(t,x) > 0$ iff $w > \rho R(t,x) (= \phi(t,x))$

Transition from unemployment to employment happens when the inidivual receives an offer(rate $\lambda(t,x)$) and accepts it($P(w > \phi(t,x)) = (1 - F(\phi(t,x)))$) , meaning that the hazard rate $\theta(t|x,v)$ can be expressed as $\theta(t|x,v) = \lambda(t,x) \cdot (1 - F(\phi(t,x)))$. This equation shows how the multiplicatitve strucutre in (8.1) is usually not possible, because $(1 - F(\phi(t,x)))$ is in general not multiplicative in$\phi(t,x)$, which in turn is not multiplicative in t and x. While there are few special specifications when such multiplicatitve structure arise (e.g., F(w) being a Pareto distribution, completely myopic agent), they are often difficult to justify in practice.

Also, a distinct but important challenge is that when one attempts semiparametric identification like leaving the distribution of unobserved heterogeneity unspecified, depending on subtle untestable assumptions(the heterogeneity dependent hazard being bounded away from 0), the nonparametric maximum likeihood estimate can have limiting information matrix to explode, leading to slower-than-cubic-root asymptotics arise, as discussed in for example (Hahn, 1994; Ridder and Woutersen, 2003). I provide partial solvency to this problem along with the curse of dimensionality consideration arising in the maximum likelihood estimation in the next secction by providing limit distribution agnostic inference for high-dimensional gaussian approximation of m-estimators, even with parameter spaces possibly non-donsker.

These considerations combined, we turn to a more structural model that can embed more theory without the need to specify the multiplicativity in the hazard.

8.2 on the job search model with endogenous effort: point identification

We turn to a canoncial model for the job training
Consider the following setup

- $F(w)$: the distribution function of the wage offer distribution evaluated at w, the reservation wage or the current employed wage

- λs : parameter of a poisson distribution which describes the random intervals at which job offers arrive, s is a measure of endogenous search effort
- $c(s)$: twice differtiable convex cost function which enters into the individual decision for search effort, with the properties $c(0) = c'(0)$
- δ: parameter of an exponential distribution which descirbes the exogenous job separation rate
- $W(w)$: the value of employment, depending on the reservation wage or current wage

Under the assumption of a forward-looking optimizing agent and some regularity conditions, $W(w)$ solves the bellman equation

$$rW(w) = \max_{s \geqslant 0}\{w - c(s) + \lambda s \int [\max\langle W(x), W(w)\rangle - W(w)]dF(x) + \delta[U - W(w)]]\} \tag{32}$$

where U is the value of non-employed search. The

Estimation Assuming that workers are identical in the sense that they face the same wage offer distribution and the same search cost, jothe labor force separation rate takes the form

$$d = \lambda s(w_i)[1 - F(w)] \tag{33}$$

, where we recall that w is the firm's wage, *delta* is the job destruction rate,

s(w) is the optimal search effort of a worke employed at wage w,

The search intensity function is the unique solution to the functional equation

$$s(w) = \phi(\lambda \int_w^{\bar{w}} \frac{[1 - F(x)]dx}{r + \delta + \lambda(x)]1 - F(x)}) \tag{34}$$

by virtue of equation (2) where $\phi(\cdot) = c'^{-1}(\cdot)$

In particular, in this section, we will model a duration model and the ATETS as the coefficient on the treatment indicator in a duration model

First we see how economic theory can play a role in the initial stage when the interest is in the job training program effect on unemployment. For unemployment, ample economic theory and evidence suggests negative duration dependence due to the so-called 'weed-out effect'. The intuition is that, as time passes, the more competent people are likely to leave first, and the people with lower latent ability remain, and they tend to keep being unemployed.

However, there are limitations to economic theory. While it provides the general declining tendency, it has been challenging to choose between different declining hazard models(e.g., PH, MPH, GAFT models) or to choose the precise distribution for the error term or the functional form, especially under selection. This variety of model choices leads to several model specifications with varying estimates, making it difficult to choose the one to report as the point estimate, which motivates the usage of experimental data.

Specifically, we propose the use of short-term experimental data as the criteria to choose among those competing models, which was filtered in the first stage based on the economics of unemployment.

Before going into the two-step procedure in detail, we shortly draw the connection with the past literature about using experimental data to evaluate econometric estimators. In the paper of, for example, (LaLonde, 1986), it illustrated when the effect of a job training program on earnings was of interest, the various econometric estimators had varied treatment effects, and it was difficult to choose between those estimators without the 'ground truth' by an RCT. While there have been several papers elaborating on the usefulness of matching estimators or model calibration (Heckman et al., 1997; Heckman et al., 1998) of nonexperimental estimators, there is a general consensus of experimental data having external validity. The novel contribution of this paper relative to that literature is twofold: (a) while the vast majority of documents taking the lalonde 1986 approach focuses on outcomes of earnings, this paper focuses on unemployment duration, which seems to be relatively few. A notable exception is by (Ham and LaLonde, 1996). Still, it focuses on noncompliance, a significant problem, but what we abstract from in this paper (b) we specifically focus on the *dynamics* of treatment effects, in which the previous sections have illustrated how short-term experimental data alone may not fully capture the dynamic confounding that is likely to occur. Therefore, the approach here is to take a two-layer approach to (i) first let the theory guide the general dynamics that is plausible in the form of shape restrictions on the hazard and then (ii) let the short-term experimental data

choose between those subtle difference in specifications to capture the baseline confounding or the distributional aspect that is arguably more stable across time. We hope this work spurs more integration of the two approaches to more effectively answer policy-relevant questions, as in (Todd and Wolpin, 2020)

8.3 specific procedure

We will briefly overview the different duration models that might be relevant to our current paper. We will focus on the arguably predominant duration model, mixed proportional hazard model (MPH), and vary various functional forms on it and see how the experimental data can aid in selecting the appropriate one.
While there may be various reasons for the popularity of Mixed proportional hazard models in economics, according to (Heckman and Singer, 1984), van2001duration, abbring2003nonparametric there are mainly two reasons. Firstly, it is the most parsimonious model that contains the three indispensible elements; baseline hazard, observed covariates, and unobserved heterogeneity in its model. Secondly, an economic interpretation of the estimates is more amenable relative to other accelerated failure time models or general transformation models that do not distinguish the general and the individual hazard rates.

9 Partial id of ATETS

In the final section, we recognize that point identification is only acheived at the limit of partially identifying assumptions, and we investigate the identifying power of exogenous variation provided by experimental data and the partially identifying assumptions motivated by economic theory.

From the perspective of the partial identification literature, potential outcomes can be seen as an interval data, with the upper and lower bound determined by the support of $Y_t(1)$, in this case $[0,1]$. Partial identification on interval outcomes have been extensively studied from manski 2003, manski and tamer 2002 to name a few. On the other hand, partial identification on interval covariate outcomes have received much less attention, and for the few research as in berst molinari 2008, they have been found to be notoriously complex to analyze. In our case , For quantities like $E[Y_2(0)|Y_1(1)]$,we have *both interval outcome and covariate*, which highlights the challenge, while the binary treatment and outcome alleviates such challenges. manski and horowitz 2000 considers a similar setting but with different assumptions like MCAR.

We provide our estimates from the worst case bounds and gradually increasing the strenght of our assumptions. We first start with assuming that we only have information on the observational data, and in the absence of any further assumptions, what can be said about ATETS $= E[Y_2(1)|Y_1(1), G=1] - E[Y_2(0)|Y_1(1), G=1]$? We prepare the following lemma from manski and horowitz 2000 .

We first begin the analysis on the ATETS assuming that only the observational data is available, and derive worst case bounds. The following proposition derives the sharp identification bounds under no additional assumptions.

定理 9.1 . the bounds on

$$\text{ATETS} = E[Y_2(1)|Y_1(1), G=1] - E[Y_2(0)|Y_1(1), G=1]$$

Assuming that only the observational data is available, absent any further information, the sharp identification bound is

$$P(Y_2=1|Y_1=0, Z_y=1, Z_x=1)P(Y_1=0|Z_y=1, Z_x=1)P(Z_y=1, Z_x=1) \tag{35}$$
$$\leqslant E[Y_2(0)=1|Y_1(1)=0, G=1] \tag{36}$$
$$\leqslant P(Y_1=0|Z_y=0, Z_x=1)P(Z_y=0, Z_x=1) + P(Z_y=0, Z_x=0) \tag{37}$$
$$+ P(Y_2=1|Z_y=1, Z_x=0)PZ_y=1, Z_x=0) \tag{38}$$

[証明]. It suffices to show that $E[Y_2(0)=1|Y_1(1)=0, G=1]$ is uninformative. Consider the missingness indicator Z_x, Z_y that is equal to 1 if and only if the covariate $Y_1(0)$ and the outcome $Y_2(1)$ is observed and 0 otherwise. Then, by an application of Bayes rule, and law of total probability

$$E[Y_2(0)=1|Y_1(1)=0, G=1] \tag{39}$$
$$= \sum_{j,k} \frac{P(Y_2=1|Y_1=0, Z_y=j, Z_x=k)P(Y_1=0|Z_y=j, Z_x=k)P(Z_y=k, Z_x=j)}{\sum_{k,j} P(Y_1=0|Z_y=j, Z_x=k)P(Z_y=j, Z_x=k)} \tag{40}$$

then by applying horowitz and manski 1995 Corollary 1.1, we obtain the following sharp bounds :

$$
\begin{aligned}
& P(Y_2 = 1|Y_1 = 0, Z_y = 1, Z_x = 1)P(Y_1 = 0|Z_y = 1, Z_x = 1)P(Z_y = 1, Z_x = 1) && (41)\\
& \leqslant E[Y_2(0) = 1|Y_1(1) = 0, G = 1] && (42)\\
& \leqslant P(Y_1 = 0|Z_y = 0, Z_x = 1)P(Z_y = 0, Z_x = 1) + P(Z_y = 0, Z_x = 0) && (43)\\
& \quad + P(Y_2 = 1|Z_y = 1, Z_x = 0)PZ_y = 1, Z_x = 0) && (44)
\end{aligned}
$$

then by horowitz and manski 2000 Corollary 1.1, we can conclude that the bounds are uninformative. ■

Next, we consider how the short-term experimental data aids us in tightening the bound. We assume external validity of the experimental data, which can be tested based on our method in Section 1.1. The following theorem indicates the sharp identification bound for this case.

The key step is to notice that

(i)
$$E[Y_2(1) = 1|Y_1(1) = 0] = \frac{E[Y_2(1)(1 - Y_1(1))|G = 1]}{E[(1 - Y_1(1))|G = 1]}$$
is point identified, and that the denominator of
$$E[Y_2(0) = 1|Y_1(0) = 1] = \frac{E[Y_2(0)(1 - Y_1(1))|G = 1]}{E[(1 - Y_1(1))|G = 1]}$$
is also point identified.

(ii) the only that needs to be bound is
$$E[Y_2(0)(1 - Y_1(1))|G = 1] = P(Y_2(0) = 1, Y_1(1) = 0|G = 1)$$
which we can get sharp bounds by an analogous procedure to the Frechet-Hoeffding, a classical sharp bound for the joint distribution whose marginals are identified.

定理 9.2 . In addition to the setting in Theorem 9.1, assume availability of the experimental data, accompanied by the internal validity (Assumption2.1) and external validity (Assumption2.2). Absent any further information, the sharp identification for ATETS is

$$
\begin{aligned}
& \frac{E[E[Y_2(1 - Y_1)|W = 1, X, G = 0]G = 1]}{E[E[(1 - Y_1)|W = 1, X, G = 0]|G = 1]}\\
& \quad - \max(0, P(Y_1|W = 1, G = 0) + P(Y_2|W = 0, G = 1)P(W = 0) + P(W = 1) - 1)\\
& \quad\quad \times E[E[(1 - Y_1)|W = 1, X, G = 0]|G = 1]\\
& \leqslant \frac{E[E[Y_2(1 - Y_1)|W = 1, X, G = 0]G = 1]}{E[E[(1 - Y_1)|W = 1, X, G = 0]|G = 1]}\\
& \quad - \min(P(Y_1|W = 1, G = 0), P(Y_2|W = 0, G = 1)P(W = 0))\\
& \quad\quad E[E[(1 - Y_1)|W = 1, X, G = 0]|G = 1]
\end{aligned}
$$

[証明]. As in the brief explanation in the body, what we need to show is that the sharp identification bound for $P(Y_2(0) = 1, Y_1(1) = 0|G = 1)$ is

$$
\begin{aligned}
\Big[&\max(0, P(Y_1 = 0|W = 1, G = 0) + P(Y_2 = 1|W = 0, G = 1)P(W = 0) + P(W = 1) - 1),\\
&\min(P(Y_1 = 0|W = 1, G = 0), P(Y_2 = 1|W = 0, G = 1)P(W = 0))\Big]
\end{aligned}
$$

Note that by the Frechet-Hoeffding theorem, the sharp identification bound when both $P(Y_2(0) = 1|G = 1)$ and $P(Y_1(1) = 0|G = 1)$ is identified is

$$
\begin{aligned}
& [\max(0, P(Y_2(0) = 1|G = 1) + P(Y_1(1) = 0|G = 1) - 1), && (45)\\
& \quad \min(P(Y_2(0) = 1|G = 1), P(Y_1(1) = 0|G = 1))] && (46)
\end{aligned}
$$

While $P(Y_1(1) = 0|G = 1)$ is point identified as $P(Y_1 = 0|W = 1, G = 0)$ from external validity and experimental internal validity, $P(Y_2(0) = 1|G = 1)$ is not because the long term outcome is not available for the experimental data. The sharp

bound for this quantity is $[P(Y_2|W=0,G=1)P(W=0), P(Y_2|W=0,G=1)P(W=0)+P(W=1)]$ by the same argument as in Section 1.2. hence taking this lower bound for the upper bound of 45, and the upper bound for the lower bound of 45 is the sharp bound. This is because the bound collapses to a point(equals to 0 a.s.) for the distribution that takes $P(Y_2=1|W=0,G=1)=0$, which is admissible under the given assumptions. ■

The above has shown the strength of experimental data even only for the short term, which tightened the bound more than half relative to the worst case. However, there is still a fundamental nonidentification for $P(Y_2(0)=1, Y_1(1)=0|G=1)$, which may possibly be still wide. We now examine structural assumptions that can be deduced by a qualitative analysis of a structural model pertaining to the unemployment dynamics which is the main interest. Notably, our modest aim of gaining reliable *partial identifying assumptions* obviates the need for 'extra-theoretical' parametric or distributional assumptions that has been the concern even for the supporters of the structural approach(Keane et al., 2011; Todd and Wolpin, 2020)

We formulate the problem as an on-the-job search with endogenous effort. We basically stick to the canonical framework from (Faberman et al., 2022; Mortensen and Pissarides, 1999; Torgovitsky, 2019). While our notation basically builds upon (Torgovitsky, 2019) , the substantive crucially differs in that (i) their potential outcome denotes the previous time employment status, in our case, it is the job training participating status, and (ii) several modifications of their theorems and proofs are necessary due to the additional job training status, as we discuss below.

Formally, consider the following setting. In stage t=0 , workers decide to join in job traiing program or not (W=1 , 0) based on the their exante evaluation of the discounted sum of reward

$$W = 1\left\{\sum_t \frac{1}{\delta^t}(Y_t(1)-Y_t(0))>0\right\}$$

worker i begins period t having either been employed or unemployed in the previous period ($Y_{i(t-1)}=1, or\, 0$).

They have exerted $E_{i(t-1)}$ units of search effort in the previous period. the worker receives a wage offe $\omega(Y_{i(t-1)}, E_{i(t-1)}, W_i, A_i, V_{it})$ depending on their work status and effort choices in the previous period (Y_2), whether they joined a job training program W_i, (time-invariant) source of (unobserved) heterogeneity A_i, and time varying wage shock V_{it}.

After observing the wage offer, the worker decides to either accept it and work in period t ($Y_{it}=1$) or the remain unemployed$Y_{it}=0$ (Not receiving an offer or being laid off corresponds to receiving an offer $-\infty$)

The criteria for this decision is based on maximization of their expected present-discounted utility using discount factor $\delta \in (0,1)$.

Specifically, under mild regularity conditions, (e.g. (Stokey, 1989; Rust, 1994)agent i's problem can be written recursively in terms of the Bellman equation, maintaining that the wage shock follows a first order Markov process throughout [†2]

$$\nu(Y_{i(t-1)}, E_{i(t-1)}, V_{it}, A_i) \tag{47}$$
$$= \max_{(y',e')\in\{0,1\}\times\mathcal{E}} \{\mu(y',e',Y_{i(t-1)},A_i,V_{it}) + \delta E[\nu(y',e',V_{i(t+1)},A_i)|Y_{i(t-1)},E_{i(t-1)},V_{it},A_i]\} \tag{48}$$
$$=: \max_{(y',e')\in\{0,1\}\times\mathcal{E}} \dot{\nu}(Y_{i(t-1)}, E_{i(t-1)}, V_{it}, A_i) \tag{49}$$

where

- ν : value function
- $\mu(y',e',Y_{i(t-1)},W_i,A_i,V_{it}) = y'\omega(Y_{i(t-1)},E_{i(t-1)},W_i,A_i,V_{it}) - \kappa(y',e',W_i,A_i)$ is the worker's flow utility
- $\omega(Y_{i(t-1)}, E_{i(t-1),W_i,A_i,V_{it}})$: wage offer at time t.
- $\kappa(y',e',A_i)$: cost of exerting e' units of search effort when making employment choice y'
- $\dot{\nu}$: short hand notation that combines flow utility and continuation value
- $S_{it} := (Y_{i(t-1),E_{i(t-1)},W_iA_i,V_{it}})$ is the state variables at time t.

Assuming there is a solution to this problem, profiling the effort decision for a fixed employment decision y' gives

$$e^\star(S_{it}||y') \tag{50}$$

[†2] This is a commonly maintained assumption in the structural literature.

$$:= \arg\max_{e' \in \mathcal{E}} \dot{\nu}(y', e', Y_{i(t-1)}, W_i, E_{i(t-1)}, V_{it}, A_i) \tag{51}$$

$$= \arg\max_{e' \in \mathcal{E}} -\kappa(y', e', W_i, A_i) + \delta E[\nu(y', e', W_i, V_{i(t+1)}, A_i)|S_{it}] \tag{52}$$

using this, we can rewrite the above by

$$\nu(S_{it}) = \max_{y' \in \{0,1\}} \dot{\nu}(y', e^\star(S_{it}||y'), S_{it}) =: \max_{y' \in \{0,1\}} \dot{\nu}(S_{it}||y')$$

The observed binary choice Y_{it} is assumed to be the optimizer of the above:

$$Y_{it} = \arg\max_{y' \in \{0,1\}} \dot{\nu}(S_{it}||y') = 1\{\Delta\dot{\nu}(S_{it}) \geqslant 0\}$$
$$\Delta\dot{\nu}(S_{it}) := \dot{\nu}(S_{it}||1) - \dot{\nu}(S_{it}||0)$$

and ties are broken in favor of $Y_{it} = 1$.

For our purposes of bounding the , we need to define the potential outcomes $Y_{it}(1)$ within this framework. We assume throughout that the wage shock $\{V_{it}\}_{t=1}^{t=T}$ follows a first order Markov process. Then, the the conditioning set for the continuation value portion of 47 can be rewritten as

$$\nu(Y_{i(t-1)}, E_{i(t-1)}, V_{it}, A_i) = \max_{(y',e') \in \{0,1\} \times \mathcal{E}} \{\mu(y', e', Y_{i(t-1)}, A_i, V_{it}) + \delta E[\nu(y', e', V_{i(t+1)}, A_i)|W_i, V_{it}, A_i]\} \tag{53}$$

Likewise, (50) can be rewritten as

$$e^\star(W_i, V_{i(t-1)}, A_i||y') := \arg\max_{e' \in \mathcal{E}} \dot{\nu}(y', e', W_i, V_{it}, A_i) \tag{54}$$

Moreover, note that the counterfactual state in period t had the worker chosen job training status $w \in \{0, 1\}$ is

$$S_{it}(w) := (Y_{i(t-1)}, e^*(w, V_{i(t-1)}, A_i||Y_{i(t-1)}), w, V_{it}, A_i) \tag{55}$$

which depends on the past employment status $Y_{i(t-1)}$, the hypothesized job training status w, the previous and current period wage shocks $(V_{i(t-1),V_{it}})$and time invariant heterogeneity A_i. The worker's present-discounted net utility from choosing employment if the job status was w can therefore be written as

$$\Delta\dot{\nu}(S_{it}(w)) = \omega(y, e^*(w, V_{i(t-1)}, A_i||Y_{i(t-1)}), A_i, V_{it}) - \Delta\kappa(V_{it}, A_i) + \Delta\gamma(w_i, V_{it}, A_i) \tag{56}$$

where $\Delta\kappa$ and $\Delta\gamma$ are shorthand for

$$\Delta\kappa(V_{it}, A_i) = [\kappa(1, e^*(w, V_{i(t-1)}, A_i||1), A_i) - \kappa(0, e^*(w, V_{i(t-1)}, A_i||0))]$$

and

$$\Delta\gamma(V_{it}, A_i) = \delta E[\nu(1, e^*(w, V_{i(t-1)}, A_i||1), V_{i(t+1)}) - \nu(0, e^*(w, V_{i(t-1)}, A_i||0), V_{i(t+1)})|W_i = w, V_{it}, A_i]$$

Then $Y_{it}(w)$ is defined as

$$Y_{it}(w) = \Delta\dot{\nu}(S_{it}(w)) = \omega(y, e^*(w, V_{i(t-1)}, A_i||Y_{i(t-1)}), A_i, V_{it}) - \Delta\kappa(V_{it}, A_i) + \Delta\gamma(w_i, V_{it}, A_i) \tag{57}$$

10 Estimation

10.1 estimation of ATETS under no state dependence

We provide the nonparametric influence function for estimating the ATETS, which coincides with the ATE . One paper we know that have derived the nonparametric influence function for this estimand is (Chen and Ritzwoller, 2021), which uses the projection upon the tangent space approach, and using an ad hoc guess and verify approach to getting the influence function. Here we provide an alternative, quicker, and mechanical approach to deriving the influence function using the path-wise derivative calculation, similar in spirit to (Ichimura and Newey, 2022) Note that in the nonparametric case, the influence function in the tangent space is unique, so the mean zero function with the appropriate inner product with the score will be what we wanted. todo

定理 10.1 . The efficient influence function for $\tau = E[Y_2(1)|Y_1(1), G=1] - E[Y_2(0)|Y_1(1)] =: \tau_1 - \tau_0$ is

$$\frac{gw}{p(G=1)}\Bigg[(y_2 - E[Y_2|W=1,Y_1,X,G=1])\frac{P(G=0|Y_1,W=1,X)}{P(G=1|Y_1,W=1,X)}\frac{1}{P(W=1|X,G=1)} \tag{58}$$

$$+ E[E[Y_2|W=1,Y_1,X,G=1|W=1,X,G=0]] - \tau_1\Bigg] \tag{59}$$

$$+ \frac{(1-g)P(G=1|X)}{P(G=1)P(G=0|X)} \tag{60}$$

$$\times \left(\frac{w(E[Y_2|Y_1,W=1,X,G=1] - E[E[Y_2|Y_1,W=1,X,G=1]|W=1,X,G=0]])}{p(W=1|X,G=0)}\right) \tag{61}$$

$$-[\frac{g(1-w)}{p(G=1)}\Bigg[(y_2 - E[Y_2|W=0,Y_1,X,G=1])\frac{P(G=0|Y_1,W=0,X)}{P(G=1|Y_1,W=0,X)}\frac{1}{P(W=0|X,G=1)} \tag{62}$$

$$+ E[E[Y_2|W=0,Y_1,X,G=1|W=0,X,G=0]] - \tau_0\Bigg] \tag{63}$$

$$+ \frac{(1-g)P(G=1|X)}{P(G=1)P(G=0|X)} \tag{64}$$

$$\times \left(\frac{(1-w)(E[Y_2|Y_1,W=0,X,G=1] - E[E[Y_2|Y_1,W=0,X,G=1]|W=0,X,G=0]])}{p(W=0|X,G=0)}\right)] \tag{65}$$

[証明]. Based on the argument in Section 6 , by symmetry, it suffices to show the influence function for$\tau_1 = E[Y_2(1)|Y_1(1), G=1] = E[E[E[Y_2|Y_1,W=1,X,G=1]|W=1,X,G=0]|G=1]$ we use $\partial_t f(t)$ to denote $\frac{\partial f(t)}{\partial_t}$ For parameter θ, let θ_t be the parameter under a regular parametric sub-model indexed by t, that includes the ground-truth model at $t=0$. Let V be the set of all observed variables. In order to obtain the influence function, we need to find a random variable ψ with mean zero, that satisfies,

$$\partial_t \psi_t = E[\psi S(V)] \tag{66}$$

where $S(V) = \partial_t log p_t(V)$. To simplify notation, we assume that all variables are discrete.

First note that

$$\partial_t \psi_t = \partial_t \sum_{y_2,y_1,x} y_2 p_t(y_2|y_1,W=1,X,G=1)p_t(y_1|W=1,x,G=0)p_t(x|G=1) \tag{67}$$

$$= \sum_{y_2,y_1,x} y_2 \partial_t p_t(y_2|y_1,W=1,X,G=1)p(y_1|W=1,x,G=0)p(x|G=1) \tag{68}$$

$$+ \sum_{y_2,y_1,x} y_2 p(y_2|y_1,W=1,X,G=1)\partial_t p_t(y_1|W=1,x,G=0)p(x|G=1) \tag{69}$$

$$+ \sum_{y_2,y_1,x} y_2 p(y_2|y_1,W=1,X,G=1)p(y_1|W=1,x,G=0)\partial_t p_t(x|G=1) \tag{70}$$

for the first term in (67), we have

$$= \sum_{y_2,y_1,x} y_2 \partial_t p_t(y_2|y_1,W=1,X,G=1)p(y_1|W=1,x,G=0)p(x|G=1)$$

$$= \sum_{y_2,y_1,x} y_2 p(y_2|y_1,W=1,X,G=1)p(y_1|W=1,x,G=0)p(x|G=1)S(y_2|y_1,W=1,x,G=1)$$

$$= \sum_{y_2,y_1,w,x,g} wgy_2 \frac{p(y_2,w|y_1,x,g)}{p(W=1|y_1(1),x,g)}$$

$$\times p(y_1|W=1,x,G=0)\frac{p(g,x)}{p(G=1)}S(y_2|y_1,W=1,x,G=1)$$

$$= \sum_{y_2,y_1,w,x,g} wgy_2 \frac{p(y_2,w|y_1,x,g}{p(W=1|y_1(1),x,g)}p(y_1|W=1,x,G=0)$$

$$\times \frac{p(g,x)}{p(G=1)}S(y_2|y_1,w,x,g) koko$$

$$= \sum_{y_2,y_1,w,x,g} wgy_2 p(y_2, w|y_1, x, g) \frac{P(G=0|Y_1, W=1, X)}{P(G=1|Y_1, W=1, X)} \frac{1}{P(W=1|X, G=1)}$$
$$\times p(y_1|W=1, x, G=0) \frac{p(g,x)}{p(G=1)} S(y_2|y_1, w, x, g) \qquad (\because \text{Bayes rule, Assumption 2.2 and 2.1})$$

$$= E\left[\frac{wg}{p(G=1)} y_2 \frac{P(G=0|Y_1, W=1, X)}{P(G=1|Y_1, W=1, X)} \frac{1}{P(W=1|X, G=1)} S(y_2|y_1, w, x, g)\right]$$
$$= E\Bigg[\frac{wg}{p(G=1)} (y_2 - E[Y_2|y_1, W=1, X, G=1]) \frac{P(G=0|Y_1, W=1, X)}{P(G=1|Y_1, W=1, X)}$$
$$\times \frac{1}{P(W=1|X, G=1)} S(y_2|y_1, w, x, g)\Bigg]$$

Note that

$$E\Bigg[\frac{wg}{p(G=1)} (y_2 - E[Y_2|y_1, W=1, X, G=1])$$
$$\times \frac{P(G=0|Y_1, W=1, X)}{P(G=1|Y_1, W=1, X)} \frac{1}{P(W=1|X, G=1)} S(y_1, w, x, g)\Bigg] = 0$$

Therefore

$$\sum_{y_2,y_1,x} y_2 \partial_t p_t(y_2|y_1, W=1, X, G=1) p(y_1|W=1, x, G=0) p(x|G=1)$$
$$= E\Bigg[\frac{wg}{p(G=1)} (y_2 - E[Y_2|y_1, W=1, X, G=1]) \frac{P(G=0|Y_1, W=1, X)}{P(G=1|Y_1, W=1, X)} \frac{1}{P(W=1|X, G=1)} S(V)\Bigg]$$

Likewise for the second term in 67

$$\sum_{y_2,y_1,x} y_2 p(y_2|y_1, W=1, X, G=1) \partial_t p_t(y_1|W=1, x, G=0) p(x|G=1)$$
$$= \sum_{y_2,y_1,x} y_2 p(y_2|y_1, W=1, X, G=1) p(y_1|W=1, x, G=0) p(x|G=1) S(y_1|W=1, x, G=0)$$

noting that

$$p(y_1|w=1, xg=0) = \frac{p(y_1, w=1, g=0|x)}{p(G=0, W=1|X)} = \frac{w(1-g)p(y_1, wg|x)}{p(g=0, w=1|x)},$$
$$p(x|G=1) = \frac{p(G=1, x)}{p(G=1)},$$

and rearranging terms,

$$\sum_{y_2,y_1,x} y_2 p(y_2|y_1, W=1, X, G=1) \partial_t p_t(y_1|W=1, x, G=0) p(x|G=1)$$
$$= \sum_{y_2,y_1,w,x,g} \frac{w(1-g) y_2 p(y_2|y_1, W=1, X, G=1) p(G=1|X)}{p(G=1) p(G=0|X) p(W=1|X, G=0)} E[Y_2|Y_1, W=1, X, G=1] S(y_1|w, x, g)$$
$$= E\left[\frac{W(1-G)p(G=1|X)}{p(G=1)p(G=0|X)p(W=1|X, G=0)} E[Y_2|Y_1, W=1, X, G=1] S(y_1|w, x, g)\right]$$
$$= E\Bigg[\frac{W(1-G)p(G=1|X)}{p(G=1)p(G=0|X)p(W=1|X, G=0)} (E[Y_2|Y_1, W=1, X, G=1]$$
$$- E[E[Y_2|Y_1, W=1, X, G=1]|W=1, X, G=0]) S(y_1|w, x, g)\Bigg]$$

Note that,

$$E\Bigg[\frac{W(1-G)p(G=1|X)}{p(G=1)p(G=0|X)p(W=1|X, G=0)} (E[Y_2|Y_1, W=1, X, G=1] -$$
$$E[E[Y_2|Y_1, W=1, X, G=1]|W=1, X, G=0]) S(W, x, G)\Bigg] = 0$$

therefore

$$\sum_{y_2,y_1,x} y_2 p(y_2|y_1, W=1, X, G=1) \partial_t p_t(y_1|W=1, x, G=0) p(x|G=1)$$

$$=E\left[\frac{W(1-G)p(G=1|X)}{p(G=1)p(G=0|X)p(W=1|X,G=0)}(E[Y_2|Y_1,W=1,X,G=1]-\right.$$
$$\left.E[E[Y_2|Y_1,W=1,X,G=1]|W=1,X,G=0]])S(V)\right]$$

for the third term

$$\sum_{y_2,y_1,x} y_2 p(y_2|y_1,W=1,X,G=1)p(y_1|W=1,x,G=0)\partial_t p_t(x|G=1)$$
$$=\sum_{y_2,y_1,x} y_2 p(y_2|y_1,W=1,X,G=1)p(y_1|W=1,x,G=0)p(x|G=1)S(x|G=1)$$
$$=\sum_{y_2,y_1,x}\frac{G}{P(G=1)} y_2 p(y_2|y_1,W=1,X,G=1)$$
$$\times p(y_1|W=1,x,G=0)p(x|G=1)S(x|G=1)$$
$$=\sum_{y_2,y_1,x,g}\frac{g}{P(G=1)}(E[E[Y_2|Y_1,W=1,X,G=1]|W=1,X,G=0]]-$$
$$\tau_1)p(x,g)S(x|g)$$

note that

$$=\sum_{y_2,y_1,x,g}\frac{g}{P(G=1)}(E[E[Y_2|Y_1,W=1,X,G=1]|W=1,X,G=0]]-$$
$$\tau_1)p(x,g)S(g)=0$$

therefore,

$$\sum_{y_2,y_1,x} y_2 p(y_2|y_1,W=1,X,G=1)p(y_1|W=1,x,G=0)\partial_t p_t(x|G=1)$$
$$=E\left[\frac{g}{P(G=1)}(E[E[Y_2|Y_1,W=1,X,G=1]|W=1,X,G=0]-\tau_1)p(x,g)S(V)\right]$$

collectivizing the three results, we have

$$\partial_t\psi_t=E\left[\frac{gw}{p(G=1)}\left[(y_2-E[Y_2|W=1,Y_1,X,G=1])\frac{P(G=0|Y_1,W=1,X)}{P(G=1|Y_1,W=1,X)}\frac{1}{P(W=1|X,G=1)}\right.\right.$$
$$\left.+E[E[Y_2|W=1,Y_1,X,G=1|W=1,X,G=0]]-\tau_1\right]+\frac{(1-g)P(G=1|X)}{P(G=1)P(G=0|X)}$$
$$\left.\times\left(\frac{w(E[Y_2|Y_1,W=1,X,G=1]-E[E[Y_2|Y_1,W=1,X,G=1]|W=1,X,G=0]])}{p(W=1|X,G=0)}\right)S(V)\right]$$

which implies that

$$\frac{gw}{p(G=1)}\left[(y_2-E[Y_2|W=1,Y_1,X,G=1])\frac{P(G=0|Y_1,W=1,X)}{P(G=1|Y_1,W=1,X)}\frac{1}{P(W=1|X,G=1)}\right.$$
$$\left.+E[E[Y_2|W=1,Y_1,X,G=1|W=1,X,G=0]]-\tau_1\right]+$$
$$\frac{(1-g)P(G=1|X)}{P(G=1)P(G=0|X)}\frac{w(E[Y_2|Y_1,W=1,X,G=1]-E[E[Y_2|Y_1,W=1,X,G=1]|W=1,X,G=0]])}{p(W=1|X,G=0)}$$

Since this is mean zero and because the nonparametric influence is unique, this is the influence function for τ_1. ■

Next, we will prove that this efficient influence function is neyman orthogonal to its nuisance parameters. We verify this by showing that the functional taylor expansion is zero at its true value.

定理 10.2 . Consider the setting above and $\tau_1-\tau_0$. the efficient influence function provided in Theorem 10.1 is Neyman-orthogonal with respect to its nuisance parameters.

10.2 estimation of the duration models

In this section, we provide the semiparametric estimation of the duration parameters provided in section todo

However, it is well known that semiparametric duration models have several nonregular characteristics that make their estimation difficult. For example, hahn 1994 proved that the mixed proportional hazard model cannot be $\sqrt{n}$ estimable under conditions by first showing how their model was part of the mixture model in chamberlain 1986 , and then indirectly using the result in Pfanzagl 2000.

However, for our purposes, the alternative impossibility result based on classical semiparametric theory is sufficient. Specifically, using van der vaart 1991, which shows that if a parameter is regular estimable,the norm of its pathwise derivative is bounded. Therefore, it is sufficient to show that the pathwise derivative of a parametric sub model has a unbounded norm in our problem which implies non $\sqrt{n}$ estimability. Below shows the formal proof todo

In later work as in ridder and weid 2003 , they show that $\sqrt{n}$ estimability for certain classed semiparametric duration models is quite sensitive to the shape of the hazard around zero, and the limiting distribution varies greatly, affecting the finite sample behavior significantly. An effective method around still seems to be an open question, which may be part of the reason for the relative lack of empirical research using such methods. Therefore, I provide non asymptotic Gaussian approximation method that is useable agnostic to the limiting distribution, as an application of the seminal works of chernozhukov , 2012

参考文献

Arellano, M. and Bonhomme, S. (2011). Nonlinear panel data analysis. Annual Review of Economics, 3(1):395–424.

Ashenfelter, O. and Card, D. (1985). susing the longitudinal structure of earnings to estimate the effect of training programs, treview of economics and statistics; 67 (4).

Athey, S., Chetty, R., and Imbens, G. (2020). Combining experimental and observational data to estimate treatment effects on long term outcomes. arXiv preprint arXiv:2006.09676.

Blundell, R., Gosling, A., Ichimura, H., and Meghir, C. (2007). Changes in the distribution of male and female wages accounting for employment composition using bounds. Econometrica, 75(2):323–363.

Chen, J. and Ritzwoller, D. M. (2021). Semiparametric estimation of long-term treatment effects. arXiv preprint arXiv:2107.14405.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014a). Anti-concentration and honest, adaptive confidence bands. The Annals of Statistics, 42(5):1787–1818.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014b). Gaussian approximation of suprema of empirical processes. The Annals of Statistics, 42(4):1564–1597.

Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2):187–202.

Faberman, R. J., Mueller, A. I., Şahin, A., and Topa, G. (2022). Job search behavior among the employed and non-employed. Econometrica, 90(4):1743–1779.

Ghanem, D., Sant'Anna, P. H., and Wüthrich, K. (2022). Selection and parallel trends. arXiv preprint arXiv:2203.09001.

Hahn, J. (1994). The efficiency bound of the mixed proportional hazard model. The Review of Economic Studies, 61(4):607–629.

Ham, J. C. and LaLonde, R. J. (1996). The effect of sample selection and initial conditions in duration models: Evidence from experimental data on training. Econometrica: Journal of the Econometric Society, pages 175–205.

Heckman, J. and Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. Econometrica: Journal of the Econometric Society, pages 271–320.

Heckman, J. J. (1981). Heterogeneity and state dependence. In Studies in labor markets, pages 91–140. University of Chicago Press.

Heckman, J. J., Ichimura, H., Smith, J. A., and Todd, P. E. (1998). Characterizing selection bias using experimental data.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. The review of economic studies, 64(4):605–654.

Heckman, J. J. and Navarro, S. (2007). Dynamic discrete choice and dynamic treatment effects. Journal of Econometrics, 136(2):341–396.

Ichimura, H. and Newey, W. K. (2022). The influence function of semiparametric estimators. Quantitative Economics, 13(1):29–61.

Imbens, G. W. and Rubin, D. B. (2015). Causal inference in statistics, social, and biomedical sciences. Cambridge University Press.

Keane, M. P., Todd, P. E., and Wolpin, K. I. (2011). The structural estimation of behavioral models: Discrete choice dynamic programming methods and applications. In Handbook of labor economics, volume 4, pages 331–461. Elsevier.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. The American economic review, pages 604–620.

Manski, C. F. (2009). Identification for prediction and decision. Harvard University Press.

Molchanov, I. (2005). Theory of random sets, volume 19. Springer.

Molchanov, I. and Molinari, F. (2018). Random sets in econometrics, volume 60. Cambridge University Press.

Mortensen, D. T. (1986). Job search and labor market analysis. Handbook of labor economics, 2:849–919.

Mortensen, D. T. and Pissarides, C. A. (1999). New developments in models of search in the labor market. Handbook of labor economics, 3:2567–2627.

Ridder, G. and Woutersen, T. M. (2003). The singularity of the information matrix of the mixed proportional hazard model. Econometrica, 71(5):1579–1589.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. Oxford economic papers, 3(2):135–146.

Rust, J. (1994). Chapter 51 Structural estimation of markov decision processes. volume 4 of Handbook of Econometrics, pages 3081–3143. Elsevier. ISSN: 1573-4412.

Stokey, N. L. (1989). Recursive methods in economic dynamics. Harvard University Press.

Todd, P. E. and Wolpin, K. I. (2020). The best of both worlds: Combining rcts with structural modeling.

Torgovitsky, A. (2019). Nonparametric inference on state dependence in unemployment. Econometrica, 87(5):1475–1505.

Van den Berg, G. J. (2001). Duration models: specification, identification and multiple durations. In Handbook of econometrics, volume 5, pages 3381–3460. Elsevier.

Van den Berg, G. J. and Ridder, G. (1998). An empirical equilibrium search model of the labor market. Econometrica, pages 1183–1221.