Identification and inference of post-treatment subgroup effects.*

Job Market Paper

Masayuki Sawada †

January 10, 2019

Click HERE for the latest version.

Abstract

In the context of a randomized experiment, I identify treatment effect heterogeneity across endogenous post-assignment decisions. Unlike existing approaches, I do not rely on any instruments or specific experimental design. Instead, I exploit a baseline survey to proxy for control outcomes. The proxy variable must be similar to the control outcome in rank orders, but need not be in their levels. I then apply this strategy to a microcredit experiment with one-sided non-compliance to identify the average treatment effect on the treated (ATT). In microcredit studies, a direct effect of the treatment assignment has been a threat to identification of the ATT based on an IV strategy. I find the IV estimate for the ATT is 2.3 times larger than my preferred estimate. I also extend this analysis to two-sided non-compliance including differential attrition problem. R package **ptse** is available for this analysis.

1 Introduction

Randomization of treatment assignment identifies the average treatment effect if everyone complies with the assigned treatment. However, randomization becomes

*I am very grateful to Edward Vytlacil, Yuichi Kitamura and Yusuke Narita for their guidance and support particularly on this paper. I am also thankful to Joseph Altonji for fruitful discussion and Mitsuru Igami for general guidance and support. I am greatly benefited by helpful comments and conversations from participants of the Yale Econometrics Workshop, Yale Labor/Public Workshop, and 2018 Kyoto Summer Workshop. All errors are my own.

†Yale

insufficient for identification when we step out of the case of perfect compliance. Under imperfect compliance, the average difference in outcomes between the treatment and control groups becomes the intention-to-treat effect (ITT). The ITT is usually not the primary parameter of interest. The ITT is not the effect of the treatment as not everyone takes the treatment. Instead, the ITT is the effect of the assignment, and is a weighted mean of heterogeneous treatment effects by the treatment take-up behavior. Alternatively, we are frequently interested in the treatment effect for those who take up the treatment, called the average treatment effect on the treated (ATT). The ATT reveals the effect of taking the treatment for the subgroup of the treated, rather than the effect of treatment assignment as the ITT identifies. However, the ATT is not identified just by randomization of treatment assignment.

I propose a general strategy to identify the treatment effect heterogeneity across endogenous post-intervention decisions. I do not rely on instruments or other specific designs which may not be available. Instead, I use an observation from a baseline survey as a proxy for the control outcome. As a particularly important case, I adopt the strategy to identify the ATT when the conventional assumption of no direct effect of treatment assignment may not hold and therefore there is no valid instrument. In the upcoming application to the experimental study of microcredit, I demonstrate that the direct effect is not just an interesting parameter in itself, but also crucial in the estimation of the magnitude of the ATT.

Development economists are interested in the effect of expanding access to microcredit opportunities in developing countries. For example, Angelucci et al. (2015), Banerjee et al. (2015) and Crépon et al. (2015) ran large-scale microcredit experiments. In these experiments, a microcredit company visits treatment villages to provide microfinance opportunities to all households in those villages. Despite new credit being available, not all people in treated villages borrow from the microcredit firms. We wish to evaluate the effects of microcredit on those who borrowed from the microcredit firms. However, the conventional IV strategy is invalid if access to the microcredit has a direct effect on business outcomes of interest. For example, availability of credit may change the behavior of the units including those who do not ex-post use the credit. The experimental randomization at the village level rather than an individual level also may generate a direct effect of access. For example, the local village economy may be altered by the equilibrium effect through the interest rate of informal lending or local sale price changes. Transfers to and from relatives or friends also generate positive or negative spillover to those who do not borrow the credit. I point out that the IV estimate of the ATT, which is consistent under the assumption of no direct effect, is subject to bias if the direct effect is non-zero, and the bias can be enormous even for a small direct effect when the take-up rate is relatively small. Using the study of Crépon et al. (2015) which collected a detailed baseline survey, I estimate the ATT as the average effect of both treatment access and microcredit take-up for those who decide to take-up the microcredit treatment, and I separately estimate the average direct effect of the microcredit access for those who do not take-up the microcredit treatment. I find that the IV point estimate of ATT is 2.3 times larger than my preferred estimate. The suggested magnitude of the ATT estimate and the sign of the estimated direct effect implies the possibility that there may be a small but positive direct effect. I find that the direct effect, while small, results in an enormous bias in the IV point estimate of the ATT.[1]

The contribution of this paper is to provide a non-parametric identification of treatment effects conditional on the treatment take-up decision. The general strategy covers identification of subgroup effects of randomized treatment conditional on subgroups generated by endogenous post-treatment variables. In this paper, I study specific cases of randomized experiments with imperfect compliance. In section 2, I show the subgroup effects conditional on treatment take-up are the ATT or the direct effect for the case of one-sided non-compliance. In section 6, I consider its extension to the Local Average Treatment Effect (LATE) with additional assumptions.

[1] Crépon et al. (2015) defines people with predicted probabilities of borrowing lower than 30% as a subgroup who would not borrow from the microcredit firms and concludes that there is no evidence of externality. For the difficulties in their approach, see the application of section 5.

In Biostatistics, studies on these subgroup effects are called principal stratification as formally defined in Frangakis and Rubin (2002). These studies rely on other assumptions which may not be plausible for some problem setup. For example, their identification strategies rely on exclusion restrictions (Imbens and Angrist, 1994 or Angrist et al., 1996) or a parametric model (Hirano et al., 2000 and Imbens and Rubin, 2015)[2]. Zhang and Rubin, 2003 and Rubin, 2006 also propose partial identification based on the monotone treatment response type of assumptions as in Manski (1997).

My approach relies on a proxy variable from a baseline survey for the control outcomes. My key assumption is the rank similarity assumption (Chernozhukov and Hansen, 2005 and Athey and Imbens, 2006)[3] in addition to other standard assumptions. Under the rank similarity assumption, the treatment take-up may be endogenous. Therefore, the proxy variable and the control outcome may be correlated with the treatment take-up even when the treatment take-up has no causal effect on the proxy variable and the control outcome. This rank similarity assumption imposes restrictions on the latent rankings of the proxy variable and the control outcome. The latent ranking of an outcome is a uniform random variable normalized to $[0, 1]$ representing underlying percentile associated with the outcome measure. The rank similarity assumes that the distributions of latent rankings for the proxy variable and for the control outcome are the same conditional on the treatment take-up.[4] With a baseline survey, the underlying data structure is related to the Change-in-Change model studied by Athey and Imbens (2006). However, the baseline survey do not generate the DiD structure as the treatment group has a

[2]More precisely, these studies employ a Bayesian approach with a parametric assumption for the relation between the potential outcomes and the post-intervention variables conditional on model parameters. At least in the frequentist point of view of the restriction, their parametric assumptions imply the conditional exogeneity of the post-intervention variable to the potential outcomes.

[3]My proposed assumption is slightly weaker than the original assumption of Chernozhukov and Hansen (2005). Nevertheless, similar lines of arguments would apply to rationalize both assumptions.

[4]If the treatment take-up is exogenous to and has no causal effects on the proxy variable and the control outcome, then this equivalence of conditional distributions is trivial as the (unconditional) distributions of the rankings are normalized to a uniform distribution on the support of $[0, 1]$.

non-compliance nature. Therefore, the DiD procedure do not apply to the problem studied in this paper unlike Athey and Imbens (2006).[5] Furthermore, the support assumption (Assumption 3.4 of Athey and Imbens, 2006) is not an additional restriction as I exploit the randomization. With the non-compliance nature, the data structure is similar to but different from the Fuzzy Differences-in-Differences model studied by de Chaisemartin and D'Haultfoeuille (2017). First, their rank imputation approach in de Chaisemartin and D'Haultfoeuille (2017) exploits the stable control group assumption (Assumption 2 in de Chaisemartin and D'Haultfoeuille, 2017) for the point-identification that excludes one-sided non-compliance problems studied in this paper. Second, their solution to two-sided non-compliance problems differs from mine as well. My strategy first identifies the heterogeneous treatment effects or the principal stratification, and then I impose a restriction on the heterogeneity to identify the LATE. On the other hand, de Chaisemartin and D'Haultfoeuille (2017) assumes the stability which excludes an example of the attrition bias as the treatment of "attrition" do not occur in the baseline survey.

Researchers have used the rank similarity assumption or its stronger concept of the rank invariance assumption in the literature of counterfactual studies that use rank imputation strategies (Juhn et al., 1993, Altonji and Blank, 1999, Machado and Mata, 2005 and Athey and Imbens, 2006). Most recently, Han (2018) uses the rank similarity assumption to identify dynamic treatment effects with non-separable models. The rank similarity assumption is generally not testable. For the quantile treatment response model (Chernozhukov and Hansen, 2005), Kim and Park (2017) proposes a testing procedure for rank similarity in the presence of overidentification. Unfortunately, their procedure does not apply to my model as their overidentification restriction arises from the exclusion restriction of multiple instruments.[6]

[5] By the non-compliance nature, the trend of the treated in the treatment group must be compared with those people who would have treated if they were in the treatment group but assigned to the control group. As we do not know the subgroup in the control group, the standard DiD is not feasible.

[6] Dong and Shen (2018) and Frandsen and Lefgren (2018) propose other testing procedures of the rank similarity assumption, but their restrictions are stronger than the original assumption

This paper also provides an estimation and an inference strategy similar to the procedures for Change-in-Change models (Athey and Imbens, 2006, Melly and Santangelo, 2015 or Callaway et al., 2018). As my primary focus is on mean effects rather than on quantile differences, I propose an estimator based on the semi-parametric distribution regression (Chernozhukov et al., 2013) instead of the quantile regressions as in Melly and Santangelo (2015) or Callaway et al. (2018).[7] I show weak convergence of the empirical process for the proposed counterfactual estimator and the validity of bootstrap inference. I also adopt the recent study on cluster robust weak convergence results by Davezies et al. (2018) to accommodate many relevant experimental studies with cluster dependencies.

The organization of this paper is as follows. In the next section, I introduce notations and parameters of interest as the subgroup effect. In section 3, I list formal identification assumptions for a general argument of the identification of the average subgroup effects. Later in the same section, I specialize the general framework to the case of one-sided non-compliance. In section 4, I develop the estimation procedure and the asymptotic properties of the estimators. I conclude that bootstrap inference is valid and the inference strategy is cluster-robust. In section 5, I apply the one-sided non-compliance procedure to the microcredit application. Finally, in section 6, I extend the framework to the case of two-sided non-compliance models including differential attrition problem, followed by a conclusion.

(Chernozhukov and Hansen, 2005) and my proposed assumption.

[7]As is discussed further in the estimation section, there are advantages and disadvantages of the quantile regression and distribution regression. While the quantile regression is straightforward to estimate the conditional quantile functions, it requires tail-trimming procedure to estimate the conditional distribution functions. For the primary purpose of the mean effect unconditional of the covariates, I need to obtain the conditional distribution functions for the whole support of the outcome of interest.

2 Parameter of Interest

2.1 General Idea

Consider a standard model of potential outcomes. Let $T \in \{0,1\}$ be the treatment assignment, and let Y be an observed outcome generated out of two potential outcomes Y_1 and Y_0 indexed by the intention to treatment $T \in \{0,1\}$[8] such that

$$Y = TY_1 + (1-T)Y_0.$$

The average effect of assignment T,

$$E[Y_1 - Y_0] = E[Y_1] - E[Y_0]$$

is called the intention to treatment effect (ITT). The ITT includes any direct effect of treatment assignment as well as any effect of treatment take-up enabled by the treatment assignment.

Let W denote pre-treatment covariates observed in a baseline survey. if the following conditional ignorability assumption (Rosenbaum and Rubin, 1983)

Assumption 2.1 (Conditional Ignorability)**.**

$$Y_t \perp\!\!\!\perp T | W, \forall t \in \{0,1\},$$

is satisfied, then the following conditional ITTs are identified:

Lemma 2.1. *If assumption 2.1 holds, then*

$$E[Y_1 - Y_0 | W = w] = E[Y|T=1, W=w] - E[Y|T=0, W=w]$$

for every $w \in \mathcal{W}$.

[8] In this paper, I consistently use one-index potential outcome Y_t as the potential outcome if the unit was given the assignment T. One may view this assignment as one of two treatments (T, D) since the assignment T would have a direct effect on the outcome Y.

Although this ITT represents the average effect of the assignment T, usually the ITT is not the ultimate goal of the study. The parameter of interest is often average effect of the treatment, not the average effect of assignment to treatment. Let D be a binary variable representing the treatment take-up that the units choose endogenously posterior to the assignment T.

One consideration is to see the heterogeneity in $E[Y_1 - Y_0]$ by the treatment take-up behavior D. For example, this paper focuses on the subgroup mean effects

$$E[Y_1 - Y_0|D = d, T = 1]$$

and the subgroup quantile differences[9]

$$Q_{Y_1|D,T}(\tau|d, 1) - Q_{Y_0|D,T}(\tau|d, 1)$$

for each value of $d \in \{0, 1\}$.

These parameters represent the effect of the treatment assignment T for those who would take $D = d, d \in \{0, 1\}$ under treatment assignment group $T = 1$. These parameters are essential to understanding the effect of the treatment take-up D and the direct effect of the assignment T separately.[10]

In particular, the subgroup mean effects defined above may be equivalent to a well-known parameter of interest. For the case of a one-sided non-compliance model, this subgroup mean effect for $D = 1, T = 1$ is equivalent to the average treatment

[9]This quantile difference is not quite a treatment effect, as this parameter does not take the form of mapping from the treatment effect $Y_1 - Y_0$. As mentioned in Abbring and Heckman (2007), one may be interested in quantiles of $Y_1 - Y_0$ rather than the quantile difference in the main text. However, the identification of the alternative parameter requires generally harder assumptions than the assumptions studied in this paper. Therefore, this paper does not study the alternative parameter.

[10]Furthermore, these parameters may be essential to the evaluation of a manipulable policy T. If policymakers face non-uniform weights on their welfare function, then the unconditional ITT should not be sufficient to decide whether the policy should be implemented. For example, the policymaker may want to avoid losses in welfare. Policymaker may want to have heterogeneous effects. The tax revenue or the economic development may be driven stronger by larger effects for a small population rather than smaller effects for the whole population. Similarly, a policymaker might also want to have homogeneous effects when the equality is an essential factor for their decisions. The ITT does not help in the decision of the policy implementation for the criterion.

effect on the treated (ATT) as parameter shown below.

2.2 ATT as the Subgroup Mean Effect under One-Sided Non-Compliance

One-sided non-compliance is a randomized experiment where not all units assigned to treatment group $T = 1$ take the treatment, i.e., have $D = 1$, but the treatment is not available for units assigned to the control $T = 0$. For one-sided non-compliance, the subgroup effect for $T = 1, D = 1$ is the average treatment effect on treated (ATT). The application to the microcredit experiment of Crépon et al. (2015) is an example of one-sided non-compliance to microcredit take-up D in response to the treatment assignment T of the microcredit promotion. T is a randomized offer and promotion to the microcredit service, and D is the realized take-up of the credit. Furthermore, the control group has no access to the credit D. Namely, $D = 0$ whenever $T = 0$.

In this case, the observed outcome is one of the three potential outcomes Y_{11}, Y_{10} or Y_0 such that

$$Y = T(DY_{11} + (1 - D)Y_{10}) + (1 - T)Y_0$$

where Y_{11} is the outcome given the offer $T = 1$ and taking the credit $D = 1$, Y_{10} is the outcome given the offer $T = 1$ but not taking the credit $D = 0$, and Y_0 is the outcome of the control group with $T = 0$ therefore $D = 0$. Then the subgroup mean effect is rewritten as

$$E[Y_1 - Y_0|T = 1, D = d] = E[Y_{1d} - Y_0|T = 1, D = d].$$

In particular,

$$E[Y_{11} - Y_0|T = 1, D = 1]$$

is the average treatment effect on the treated, the ATT, and

$$E[Y_{10} - Y_0|T = 1, D = 0]$$

is the direct effect of the assignment T for those who do not take-up the credit, i.e., $D = 0$. The random variable $Y_{10} - Y_0$ is refereed to as a direct effect in the Mediation analysis literature (Imai et al., 2010, Pearl, 2014), and I will call the average of this effect for those assigned to treatment but who do not take-up the treatment, $E[Y_{10} - Y_0 | T = 1, D = 0]$, the direct effect of treatment assignment on the non-treated.

2.3 How the Identification Fails: Bad Control Problem

In general, we cannot achieve identification of these parameters by merely running a regression on the observed treatment take-up variable D.

For usual pre-assignment covariates W such as strata of randomization, observed conditional moments reveal the ITT:

$$E[Y|T = 1, W = w] - E[Y|T = 0, W = w] = E[Y_1 - Y_0|W = w].$$

One may wonder whether an analogue to the strategy of lemma 2.1 works for the post treatment assignment variable D, i.e., whether

$$E[Y|T = 1, D = d] - E[Y|T = 0, D = d]$$

identifies the ATT. However, this quantity is not interpretable. This is known as the bad control problem (Angrist and Pischke, 2009).

Consider the case of one-sided non-compliance. First note that the above formula is not even well-defined for $D = 1$ as $E[Y|T = 0, D = 1]$ does not exist. It is also generally invalid for $D = 0$ as

$$
\begin{aligned}
&E[Y|T=1,D=0] - E[Y|T=0,D=0] \\
=&E[Y_{10}|T=1,D=0] - E[Y_0|T=0,D=0] \\
=&\underbrace{E[Y_{11}-Y_0|T=1,D=0]}_{\text{a causal effect}} + \underbrace{E[Y_0|T=1,D=0] - E[Y_0|T=0]}_{\text{selection bias}}
\end{aligned}
$$

when D is endogenous to Y_0.

2.4 How the Identification Fails: T as an Invalid Instrument

Concerning the bad control problem due to the endogeneity of D, researchers often use T as an instrument of D. For the purpose of identifying the ATT, the instrument T must satisfy three conditions: (a) independence: $(Y_{11}, Y_{10}, Y_0) \perp\!\!\!\perp T$, (b) relevancy: $D \not\perp\!\!\!\perp T$, and (c) exclusion $Y_{10} = Y_0$ almost surely. While the randomization implies that the first independence assumption holds, and we can verify the second relevancy assumption, the third condition of the exclusion will not hold if the treatment assignment T has a direct impact on the outcome Y.

If the three conditions are satisfied, then the ATT is identified by the IV strategy. Using random assignment as an instrument for treatment take-up, the probability limit of the IV estimator can be shown to equal $\frac{E[Y_1-Y_0]}{P(D=1|T=1)}$.[11] By the total law of expectation,

$$
\begin{aligned}
&\frac{E[Y_1-Y_0]}{P(D=1|T=1)} \\
&= E[Y_1-Y_0|T=1,D=1]\frac{P(D=1|T=1)}{P(D=1|T=1)} + E[Y_1-Y_0|T=1,D=0]\frac{P(D=0|T=1)}{P(D=1|T=1)} \\
&= E[Y_{11}-Y_0|T=1,D=1] + E[Y_{10}-Y_0|T=1,D=0]\frac{P(D=0|T=1)}{P(D=1|T=1)}.
\end{aligned}
$$

When the exclusion condition holds so that $Y_{10} - Y_0 = 0$ almost surely, the above

[11]Note that the probability limit of the IV estimator $\frac{Cov(Y,T)}{Cov(D,T)}$, which can be shown to equal $\frac{E[Y_1-Y_0]P(T=1)P(T=0)}{P(D=1|T=1)P(T=1)P(T=0)} = \frac{E[Y_1-Y_0]}{P(D=1|T=1)}$.

expression equals the ATT, $E[Y_{11} - Y_0|T = 1, D = 1]$.

What if the average direct effect of T, $E[Y_{10} - Y_0|T = 1, D = 0]$, does not equal zero? In this case, the IV estimator is biased by the term

$$E[Y_{10} - Y_0|T = 1, D = 0]\frac{P(D = 0|T = 1)}{P(D = 1|T = 1)}.$$

It is important to note that the magnitude of the bias is not just the size of the direct effect, $E[Y_{10} - Y_0|T = 1, D = 0]$, but the multiplicative form with $\frac{P(D=0|T=1)}{P(D=1|T=1)}$. This form is less of a concern if the take-up probability $P(D = 1|T = 1)$ is high as it reduces the bias if $P(D = 1|T = 1) > 0.5$. However, this form becomes a serious concern when the take-up probability is low.

The microcredit applications are one of these concerning experiments as the take-up probability usually is as small as 15%. Therefore, the IV estimator is subject to an inflated bias from the direct effect by $0.85/0.15 \approx 5.6$, and the small violation to the exclusion restriction may generate enormous bias in the estimate of the ATT.

3 Identification

3.1 Identification Assumption

The primary goal of this paper is to provide an identification strategy with endogenous post-treatment covariates D without relying on instruments or specific designs. The key idea is the use of an additional variable Y_b from a baseline survey as a proxy for the control outcome Y_0. The baseline survey is data collected before the experiment starts. Collecting a baseline survey is a common practice primarily for attaining more precise estimates or studying the subgroup effects with the baseline covariates W. Although I require certain similarity in the proxy variable Y_b and the control outcome Y_0, these two random variables Y_b and Y_0 may have entirely different distribution functions. Instead, I impose a restriction that the rankings of the two random variables Y_b and Y_0 to be similar. First, let me introduce the concept of the rankings.

Definition 3.1 (Latent Rank Variables)**.** *Let W be a vector of baseline covariates. A random variable $U_w \sim U[0,1]$ indexed by each $W = w$ is called (conditional) latent rank variable for a random variable Y if*

$$Y = Q_{Y|W}(U_W|W)$$

where $Q_{Y|W}(u|w) = \inf\{y : F_{Y|W}(y|w) \leq u\}$.

Remark. *Note that the existence of the latent ranking variable U_w is not an assumption, such a variable exists whether Y is finitely supported or continuous.*

We can always construct such a conditional latent variable U_w as

$$U_w = F_{Y|W=w}(Y-) + V \cdot (F_{Y|W=w}(Y) - F_{Y|W=w}(Y-))$$

where $V \sim U[0,1]$ and $V \perp\!\!\!\perp W$. The existence of such a conditional latent variable can be shown as an extension to the unconditional latent variable existence of Proposition 2.1 in Rüschendorf (2009).

For the identification, I need two requirements on the relation between the proxy variable Y_b and the control outcome Y_0. First, I need to learn the complete latent ranking of Y_b, $U_{b,w}$, over the whole support of $[0,1]$. In order to achieve the requirement, I assume Y_b has a strictly increasing distribution function over its support.

Assumption 3.1 (Unique quantile and random assignment to the proxy)**.** *Assume for every $w \in \mathcal{W}$, $F_{Y_b|W}(\cdot|w)$ is strictly increasing over its conditional support.*[12] *Furthermore, I assume the treatment assignment is also independent from the proxy variable*

$$Y_b \perp\!\!\!\perp T|W.$$

[12] This condition is equivalent to the conditional quantile function satisfies $Q_{Y_b|W}(F_{Y_b|W}(y|w)|w) = y$ for every value of y in the support of Y_b conditional on $W = w$.

This assumption is required for the point identification of the parameter of interest.[13] Although I rely on a continuous proxy variable Y_b, I do not restrict the nature of potential outcomes Y_1 and Y_0 and they are allowed to be finitely supported. The latter conditional independence is not essential, but the interpretation of the identification assumption becomes straightforward.

Second, I assume the latent ranking of Y_b has the same distribution as the latent ranking of Y_0 conditional on $W = w, T = 1, D = d$. This restriction is called the rank similarity assumption.

Assumption 3.2 (Conditional Rank Similarity)**.** *Let $U_{b,w}$ be the latent ranking of Y_b, and $U_{0,w}$ be the latent ranking of Y_0 as defined in definition 3.1. Suppose*

$$U_{b,w} \sim U_{0,w} | W = w, T = 1, D = d$$

for each $d \in \{0, 1\}, w \in \mathcal{W}_1$, where $\mathcal{W}_1$ is the support of W conditional on T.

This assumption says that the conditional latent rankings of Y_b and Y_0 have the same distribution for the subgroup which was assigned to treatment ($T = 1$) and chose $D = d$. It is important to note that the distribution of the levels of the proxy Y_b and the control outcome Y_0 may differ arbitrarily. After the discussion of the identification formula, I will revisit this assumption in the case of one-sided non-compliance as it has a straightforward interpretation.

In addition to these two requirements, I assume the randomization assumption 2.1, as well as following assumption on the support of the covariates W,[14]

Assumption 3.3. *Let $\mathcal{W}_1$ be the support of W conditional on $T = 1$, and $\mathcal{W}_0$ be the support of W conditional on $T = 0$. Suppose $\mathcal{W}_1 = \mathcal{W}_0 \equiv \mathcal{W}$.*

[13]The partial identification can be possible with a finitely supported proxy variable Y_b, but I do not discuss in this paper.

[14]This assumption is stronger than necessary as I only require that the support $\mathcal{W}_1$ is a subset of the support $\mathcal{W}_0$.

3.2 Identification Result

The parameter of interest is the counterfactual distribution $F_{Y_0|T,D}(y|1,d)$ and the counterfactual mean computed from the distribution.

Theorem 3.1. *If the assumptions 2.1, 3.1, 3.2, and 3.3 hold, then*

$$F_{Y_0|W,T,D}(y|w,1,d) = F_{Y_b|W,T,D}(Q_{Y_b|W,T}(\tau_{y,w}|w,1)|w,1,d)$$

for every $d \in \{0,1\}, w \in \mathcal{W}$ and $y \in \mathcal{Y}_0$ where

$$\tau_{y,w} \equiv F_{Y_0|W,T}(y|w,0).$$

The mean of the post-treatment subgroup effect is identified as

$$E[Y_1|T=1,D=d] - E[Y_0|T=1,D=d]$$

as well as the subgroup quantile difference is identified as

$$Q_{Y_1|T,D}(\tau|1,d) - Q_{Y_0|T,D}(\tau|1,d)$$

for every $d \in \{0,1\}$ where

$$E[Y_0|T=1,D=d] = \int y dF_{Y_0|T,D}(y|1,d),$$

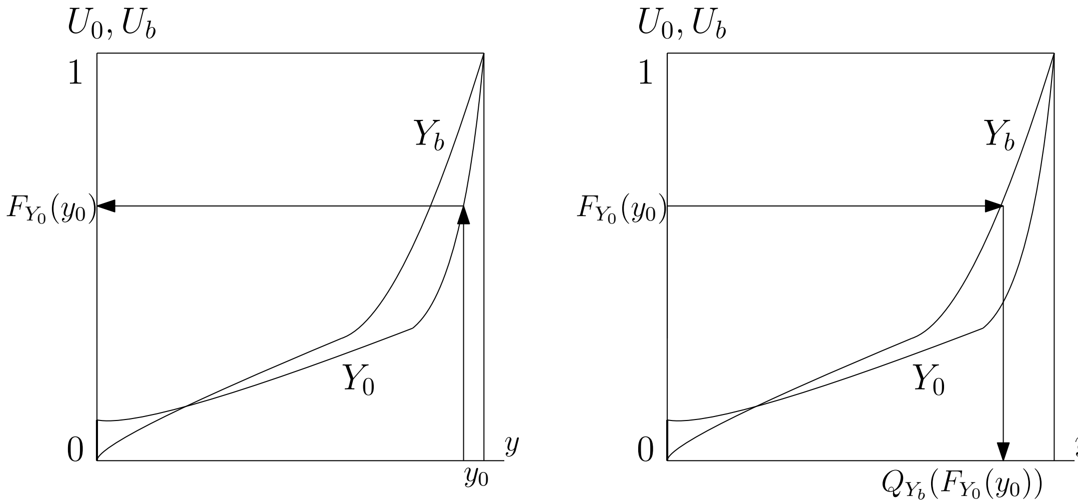
$$F_{Y_0|T,D}(y|1,d) = \int F_{Y_0|W,T,D}(y|w,1,d)dF_{W|T,D}(w|1,d),$$

and

$$Q_{Y_t|T,D}(\tau|1,d) = \inf\{y \in \mathcal{Y}_t : F_{Y_t|T,D}(y|1,d) \geq \tau\}$$

for every $\tau \in [0,1]$.

The idea of the formula is summarized in the following figure 3.1.



Panel A Panel B

Figure 3.1: Graphical representation of the formula

For the simplicity, consider the case without covariates W. As in the panel A, first note that the events $\{Y_0 \leq y\}$ and $\{U_0 \leq F_{Y_0}(y_0)\}$ are equivalent almost surely for every y_0 in the support of $\mathcal{Y}_0$. [15] Rank similarity equates the events $\{U_0 \leq F_{Y_0}(y_0)\}$ and $\{U_b \leq F_{Y_0}(y_0)\}$ in expectation conditional on $\{T = 1, D = d\}$. As in the panel B, we may find the quantile of Y_b, $Q_{Y_b}(F_{Y_0}(y_0))$, which corresponds to the value of the random variable U_b, $F_{Y_0}(y_0)$, such that the events $\{U_b \leq F_{Y_0}(y_0)\}$ $\{Y_b \leq Q_{Y_b}(F_{Y_0}(y_0))\}$ have the same probability conditional on $\{T = 1, D = d\}$. Such a quantile $Q_{Y_b}(u)$ can be found for every $u \in [0, 1]$ from the continuity of Y_b, and the final procedure is the formula for the identification shown in the theorem.

Proof. The counterfactual cdf $F_{Y_0|W,T,D}(y|w, 1, d)$ is expressed as

$$
\begin{aligned}
F_{Y_0|W,T,D}(y|w,1,d) =& E[1\{Y_0 \leq y\}|W = w, T = 1, D = d] \\
=& E[1\{F_{Y_0|W,T}(Y_0|w,0) \leq F_{Y_0|W,T}(y|w,0)\}|W = w, T = 1, D = d] \\
=& E[1\{F_{Y_0|W}(Y_0|w) \leq F_{Y_0|W}(y|w)\}|W = w, T = 1, D = d] \\
=& E[1\{U_{0,w} \leq F_{Y_0|W}(y|w)\}|W = w, T = 1, D = d]
\end{aligned}
$$

[15] Y_0 may have positive mass in the support of Y_0 for this relation as long as y_0 is evaluated in the support of Y_0.

in terms of the conditional latent rank $U_{0,w}$. In the third equality, I use the assumption that $Y_0 \perp\!\!\!\perp T|W$. The conditional rank similarity assumption implies

$$
\begin{aligned}
&E[1\{U_{0,w} \leq F_{Y_0|W}(y|w)\}|W = w, T = 1, D = d] \\
=&E[1\{U_{b,w} \leq F_{Y_0|W}(y|w)\}|W = w, T = 1, D = d] \\
=&E[1\{F_{Y_b|T,W}(Y_b|1, w) \leq F_{Y_0|W}(y|w)\}|W = w, T = 1, D = d].
\end{aligned}
$$

In the last equality, I use the assumption $Y_b \perp\!\!\!\perp T|W$. From the unique quantile transformation with Y_b, we have

$$
\begin{aligned}
&E[1\{F_{Y_b|T,W}(Y_b|1, w) \leq F_{Y_0|W}(y|w)\}|W = w, T = 1, D = d] \\
=&E[1\{Y_b \leq Q_{Y_b|T,W}(F_{Y_0|W}(y|w)|1, w)\}|W = w, T = 1, D = d] \\
=&F_{Y_b|W,T,D}(Q_{Y_b|T,W}(F_{Y_0|T,W}(y|0, w)|1, w)|w, 1, d)
\end{aligned}
$$

□

One may think of this use of the proxy variable Y_b in an analogy to the control function approach. (For example, Imbens and Newey, 2009, D'Haultfoeuille and Février, 2015, Torgovitsky, 2015, and Ishihara, 2017). I use the reduced form variation in Y_b conditional on W to back out the scalar unobservable $U_{b,w}$ in the reduced form equation. Unlike the triangular equations model, the structural error $U_{0,w}$ has, by assumption, a direct relation to the reduced form error $U_{b,w}$. From the direct connections in the errors, we can achieve the shape of the conditional quantile function of Y_0 as a function of the ranking $U_{0,w}$ without employing an excluded instrument.

3.3 The Case of One-Sided Non-Compliance

The one-sided non-compliance structure reduces the rank similarity to a fairly straightforward restriction. To fix ideas, consider the case of the microcredit experiment. Let Y be the sales value of the production output two years after the experiment. Y_b

is the same sales value of the production output but measured before the experiment.

To rationalize the rank similarity between Y_b and Y_0, consider there are scalar latent productivity $U_{b,w}$ and $U_{0,w}$ which determine the rankings of the variables Y_b and Y_0 conditional on other baseline exogenous determinants W. Such scalar rankings exist by definition 3.1. I need to assume that the rankings $U_{b,w}$ and $U_{0,w}$ have the same meaning as the productivity. In other words, people with higher $U_{b,w}$ and $U_{0,w}$ should represent people with higher latent productivity.

This increasing property helps the justification of the rank similarity assumption. The rank similarity assumption holds if the sorting of the two latent productivity $U_{b,w}$ and $U_{0,w}$ conditional on the take-up decision D are the same. In order to further understand the restriction, consider that there is common latent productivity U such that $U_{b,w} = U$ and $U_{0,w} = U$ almost surely, then the rank similarity trivially holds. This extreme example is called the rank invariance restriction.[16] While this specification is an extreme example, this is a fair starting point to relax the restriction. In particular, I may accept the random permutations to the rankings $U_{b,w}$ and $U_{0,w}$ called "slippages" (Heckman et al., 1997). For example, let $\tilde{U}_{b,w}$ and $\tilde{U}_{0,w}$ be a pair of identically distributed random variables. Let $D \equiv g(W, V)$ for some measurable function g and the vector of unobserved determinants of the post-treatment choice D. This unobserved determinants V may include U. If the slippages $(\tilde{U}_{b,w}, \tilde{U}_{0,w})$ are independent of V, then the rank similarity may be maintained for the latent productivity of the forms

$$U_{b,w} \equiv U + \tilde{U}_{b,w},\ U_{0,w} \equiv U + \tilde{U}_{0,w}. \tag{1}$$

This form shares similarity to a factor model where a single factor of productivity and independent shocks determine the two latent productivity.

[16]While the rank invariance assumption restricts the joint copula of two random variables, the rank similarity imposes assumptions on the marginal distributions of two random variables. In particular, $U_{b,w}$ and $U_{0,w}$ may have arbitrary positive correlation under the rank similarity, while $U_{b,w}$ and $U_{0,w}$ must have correlation of 1 under the rank invariance.

Let $F(U_{b,w})$ and $F(U_{0,w})$ be the latent rankings where F is the common distribution function of $U + \tilde{U}_{b,w}$ and $U + \tilde{U}_{0,w}$.

Proposition 3.2. *Let $F(U_{b,w})$ and $F(U_{0,w})$ be the latent rank variables for Y_b and Y_0 conditional on W constructed as in (1). Let V be the vector of unobserved determinants of D. If $(\tilde{U}_{b,w}, \tilde{U}_{0,w}) \perp\!\!\!\perp (V, T)|W$ then the rank similarity assumption 3.2 holds.*

Proof. Statement immediately follows from the fact that the determinants of the latent rankings are either $\tilde{U}_{b,w}, \tilde{U}_{0,w}$, which are independent of unobserved determinants of D, or the common term U.

$$
\begin{aligned}
&P(F(U_{b,w}) \leq \tau | W = w, D = d, T = 1) = P(F(U + \tilde{U}_{b,w}) \leq \tau | W = w, D = d, T = 1) \\
=&P(F(U + \tilde{U}_{b,w}) \leq \tau | W = w, V \in \mathcal{V}_{D=d,T=1,W=w}, T = 1) \\
=&P(F(U + \tilde{U}_{0,w}) \leq \tau | W = w, V \in \mathcal{V}_{D=d,T=1,W=w}, T = 1) \\
=&P(F(U_{0,w}) \leq \tau | W = w, D = d, T = 1)
\end{aligned}
$$

where $\mathcal{V}_{D=d,T=1,W=w}$ is the set of V which is compatible with $D = d$ conditional on $T = 1$ and $W = w$. □

3.4 Flexibility of the Rank Similarity Assumption on D, Y_1

The rank similarity assumption imposes restrictions on the marginal distributions of Y_0 and Y_b. On the other hand, the remaining observed variables (D, Y_1) are left flexible. For example, the common latent ranking U may arbitrarily correlate with (D, Y_1) as long as the slippages $(\tilde{U}_{b,w}, \tilde{U}_{0,w})$ are independent of (D, Y_1). This flexibility implies that D may be endogenous to the potential outcomes Y_1 and Y_0 as well as D may be exogenous to these potential outcomes. The conditional exogeneity may be possible under the rank similarity assumption, although they are not nested.

Furthermore, the flexibility implies that Y_1 may be left unrestricted as long as the slippages of Y_0 and Y_b are independent of Y_1. It is worth noting two desirable facts on this flexibility. First, this rank similarity does not necessarily restrict the form of the gains from T, $Y_1 - Y_0$. This absence of the restriction on the gain is a sharp distinction from other bounding approaches based on restrictions such as monotone treatment response assumptions $Y_1 \geq Y_0$ such as Manski (1997).

Second, this model flexibly accepts peer effects or equilibrium effects due to the treatment exposure $T = 1$ as in the following example.

Example 3.1. *Consider every unit i faces his/her reference group N_i which would affect his/her outcome as an equilibrium response to the common treatment assignment $T = 1$ within N_i.*

For every unit i, suppose that there are underlying potential outcomes $Y_{11,i}, Y_{10,i}$ such that

$$Y_{1i} = D_i Y_{11,i} + (1 - D_i) Y_{10,i}$$

whereas D_i is determined as a measurable function of $Y_{11,i}, Y_{1,0i}$ and a vector of other unobservables V_i.

The reference group may be observed or unobserved. In either cases, his/her treatment response outcome $Y_{11,i}, Y_{10,i}$ are determined not just by his/her own base latent ranking U_i and shocks $\tilde{U}_i$, but also affected through these peers latent rankings $\{U_j\}_{j \in N_i}$ and shocks $\{\tilde{U}_{1j}\}_{j \in N_i}$.

In summary, such structure reduces the post-intervention choice treatment as a measurable function of all the unobservables

$$D_i = \delta(U_i, \tilde{U}_{1i}, \{U_j, \tilde{U}_{1j}\}_{j \in N_i}, V_i).$$

Nevertheless, the proposition 3.2 applies to maintain the rank similarity as long as

$$(U_{0i}, U_{bi}) \perp\!\!\!\perp (U_i, \tilde{U}_{1i}, V_i, \{U_j, \tilde{U}_{1j}\}_{j \in N_i}).$$

4 Estimation

4.1 Estimation for discrete post-treatment covariates D

For justifying the conditional rank similarity, it is desirable to condition on pre-treatment covariates W. The curse of dimensionality becomes a serious issue when we estimate conditional quantities for each subsample $\{T = 1, D = d, W = w\}$ non-parametrically. Melly and Santangelo (2015) and Callaway et al. (2018) proposes an extension to the Athey and Imbens (2006) estimator allowing us to incorporate covariates as a semi-parametric model. As my major focus is the mean effects rather than the quantile effects, I consider a distribution regression-based approach rather than following a quantile regression approach of Melly and Santangelo (2015). Furthermore, a different data generating process in my model relative to the Change-in-Change needs a modified theory on the inference. Nevertheless, my inference strategy closely follows the strategies in Melly and Santangelo (2015) as well as Chernozhukov et al. (2013).

First consider an estimation of semi-parametric conditional distribution functions. For the estimation of the parameter of interest, I need the distribution functions $F_{Y_b|T=1,W}, F_{Y_b|T=0,W}, \{F_{Y_b|T=1,D=d,W}\}_{d\in\{0,1\}}$. From now on, let W be a vector of transformation of the original pre-treatment covariates such as polynomials or B-splines. Following Foresi and Peracchi (1995) and Chernozhukov et al. (2013), I estimate conditional distribution functions for Y conditional on a subgroup k out of K and W as

$$\hat{F}_{Y|W,K}(y|w, k) = \Lambda(w'\hat{\beta}^k(y))$$

for some known link function $\Lambda(\cdot)$[17] and

$$\hat{\beta}^k(y) = \arg\max_{b\in\mathbb{R}^{dW}} \sum_{i=1}^{n} \{[1\{Y_i \leq y\} \log[\Lambda(w'b)]] + [1\{Y_i > y\} \log[1 - \Lambda(w'b)]]\} I\{K_i = k\}$$

[17] Theoretically, the link functions can be different across subgroups k. I use logit link function throughout the application but the robustness to other choice of link functions such as probit link or complementary log-log link are shown in the appendix.

for each $y \in \mathcal{Y}^k$ where $\mathcal{Y}^k$ is the support of Y conditional on the subgroup k, and dW is the dimension of W. The subgroup can be either $\{T = 1\}, \{T = 1, D = d\}$ or $\{T = 0\}$.

Once these estimators are obtained, the conditional counterfactual distribution is obtained as

$$\hat{F}_{Y_0|W,T,D}(y|w, 1, d) = \hat{F}_{Y_b|W,T,D}(\hat{Q}_{Y_b|W,T}(\hat{F}_{Y_0|W,T}(y|w, 0)|w, 1)|w, 1, d)$$

where

$$\hat{Q}_{Y_b|W,T}(\tau|w, 1) = \inf\left\{y \in \mathcal{Y}_b^{1,w} : \hat{F}_{Y_b|W,T}(y|w, 1) \geq \tau\right\},$$

where $\mathcal{Y}_b^k$ is the support of Y_b conditional on a subgroup k, and therefore the unconditional distribution is obtained by

$$\hat{F}_{Y_0|T,D}(y|1, d) = n_{1,d}^{-1} \sum_{i=1}^{n} \hat{F}_{Y_0|W,T,D}(y|W_i, 1, d) I\{T_i = 1, D_i = d\}$$

where $n_{1,d} \equiv \sum_i 1\{T_i = 1, D_i = d\}$. The mean effect of interest is then obtained as follows

$$\hat{\mu}_d = \frac{1}{n_{1,d}} \sum_{i:T_i=1,D_i=d} Y_i - \int_{\mathcal{Y}^0} y d\hat{F}_{Y_0|T,D}(y|1, d)$$

and the quantile difference is obtained by inverting the distribution functions

$$\hat{Q}_{Y_1|T,D}(\tau|1, d) - \hat{Q}_{Y_0|T,D}(\tau|1, d).$$

4.2 Asymptotic Normality and Bootstrap Validity

Assume following data generating process,

Assumption 4.1 (DGP)**.** *The sample $\{Y_i, Y_{b,i} D_i, W_i, T_i\}_{i=1}^n$ is an iid draw from the probability law P over the support $\{\mathcal{Y} \times \mathcal{Y}_b \times \{0, 1\} \times \mathcal{W} \times \{0, 1\}\}$.*

Let $\mathcal{Y}^0$ be a support of Y conditional on $T = 0$, and let $\mathcal{Y}_b^{1,d}$ and $\mathcal{W}_d$ be supports of Y_b and W conditional on $T = 1$ and $D = d$ for each $d \in \{0, 1\}$. Suppose $\mathcal{Y}^0 \times \mathcal{W}$

and $\mathcal{Y}_b^{1,d} \times \mathcal{W}_d$ *are compact subsets of* $\mathbb{R}^{1+d_w}$ *for each* $d \in \{0,1\}$. *If* Y_0 *is absolutely continuous with respect to the Lebesgue measure, then suppose the conditional density* $f_{Y|W,T}(y_0|w,0)$ *is uniformly bounded and uniformly continuous in* $(y_0,w) \in \mathcal{Y}^0 \times \mathcal{W}$. *Also suppose that* $f_{Y_b|W,T}(y_b|w,1)$ *and* $f_{Y_b|W,T,D}(y_b|w,1,d)$ *are uniformly bounded, and uniformly continuous in and* $(y_b,w) \in \mathcal{Y}_b \times \mathcal{W}$ *for each* $d \in \{0,1\}$. *Furthermore,* $\frac{n_{1,d}}{n} \equiv \frac{1}{n}\sum_i 1\{T_i = 1, D_i = d\} \to^p \alpha_{1,d} \equiv Pr(T_i = 1, D_i = d) > 0$ *for each* $d \in \{0,1\}$, *and* $\frac{n_0}{n} = \frac{1}{n}\sum_i 1\{T_i = 0\} \to^p \alpha_0 \equiv Pr(T_i = 0) > 0$.

Assume also that the conditional distribution functions have the following semi-parametric forms

Assumption 4.2 (Distribution Regression)**.** *Suppose we have*

$$F_{Y|W,T}(y|w,0) = \Lambda(w'\beta^0(y)),$$

for some link function $\Lambda(\cdot)$ *for all* y, w. *Assume also that the minimal eigenvalue of* $J_0(y) \equiv E\left[\frac{\lambda(W'\beta^0(y))^2}{\Lambda(W'\beta^0(y))[1-\Lambda(W'\beta^0(y))]}WW'\right]$ *is bounded away from zero uniformly over* y. *And the analogous restriction holds for*

$$F_{Y_b|W,T}(y|w,1) = \Lambda(w'\beta^1(y))$$

and

$$F_{Y_b|W,T,D}(y|w,1,d) = \Lambda(w'\beta^{1,d}(y))$$

for every $d \in \{0,1\}$.

Assume further that $E\|W\|^2 < \infty$.

This is a standard regularity condition for distribution regression models (Chernozhukov et al., 2013). Under these assumptions, these conditional distribution functions weakly converge jointly. Let

$$\hat{G}^{1,d}(y_b^{1,d}, w) = \sqrt{n}\left(\hat{F}_{Y_b|W,T,D}(y_b^{1,d}|w,1,d) - F_{Y_b|W,T,D}(y_b^{1,d}|w,1,d)\right), \forall y_b^{1,d} \in \mathcal{Y}_b^{w,\{1,d\}}$$
$$\hat{G}^{1}(y_b^{1}, w) = \sqrt{n}\left(\hat{F}_{Y_b|W,T}(y_b^{1}|w,1) - F_{Y_b|W,T}(y_b^{1}|w,1)\right), \forall y_b^{1} \in \mathcal{Y}_b^{w,1}$$
$$\hat{G}^{0}(y_0, w) = \sqrt{n}\left(\hat{F}_{Y_0|W,T}(y_0|w,0) - F_{Y_0|W,T}(y_0|w,0)\right), \forall y_0 \in \mathcal{Y}^{w,0},$$

for every $w, d \in \mathcal{W} \times \{0,1\}$.

Lemma 4.1. *Under assumptions 2.1, 3.1, 3.2, 3.3, 4.1 and 4.2,*

$$\left(\hat{G}^1(y_b^1, w), \hat{G}^{1,d}(y_b^{1,d}, w), \hat{G}^0(y_0, w)\right) \rightsquigarrow \left(\mathbb{G}^1(y_b^1, w), \mathbb{G}^{1,d}(y_b^{1,d}, w), \mathbb{G}^0(y_0, w)\right)$$

in $l^\infty(\mathcal{Y}_b \times \mathcal{W} \times \mathcal{Y}_b^{1,d} \times \mathcal{W}_d \times \mathcal{Y}^0 \times \mathcal{W})$, *where* $\mathbb{G}^k(y,w)$ *for every* $k \in \{1, \{1,d\}_{d\in\{0,1\}}, 0\}$ *are tight zero-mean Gaussian processes with each covariance function of the form*

$$V_{k,k}(y,w,\tilde{y},\tilde{w}) = \alpha_k^{-1} w' J_k^{-1}(y) \lambda_k(w'\beta^k(y)) \Sigma_k(y,\tilde{y}) \lambda_k(\tilde{w}'\beta^k(\tilde{y})) J_k^{-1}(\tilde{y}) \tilde{w}$$
$$V_{1,\{1,d\}}(y,w,\tilde{y},\tilde{w}) = \alpha_{1,d}^{-1} w' J_1^{-1}(y) \lambda_1(w'\beta^1(y)) \Sigma_{1,\{1,d\}}(y,\tilde{y}) \lambda_{1,d}(\tilde{w}'\beta^{1,d}(\tilde{y})) J_{1,d}^{-1}(\tilde{y}) \tilde{w}$$

where

$$\begin{aligned}
\Sigma_k(y,\tilde{y}) = & E[I\{K=k\} W H(W'\beta^k(y)) \\
& \times \{\min\{\Lambda(W'\beta^k(y)), \Lambda(W'\beta^k(\tilde{y}))\} - \Lambda(W'\beta^k(y))\Lambda(W'\beta^k(\tilde{y}))\} \\
& \times H(W'\beta^k(\tilde{y}))] W', \\
\Sigma_{1,\{1,d\}}(y,\tilde{y}) = & E[I\{T=1, D=d\} W H(W'\beta^1(y)) \\
& \times \{\min\{\Lambda(W'\beta^1(y)), \Lambda(W'\beta^{1,d}(\tilde{y}))\} - \Lambda(W'\beta^1(y))\Lambda(W'\beta^{1,d}(\tilde{y}))\} \\
& \times H(W'\beta^{1,d}(\tilde{y})) W']
\end{aligned}$$

for each $k \in \{1, \{1,d\}_{d\in\{0,1\}}, 0\}$ *and* $V_{1,0}(y,w,\tilde{y},\tilde{w}) = V_{\{1,d\},0}(y,w,\tilde{y},\tilde{w}) = 0$.

Given the weak convergence of the distribution regressions, the conditional estimator

$$\hat{F}_{Y_b|W,T,D}(\hat{Q}_{Y_b|W,T}(\hat{F}_{Y_0|W,T}(y|w,0)|w,1)|w,1,d)$$

is in the form of following map: for distribution functions $F_{1,d}, F_1, F_0$,

$$m(F_{1,d}, F_1, F_0) = F_{1,d} \circ Q_1 \circ F_0.$$

where Q_1 is the quantile function from F_1. In a parallel argument to Melly and Santangelo (2015) based on quantile regressions, the above map is Hadamard differentiable from the Hadamard-differentiability of the quantile function (Lemma 21.4 (ii), van der Vaart, 1998) and the chain rule of the Hadamard-differentiable maps (Lemma 20.9, van der Vaart, 1998).

Lemma 4.2. *Let F_1 and $F_{1,d}$ be uniformly continuous and differentiable distribution functions with uniformly bounded densities f_1 and $f_{1,d}$. Let F_0 be also a distribution function. Suppose F_1 has a support $[a,b]$ as a bounded subset of real line, and $F_1 \circ Q_1(p) = p$ for every $p \in [0,1]$.*

Then the map $m(F_{1,d}, F_1, F_0)$ is Hadamard differentiable at $(F_{1,d}, F_1, F_0)$ tangentially to a set of functions $h_{1,d}, h_1, h_0$ with the derivative map

$$h_{1,d} \circ Q_1 \circ F_0 - f_{1,d}(Q_1 \circ F_0)\frac{h_1 \circ Q_1 \circ F_0}{f_1(Q_1 \circ F_0)} + f_{1,d}(Q_1 \circ F_0)\frac{h_0}{f_1(Q_1 \circ F_0)}.$$

Proof. See the appendix. □

Next, let

$$\hat{G}(f) = \sqrt{n}\left(\int f d\hat{F}_{W,T,D} - \int f dF_{W,T,D}\right)$$

where $\hat{F}_{W,T,D}(w,t,d) = n^{-1}\sum_{i=1}^{n} 1\{W_i \le w, T = t, D = d\}$ for $f \in \mathcal{F}$ where $\mathcal{F}$ is a class of suitably measurable functions[18] including

$$\{F_{Y|W,K}(y|\cdot,k), y \in \mathcal{Y}^k, k \in K \equiv \{\{T=1\},\{T=0\},\{T=1,D=d\}_{d\in\{0,1\}}\}$$

and all the indicators of the rectangles in $\bar{\mathbb{R}}^{dW}$.

From the derivative expression in the lemma above and the joint weak convergence of the empirical processes, the counterfactual conditional distribution weakly converges.

Theorem 4.3. *Under assumptions 2.1, 3.1, 3.2, 3.3 , 4.1 and 4.2,*

$$\hat{G}(f) \rightsquigarrow \mathbb{G}(f)$$

in $l^\infty(\mathcal{F})$ for $\mathcal{F}$ specified earlier where $\mathbb{G}(f)$ is a Brownian bridge, and

$$\sqrt{n}\left(\hat{F}_{Y_0|W,T,D}(y|w,1,d) - F_{Y_0|W,T,D}(y|w,1,d)\right) \rightsquigarrow \mathbb{G}^{WCF}_{1,d}(y,w),\ \ in\ l^\infty(\mathcal{Y}^0\times\mathcal{W})$$

[18]Suitably measurability or P-measurability can be verified by showing the class is pointwise measurable.

A class $\mathcal{F}$ of measurable functions is pointwise measurable if there is a countable subset $\mathcal{G}\subset\mathcal{F}$ such that for every $f\in\mathcal{F}$ there is a sequence $\{g_m\}\in\mathcal{G}$ such that

$$g_m(x)\to f(x)$$

for every $x\in\mathcal{X}$.

For example, the class $\mathcal{F}\equiv\{1\{x\le t\}, t\in\mathbb{R}\}$ is pointwise measurable. This is because we can take

$$\mathcal{G}=\{1\{x\le t\}: t\in\mathbb{Q}\}$$

which is countable, and for arbitrary $t_0\in\mathbb{R}$ which characterise the arbitrary function $f(x) = \{1\{x\le t_0\}\}$, and the sequence of functions

$$g_m(x)=1\{x\le t_m\}$$

such that $t_m\ge t_0, t_m\to t_0$ converges to $f(x)$. Therefore, the relevant class used here is shown to be pointwise measurable.

where $\mathbb{G}^{WCF}_{b,d}(y,w)$ *is a tight zero mean Gaussian process indexed by* (y,w) *such that*

$$\begin{aligned}
&\mathbb{G}^{WCF}_{1,d}(y,w)\\
&\quad=\mathbb{G}^{1,d}(Q_{Y_b|W,T}(F_{Y_0|W,T}(y|w,0)|w,1),w)\\
&\quad-\frac{f_{1,d}(y,w)}{f_1(y,w)}\left(\mathbb{G}^1(Q_{Y_b|W,T}(F_{Y_0|W,T}(y|w,0)|w,1),w)+\mathbb{G}^0(y,w)\right).
\end{aligned}$$

where

$$\frac{f_{1,d}(y,w)}{f_1(y,w)}\equiv\frac{f_{Y_b|W,T,D}(Q_{Y_b|W,T}(F_{Y_0|W,T}(y|w,0)|w,1)|w,1,d)}{f_{Y_b|W,T}(Q_{Y_b|W,T}(F_{Y_0|W,T}(y|w,0)|w,1)|w,1)}$$

Proof. From lemma A.2 in Appendix, we can choose $\mathcal{F}$ satisfying the requirement defined earlier so that $\mathcal{F}$ satisfies the DKP condition (Chernozhukov et al., 2013, Appendix A.). Then assumptions of Lemma E.4 in Chernozhukov et al. (2013) is satisfied to conclude the first statement.

For the second statement, note that

$$\begin{aligned}
&\sqrt{n}\left(\hat{F}_{Y_0|T,W,D}(y|1,w,d)-F_{Y_0|T,W,D}(y|1,w,d)\right)\\
&\quad=\sqrt{n}\left(m(\hat{F}_{Y_b|W=w,D=d},\hat{F}_{Y_b|W=w},\hat{F}_{Y_0|W=w})-m(F_{Y_b|W=w,D=d},F_{Y_b|W=w},F_{Y_0|W=w})\right).
\end{aligned}$$

Therefore, the functional delta method and the Hadamard differentiability of the transformation $m(\cdot,\cdot,\cdot)$ implies the above process weakly converges to the process shown in the statement. □

The unconditional counterfactual distribution is attained by applying the Lemma D.1 from Chernozhukov et al. (2013) showing the Hadamard differentiability of the counterfactual operator

$$\phi^C(F,G)=\int F(y|w)dG(w)$$

Theorem 4.4. *Under assumptions 2.1, 3.1, 3.2, 3.3, 4.1 and 4.2,*

$$\sqrt{n}\left(\hat{F}_{Y_0|T,D}(y|1,d) - F_{Y_0|T,D}(y|1,d)\right) \rightsquigarrow \mathbb{G}^{CF}_{1,d}(y), \text{ in } l^\infty(\mathcal{Y}_0).$$

where $\mathbb{G}^{CF}_{1,d}(y)$ *is a tight mean zero Gaussian process such that*

$$\begin{aligned}\mathbb{G}^{CF}_{1,d}(y) \equiv & \alpha^{-1}_{1,d}\int \mathbb{G}^{WCF}_{1,d}(y,w)I\{T=1,D=d\}dF_{W,T,D}(w,1,d)\\ & + \alpha^{-1}_{1,d}\mathbb{G}(F_{Y_0|W,T,D}(y|w,1,d)I\{T=1,D=d\}).\end{aligned}$$

Proof. Note that

$$\sqrt{n}\hat{F}_{Y_0|T,D}(y|1,d) = \sqrt{n}\frac{\hat{F}_{Y_0,T,D}(y,1,d)}{n_{1,d}/n}$$

and

$$\sqrt{n}F_{Y_0|T,D}(y|1,d) = \sqrt{n}\frac{F_{Y_0,T,D}(y,1,d)}{\alpha_{1,d}}.$$

Also we have,

$$\begin{aligned}&\sqrt{n}(\hat{F}_{Y_0,T,D}(y,1,d) - F_{Y_0,T,D}(y,1,d))\\ =&\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{F}_{Y_0|W,T,D}(y|W_i,1,d)1\{T_i=1,D_i=d\} - \int F_{Y_0|W,T,D}(y|W,1,d)1\{T=1,D=d\}dP\right)\\ =&\sqrt{n}\left(\int \hat{F}_{Y_0|W,T,D}(y|W,1,d)1\{T=1,D=d\}dP_n - \int F_{Y_0|W,T,D}(y|W,1,d)1\{T=1,D=d\}dP\right).\end{aligned}$$

From the Slutzky lemma (Theorem 18.10 (v) in van der Vaart, 1998), Functional delta method, and the Hadamard differentiability of the operator $\phi^C(F,G)$ (Chernozhukov et al., 2013, Lemma D.1), we have

$$\left(\sqrt{n}(\hat{F}_{Y_0,T,D}(y,1,d) - F_{Y_0,T,D}(y,1,d)), n_{1,d}/n\right) \rightsquigarrow \left(\tilde{\mathbb{G}}(y,w,d), \alpha_{1,d}\right)$$

where

$$\tilde{\mathbb{G}}(y,w,d) \equiv \left(\int \mathbb{G}^{WCF}_{1,d}(y,w)dF_{W,T,D}(w,1,d) + \mathbb{G}(F_{Y_0|W,T,D}(y|\cdot,1,d)\right)I\{T=1,D=d\})$$

Therefore, continuous mapping theorem (Theorem 18.11 in van der Vaart, 1998) implies the statement. ☐

Corollary 4.5. *Under assumptions 2.1, 3.1, 3.2, 3.3, 4.1 and 4.2,*

$$\sqrt{n}\left(\hat{Q}_{Y_0|T,D}(y|1,d) - Q_{Y_0|T,D}(y|1,d)\right) \rightsquigarrow -\frac{\mathbb{G}^{CF}_{1,d}(Q_{Y_0|T,D}(y|1,d))}{f_{Y_0|T,D}(Q_{Y_0|T,D}(y|1,d)|1,d)}, \ \textit{in } l^{\infty}(\mathcal{Y}_0)$$

for every $d \in \{0,1\}$.

Proof. Immediate from the Hadamard-differentiability of the quantile function (Lemma 21.4 (ii), van der Vaart, 1998). ☐

Corollary 4.6. *Under assumptions 2.1, 3.1, 3.2, 3.3, 4.1 and 4.2,*

$$\sqrt{n}\left(\int yd\hat{F}_{Y_0|T,D}(y|1,d) - \int yF_{Y_0|T,D}(y|1,d)\right) \rightsquigarrow \int yd\mathbb{G}^{CF}_{1,d}(y)$$

Proof. Let $\mu(F) = \int ydF(y)$ be the mapping $\mu : F \to \mathbb{R}$. Let $h_t \to h$ as $t \to 0$ and let $F_t = F + th_t$. Then it is Hadamard differentiable at F tangentially to a set of functions h such that

$$\frac{1}{t}\left(\int ydF_t - \int ydF\right) = \int ydh_t \to_{t\to\infty} \int ydh.$$

Since

$$\sqrt{n}(\hat{F}_{Y_0|T,D}(y|1,d) - F_{Y_0|T,D}(y|1,d)) \rightsquigarrow \mathbb{G}^{CF}_{1,d}(y),$$

the statement holds. ☐

Given the asymptotic normality, I propose an inference based on a bootstrap procedure. Suppose the bootstrap draws are exchangeable.

Assumption 4.3 (Exchangeable bootstrap)**.** *Let* $(w_1, \ldots, w_n)$ *is an exchangeable, non-negative random vector independent of the data* $\{Y_i, Y_{b,i}, D_i, W_i, T_i\}_{i=1}^n$ *such that* *for some* $\epsilon > 0$,

$$E[w_1^{2+\epsilon}] < \infty, n^{-1}\sum_{i=1}^{n}(w_i - \bar{w})^2 \to^{\mathbb{P}} 1, \bar{w} \to^{\mathbb{P}} 1$$

where $\bar{w} = n^{-1}\sum_{i=1}^{n} w_i$, *and* $\mathbb{P}$ *is an outer probability measure with respect to* P.[19]

Let $\hat{F}^*_{Y_b|W,T,D}(y|w,1,d), \hat{F}^*_{Y_b|W,T}(y|w,1), \hat{F}^*_{Y_0|W,T}(y|w,0)$ be the bootstrapped version of the estimators using

$$\hat{\beta}^{*,k}(y) = \arg\max_{b\in\mathbb{R}^{dW}}\sum_{i=1}^{n} w_i I\{K_i = k\}\left[1\{Y_i \le y\}\log\Lambda(W_i'b) + 1\{Y_i > y\}\log(1-\Lambda(W_i'b))\right]$$

and let

$$\hat{F}^*_{W,T,D}(w,1,d) = (n^*)^{-1}\sum_{i=1}^{n} w_i 1\{W_i \le w, T_i = 1, D_i = d\}, w \in \mathcal{W}$$

where $n^* = \sum_i^n w_i$.

Corollary 4.7. *Let*

$$\hat{G}^*(f) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(w_i - \bar{w})f$$

for $f \in \mathcal{F}$, *and let*

$$\hat{G}^{*,k}(y,w) = \sqrt{n}\left(\hat{F}^*_{Y|W,K}(y|w,k) - \hat{F}_{Y|W,K}(y|w,k)\right)$$

Then under assumptions 2.1, 3.1, 3.2, 3.3, 4.1, 4.2, and 4.3

$$\left(\hat{G}^{*,1,d}(y,w), \hat{G}^{*,1}(y,w), \hat{G}^{*,0}(y,w), \hat{G}^*(f)\right) \rightsquigarrow^{\mathbb{P}} \left(\mathbb{G}^{1,d}(y,w), \mathbb{G}^1(y,w), \mathbb{G}^0(y,w), \mathbb{G}(f)\right)$$

in $l^\infty(\mathcal{Y}_b^{1,d} \times \mathcal{W}_d \times \mathcal{Y}_b^1 \times \mathcal{W} \times \mathcal{Y}_0 \times \mathcal{W}) \times l^\infty(\mathcal{F})$. *Therefore,*

$$\sqrt{n}\left(\hat{F}^*_{Y_0|T,D}(y|1,d) - \hat{F}_{Y_0|T,D}(y|1,d)\right) \rightsquigarrow^{\mathbb{P}} \mathbb{G}^{CF}_{1,d}(y).$$

[19] For an arbitrary maps $D : \Omega \mapsto \mathbb{D}$ on a metric space $\mathbb{D}$ and a bounded function $f : \mathbb{D} \mapsto \mathbb{R}$, $\mathbb{P}f(D) = \inf\{PU : U : \Omega \mapsto \mathbb{R}$ is measurable $, U \ge f(D), PU$ exists.$\}$

Proof. Argument follows from Theorem 3.6.13 of van der Vaart and Wellner (1996) as employed in Chernozhukov et al. (2013), Theorem 5.1 and 5.2. □

4.3 Cluster-robust results

Assumption 4.4 (Cluster DGP). *Let $j = 1, \dots, \bar{C}$ denote clusters, and an index (j, i) denote individual i's observation in a cluster j. Let $\{N_j, \{Y_{j,i}, Y_{b,j,i}, D_{j,i}, W_{j,i}, T_{j,i}\}_{i\geq 1}\}_{j\geq 1}$ be exchangeable, namely, for any permutation π of $\mathbb{N}$,*

$$\{N_j, \{Y_{j,i}, Y_{b,j,i}, D_{j,i}, W_{j,i}, T_{j,i}\}_{i\geq 1}\}_{j\geq 1} \sim \{N_{\pi(j)}, \{Y_{\pi(j),i}, Y_{b,\pi(j),i}, D_{\pi(j),i}, W_{\pi(j),i}, T_{\pi(j),i}\}_{i\geq 1}\}_{j\geq 1}.$$

and clusters $\{N_j, \{Y_{j,i}, Y_{b,j,i}, D_{j,i}, W_{j,i}, T_{j,i}\}$ are independent across j. Suppose further that $E[N_1] > 0, E[N_1^2] < \infty$ and $\bar{C} \to \infty$. As before, assume T is supported for $\{0, 1\}$ and D has a finite support $\{0, 1\}$.

Suppose also that the same support and density conditions as in assumption 4.1. Furthermore, $\hat{\alpha}_{1,d} \equiv \frac{1}{\bar{C}} \sum_{1\leq j\leq \bar{C}} \frac{1}{N_j} \sum_{i=1}^{N_j} 1\{T_i = 1, D_i = d\} \to \alpha_{1,d}$ where $\alpha_{1,d} \equiv E\left[\frac{1}{N_1} \sum_{i=1}^{N_1} 1\{T_i = 1, D_i = d\}\right] > 0$ for each $d \in \{0, 1\}$, and $\hat{\alpha}_0 = \frac{1}{\bar{C}} \sum_{1\leq j\leq \bar{C}} \frac{1}{N_j} \sum_{i=1}^{N_j} 1\{T_i = 0\} \to \alpha_0 \equiv E[\frac{1}{N_1} \sum_{i=1}^{N_1} 1\{T_i = 0\}] > 0$.

Assumption 4.5 (Distribution Regression and Cluster Moment Condition). *Suppose for each $k \in \mathcal{K}$,*

$$F_{Y^k|W,K}(y|w, k) = \Lambda(w'\beta^k(y)),$$

for some link function $\Lambda(\cdot)$ for all y, w. Assume also that the minimal eigenvalue of $J_k(y) \equiv E\left[\sum_{i=1}^{N_1} \frac{\lambda(W_i'\beta^k(y))^2}{\Lambda(W_i'\beta^k(y))[1-\Lambda(W_i'\beta^k(y))]} W_i W_i'\right]$ is bounded away from zero uniformly over y respectively for $k \in \{0, 1, \{1, d\}_{d\in\{0,1\}}\}$, where λ is the derivative of Λ. Assume further that $E\left[N_1 \sum_{i=1}^{N_1} \|W_i\|^2\right] < \infty$.

Remark. *The additional assumption of $E[N_1 \sum_{i=1}^{N_1} \|W_i\|^2] < \infty$ is replacing the square integrability of the envelope function const $\cdot \|W_i\|$ for a class of Z-maps as the first order conditions of the semiparameteric distribution regressions. See the appendix for the exact argument of the envelope function of the class of Z-maps.*

From the linearity of the expectation operator, it is sufficient to have the cluster size finite $N_1 < \infty$ as well as the previous moment condition $E[\|W_i\|^2] < \infty$ for every $i \in N_1$.

Let

$$\hat{G}^C(f) = \sqrt{\bar{C}} \left(\frac{1}{\bar{C}} \sum_{1 \leq j \leq \bar{C}} \sum_{i=1}^{N_j} f(W_{j,i}, T_{j,i}, D_{j,i}) - E\left[\sum_{i=1}^{N_1} f(W_{1,i}, T_{1,i}, D_{1,i}) \right] \right)$$

and

$$\begin{aligned}
\hat{G}^{C,1,d}(y_{1,d}w) =& \sqrt{\bar{C}} \left(\hat{F}^C_{Y_b|W,T,D}(y^{1,d}|w,1,d) - F_{Y_b|W,T,D}(y^{1,d}|w,1,d) \right), \forall y^{1,d} \in \mathcal{Y}_b^{w,\{1,d\}} \\
\hat{G}^{C,1}(y^1, w) =& \sqrt{\bar{C}} \left(\hat{F}^C_{Y_b|W,T}(y^1|w,1) - F_{Y_b|W,T}(y^1|w,1) \right), \forall y^1 \in \mathcal{Y}_b^{w,1} \\
\hat{G}^{C,0}(y_0, w) =& \sqrt{\bar{C}} \left(\hat{F}^C_{Y_0|W,T}(y^0|w,0) - F_{Y_0|W,T}(y^0|w,0) \right), \forall y^0 \in \mathcal{Y}^{w,0},
\end{aligned}$$

where

$$\hat{F}^C_{Y|W,K}(y|w,k) = \Lambda(w'\hat{\beta}^{C,k}(y))$$

and

$$\begin{aligned}
\hat{\beta}^{C,k}(y) = \arg\max_{b \in \mathbb{R}^{dW}} \sum_{j=1}^{\bar{C}} \sum_{i=1}^{N_j} \{ & [1\{Y_{j,i} \leq y\} \log[\Lambda(w'b)]] \\
& + [1\{Y_{j,i} > y\} \log[1 - \Lambda(w'b)]]\} \, I\{K_{j,i} = k\}
\end{aligned}$$

for each $y \in \mathcal{Y}_k$.

Corollary 4.8. *Under assumptions 2.1, 3.1, 3.2, 3.3, 4.4 and 4.5, we have*

$$\left(\hat{G}^{C,1}(y,w), \hat{G}^{C,1,d}(y,w), \hat{G}^{C,0}(y,w) \right) \rightsquigarrow \left(\mathbb{G}^{C,1}(y,w), \mathbb{G}^{C,1,d}(y,w), \mathbb{G}^{C,0}(y,w) \right)$$

in $l^\infty(\mathcal{Y}_b \times \mathcal{W} \times \mathcal{Y}_b^{1,d} \times \mathcal{W}_d \times \mathcal{Y}^0 \times \mathcal{W})$, where $\mathbb{G}^k(y,w)$ for every $k \in \{\{1,d\}_{d \in \{0,1\}}, 1, 0\}$

are tight zero-mean Gaussian processes with each covariance function of the form

$$V_{k,k}^C(y,w,\tilde{y},\tilde{w}) = \alpha_k^{-1} w' J_k^{-1}(y)\lambda_k(w'\beta^k(y))\Sigma_k(y,\tilde{y})\lambda_k(\tilde{w}'\beta^k(\tilde{y}))J_k^{-1}(\tilde{y})\tilde{w}$$

$$V_{1,\{1,d\}}(y,w,\tilde{y},\tilde{w}) = \alpha_{1,d}^{-1} w' J_1^{-1}(y)\lambda_1(w'\beta^1(y))\Sigma_{1,\{1,d\}}(y,\tilde{y})\lambda_{1,d}(\tilde{w}'\beta^{1,d}(\tilde{y}))J_{1,d}^{-1}(\tilde{y})\tilde{w}$$

where

$$\begin{aligned}
\Sigma_k^C(y,\tilde{y}) =& E[\sum_{i=1}^{N_1} I\{K_{1,i}=k\}W_{1,i}H(W_{1,i}'\beta^k(y)) \\
&\times\{\min\{\Lambda(W_{1,i}'\beta^k(y)),\Lambda(W_{1,i}'\beta^k(\tilde{y}))\} - \Lambda(W_{1,i}'\beta^k(y))\Lambda(W_{1,i}'\beta^k(\tilde{y}))\} \\
&\times H(W_{1,i}'\beta^k(\tilde{y}))]W_{1,i}', \\
\Sigma_{1,\{1,d\}}^C(y,\tilde{y}) =& E[\sum_{i=1}^{N_1} I\{T_{1,i}=1, D_{1,i}=d\}W_{1,i}H(W_{1,i}'\beta^1(y)) \\
&\times\{\min\{\Lambda(W_{1,i}'\beta^1(y)),\Lambda(W_{1,i}'\beta^{1,d}(\tilde{y}))\} - \Lambda(W_{1,i}'\beta^1(y))\Lambda(W_{1,i}'\beta^{1,d}(\tilde{y}))\} \\
&\times H(W_{1,i}'\beta^{1,d}(\tilde{y}))W_{1,i}']
\end{aligned}$$

for each $k \in \{\{1,d\}_{d\in\{0,1\}}, 1, 0\}$ *and* $V_{1,0}^C(y,w,\tilde{y},\tilde{w}) = V_{\{1,d\},0}^C(y,w,\tilde{y},\tilde{w}) = 0$*, and we have*

$$\hat{G}^C(f) \rightsquigarrow \mathbb{G}^C(f)$$

in $l^\infty(\mathcal{F})$ *for suitably measurable* $\mathcal{F}$ *specified earlier where* $\mathbb{G}^C(f)$ *is a tight zero-mean Gaussian process with the covariance kernel*

$$V^C(f_1,f_2) = Cov\left(\sum_{i=1}^{N_1} f_1(W_{1,i},D_{1,i},T_{1,i}), \sum_{i=1}^{N_1} f_2(W_{1,i},D_{1,i},T_{1,i})\right)$$

Proof. It is sufficient to verify conditions for Theorem 3.1 in Davezies et al. (2018). Assumption 1 is guaranteed by the exchangeable cluster assumption. Assumption 2 is assumed and can be verified to be pointwise measurable. For the assumption 3 in Davezies et al. (2018), it is sufficient to show that the envelope function of the class of Z-functions F satisfies

$$E[N_1 \sum_{i=1}^{N_1} F(Y_{1,i}, W_{1,i}, D_{1,i}, T_{1,i})^2] < \infty$$

In fact, the envelope function is $const \cdot \|W_i\|$ and therefore it is sufficient to have

$$E[N_1 \sum_{i=1}^{N_1} \|W_{1,i}\|^2] < \infty$$

which is guaranteed by the assumption. Finally, the finiteness of the uniform entropy integral is also guaranteed by the lemma A.1 so that all the assumptions hold.

As the weak convergences of the Z-function processes are guaranteed, Functional Delta method guarantees the statement to hold. □

5 Application

5.1 Background of the Microcredit Experiment in Morocco

Crépon et al. (2015) ran an experiment in rural areas of Morocco. Authors choose targeted areas so that villagers in the areas had not experienced the microcredit services before this experiment. This location choice is an innovative feature of this microcredit study that allowed the authors to successfully estimate the effect of the access to the microcredit services rather than the effect of the expansion of the microcredit services. As is introduced in the identification section, let T denote the binary access to the microcredit service, and let D be the binary take-up of the credit. By construction, the units in the control villages $T = 0$ have $D = 0$ automatically. The administrative observation of the take-up decision verifies the successful implementation of this procedure.

The most feature of Crépon et al. (2015) that is most relevant to the analysis of this paper is that they collect a detailed baseline survey before the experimental intervention. As a result, the dataset contains the market values of the production output before and after the experiment. More precisely, the study collects the output measure in the baseline survey before the experiment as well as in the follow-up survey two years after the experiment. Let Y_b denote the baseline sales value output and $Y = TY_1 + (1-T)Y_0 = T(DY_{11} + (1-D)Y_{10}) + (1-T)Y_0$ denote the endline sales value output.

In the experiment, a local microfinance institution called Al Amana entered randomly selected treatment villages. The flow of the experiment is as follows. Initially, Al Amana opened new branches at the beginning of the study. After the opening of the branches, the study took a baseline survey from villages. This baseline survey contains all the outcome measures that the study is interested in as the final outcome measures. Once the study completed the baseline survey, randomization separated 162 villages into 81 pairs of similar villages in observed characteristics and one of each pair was randomly assigned to be a treatment village and the other to be a control villages. For the treatment villages, Al Amana agents visited the villages and promoted the microcredit participation. At the time of the intervention, all the new branches were fully functional in their services. On the other hand, the control villages had no access to the microcredit. We can verify this feature of the one-sided non-compliance as the administrative report of the fraction of Al Amana clients is zero for the control. The study collected baseline survey observations of 4465 households from 162 villages.

The table 5.1 below shows the ITT estimates using regression analysis to control for covariates. The covariates include number of household members, number of adults, head age, does animal husbandry, does other non-agricultural activity, had an outstanding loan over the past 12 months, HH spouse responded to the survey, and other HH member responded to the survey.[20] As the column (1) shows, there is a positive and significant effect of the microcredit intervention on the sales value of the production outputs. As my method requires a continuous Y_b, I need to restrict

[20] These linear regression analysis also condition on strata dummies (paired villages). For my procedures, these paired village dummies are not included in order to prevent the incidental parameter problem for non-linear estimators.

Table 5.1: Intention to Treatment Effects on Business Outcomes

Outcome:	output	output	log output †
	(1)	(2)	(3)
Treatment	6061.3***	5888.8*	0.3215***
	(2166.8)	(3038.7)	(0.1167)
		self-employed at baseline	self-employed at baseline
control mean	30,450	33,554	8.7079
Obs	4,934	2,453	2,453

Note: Standard errors reported in parenthesis are clustered in village levels. *, ** and *** indicate statistical significance of 10%, 5%, 1% sizes respectively. Units in levels are Moroccan Dirham, 1MAD ≈ 0.106 USD.

† : inverse hyperbolic sine transformation, $log(x + \sqrt{x^2 + 1})$, is applied instead of the log to prevent ill-defined value for output value equals to zero.

the study sample to the individuals who have positive sales values in the baseline survey.[21] The columns (2) and (3) show the same estimates but for the self-employed individuals at the baseline. This procedure reduces the sample to about a half of the original sample. The column (2) shows the effect on the level of the output and the column (3) shows the effect on the log of the output. [22]Both effects are positive and

Table 5.2: Quantiles of the outcome measures

	output whole	output subsample	log output subsample
min	0	0	0
25%	505	1,816	8.198
median	6,960	9,706	9.883
75%	26,412	30,238	11.010
max	1,387,053	1,231,900	14.717

precisely measured.

5.2 Original Use of the Baseline Survey

Crépon et al. (2015) collected the baseline survey for the similar purpose to mine.

[21] As emphasized earlier, Y_0 and Y_1 do not need to be continuous. Conditional on $Y_b > 0$, there are observations with $Y = 0$ which represents exit behaviors during the study period.

[22] As the output values contain zero, the inverse hyperbolic sine transformation, $log(x+\sqrt{x^2 + 1})$, is applied instead of the natural log. Therefore, the estimates have an interpretation of the approximated semi-elasticities for small effects. Although we need to convert larger effects through hyperbolic sine formula, standard exponential approximation works well for the evaluation at large mean, which is the case for this study. Bellemare and Wichman (2018) explores detailed discussion.

The purpose is to assess the heterogeneity in the ITT of the microcredit access by its take-up decision. In particular, our shared concern is the possibility of the direct effect of the microcredit access for those who do not take-up the credit.

There are many reasons why the access to the microcredit may have a direct effect even though units do not take-up the credit. For example, promotion of the credit company may encourage small business owners to continue their business. Because of the new access, the owners expect to have an additional credit available in the future even though they are not borrowing right now. The nature of the village level assignment also generates equilibrium effects and peer effects. The expanded credit use in the treatment village would alter the interest rates of informal lending and product prices within the village as an equilibrium realization. Furthermore, non-borrowing units may receive transfers from relatives and friends who succeed in their business due to borrowing, while the units may also need to cover the payment for the debt of failing relatives.

The idea of the original study is to estimate the propensity to borrow the credit using covariates from a baseline survey. As the observations from the baseline survey are not affected by the randomization, there is no complication to the identification of the treatment effect conditional on the propensity to borrow. In other words, the parameter

$$E[Y_1 - Y_0|P(D = 1|W = w) = p] = E[Y_1 - Y_0|W \in W_p]$$

is identified, where $W_p \equiv \{w \in \mathcal{W} : P(D = 1|W = w) = p\}$. With this parameter, one may test the following hypotheses

$$H_0^{High} : E[Y_1 - Y_0|P(D = 1|W = w) \geq p_H] = 0, H_1^{High} : E[Y_1 - Y_0|P(D = 1|W = w) \geq p_H] \neq 0,$$

and

$$H_0^{Low}: E[Y_1 - Y_0 | P(D=1|W=w) < p_L] = 0, H_1^{Low}: E[Y_1 - Y_0 | P(D=1|W=w) < p_L] \neq 0.$$

Table 5.3 shows the results in the original article based on a regression analysis. The results suggest that the units with top 30% of the propensity score have a more substantial and significant effect, while the units with bottom 30% have a small and insignificant effect.

Table 5.3: Heterogeneity by High/Low Propensities

Outcome:	output (1)	output (2)	log output (3)
Treatment High 30%	15773.7*** (4153.6)	13647.2** (6095.3)	0.4736** (0.1985)
Treatment Low 30%	646.6 (2701.1)	1818.3 (4088.1)	0.2739 (0.1900)
		self-employed at baseline	self-employed at baseline
control mean	30,450	33,554	8.7079
Obs	4,934	2,453	2,453

Note: Standard errors reported in parenthesis are clustered in village levels. *, ** and *** indicate statistical significance of 10%, 5%, 1% sizes respecively.

There are two concerns with this procedure as the primary interest is the test of a different null hypothesis

$$H_0: E[Y_1 - Y_0 | D=0, T=1] = E[Y_{10} - Y_0 | D=0, T=1] = 0$$

which is the sufficient condition for the average treatment effect on treated (ATT)

$$E[Y_1 - Y_0 | D=1, T=1] = E[Y_{11} - Y_0 | D=1, T=1]$$

to be identifiable from the IV strategy with T as an instrument for D.

There are two concerns with testing H_0^{High} and H_0^{Low} in order to evaluate H_0. First, the event of the low propensity score $\{P(D=1|W=w) < p_L\}$ is different from the event of no take-up under treatment $\{D = 0, T = 1\}$. One needs to pick an arbitrary threshold for the low propensity score p_L which is set to the bottom 30% of the distribution. However, there is no reasonable threshold to conclude the conditioning group of low propensity is similar to the subgroup of non-borrowers in the treatment village, which consists of 85% of the treated individuals. Furthermore, the subgroup of non-borrowers $(D = 0, T = 1)$ are sorted on unobservables by D. On the other hand, the above procedure sorts the individuals only by observed characteristics W.

Second, as we saw in the discussion of the identification section, a small violation to the exclusion restriction may result in a large magnitude of the bias in the ATT when the take-up probability is low. Therefore, a small violation to the null H_0, which is not detectable, may generate a huge bias in the ATT.

Following table 5.4 shows the original and additional IV estimates which are valid only if the direct effect is zero. Under the conventional assumption, the estimated ATT may overestimate the effect for those who took the credit, and the policymakers may have been overly encouraged to promote the entry of the microcredit services. The main question of this section is if these numbers are valid.

Table 5.4: ATT under assumption of no direct effect

Outcome:	output (1)	output (2)	log output (3)
Treatment	36252.6***	42025.7*	2.2949***
	(12494.2)	(21525.0)	(0.8494)
		self-employed at baseline	self-employed at baseline
control mean	30,450	33,554	8.7079
Obs	4,934	2,453	2,453

Note: Standard errors reported in parenthesis are clustered in village levels. *, ** and *** indicate statistical significance of 10%, 5%, 1% sizes respecively.

5.3 Estimation of Direct and ATT of the Microcredit

With the baseline outcome Y_b being the proxy for the control outcome Y_0, we may identify the counterfactual distribution of Y_0 conditional on the endogenous subgroup $\{T = 1, D = d\}$ for each $d \in \{0, 1\}$. See the detail of this procedure for section 3.

Outcomes Y_b and Y_0 are sales values of production outputs from small business activities. By the random assignment of the credit access T and the fact that the microcredit had not been available for the baseline period and the control villages, these two outcomes Y_b and Y_0 should be similar but the random shocks realized over two years of the study. If we assume that the changes over time, the slippages, in the rank orders from $U_{b,w}$ to $U_{0,w}$ do not depend on their counterfactual credit take-up decision D, then we may justify the rank similarity assumption.

Table 5.5 shows the estimates of the unconditional subgroup effects. Column (1) is for the direct effect of not taking the credit $D = 0$ in the treatment village $T = 1$. Column (2) is for the ATT as the combined effect of the treatment take-up $D = 1$ and the access $T = 1$.

Table 5.5: Estimates with baseline proxy Y_b

Method:	Rank Similar		2SLS	Unconditional
Label:	Direct	ATT	ATT	ITT
	(1)	(2)	(3)	(4)
Treatment	0.2933	0.9871**	2.2949**	0.4569
Effect	(0.2595)	(0.4248)	(0.8494)	
exp(TE) - 1	0.3408	1.6834	8.9234	0.5792
Obs	2453	2453	2453	2453

Note: Standard errors reported in parenthesis are generated from 300 bootstrap draws clustered in village levels for (1)-(4). *,**,*** indicates statistical significance of 10%,5% and 1% sizes respectively. Logit link is used for (1) and (2). (3) is the 2SLS estimate of the ATT under the assumption of no direct effect.

The ATT of the microcredit use and access for those who take-up the credit is shown in column (2). The ATT is strongly positive and significant. However, the magnitude of the ATT is less than 40% of the 2SLS estimate which should be comparable if the direct effect is zero.[23] The direct effect of the microcredit for those

[23] They are not precisely comparable to the estimate in (2) is the unconditional subgroup effect,

who do not borrow the credit shown in column (1). This direct effect estimate itself is similar to the magnitude of the low propensity group estimate in Table 5.3 of column (3). Although the estimated direct effect is not significant, the magnitude of the ATT (2) indicates that the direct effect must be positive.

It is worth emphasizing that the findings in Table 5.5 may change the decision of the policymakers. In order to better inform the policy makers, it is encouraged to use this approach when the direct effect is of direct interest. For the purpose, I encourage the experimental designers to collect a detailed baseline survey.

6 Extensions

6.1 The Case of Two-Sided Non-Compliance

I have demonstrated the usefulness of this methodology for an experiment with one-sided non-compliance. With an additional assumption, I show that this approach also applies to an experiment with two-sided non-compliance.

Consider a two-sided incompliance case where the treatment D may be available before the experimental assignment of the treatment T. Let D_1 and D_0 denote the potential choice under the treatment assignment T so that $D = TD_1 + (1-T)D_0$. As is usually assumed, suppose the monotonicity holds. Suppose $D_1 \geq D_0$ almost surely as the treatment assignment let more people taking-up the treatment D.

In this case, the outcome of interest takes the form

$$Y = T(D_1 Y_{11} + (1-D_1)Y_{10}) + (1-T)(D_0 Y_{01} + (1-D_0)Y_{00})$$

where Y_{td} represents the outcome if the treatment assignment is $T = t$ and the treatment take-up is $D = d$.

while the estimate in (3) is the conditional subgroup effect from the regression analysis. Nevertheless, the unconditional ATT by integrating out the covariates produces higher magnitudes which instead reinforces the issue I raise here.

The subgroup effects are now harder to interpret without additional assumptions. Here I propose a weaker version of no direct effect assumption.

Assumption 6.1 (Homogeneous direct effect)**.** *Suppose*

$$Y_{11} - Y_{01} = Y_{10} - Y_{00}$$

almost surely, and

$$E[Y_{11} - Y_{01}|T = 1, D = 1] = E[Y_{11} - Y_{01}|T = 1, D = 0]$$

This restriction assumes the direct effect of having the treatment assignment is homogeneous with or without the actual treatment take-up $D = 1$ or $D = 0$. This assumption is strong. However, the homogeneity restriction is substantially weaker than the no direct effect assumption, which assumes the direct effect is constant and the constant is zero. This feature is essential to the case when the direct effect is concerning and compatible with behavioral and equilibrium effects to generate the direct effect.

Let

$$DE \equiv Y_{11} - Y_{01} = Y_{10} - Y_{00}.$$

Let me also specify

$$\tilde{Y}_1 \equiv Y_{01}, \tilde{Y}_0 \equiv Y_{00}$$

so that

$$Y_{11} = \tilde{Y}_1 + DE, Y_{10} = \tilde{Y}_0 + DE.$$

Here $\tilde{Y}_d$ represents the outcome of taking the treatment D or not.

Under this assumption, the subgroup effects take following forms

$$\begin{aligned} E[Y_1 - Y_0|T=1, D=1] =& E[(1-D_0)(\tilde{Y}_1 - \tilde{Y}_0) + DE|T=1, D_1=1] \\ =& E[(1-D_0)(\tilde{Y}_1 - \tilde{Y}_0)|T=1, D_1=1] + E[DE|T=1], \end{aligned}$$

and

$$E[Y_1 - Y_0|T=1, D=0] = E[DE|T=1, D_1=0] = E[DE|T=1].$$

Note that

$$\begin{aligned} & E[(1-D_0)(\tilde{Y}_1 - \tilde{Y}_0)|T=1, D_1=1] \\ =& E[1 * (\tilde{Y}_1 - \tilde{Y}_0)|T=1, D_1=1, D_0=0]P(D_0=0|D_1=1) \\ & + E[0 * (\tilde{Y}_1 - \tilde{Y}_0)|T=1, D_1=1, D_0=1]P(D_0=1|D_1=1) \\ =& E[\tilde{Y}_1 - \tilde{Y}_0|T=1, D_1 > D_0]\frac{P(D_1 > D_0)}{P(D_1=1)} \end{aligned}$$

where $P(D_1 > D_0) = 1 - P(D_1 = 0) - P(D_0 = 1)$, and

$$E[\tilde{Y}_1 - \tilde{Y}_0|D_1 > D_0]$$

is the local average treatment effect (LATE).

The subgroup effects themselves have an interpretation, but it is more straightforward to estimate the direct effect

$$E[Y_1 - Y_0|T=1, D=0] = E[DE|T=1]$$

as well as the LATE

$$E[\tilde{Y}_1 - \tilde{Y}_0|D_1 > D_0] = \frac{E[Y_1 - Y_0|T=1, D=1] - E[DE]}{P(D_0=0|D_1=1)}.$$

For the identification, rank similarity of Y_0 and Y_b, may be reasonable if they are comparing the same kinds of outcomes. Here suppose we have proxy Y_b which is similar to the control outcome without treatment take-up Y_{00}.

One possibility is that the treatment is not available before the experiment. Therefore, Y_b is similar to the outcome under no treatment is taken under the control group. Namely use Y_b as the proxy for the outcome Y_{00}.

Theorem 6.1. *Consider two-sided non-compliance model* $(Y_{11}, Y_{10}, Y_{01}, Y_{00}, D_1, D_0, W, Y_b)$ *as specified above. Suppose assumption 2.1 holds and, assumptions 3.1 and 3.3 hold for a baseline variable* Y_b*. Let* ν_{b0} *be the latent ranking of* Y_b *conditional on* $T = 0, D = 0, W = w$ *and assume* ν_{b0} *accepts strictly increasing distribution function as well.*

Suppose further that assumption 3.2 holds as $\nu_{b0}|D_1 = 0 \sim \nu_{00}|D_1 = 0$*, and the monotonicity in the form of* $D_1 \geq D_0$ *almost surely holds. Then*

$$F_{Y_0|D_1=0}(y) = \int F_{Y_b|D_1=0,W}(Q_{Y_b|D_0=0,W}(F_{Y_{00|W}(y)}))dF_{W|D_1=0}$$

Proof. For the proof, we omit the covariates as the same argument goes through under the common support assumption.

The formula is achieved in the following steps:

$$\begin{aligned}
F_{Y_0|D_1=0}(y) = P(Y_0 \leq y|D_1 = 0) &= P(Y_{00} \leq y|D_1 = 0, D_0 = 0) \\
=&P(\nu_{00} \leq F_{Y_{00}}(y)|D_1 = 0, D_0 = 0) = P(\nu_{b0} \leq F_{Y_{00}}(y)|D_1 = 0, D_0 = 0) \\
=&P(Y_b \leq Q_{Y_b|D_0=0}(F_{Y_{00}}(y))|D_1 = 0, D_0 = 0) = P(Y_b \leq Q_{Y_b|D_0=0}(F_{Y_{00}}(y))|D_1 = 0) \\
=&P(Y_b \leq Q_{Y_b|D_0=0}(F_{Y_{00}}(y))|D = 0, T = 1).
\end{aligned}$$

Let me walk through each equality. The second equality is by the definition of $Y_0 \equiv D_0Y_{01} + (1 - D_0)Y_{00}$ and the monotonicity. The monotonicity implies that every unit under $T = 0$ with $D_1 = 0$ must not receive the treatment, namely $D_0 = 0$. This fact is shown in the following lemma A.4 in the appendix. The third equality is from the definition of ν_{00} as a ranking within $D_0 = 0$. For the fourth equality, I applied the rank similarity conditional on $D_1 = 0$ between ν_{b0} and ν_{00}. The fifth inequality converts ν_{b0} back into the level. In order to do so, we need strictly monotone $Q_{Y_b|D_0=0}$ which is the quantile function for ν_{b0}. Finally, monotonicity is applied again to conclude that we observe the quantity as $D_1 = 0$. □

Remark. *It is worth noting the identification itself does not require the homogeneous direct effect assumption. Also, the monotonicity itself is not used, but the fact that*

$$P(Y_b \leq y|D_1 = 0) = P(Y_b \leq y|D_1 = 0, D_0 = 0)$$

and

$$P(Y_0 \leq y|D_1 = 0) = P(Y_{00} \leq y|D_1 = 0, D_0 = 0)$$

while the latter may be rationalized well only in the monotonicity, as otherwise it requires equivalence in distributions of Y_{01} and Y_{00}. See these statements are in fact necessary conditions for the monotonicity in the appendix lemma A.4.

Once we get

$$E[Y_{00}|D_1 = 0, T = 1],$$

we have

$$E[Y_{10} - Y_{00}|D_1 = 0, T = 1] = E[DE]$$

so that

$$LATE = \frac{ITT - DE}{P(D_0 = 0|D_1 = 1)}$$

where $ITT = E[Y_1 - Y_0] = E[Y|T = 1] - E[Y|T = 0]$.

6.2 Differential Attrition Problem

The similar argument may apply to the differential attrition. Consider the case of potential outcomes Y_1, Y_0 are sample selected values of true potential outcomes

Y_1^*, Y_0^* so that

$$Y_1 = D_1 Y_1^*, Y_0 = D_0 Y_0^*$$

Let $Y_1^*, Y_0^* > 0$ so that 0 is taken only by the missing values. This definition is a normalization available for the bounded supported Y_t^* by adding the minimum of Y_t^* to make it positive.

Suppose the case where $D_1 \leq D_0$ almost surely. In other words, the treatment T requires some actions which may reduce the take-up rate for the survey D_1 in treatment relative to the control D_0.

Contrary to the previous argument of general case of two-sided non-compliance, we are interested primary in

$$E[Y_1 - Y_0 | T = 1, D_1 = 1] = E[Y_1 - Y_0 | T = 1, D_1 = 1, D_0 = 1] = E[Y_1^* - Y_0^*]$$

from the monotonicity, and there is no role of the homogeneous direct effect. Finally, assume we have baseline observation Y_b for the proxy of Y_0.

Corollary 6.2. *Consider the differential attrition model* $(Y_1^*, Y_0^*, D_1, D_0, W, Y_b)$ *as specified above. Suppose assumption 2.1 holds and, assumptions 3.1 and 3.3 hold for a baseline variable* Y_b*. Let* ν_{b1} *be the latent ranking of* Y_b *conditional on* $T = 0, D = 1, W = w$ *and assume* ν_{b1} *accepts strictly increasing distribution function as well.*

Suppose further that assumption 3.2 holds as $\nu_{b1} | D_1 = 1 \sim \nu_0^* | D_1 = 1$*, and the monotonicity in the form of* $D_0 \geq D_1$ *almost surely holds. Then*

$$F_{Y_0^* | D_1 = 1}(y) = \int F_{Y_b | D_1 = 1, W}(Q_{Y_b | D_0 = 1, W}(F_{Y^*_{0|W}(y)})) dF_{W | D_1 = 1}$$

so that we have identification of

$$E[Y_1^* - Y_0^* | D_1 = 1].$$

Proof. The proof immediately follows as a special case of theorem 6.1 by flipping the order of the monotonicity from $D_1 \geq D_0$ to $D_0 \geq D_1$, and focused on the case of $D_1 = 1$ which implies $D_0 = 1$ almost surely. □

Remark. *As in the case of two-sided non-compliance, the conditional independence statements*

$$P(Y_b \leq y | D_1 = 1) = P(Y_b \leq y | D_1 = 1, D_0 = 1)$$

and

$$P(Y_0 \leq y | D_1 = 1) = P(Y_0^* \leq y | D_1 = 1, D_0 = 1)$$

are sufficient for the identification. In fact, for the differential attrition, these restrictions may be substantially weaker than the monotonicity as the former may allow shocks independent from Y_b, Y_0^* conditional on $D_1 = 1$ to deviate D_0 from 1 and it does not generate the complication in the case of the general two-sided non-compliance as the missing value has the known ranking of the bottom.

The same approach may not work if the differential attrition works in the other way, namely $D_1 \geq D_0$. Nevertheless, if we may assume the rank similarity in ν_b and ν_1^* conditional on $T = 1, D = 1, W = w$, then the same argument works to identify

$$E[Y_1^* - Y_0^* | D_0 = 1].$$

7 Conclusion

This paper presents a method to identify treatment effect heterogeneity across endogenous decisions in a randomized experiment that does not require any additional instruments or specific experimental design. Instead, I use a variable from a baseline survey to proxy for control outcomes. This method only requires that the proxy variable and the control outcome to be similar in rank order.

For one-sided non-compliance case of the microcredit experiment, I identify the average treatment effect on the treated (ATT) while allowing access to the microcredit to have a direct effect (DE). The identification of the ATT uses the proxy variable of the same outcome measure in the baseline survey. I find that the ATT estimate under the conventional assumption of no direct effect produces estimates 2.3 times larger than that of my preferred estimate.

It is worth noting that this subgroup effects and the procedure for the identification of the subgroup effect apply flexibly to a variety of randomized policy evaluation problems. First flexibility appears in the type of outcome measures. Although the most natural proxy variable is the baseline survey version of the endline outcomes of interest, the proxy variable is not necessarily the same measure of the outcome of interest. This feature expands the flexibility of the application to the binary outcome measures such as an event of survival or attainment of degrees. Second flexibility is the applicability to other forms of experiments. This paper demonstrates the use of the strategy to the two-sided non-compliance problem. For the general case of the two-sided non-compliance, I show the identification of the LATE and the DE additional assumptions of monotonicity and homogeneous direct effect. The differential attrition is an important special case where I show the identification of the attrition free average effect under the monotonicity assumption. Both cases of the two-sided non-compliance process the same argument under the similar type of proxy variables for the control outcomes without any taking-up the treatment.

As the microcredit application suggests an enormous bias in the ATT due to a failure of the no DE assumption, this paper shows the importance of identifying the treatment effect heterogeneity directly. This study enables such a direct procedure with the help of proxy variables from the baseline survey. For better policy guidance and better understanding of the treatment, applied researchers should use the proposed methods, and this study encourages the baseline data collection for this vital purpose.

Although the identification is fully non-parametric, the estimation procedure follows a specific semi-parametric distribution regression. The parametric assumption imposed may not be an innocuous restriction as well as the proposed procedure may have a small sample problem when there are many discrete covariates. The current procedure excludes unbounded support of the outcome variables which may be not be appreciated. Tackling these estimation issues are future research interest.

References

Abbring, J. H. and J. J. Heckman (2007): "Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation," in *Handbook of Econometrics*, Elsevier, vol. 6B, chap. 72.

Altonji, J. G. and R. M. Blank (1999): "Race and Gender in the Labor Market," in *Handbook of Labor Economics*, Elsevier, vol. 3, chap. 48.

Angelucci, M., D. Karlan, and J. Zinman (2015): "Microcredit Impacts: Evidence from a Randomized Microcredit Program Placement Experiment by Compartamos Banco," *American Economic Journal: Applied Economics*, 7, 151–182.

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

Angrist, J. D. and J.-S. Pischke (2009): *Mostly Harmless Econometrics:an empiricist's companion*, Princeton University Press.

Athey, S. and G. W. Imbens (2006): "Identification and Inference in Nonlinear Difference-In-Differences Models," *Econometrica*, 74, 431–497.

Banerjee, A., D. Karlan, and J. Zinman (2015): "Six Randomized Evaluations of Microcredit: Introduction and Further Steps," *American Economic Journal: Applied Economics*, 7, 1–21.

Bellemare, M. F. and C. J. Wichman (2018): "Elasticities and the Inverse Hyperbolic Sine Transformation," *Working paper*.

Callaway, B., T. Li, and T. Oka (2018): "Quantile Treatment Effects in Difference in Differences Models under Dependence Restrictions and with only Two Time Periods," *Journal of Econometrics*, 206, 395–413.

Chernozhukov, V., I. Fernández-Val, and B. Melly (2013): "Inference on Counterfactual Distributions," *Econometrica*, 81, 2205–2268.

Chernozhukov, V. and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–261.

Crépon, B., F. Devoto, E. Duflo, and W. Parienté (2015): "Estimating the Impact of Microcredit on Those Who Take It Up: Evidence from a Randomized Experiment in Morocco," *American Economic Journal: Applied Economics*, 7, 123–50.

Davezies, L., X. D'Haultfoeuille, and Y. Guyonvarch (2018): "Asymptotic results under multiway clustering," *Working Paper*.

de Chaisemartin, C. and X. D'Haultfoeuille (2017): "Fuzzy Differences-in-Differences," *Review of Economic Studies*, 85, 999–1028.

D'Haultfoeuille, X. and P. Février (2015): "Identification of nonseparable triangular models with discrete instruments," *Econometrica*, 83, 1199–1210.

Dong, Y. and S. Shen (2018): "Testing Rank Invariance or Similarity in Program Evaluation," *Review of Economics and Statistics*, 100, 79–85.

Foresi, S. and F. Peracchi (1995): "The Conditional Distribution of Excess Returns: an Empirical Analysis," *Journal of the American Statistical Association*, 87, 869–877.

Frandsen, B. and L. Lefgren (2018): "Testing Rank Similarity," *Review of Economics and Statistics*, 100, 86–91.

Frangakis, C. E. and D. B. Rubin (2002): "Principal stratification in causal inference," *Biometrics*, 58, 21–29.

Han, S. (2018): "Identification in Nonparametric Models for Dynamic Treatment Effects," *Working Paper*.

Heckman, J. J., J. Smith, and N. Clements (1997): "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *Review of Economic Studies*, 64, 487–535.

Hirano, K., G. W. Imbens, D. B. Rubin, and X.-H. Zhou (2000): "Assessing the effect of an influenza vaccine in an encouragement design," *Biostatistics*, 1, 69–88.

Imai, K., L. Keele, and T. Yamamoto (2010): "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects," *Statistical Science*, 25, 51–71.

Imbens, G. and J. D. Angrist (1994): "Identification and estimation of local average treatment effects," *Econometrica*, 62, 467–475.

Imbens, G. W. and W. K. Newey (2009): "Identification and estimation of triangular simultaneous equations models without additivity," *Econometrica*, 77, 1481–1512.

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

Ishihara, T. (2017): "Partial Identification of Nonseparable Models using Binary Instruments," *Working Paper*.

Juhn, C., K. M. Murphy, and B. Pierce (1993): "Wage Inequality and the Rise in Returns to Skill," *Journal of Political Economy*, 101, 410–442.

Kim, J. H. and B. G. Park (2017): "Testing Rank Similarity in the Heterogeneous Treatment Effect Models," *Working Paper*.

Machado, J. F. and J. Mata (2005): "Counterfactual Decomposition of Changes in Wage Distribution Using Quantile Regression," *Journal of Applied Econometrics*, 20, 445–465.

Manski, C. F. (1997): "Monotone Treatment Response," *Econometrica*, 65, 1311–1334.

Melly, B. and G. Santangelo (2015): "The Change-in-Changes Model with Covariates," *Working Paper*.

Pearl, J. (2014): "Interpretation and Identification of Causal Mediation," *Psychological Methods*, 19, 459–81.

Rosenbaum, P. R. and D. B. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrica*, 70, 41–55.

Rubin, D. B. (2006): "Causal Inference Through Potential Outcomes and Principal Stratification: Application to Studies with Censoring Due to Death," *Statistical Science*, 21, 299–309.

Rüschendorf, L. (2009): "On the distributional transform, Sklar's Theorem, and the empirical copula process," *Journal of Statistical Planning and Inference*, 139, 3921–3927.

Torgovitsky, A. (2015): "Identification of nonseparable models using instruments with small support," *Econometrica*, 83, 1185–1197.

van der Vaart, A. W. (1998): *Asymptotic Statistics*, Cambridge University Press, 2nd ed.

van der Vaart, A. W. and J. Wellner (1996): *Weak Convergence and Empirical Processes: With Applications to Statistics*, New York:Springer.

Zhang, J. L. and D. B. Rubin (2003): "Estimation of Causal Effects via Principal Stratification When Some Outcomes are Truncated by Death," *Journal of Educational and Behavioral Statistics*, 28, 353–368.

A Appendix

A.1 Proofs of main results

Proof of Lemma 4.1. Step 1: Weak convergences of Z-functions

For the proof of the lemma 4.1, we would like to introduce approximate Z-map notations.

For every $y \in \mathcal{Y}$, let $\Psi(y, \beta)$ be dW-vector of population moment equations such that the true parameter $\beta^* \in \mathbb{R}^{dW}$ solves the moment condition $\Psi(y, \beta) = 0$. Let $\hat{\Psi}(y, \beta)$ be an empirical analogue. Let an estimator $\hat{\beta}(y)$ satisfies

$$\|\hat{\Psi}(y, \hat{\beta}(y))\|^2 \leq \inf_{\beta \in \mathbb{R}^{dW}} \|\hat{\Psi}(y, \beta)\|^2 + \hat{r}(y)^2$$

where $\hat{r}(y)$ is a numerical tolerance parameter with $\|\hat{r}\|_{\mathcal{Y}} = o_p(n^{-1/2})$.

Let $\phi(\Psi(y, \cdot), r(y)) : l^{\infty}(\mathbb{R})^{dW} \times \mathbb{R} \mapsto \mathbb{R}^{dW}$ be an approximate Z-map which assigns one of its $r(y)$-approximate zeros to each element $\Psi(y, \cdot)$ so that

$$\hat{\beta}(\cdot) = \phi(\hat{\Psi}(\cdot, \cdot), \hat{r}(\cdot)), \beta^*(\cdot) = \phi(\Psi(\cdot, \cdot), 0).$$

For each $q = 1, \ldots, dW$, let

$$\psi^{0,q}_{y,\beta^0}(Y, W, T) = I\{T = 0\} \left[\Lambda(W'\beta^0) - 1\{Y \leq y\}\right] H(W'\beta^0) W_q,$$

$$\psi^{1,q}_{y,\beta^1}(Y, W, T) = I\{T = 1\} \left[\Lambda(W'\beta^1) - 1\{Y_b \leq y\}\right] H(W'\beta^1) W_q,$$

$$\psi^{\{1,d\},q}_{y,\beta^{1,d}}(Y, W, T, D) = I\{T = 1, D = d\} \left[\Lambda(W'\beta^{1,d}) - 1\{Y_b \leq y\}\right] H(W'\beta^{1,d}) W_q,$$

and for each $d \in \mathcal{D}_1$. Also let ψ^0_{y,β^0}, ψ^1_{y,β^1} and $\psi^{1,d}_{y,\beta^{1,d}}$ be dW-vector valued functions with each q-th coordinate being $\psi^{0,q}_{y,\beta^0}$, $\psi^{1,q}_{y,\beta^1}$ and $\psi^{\{1,d\},q}_{y,\beta^{1,d}}$.

In the lemma A.1 below, it is shown that the union of classes of functions

$$\begin{aligned}\cup_{d\in\mathcal{D}_1, q\in\{1,\dots,dW\}}\{\psi^{0,q}_{y,\beta^0}(Y,W,T) : (y,\beta^0)\in\mathcal{Y}\times\mathbb{R}^{dW})\} \\ \cup\,\{\psi^{1,q}_{y,\beta^1}(Y,W,T) : (y,\beta^1)\in\mathcal{Y}\times\mathbb{R}^{dW})\} \\ \cup\,\{\psi^{\{1,d\},q}_{y,\beta^{1,d}}(Y,W,T) : (y,\beta^{1,d})\in\mathcal{Y}\times\mathbb{R}^{dW})\},\end{aligned}$$

is P-Donsker with a square-integrable envelope function. Let $\hat{\Psi}^t = P_n\psi^t_{y,\beta^t}$, $\hat{\Psi}^{1,d} = P_n\psi^{1,d}_{y,\beta^{1,d}}$ and $\Psi^t = P\psi^t_{y,\beta^t}$, $\Psi^{1,d} = P\psi^{1,d}_{y,\beta^{1,d}}$ for each $t\in\{0,1\}$, then the Donskerness implies

$$\left(\sqrt{n}(\hat{\Psi}^1-\Psi^1), \sqrt{n}(\hat{\Psi}^0-\Psi^0), \sqrt{n}(\hat{\Psi}^{1,d}-\Psi^{1,d})\right) \rightsquigarrow (\mathbb{G}(\psi^1_{y^1,\beta^1}), \mathbb{G}(\psi^0_{y^0,\beta^0}), \mathbb{G}(\psi^{1,d}_{y^{1,d},\beta^{1,d}}))$$

in $l^\infty(\mathcal{Y}^1_b\times\mathbb{R}^{dW})^{dW}\times l^\infty(\mathcal{Y}^0\times\mathbb{R}^{dW})^{dW}\times l^\infty(\mathcal{Y}^{1,d}_b\times\mathbb{R}^{dW})^{dW}$ where $\mathbb{G}(\psi^k_{y,\beta^k})$ for each $k\in\{1,0,\{1,d\}\}$ are P-Brownian bridges.

Step 2: Applying Functional Delta method through Stacking rule

From the first order conditions, $\hat{\beta}^k(y) = \phi^k(\hat{\Psi}^k(y,\cdot), \hat{r}^k(y))$, $\hat{r}^k(y) = \max_{1\le i\le n}\|W_i\|dW/n$ for each $y\in\mathcal{Y}^k$ and $n^{1/2}\|\hat{r}\|_{\mathcal{Y}^k}\to^{\mathbb{P}} 0$, and $\beta_k(y) = \phi^k(\Psi^k(y,\cdot),0)$ for each $y\in\mathcal{Y}^k$ for every $k\in\{1,0,\{1,d\}\}$.

Following the argument of Chernozhukov et al. (2013), the three kinds of approximate Z-maps $\phi^1,\phi^0,\phi^{1,d}$ are Hadamard differentiable for each case, and from the Stacking Rule as in Lemma B.2 of Chernozhukov et al. (2013), we have

$$\begin{aligned}&\left(\sqrt{n}(\hat{\beta}^1(y^1)-\beta^1(y^1)), \sqrt{n}(\hat{\beta}^0(y^0)-\beta^0(y^0)), \sqrt{n}(\hat{\beta}^{1,d}(y^{1,d})-\beta^{1,d}(y^{1,d}))\right) \\ &\rightsquigarrow \left(-J_1^{-1}\mathbb{G}(\psi^1_{y^1,\beta^1(y^1)}), -J_0^{-1}\mathbb{G}(\psi^0_{y^0,\beta^0(y^0)}), -J_{1,d}^{-1}\mathbb{G}(\psi^{1,d}_{y^{1,d},\beta^{1,d}(y^{1,d})})\right)\end{aligned}$$

in $l^\infty(\mathcal{Y}^1_b)^{dW}\times l^\infty(\mathcal{Y}^0)^{dW}\times l^\infty(\mathcal{Y}^{1,d}_b)^{dW}$ by the functional delta method.

Step 3: Applying another Hadamard differentiable map to conclude the statement

Finally, consider the mapping $\nu^k : \mathbb{D}_{\nu^k} \subset l^\infty(\mathcal{Y}^k)^{dW} \mapsto l^\infty(\mathcal{Y}^k \times \mathcal{W})$ such that

$$b \mapsto \nu^k(b), \nu^k(b)(w, y) = \Lambda(w'b(y))$$

for every $k \in \{1, 0, \{1, d\}\}$. From the Hadamard differentiability of ν^k at $b^k(\cdot) = \beta^k(y)$ tangentially to $C(\mathcal{Y}^k)^{dW}$ with the derivative map $\alpha \mapsto \nu'_{\beta^k(\cdot)}(\alpha)(w, y) = \lambda(w'\beta^k(y))w'\alpha(y)$. From the stacking rule, applying the mapping for each process, the statement of the lemma holds. □

Lemma A.1. *Under the assumptions of the lemma 4.1, the class of functions*

$$\begin{aligned}
\cup_{d\in\mathcal{D}_1, q\in\{1,\dots,dW\}} \{\psi^{0,q}_{y,\beta^0}(Y, W, T) : (y, \beta^0) \in \mathcal{Y}^0 \times \mathbb{R}^{dW})\} \\
\cup \{\psi^{1,q}_{y,\beta^1}(Y, W, T) : (y, \beta^1) \in \mathcal{Y}^1_b \times \mathbb{R}^{dW})\} \\
\cup \{\psi^{\{1,d\},q}_{y,\beta^{1,d}}(Y, W, T) : (y, \beta^{1,d}) \in \mathcal{Y}^{1,d}_b \times \mathbb{R}^{dW})\},
\end{aligned}$$

is P-Donsker with a square-integrable envelope.

Proof. From Theorem 19.14 in van der Vaart (1998), a suitable measurable class of measurable functions $\mathcal{G}$ is P-Donsker if the uniform entropy integral with respect to an envelope function G

$$J(1, \mathcal{G}, L_2) = \int_0^1 \sqrt{\log \sup_Q N_{[]}(\epsilon \|G\|_{Q,2}, \mathcal{G}, L_2(G))} d\epsilon$$

is finite and the envelope function G satisfies $\mathbb{P}_1 G^2 < \infty$.

Consider classes of functions

$$\mathcal{F}_1 = \{W'\beta : \beta \in \mathbb{R}^{dW}\}, \mathcal{F}_{2,k} = \{1\{Y \leq y\}, y \in \mathcal{Y}^k\}, \{W_q : q = 1, \dots, d_w\}$$

which are VC classes of functions. Note that the target class of functions is the union of

$$\{I\{T = t\}(\Lambda(\mathcal{F}_1) - \mathcal{F}_{2,t})H(\mathcal{F}_1)W_q : q = 1, \dots, d_w\}$$

for each $t \in \{0,1\}$ and

$$\{I\{T=1,D=d\}(\Lambda(\mathcal{F}_1)-\mathcal{F}_{2,\{1,d\}})H(\mathcal{F}_1)W_q : q=1,\dots,d_w\}$$

for each $d \in \mathcal{D}_1$. These are Lipschitz transformation of VC-class of functions and finite set of functions $I\{T=t\}$ and $I\{T=1,D=d\}$ where the Lipschitz coefficients bounded by $const \cdot \|W\|$. Therefore, from Example 19.19 of van der Vaart (1998), the constructed class of functions has the finite uniform entropy integral relative to the envelope function $const \cdot \|W\|$ which is square-integrable from the assumption. Suitable measurability is granted as it is a pointwise measurable class of functions. Thus, the class of functions is Donsker. □

Lemma A.2. *Under assumptions 3.1, 3.3 and 4.1, it is possible to construct a class of measurable functions $\mathcal{F}$ including*

$$\{F_{Y|W,K}(y|\cdot,k), y \in \mathcal{Y}^k, k \in K \equiv \{\{T=1\},\{T=0\},\{T=1,D=d\}_{d\in\mathcal{D}_1}\}$$

and all the indicators of the rectangles in $\bar{\mathbb{R}}^{dW}$ such that $\mathcal{F}$ is DKP class (Chernozhukov et al., 2013, Appendix A.).

Proof. As in Step 2 in the proofs of Theorem 5.1 and 5.2 in Chernozhukov et al. (2013), $\mathcal{F}_1 = \{F_{Y|W,T}(y|\cdot,1), y \in \mathcal{Y}_b^1\}, \mathcal{F}_0 = \{F_{Y|W,T}(y|\cdot,0), y \in \mathcal{Y}^0\}$ and $\mathcal{F}_{1,d} = \{F_{Y|W,T,D}(y|\cdot,1,d), y \in \mathcal{Y}_b^{1,d}\}$ are uniformly bounded "parametric" family (Example 19.7 in van der Vaart (1998)) indexed by $y \in \mathcal{Y}^k$ for each $k \in \{1,0,\{1,d\}\}$ respectively. From the assumption that the density function $f_{Y|W,K}(y|\cdot,k)$ is uniformly bounded,

$$|F_{Y|W,K}(y|\cdot,k)I\{K=k\} - F_{Y|W,K}(y'|\cdot,k)I\{K=k\}| \le L|y-y'|$$

for some constant L for every $k \in \{1,0,\{1,d\}\}$. The compactness of $\mathcal{Y}^k$ implies the uniform ϵ-covering numbers to be bounded by $const/\epsilon$ independent of F_W so that the Pollard's entropy condition is met. As noted in footnote in the section 4, the class of $\mathcal{F}$ is suitably measurable as well. As indicator functions of all rectangles in $\bar{R}^{dW}$ form a VC class, we can construct $\mathcal{F}$ that contains union of all the families $\mathcal{F}_0, \mathcal{F}_1$ and $\mathcal{F}_{1,d}$ and the indicators of all the rectangles in $\bar{\mathbb{R}}^{dW}$ that satisfies DKP condition.

□

For the proof of the Hadamard derivative expression, I show the following lemma.

Lemma A.3. *Suppose F_1 is a uniformly continuous and differentiable distribution function with uniformly bounded density function f_1 with its support $[a,b]$ where $0 < a < b < 1$. Let $\psi_{F_1}(p) = Q_1(p)$ for all $p \in [0,1]$.*

Let F_0 be a distribution function over a support $\mathcal{Y}^0$, then ψ_{F_1} is Hadamard differentiable at F_0 tangentially to a set of function h_0 with the derivative map

$$\frac{h_0}{f_1 \circ Q_1 \circ F_0}.$$

Proof. From the assumption, we have

$$F_1 \circ Q_1 \circ (F_0 + th_{0t}) - F_1 \circ Q_1 \circ F_0 = (F_0 + th_{0t}) - F_0 = th_{0t}.$$

Let $h_{0t} \to h_0$ uniformly in $\mathcal{Y}_0$ as $t \to 0$, and let $g_t \equiv \psi_{F_1} \circ (F_0 + h_{0t})$ and $g \equiv \psi_{F_1} \circ F_0$. Then $g_t \to g$ uniformly in $\mathcal{Y}_0$ as $t \to 0$ by the assumption.

Thus, we have

$$\frac{F_1(g_t) - F_1(g)}{g_t - g} \frac{g_t - g}{t} = h_{0t}$$

so that we have

$$\frac{g_t - g}{t} = \frac{h_{0t}}{\frac{F_1(g_t) - F_1(g)}{g_t - g}}$$

whereas the RHS has a limit

$$\frac{h_{0t}}{\frac{F_1(g_t)-F_1(g)}{g_t-g}} \to^{t\to 0} \frac{h_0}{f_1 \circ Q_1 \circ F_0}$$

by the uniform differentiability of F_1.

□

Proof of Lemma 4.2. First, let $\phi(F_1, F_0) \equiv Q_1 \circ F_0$ so that $m(F_{1,d}, F_1, F_0) = F_{1,d} \circ \phi(F_1, F_0)$. From the lemma A.3 and the lemma 21.4 (ii) from van der Vaart (1998), ϕ is Hadamard differentiable at (F_1, F_0) tangentially to the set of functions (h_1, h_0) with the derivative map

$$-\frac{h_1 \circ Q_1 \circ F_0}{f_1 \circ Q_1 \circ F_0} + \frac{h_0}{f_1 \circ Q_1 \circ F_0}.$$

Therefore, from the Chain rule of the Hadamard differentiability (lemma 20.9, van der Vaart, 1998), the map m is Hadamard differentiable with the derivative map shown in the lemma. □

A.2 Monotonicity Fact

Lemma A.4. *Suppose $D_0 \leq D_1$ almost surely and $(Y_{01}, Y_{00}, Y_b, D_1, D_0)$ are measurable with respect to the same probability space $(\Omega, \mathcal{A}, P)$. Then*

$$P(Y_b \leq y | D_1 = 0) = P(Y_b \leq y | D_1 = 0, D_0 = 0)$$

and

$$P(Y_0 \leq y | D_1 = 0) = P(Y_{00} \leq y | D_1 = 0, D_0 = 0)$$

where $Y_0 = D_0 Y_{01} + (1 - D_0) Y_{00}$.

Proof. From the assumption, for P-almost every $\omega \in \Omega$,

$$D_0(\omega) \leq D_1(\omega)$$

therefore if $D_1(\omega) = 0$ then it must be $D_0(\omega) = 0$ for P-almost every ω. Thus, for $\Omega_0 \equiv \{\omega : D_1(\omega) = 0\}$, $\Omega_{00} \equiv \{\omega : D_1(\omega) = 0, D_0(\omega) = 0\}$, $P(\Omega_0 \backslash \Omega_{00}) = 0$. Let $\tilde{\Omega} \equiv \Omega_0 \backslash \Omega_{00}$.

As a result,

$$\begin{aligned}
\frac{P(\{\omega : Y_b(\omega) \leq y\} \cap \Omega_0)}{P(\Omega_0)} =& \frac{P(\{\omega : Y_b(\omega) \leq y\} \cap (\Omega_{00} \cup \tilde{\Omega}))}{P(\Omega_0)} \\
=& \frac{P(\{\omega : Y_b(\omega) \leq y\} \cap \Omega_{00})}{P(\Omega_0)} + \frac{P(\{\omega : Y_b(\omega) \leq y\} \cap \tilde{\Omega})}{P(\Omega_{00})} \\
=& \frac{P(\{\omega : Y_b(\omega) \leq y\} \cap \Omega_{00})}{P(\Omega_{00})}
\end{aligned}$$

where the second equality is by the disjointness of Ω_{00} and $\tilde{\Omega}$, and finally equality is from the fact that the second term has measure zero as well as $P(\Omega_0) = P(\Omega_{00})$.

Similarly,

$$\begin{aligned}
\frac{P(\{\omega : Y_0(\omega) \leq y\} \cap \Omega_0)}{P(\Omega_0)} =& \frac{P(\{\omega : Y_0(\omega) \leq y\} \cap (\Omega_{00} \cup \tilde{\Omega}))}{P(\Omega_0)} \\
=& \frac{P(\{\omega : Y_{00}(\omega) \leq y\} \cap \Omega_{00})}{P(\Omega_0)} + \frac{P(\{\omega : Y_{01}(\omega) \leq y\} \cap \tilde{\Omega})}{P(\Omega_{00})} \\
=& \frac{P(\{\omega : Y_{00}(\omega) \leq y\} \cap \Omega_{00})}{P(\Omega_{00})}.
\end{aligned}$$

□

A.3 Quantile Difference Plots

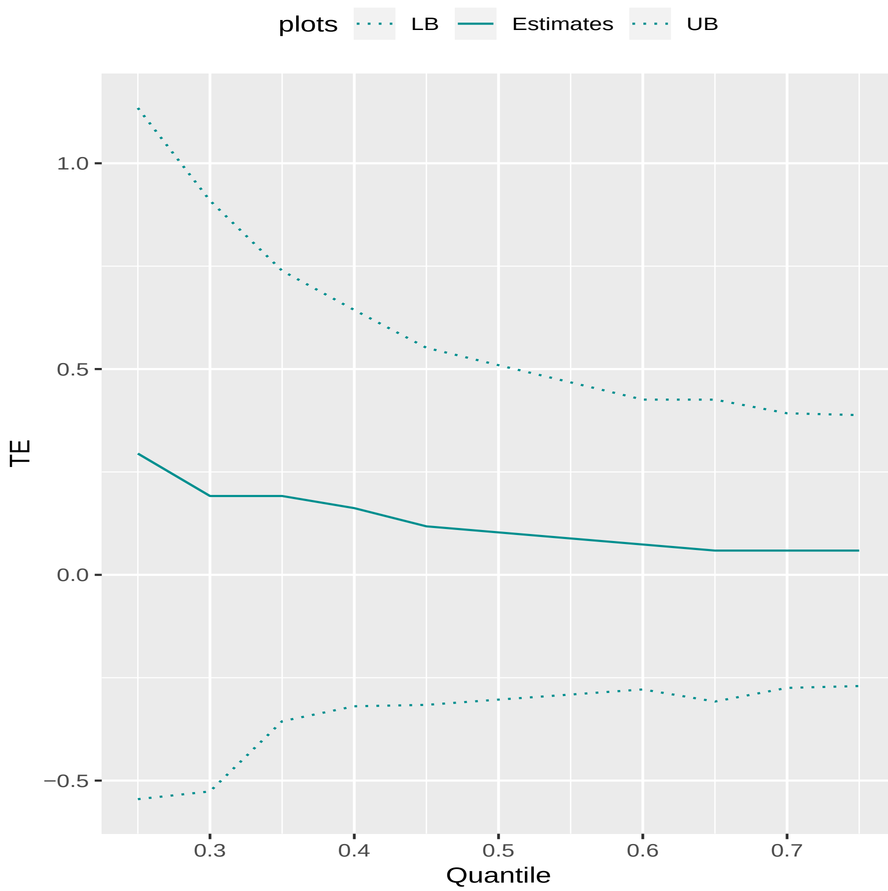


Figure A.1: Quantile difference for D = 0 with Uniform 95% CI

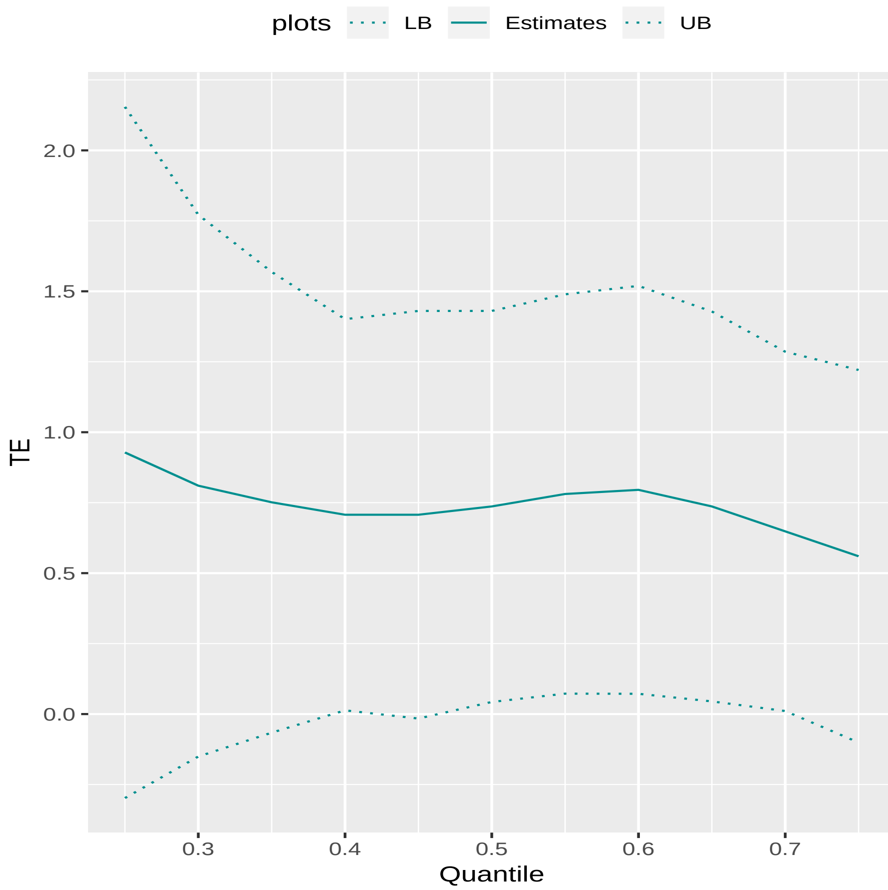


Figure A.2: Quantile difference for D = 1 with Uniform 95% CI

A.4 Robustness to other link functions

Table A.1: ATT estimates with other link functions

Link:	Logit	Probit	comp LogLog	Cauchy
	(1)	(2)	(3)	(4)
Treatment	0.9871**	0.9954**	0.9895**	0.8722*
	(0.4248)	(0.4215)	(0.4569)	(0.4508)
Obs	2453	2453	2453	2453

Note: Standard errors reported in parenthesis are generated from 300 bootstrap draws clustered in randomization cluster levels for (1)-(4). *,**,*** indicates statistical significance of 10%,5% and 1% sizes respectively.

Table A.2: Direct effect estimates with other link functions

Link:	Logit	Probit	comp LogLog	Cauchy
	(1)	(2)	(3)	(4)
Treatment	0.2933	0.2926	0.2913	0.2693
	(0.2593)	(0.2599)	(0.2623)	(0.2709)
Obs	2453	2453	2453	2453

Note: Standard errors reported in parenthesis are generated from 300 bootstrap draws clustered in randomization cluster levels for (1)-(4). *,**,*** indicates statistical significance of 10%,5% and 1% sizes respectively.

A.5 Map of Al Amana Branches

Source: Al Amana and Crépon, Devoto, Duflo and Parienté (2015)