Three essays on econometrics of heterogeneous effects.

by

Santiago Acerenza Fleitas

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Economics

Program of Study Committee:
Otavio Bartalotti, Co-major Professor
Désiré Kédagni, Co-major Professor
Quinn Weninger
Peter F. Orazem
Jae-Kwang Kim

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2021

DEDICATION

I would like to dedicate this thesis to my parents Pablo and Anabel. To grandparents, Nene and Felipe. Also to my friends Carlos, Matias, Santiago, Eseul, Raul and Hannah. Without these people support I would not have been able to reach this particular moment of my life. Finally I would also like to dedicate this to the memory of Luis Alberto Spinetta, without his music I would have not made it to the end.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

ACKNOWLEDGMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this thesis. First and foremost my advisors Désiré Kédagni and Otavio Bartalotti for their guidance, patience and support throughout this research and the writing of this thesis. Their insights and words of encouragement have often inspired me and renewed my hopes for completing my graduate education. I would also like to thank my committee members for their efforts and contributions to this work: Quinn Weninger, Peter F. Orazem and Jae-Kwang Kim. I would also like to thank Brent Kreider and Kyunghoon Ban for useful comments. I would additionally like to thank Nestor Gandelman, Flavia Roldan, Francisco Rosas and Juan Jose Barrios for their guidance throughout the initial stages of my career.

ABSTRACT

In this dissertation, econometric methods with heterogeneity are the focus. Unobserved heterogeneity affects both inference and identification of treatment effects of interest. This dissertation focuses on how to identify and do inference on parameters of interest when there is heterogeneity. This heterogeneity may affect the outcome and the variable of interest jointly, as well as it can affect the process by which a variable or an individual is observed.

CHAPTER 1. GENERAL INTRODUCTION

This dissertation comprises three papers, each connected by a shared interest in identification and inference in the presence of heterogeneity. They are ultimately concerned with understanding and developing methods that applied researchers can use in settings with heterogeneity. The first paper provides novel identification results relying on shape restrictions of marginal treatment effects in the presence of a misclassified treatment and self-selection into treatment with discrete instruments. The second provides a new method to conduct a joint hypothesis test of the identifying assumptions of bivariate probit models, a workhorse model when there is an endogenous treatment, and heterogeneous effects of such treatment on individuals. The final chapter proposes a method for identifying and doing inference correctly on a parameter of interest in the presence of a non-random sample of individuals which are connected by some unobserved shared heterogeneity across clusters.

The overarching questions for these papers are what we can do to properly estimate the effects of policies of interest when some unobserved factor makes individuals inherently different? How do we know which model to use in the presence of heterogeneity without being overly restrictive? The first and second papers tackle the second question. The first paper focuses on relaxing functional form assumptions. The second paper by testing well-established parametric assumptions. The first and third papers tackle the first question. The first paper by providing novel identification results in the presence of endogeneity and heterogeneity. The third one is extending existing methods to account for heterogeneity among individuals in certain regions in the selection of being on the sample.

CHAPTER 2. PARTIAL IDENTIFICATION OF MARGINAL TREATMENT EFFECTS WITH DISCRETE INSTRUMENTS AND MISREPORTED TREATMENT

Santiago Acerenza

Department of Economics, Iowa State University, Ames, IA, 50011, USA

Modified from a manuscript to be submitted to *Econometric Theory*

2.1 Abstract

This paper provides partial identification results for the marginal treatment effect (MTE) when the binary treatment variable is potentially misreported, and the instrumental variable is discrete. Identification results are derived under different sets of nonparametric assumptions. The identification results are illustrated in identifying the marginal treatment effects of food stamps on health.

2.2 Introduction

This paper provides partial identification results for the Marginal Treatment Effect (MTE) in the presence of measurement error in the treatment variable when only a discrete instrument is available. The discrete instrument case is relevant as many applications in the literature rely on binary instruments. The discrete nature of the instrument requires identification strategies to recover the MTE that differ from those explored in the previous literature with continuous instruments.

Researchers mostly work with self-reported data from surveys; such data systematically present reporting problems that lead to measurement error of the treatment status and, consequently, to bias in the treatment effect of interest. The combination of measurement error

with discrete instruments has not been explored in the literature, and it is a fairly common situation to encounter. The results in this paper are useful for identifying MTE (which can be used to recover average effects or policy-relevant effects) in the presence of the two previously mentioned problems for identification.

In most cases, researchers observe a discrete (often binary) instrument such as assignment to treatment. In these cases, point identification of the MTE (even without measurement error) is not possible, relying only on the standard assumptions of instrument exogeneity and relevance (See for example Brinch et al. (2017)). In this paper, under a set of restrictions on the severity of measurement error and different type of shape and support restrictions, we provide partial identification results for the MTE in the presence of measurement error when a discrete instrument is available.

The MTE can help reveal the heterogeneity in the treatment effect. The usual parameters of interest can be recovered as weighted averages of MTE. In particular, the MTE is relevant in recovering Policy Relevant Treatment Effect parameters ($PRTE$s), Average Treatment Effect (ATE), Average Treatment on the Treated (ATT), Average Treatment on the Untreated (ATU), Local Average Treatment Effects ($LATE$), etc. See Heckman and Vytlacil (2005), Heckman et al. (2006), who show the link between the MTE and those parameters via properly weighting the MTE.

To explain the relevance and intuition of this parameter, consider the example in Carneiro et al. (2011) about college attendance and wages. The MTE reflects the return to college attendance across different latent levels of the cost of going to college. Consequently, this parameter can be used to understand the heterogeneity of college effects on wages and employment; this knowledge can then be used to optimally design policies focusing on college affordability. In contrast with the standard model of education returns, the MTE allows returns to schooling to vary across individuals and to be correlated with the amount of schooling the individual takes. In terms of the traditional Mincer equation, $Y = a + bD + e$ (where Y is log wage and D is college attendance), in the MTE context, b can be interpreted as a random

coefficient potentially correlated with D in contrast with an exogenous fixed parameter. These changes have consequences for the way we conduct policy analysis. In the model with varying returns, no single average return summarizes the distribution of returns to schooling in the population. For example, the individual at the margin between schooling or not may have very different returns from all the infra-marginal individuals. In this context, standard instrumental variables estimates of the returns to schooling estimate $LATE$, which does not, in general, correspond to the return to the marginal person (who is more likely to be affected by policies on schooling than anyone else in the economy). Furthermore, different policies may affect different groups of individuals. The MTE allows b to vary across individuals and to be correlated with D[1] since individuals self select into college based on their own perceived benefit of attending college and on costs (V). It is possible that individuals who find college attendance costly are the ones that get fewer benefits from education since this cost may be correlated with their underlying ability (U). This reflects heterogeneity, and we might expect agents with the highest returns to schooling in their wages ($Y_1 - Y_0$) to be more likely to enroll in higher levels of schooling.

To achieve partial identification, we introduce different ways of recovering the MTE using nonparametric restrictions. Among the nonparametric restrictions, we present two identification routes i) monotonicity of the marginal treatment effects ($E[Y_1 - Y_0|V = v]$) in the unobservable (v) and support conditions on Y, and, ii) smoothness conditions on the marginal treatment responses ($E[Y_d|V = v]$). To deal with the misreporting of the binary treatment, the analysis relies on treating the unconditional probability of misreporting as given.[2] This can be either interpreted as the researcher having prior knowledge on the possible value of the misclassification rates or as a sensitivity analysis tool where the researcher allows for the possibility of misclassification up to a certain level. In the different identification strategies used, relevance and independence of the instrument are required.

[1] To be more specific, the determination of wages (Y or in potential notation Y_d) depends on observable covariates. Additionally, individuals with different levels of education have different wages. Also, wages are determined by unobservable components (say U). Then, we can say wages are a function of education, observed covariates, and unobservable components like ability that are captured by U.

[2] One could alternatively take the results from this paper and assume a known upper bound of this probability and take the union of the bounds derived here.

The results of this paper are relevant since it is often true that researchers have access to an instrument with discrete variation (for example, assignment to treatment via an institutional rule), and it is also true that misreporting is a common problem in survey data which is one of the main sources of empirical research.

Empirical research often entails a measurement error problem combined with endogeneity and heterogeneity. Ura (2018) documents in his work, as an example of this, that there is a substantial measurement error in educational attainments in the 1990 U.S. Census. At the same time, educational attainments are endogenous as treatment variables in return to schooling analyses because, among other possibilities, unobserved individual ability affects both schooling decisions and wages. Labor supply response to welfare program participation, in which the outcome is employment status, and the treatment is welfare program participation is subject to similar issues. Self-reported program participation in survey datasets can be misreported as stated by Hernandez and Pudney (2007). The psychological cost of welfare program participation affects job search behavior and welfare program participation simultaneously. Welfare stigma may discourage individuals from participating in a welfare program and, at the same time, affect an individual's effort in the labor market. Moreover, the welfare stigma gives welfare recipients some incentive not to reveal their participation status to the surveyor, which causes endogenous measurement error. In a job training program setting, something similar to the returns to schooling case happens. Self-reported completion of a job training program is subject to measurement error.

2.2.1 Related literature

This subsection lists some relevant papers related to the current research paper based on their connections to different aspects of the problem. Namely, misreporting, partial identification, endogeneity, marginal treatment effects.

2.2.1.1 Partial identification of $LATE, MTE$ and ATE with endogenous misreported binary treatments and heterogeneous effects

Ura (2018) using a binary instrumental variable, derives bounds for $LATE$ with a binary misreported treatment when an instrument is available, and monotonicity of the true (not observed) treatment in the instrument holds. Identification is achieved by exploiting the relationship between the probability of being a complier and the total variation distance[3] between people assigned to treatment and the ones that are not. The under-identification for $LATE$ is a consequence of the under-identification for the size of compliers; with no measurement error, one could compute the size of compliers based on the measured treatment and, therefore, $LATE$ would be the Wald estimand. The total variation distance plays a key role in determining the sharp identified set in Ura (2018). First, it measures the strength of the instrumental variable; when the total variation distance is positive, the identified set of $LATE$ is a strict subset of the whole parameter space, which implies that Z has some identifying power. Secondly, as shown in Ura (2018) lemma 3, the total variation distance is a lower bound for the proportion of compliers which is the under-identified element in the presence of measurement error. Tommasi and Zhang (2020) extend Ura (2018)'s results for the case where the instrument can take multiple discrete values. Acerenza et al. (2021) focuses on bounding the marginal treatment effects when there is a continuous instrument. Kreider et al. (2012) using auxiliary information about the possibility of misreporting and under different combinations of the outcome, treatment, and instrumental monotonicity bounds the ATE for a binary outcome. Possebom (2021) focuses on partially identifying the MTE with a continuous instrument and imposing sign and functional relationships between the derivatives of the true propensity score and the observed one with respect to the continuous instrument.

[3]The total variation distance between two probability measures P and Q on a sigma-algebra $\mathcal{F}$ of subsets of the sample space Ω is defined via $\delta(P,Q) = \sup_{A\in\mathcal{F}} |P(A) - Q(A)|$. It can alternatively be defined for probability measures that have densities to be $\frac{1}{2}\int |p-q| d\nu$ where ν is a measure dominating both probability measures. In the context of Ura (2018) paper the total variation distance calculated is $TV_{Y,D} = \frac{1}{2}\int \Big(\sum_{d=0,1} |f_{y,d|Z=1}(y,d) - f_{y,d|Z=0}(y,d)| \Big) d\nu$ where $f_{y,d|z}$ is the joint density of the observed outcome variable and the observed treatment variable conditional on the value of the instrument.

This current paper complements the previously mentioned papers. Fundamentally this paper focuses on identifying the MTE when discrete instruments are available. Such a task requires a different set of assumptions than the ones used to recover directly ATE, $LATE$, or MTE with continuous instruments. We complement Ura (2018), Kreider et al. (2012) and Tommasi and Zhang (2020) because we are interested in identifying MTE (which can then be used to achieve identification of $LATE$ and ATE) instead of the $LATE$ and ATE. It is also complementing Acerenza et al. (2021) since their analysis relies on the continuity of the instrument. It is worth noticing that it is more common to observe discrete (mostly binary) instruments such as random selection to receive treatment like in medical studies or random selection to receive a treatment conditional on covariates in social sciences (e.g., Supplemental Nutrition Assistance Program, SNAP). We complement Possebom (2021) since we provide an alternative set of assumptions to identify the MTE, and also, we are focusing on a discrete instrument.

2.2.1.2 Identification of homogeneous treatment effects with exogenous misreported treatments

This paper is also related and complements existing papers that focus on identification in the absence of unobserved heterogeneity. Lewbel (2007), Hu (2008) and Mahajan (2006) used an instrumental variable to correct for measurement error in discrete homogeneous treatment framework and achieve point identification of the average treatment effect while assuming true unobserved treatment exogeneity. For an exogenous misreported treatment Aigner (1973) used an instrumental variable to achieve identification. Similarly, but with an endogenous treatment, there is the work of DiTraglia and García-Jimeno (2019). For more examples on this framework as well as with continuous treatments instead of binary ones, see Ura (2018) paper and references therein. The current paper complements this strand of the literature by focusing on unobserved heterogeneity and relaxing the assumption of non-differential measurement error (more on this assumption below).

2.2.1.3 Other results from the partial identification literature

Kreider and Pepper (2007) and Imai and Yamamoto (2010) get bounds for the average treatment effect in the case of a misreported binary regressor utilizing the knowledge of the marginal distribution for the true treatment. Those papers obtain from auxiliary data sets the marginal distribution for the true treatment. Molinari (2008) addresses the problem of data errors in discrete variables. When data errors occur, the observed variable is a misreported version of the variable of interest, whose distribution is not identified. The approach is based on the observation that in the presence of misreporting errors, the relation between the distribution of the true but unobservable variable and its misreported representation is given by a linear system of simultaneous equations; the coefficient matrix is the matrix of misreporting probabilities. Another strand of related papers focuses on misreporting of the instrument. Chalak (2017) investigated the consequences of measurement error in the instrumental variable instead of the treatment and discussed the interpretation of various estimands (Wald, Instrumental variable: IV, local instrumental variable: LIV) when the true instrument is valid but misreported. In this setting, the MTE is not identified. Under distributional assumptions between the true instrument and the proxy, Chalak (2017) can recover the sign of the MTE. In the context of the true instrument being invalid and misreported, Kédagni (2019) derive sharp bounds on the local average treatment effect using results from Henry et al. (2014).

2.2.1.4 Identifying marginal treatment effects with discrete instruments

Brinch et al. (2017) show how a discrete instrument can be used to identify the marginal treatment effects under a functional structure that allows for treatment heterogeneity among individuals with the same observed characteristics and self-selection based on the unobserved gain from treatment. This paper builds upon Brinch et al. (2017) results by considering the case with (endogenous) misreporting and more flexible restrictions (such as shape restrictions instead of parametric assumptions) at the cost of losing point identification. The second one is Mogstad et al. (2018) which using the observed instrumental variables estimates, develops a linear

programming approach to recover policy-relevant treatment effects such as the MTE. This paper differs from it by finding analytical bounds under the different smoothness and shape restrictions. Such bounds permit one to have a first-hand insight into how the assumptions are aiding identification. Estimation of the analytical bounds is simple since it can be performed using their respective sample analogs. Additionally, Mogstad et al. (2018) does not allow for the possibility of the treatment to be misreported while here is allowed. In the presence of misreporting, the results from Mogstad et al. (2018) do not apply directly while the ones derived here do. In the case of no misreporting, our bounds remain valid; in that sense, our results complement the ones from Mogstad et al. (2018) and Brinch et al. (2017).

2.2.2 Outline of the paper

The rest of the paper is organized as follows, section 2.3 introduces the main framework and assumptions. Section 2.4 shows the main identification results. Section 2.5 illustrates the identification results with a numerical example. Section 3.8 has an application of the identification results to Kreider et al. (2012). Section 2.7 concludes. Additional non analytical results on partial identification without additional shape restrictions extending Mogstad et al. (2018) are included in section 2.9. Sections 2.10 and 2.11 discuses asymptotic inference for the ATE. Section 2.12 shows in detail how to partially identify the propensity score as in Acerenza et al. (2021).

2.3 Analytical Framework

Consider the following framework (Acerenza et al. (2021), Heckman et al. (2006) and Heckman and Vytlacil (1999)):

$$\begin{cases} Y &= Y_1 D + Y_0(1-D) \\ D &= \mathbb{1}\{p(Z) - V \geq 0\} \\ D^* &= D(1-\varepsilon) + (1-D)\varepsilon \end{cases} \tag{2.1}$$

Where Y is an outcome variable that can be discrete, continuous, or mixed, the potential outcomes are denoted by Y_d, which is the outcome realization for when treatment $D = d$,

$D = \{0, 1\}$ is a binary unobserved endogenous treatment. Let $Z \in \mathcal{Z} = \{z_0, z_1, ..., z_k\}$ be a discrete instrument,[4] V is a latent scalar random variable normalized to be uniformly distributed between $(0, 1)$. D^* is a misreported binary proxy of D, the true unobserved treatment status. $\varepsilon \in \{0, 1\}$ is a random variable indicating the presence of misreporting or not. The vector (Y, D^*, Z) is the observed data while $(Y_1, Y_0, D, \varepsilon, V)$ are latent (unobserved). In the rest of the document, small case letters denote realizations of the respective random variables.

Object of interest: In this paper, we care about identifying the $MTE(v^*)$ which is the marginal treatment effect at a particular level $V = v^*$, more precisely, it is defined as $E[Y_1 - Y_0|V = v^*]$.

To identify the MTE in this context, we introduce baseline assumptions that additional assumptions will aid. The baseline assumptions are:

Assumption 1 (Random Assignment and Absolute Continuity)**.** *The following two conditions hold:*

1. *Z is independent of (Y_d, V, ε) for all $d = (0, 1)$.*

2. *The distribution of V is absolutely continuous.*

The previous assumption and the model structure makes innocuous to say that V is uniform between $[0, 1]$ and that $p(Z) = P(D = 1|Z)$.

Assumption 2 (Relevance)**.** *Let Z be such that for any $z \in \mathcal{Z}$:*

1. *$1 > p(z) > 0$.*

2. *$p(z) \neq p(z\prime)$ for any $z, z\prime \in \mathcal{Z}$.*

Assumption 1-2 include the instrument independence and validity assumption as in Heckman and Vytlacil (1999), Heckman et al. (2006) among others. Assumption 1 requires that Z be a valid instrument, in the sense that it is statistically independent of the unobservables in the

[4]The results will focus on the binary z case but the generalization is natural for more than two values of z.

selection equation and the outcome equation. The assumption also implies that misreporting is independent of the outcome conditional on the true treatment. This assumption was used in Acerenza et al. (2021). This assumption does not require the measurement error to be non-differential; that is, conditional on the unobserved heterogeneity that drives the selection into treatment, misreporting is independent of the potential outcomes. Non-differential measurement error combined with assumption 1 implies that misreporting is independent of the outcome conditional on the true treatment, which is in general restrictive. Note that the measurement error can still depend on z, but this is through the true treatment since $D^* = D + (1 - 2D)\varepsilon$. Assumption 1 is restricting the indicator of the existence of measurement error to be independent of z but not the measurement error itself. Assumption 2 requires the existence of an instrument that shifts the probability of selection into treatment.

We introduce the working example that will help interpret the assumptions and results through the rest of the document.

Example 1. *The researcher is interested in measuring marginal returns of receiving SNAP (see details in 3.8) on food security. It is well documented that underreporting of SNAP exists. In this case, the variable Y is a binary outcome of being food secure, and D is the true indicator for being a SNAP recipient. The variable Z is the indicator of having certain assets in the household or having cars exempt from an asset test that recipients have to complete (see Kreider et al. (2012) and references therein).*

The latent variable V could be interpreted as the stigma cost of SNAP as in Moffitt (1983). As stated by Contini and Richiardi (2012) stigma is acknowledged as one of the determinants of welfare participation, and there is wide evidence that it negatively affects take-up rates.

Let Y_1 be the potential food security status for someone on SNAP, and Y_0 when the same individual does not receive it. Y_d can be correlated with the stigma cost V. As noted by Palar et al. (2018), internalized stigma may lead to food insecurity if it causes or intensifies isolation from social support systems that would allow access to food. Additionally, as stated by Earnshaw and Karpyn (2020) stigma manifestations lead to food inequities through a series of mediating

mechanisms experienced and enacted by targets of the stigma that undermine healthy food consumption, contribute to food insecurity, and ultimately impact diet quality. In that sense, psycho-social processes represent how individuals respond to stigma, which ultimately shapes their food selection, purchasing, and consumption behaviors. Enacted and anticipated stigma are characterized as significant stressors, and individuals may cope with these stressors through unhealthy eating behaviors or irrational choices that increase the likelihood of food insecurity. This is then implicitly saying that stigma could be correlated with the potential outcomes.

The variable D^* is the individual's reported (observed) indicator for SNAP recipiency. In this context, the last part of assumption 1 is consistent with saying that the stigma cost is also determining the misreporting behavior of the individual says $\varepsilon = \mathbb{1}\{f(V) \geq e\}$, if the function of the stigma cost is big enough to pass some threshold e the individual chooses to misreport consistent with Hernandez and Pudney (2007). Additionally, the assumption is consistent with random misreporting; one could think that individuals make errors when answering the survey question about SNAP recipiency with no intention. In such case $\varepsilon = \mathbb{1}\{f(\eta) \geq 0\}$ where $\eta V, Y_d$.

Besides the previously mentioned baseline assumptions, the following two different set of assumptions are introduced.

2.3.1 Monotonicity and Bounded support

Assumption 3 (Regression dependence/ Monotonicity)**.** *$E[Y_1 - Y_0|V = v]$ is either positive regression dependent or negative regression dependent. Which means, for any $v_1 > v_2$ either $E[Y_1 - Y_0|V = v_1] \geq E[Y_1 - Y_0|V = v_2]$ or $E[Y_1 - Y_0|V = v_1] \leq E[Y_1 - Y_0|V = v_2]$.*

Assumption 4 (Bounded Support)**.** *The support of Y_d is bounded above and bellow by known elements Y_{ud}, Y_{ld} respectively.*

Assumption 3 imposes constraints on the behavior of the potential outcomes and the unobserved element V by specifying the direction of the unobserved component with the treatment effect. Similar assumptions were used in Jun et al. (2011), Kédagni and Mourifié (2014), Chesher (2005) but for potential outcomes instead of the treatment effect. The concept

was also discussed by Lehmann (1966) and Balkrishnan and Lai (2009). It can be assumed the regression dependence is positive. The results for negative regression dependence are symmetric.

Such assumption is consistent for example with linear specifications of the MTE such as $E[Y_d|V=v] = \mu_d + a_d v$ and thus $MTE(v) = \mu + av$ or quadratic ones such as $E[Y_d|V=v] = \mu_d + a_d v + \beta_d v^2$ and thus $MTE(v) = \mu + av + \beta v^2$ where either $a \geq -2\beta v$ or $a \leq -2\beta v$ for every v.

Example 2 (Continued)**.** *In the context of SNAP, this would mean that the underlying cost of receiving it (the stigma, v) is related to the underlying unobservable part of my health in a monotonic way. In this sense, we could say that if taking SNAP has a psychological cost due to stigma, such costs should be monotonically related to the effect SNAP has on food insecurity.*

Assumption 4 can be seen as a generalization of the one implicitly used in Kreider et al. (2012) where they bound the average treatment effects for a binary Y with an endogenous and misreported treatment. In some settings, this assumption is trivial (for example, a binary Y like in the food security case or a discrete ordered Y such as a 0-10 happiness index).

2.3.2 Smoothness

Kim et al. (2018) introduces smoothness conditions for $E[Y_d|D=d]$ to bound the ATE without an instrument and treatment exogeneity; this approach has the same spirit. In this case, we can build on their insight to provide bounds for the MTE using similar smoothness conditions.

Assumption 5 (Smoothness)**.** *There exists known constants, $b_1, b_0 > 0$ such that for any pairs $v_1 \neq v_2$ in the support of V:*

$$-b_1|v_1 - v_2| \leq E[Y_1|V=v_1] - E[Y_1|V=v_2] \leq b_1|v_1 - v_2| \tag{2.2}$$

$$-b_0|v_1 - v_2| \leq E[Y_0|V=v_1] - E[Y_0|V=v_2] \leq b_0|v_1 - v_2| \tag{2.3}$$

Or more generally:

$$-b|v_1 - v_2| \leq E[Y_d|V=v_1] - E[Y_d|V=v_2] \leq b|v_1 - v_2| \tag{2.4}$$

For $d = 0, 1$. *Where* $b = \max\{b_0, b_1\}$

The previous assumption states the degree of smoothness of the marginal treatment responses ($E[Y_d|V = v]$). Generally speaking, we may interpret our identification analysis in this section as a conditional one indexed by b. Furthermore, we may conduct a sensitivity analysis by looking at different values of b. The parameter b is the Lipschitz constant which serves as a measure of smoothness. In this case we are assuming a maximum level of smoothness b.

Assumption 5 is restricting the functional form for the marginal responses, but considering all possible functionals in the lipschitz family with smoothness parameter b or smaller instead of a particular parametric family (like for example linear functions). It is stating the degree of smoothness of the potential responses without assuming a particular functional form of it. In this sense $E[Y_d|V = v]$ could be for example linear $E[Y_d|V = v] = \mu_d + a_d v$ (in which case $b = a_d$) or quadratic $E[Y_d|V = v] = \mu_d + a_d v + c_d v^2$ (in which case $b = a_d + 2c_d$) among different possibilities.

The following remark adapted from Kim et al. (2018) is relevant to understand what this type of assumption is imposing on the marginal treatment responses.

Remark 1. *An alternative way of bounding the rate of change in the marginal treatment responses is to impose further global restrictions in addition to monotonicity such as concavity. The approach used in this paper imposes restrictions directly on the rate of change in its nature, whereas the combination of concavity and monotonicity restricts the rate of change indirectly. There is no clear dominance between each of these ways of imposing restrictions except the belief the researcher has on the behaviour of the marginal treatment responses.*

More generally one could say $b'|v_1 - v_2| \leq E[Y_d|V = v_1] - E[Y_d|V = v_2] \leq b|v_1 - v_2|$ as stated by Kim et al. (2018), furthermore, letting $b' = 0$ and saying
$E[Y_d|V = v_1] - E[Y_d|V = v_2] \leq b|v_1 - v_2|$ for $v_1 > v_2$ would be combining monotonicity of the treatment responses with assumption 5. More precisely:

Assumption 6. *There exists known constants,* $b_1, b_0 > 0$ *such that for any pairs* $v_1 \geq v_2$ *in the support of* v*:*

$$0 \leq E[Y_1|V = v_1] - E[Y_1|V = v_2] \leq b_1(v_1 - v_2) \tag{2.5}$$

$$0 \leq E[Y_0|V = v_1] - E[Y_0|V = v_2] \leq b_0(v_1 - v_2) \tag{2.6}$$

Or more generally:

$$0 \leq E[Y_d|V = v_1] - E[Y_d|V = v_2] \leq b(v_1 - v_2) \tag{2.7}$$

Where $b = \max\{b_0, b_1\}$

Example 3 (Continued)**.** *In the context of SNAP, a binary treatment, and food security, a binary outcome, one could model the relationship using a bivariate probit model. Nevertheless, this can be restrictive since it implies a known joint distribution of the unobservables and a parametric index structure. Alternatively, one could choose to allow for all the models with* $b \leq 0.5$*. This is consistent with the bivariate probit models and allows for more generality by relaxing the normality assumption.*

2.3.3 Identification breakdown and connection to previous work.

Note that following Heckman et al. (2006) and their standard assumptions (1-2 above), without further restrictions the MTE at the level of heterogeneity v^* ($E[Y_1 - Y_0|V = v^*]$) is not identified in this setting with discrete instruments and a misreported treatment.

From standard results, we get:

$$E[Y|Z = z] = \int_0^{p(z)} E[Y_1|V = v]dv + \int_{p(z)}^1 E[Y_0|V = v]dv$$

$$E[Y|Z = z] - E[Y|Z = z\prime] = \int_{p(z\prime)}^{p(z)} E[Y_1|V = v] - E[Y_0|V = v]dv$$

The first equation shows that the observed conditional expectation of the outcome variable of interest is a combination of the share of individuals from which at that particular level of $p(z)$

(and thus of z) decided to take the treatment (all those with v between 0 and $p(z)$, recall V is normalized to be uniform) and those who decided not to take the treatment (all those with v between $p(z)$ and 1). This reveals how the observed effect is a combination of marginal responses since individuals with different levels of v will decide to self-select into taking the treatment and others not to take it. The second equation takes the difference of the first equation for any two values of the instrument connecting the observed shift in Y caused by changes in z and the underlying treatment effect for all the individuals affected by such a change of the instrument.

For any given z, say $z\prime$ we can get:

$$\begin{aligned} P[D^*=1|Z=z\prime] &= P[D=1|D^*=D,Z=z\prime]P[D^*=D|Z=z\prime] \\ &+ P[D^*=1|D^*\neq D,Z=z\prime]P[D^*\neq D|Z=z\prime]. \end{aligned} \tag{2.8}$$

Where the first equality is because we are conditioning on $D^*=D, D^*\neq D$ and applying the properties of probabilities. This last equation reflects that the observed propensity score for the proxy of the true treatment variable conditional on $z\prime$ equals the share of treated individuals who at that particular $z\prime$ report treatment status correctly multiplied by the probability of reporting correctly, plus the share of not treated individuals who at that particular $z\prime$ report treatment status incorrectly multiplied by the probability of reporting incorrectly.

The previous two expressions depend on unobserved components. While $E[Y|Z=z\prime], P[D^*=1|Z=z\prime]$ are observed, $P[D^*=D|Z=z\prime], P[D=1|D^*=D,Z=z\prime]$ and $p(Z)$ are not, which without further assumptions do not allow for identification of the true propensity score and also of the MTE.

If $p(z)$ was observed and z was continuous, then the $MTE(v^*)$ would be identified as $\frac{\partial E[Y|P(Z)=v^*]}{\partial v^*}=\frac{\int_0^{v^*}E[Y_1|V=v]dv+\int_{v^*}^1 E[Y_0|V=v]}{\partial v^*}=E[Y_1|V=v^*]-E[Y_0|V=v^*]$. So this displays the two main identification challenges, the non-continuity of z and the fact that $p(z)$ is not observed. MTE is not identified because we don't know $p(z)$ and even if we did is not continuous (since Z is discrete).

Before proceeding to the identification results, it is worth showing the main elements of the current work and how they differentiate from previous identification results of MTE with discrete instruments. It is also relevant to show the role of misreporting.

From the observed data if there is no misreporting from assumptions 1-2 one can identify:

$$E[YD|Z=z] = E[Y_1 \mathbb{1}\{p(Z) - V \geq 0\}|Z=z] = E[Y_1|p(z) - V \geq 0, Z=z]p(z)$$
$$= E[Y_1|p(z) - V \geq 0]p(z) = \int_0^{p(z)} E[Y_1|V=v]dv$$

The first equality comes from the definition of the model, the second one from the laws of probability, the third one by the independence of Z from Y_1, V and the last one from the properties of conditional expectations and the normalization that V is marginally uniform.

Similarly, we have:

$$E[Y(1-D)|Z=z] = \int_{p(z)}^1 E[Y_0|V=v]dv$$

This then implies the equality expressed at the beginning of this subsection:

$$E[Y|Z=z] = \int_0^{p(z)} E[Y_1|V=v]dv + \int_{p(z)}^1 E[Y_0|V=v]dv \tag{2.9}$$

Without misreporting the propensity score $p(z)$ is identified and, given assumptions 1-2 index sufficiency holds and thus $E[YD|Z=z] = E[YD|p(Z)=p]$. So we can rewrite the previous equalities as functions of p instead of z. Where $p = p(z)$.

In this context without differentiability of p the key insight from Brinch et al. (2017) is to introduce parametric restrictions that for example say that $E[Y_d|V=v] = \mu_d + a_d v$ and thus $E[Y_1 - Y_0|V=v] = \mu + av$, where $\mu \equiv \mu_1 - \mu_0$ and $a \equiv a_1 - a_0$. Additionally define $c = \mu_0 + \frac{a_0}{2}$. In this case

$$E[YD|p(Z)=p] = \mu_1 p + \frac{a_1}{2}p^2$$
$$E[Y(1-D)|p(Z)=p] = \mu_0(1-p) + \frac{a_0}{2}(1-p^2)$$
$$E[Y|p(Z)=p] = c + \mu p + \frac{a}{2}p^2$$

Then from $E[YD|p(Z)=p]$ for different values of p (at least two which is enough with a binary instrument) we can solve for μ_1, a_1. Similarly for a_0, μ_0 from $E[Y(1-D)|p(Z)=p]$.

Note that then given the marginal treatment responses, $E[Y_d|V=v]$ is identified, so it is the MTE as their difference.

One might not be willing to assume particular parametric specifications for the conditional expectations of the potential outcomes since they are restrictive. One of the contributions of the current work is relaxing such restrictions and still recovering analytically tractable expression for the bounds of the $MTE(v^*)$.

The current work relates Mogstad et al. (2018) in the following way. Mogstad et al. (2018) relies on recovering the set of marginal treatment responses consistent with observed IV-like estimands. In this setting, such strategy would rely on finding all the candidates $E[Y_d|V=v]$ functions consistent with:

$$\begin{aligned}
\frac{E[Y|Z=z]-E[Y|Z=z\prime]}{p(z)-p(z\prime)} &= \int_{p(z\prime)}^{p(z)} \frac{E[Y_1|V=v]}{p(z)-p(z\prime)}dv + \int_{p(z\prime)}^{p(z)} \frac{-E[Y_0|V=v]}{p(z)-p(z\prime)}dv \\
&= \int_0^1 \frac{E[Y_1|V=v]}{p(z)-p(z\prime)} 1\{v \in (p(z\prime), p(z))\}dv \\
&+ \int_0^1 \frac{-E[Y_0|V=v]}{p(z)-p(z\prime)} 1\{v \in (p(z\prime), p(z))\}dv \\
&\equiv \int_0^1 E[Y_1|V=v]w_1 dv + \int_0^1 E[Y_0|V=v]w_0 dv
\end{aligned}$$

Their strategy relies on the fact that $w_1 w_0$ are known (or identified). In the case of misreporting, where we do not know exactly the rate of false positives and false negatives for every value of z, we have that $p(z)$, and thus, the weights are not identified. This makes the current work to differ from the existing literature since developed computational methods rely on the weights being known or identified. In this context one of the main contributions of the current paper is working in the context where the weights are not identified but actually can be partially identified.

One could try to use the estimand that is identified in the data which implies replacing the denominator by the observed propensity score. In such a case one would get:

$$\begin{aligned}
\frac{E[Y|Z=z]-E[Y|Z=z\prime]}{P(D^*=1|Z=z)-P(D^*=1|Z=z\prime)} &= \int_{p(z\prime)}^{p(z)} \frac{E[Y_1|V=v]}{P(D^*=1|Z=z)-P(D^*=1|Z=z\prime)}dv \\
&+ \int_{p(z\prime)}^{p(z)} \frac{-E[Y_0|V=v]}{P(D^*=1|Z=z)-P(D^*=1|Z=z\prime)}dv \\
&= \int_0^1 \frac{E[Y_1|V=v]}{P(D^*=1|Z=z)-P(D^*=1|Z=z\prime)} 1\{v\in(p(z\prime),p(z))\}dv \\
&+ \int_0^1 \frac{-E[Y_0|V=v]}{P(D^*=1|Z=z)-P(D^*=1|Z=z\prime)} 1\{v\in(p(z\prime),p(z))\}dv \\
&\equiv \int_0^1 E[Y_1|V=v]w_1^*dv + \int_0^1 E[Y_0|V=v]w_0^*dv
\end{aligned}$$

But note that the weights w_d^* depend on unknown quantities and are thus not identified.

In section 2.4 we start from the same insight as Mogstad et al. (2018), but instead of solving a linear problem, we aid identification with shape restrictions to get analytical bounds on the MTE. In appendix 2.9 an extension of Mogstad et al. (2018) is discussed without aiding identification with shape restrictions by solving the same linear problem as in Mogstad et al. (2018) but for different values of $p(z)$ score in the identified set. As the identified set of $p(z)$ is not finite, the solution can only be approximated.

2.4 Identification results

In subsection 2.4.1 identification without misreporting will be discussed. Subsection 2.4.2 incorporates misreporting.

2.4.1 Identification without misreporting

Assumptions 3 (monotonicity) and 4 (bounded support) can aid identification of the MTE. To show their usefulness suppose we are interested in identifying the MTE at a v^* such that

$p(z\prime) \leq v^* \leq p(z)$. Then from taking the difference of equation 2.9 for two values of z:

$$\begin{aligned} E[Y|Z=z] - E[Y|Z=z\prime] &= \int_{p(z\prime)}^{p(z)} E[Y_1|V=v] - E[Y_0|V=v]dv \\ &= \int_{p(z\prime)}^{v^*} E[Y_1|V=v] - E[Y_0|V=v]dv \\ &+ \int_{v^*}^{p(z)} E[Y_1|V=v] - E[Y_0|V=v]dv \end{aligned} \tag{2.10}$$

Where the second equality is due to partitioning the integration region. Note that for every v between $(p(z\prime), v^*)$ if we assume $E[Y_1|V=v] - E[Y_0|V=v]$ is positive monotonic in v it is true that $E[Y_1|V=v] - E[Y_0|V=v] \leq E[Y_1|V=v^*] - E[Y_0|V=v^*]$. Also since Y_d is bounded, for any v between $(v^*, p(z))$ we have $E[Y_1|V=v] - E[Y_0|V=v] \leq Y_{1u} - Y_{0l}$. Then we get:

$$\begin{aligned} E[Y|Z=z] - E[Y|Z=z\prime] &= \int_{p(z\prime)}^{p(z)} E[Y_1|V=v] - E[Y_0|V=v]dv \\ &= \int_{p(z\prime)}^{v^*} E[Y_1|V=v] - E[Y_0|V=v]dv \\ &+ \int_{v^*}^{p(z)} E[Y_1|V=v] - E[Y_0|V=v]dv \\ &\leq (E[Y_1|V=v^*] - E[Y_0|V=v^*])[v^* - p(z\prime)] \\ &+ (Y_{1u} - Y_{0l})[p(z) - v^*]. \end{aligned}$$

Then we get the bound:

$$E[Y_1|V=v^*] - E[Y_0|V=v^*] \geq \frac{E[Y|Z=z] - E[Y|Z=z\prime] - (Y_{1u} - Y_{0l})[p(z) - v^*]}{v^* - p(z\prime)}.$$

Similarly:

$$\begin{aligned} E[Y|Z=z] - E[Y|Z=z\prime] &\geq (Y_{1l} - Y_{0u})[v^* - p(z\prime)] \\ &+ (E[Y_1|V=v^*] - E[Y_0|V=v^*])[p(z) - v^*]. \end{aligned}$$

Then we get:

$$E[Y_1|V=v^*] - E[Y_0|V=v^*] \leq \frac{E[Y|Z=z] - E[Y|Z=z\prime] - (Y_{1l} - Y_{0u})[v^* - p(z\prime)]}{p(z) - v^*}$$

This same logic can be applied for different positions of v^* relative to different $p(z)$'s. The key insight is to partition the integral accordingly to the position of the v^* of interest. This insight extends in the case of misreporting. The previous bounds depend on the relative position of v^* with respect to $p(z)$ as well as $p(z)$ itself, which makes clear the role misreporting since, in that case, $p(z)$ and the relative positions are no longer identified. Nevertheless, partially identifying $p(z)$ will provide a way of bounding the MTE even in the presence of misreporting following a similar logic of the previous display.

Instead of assuming monotonicity, one can introduce restrictions on the degree of differentiability of the marginal treatment response functions ($E[Y_d|V=v]$) such as in assumption 5 and without the need to introduce assumption 4 (bounded support). Note than since

$$E[Y|Z=z] = \int_0^{p(z)} E[Y_1|V=v]dv + \int_{p(z)}^1 E[Y_0|V=v]dv.$$

We can for any two values of Z:

$$\begin{aligned}
E[Y|Z=z] - E[Y|Z=z\prime] &= \int_{p(z\prime)}^{p(z)} E[Y_1|V=v] - E[Y_0|V=v]dv \\
&= \int_{p(z\prime)}^{p(z)} \Big(E[Y_1|V=v] - E[Y_1|V=v^*] - E[Y_0|V=v] \\
&+ E[Y_0|V=v^*] + E[Y_1|V=v^*] - E[Y_0|V=v^*] \Big) dv \\
&\leq \int_{p(z\prime)}^{p(z)} 2b|v-v^*|dv + \int_{p(z\prime)}^{p(z)} MTE(v^*)dv \\
&= \int_{p(z\prime)}^{p(z)} 2b|v-v^*|dv + MTE(v^*)(p(z)-p(z\prime)).
\end{aligned}$$

Where we are adding and subtracting the marginal treatment responses ($E[Y_d|V=v^*]$) related to the marginal treatment effect at the v^* of interest ($E[Y_1|V=v^*] - E[Y_0|V=v^*]$), then using $E[Y_d|V=v] - E[Y_d|V=v^*] \leq b|v-v^*|$ twice and also the definition of MTE. Similarly we can get:

$$E[Y|Z=z] - E[Y|Z=z\prime] \geq \int_{p(z\prime)}^{p(z)} -2b|v-v^*|dv + MTE(v^*)(p(z)-p(z\prime)).$$

Then:

$$\frac{E[Y|Z=z] - E[Y|Z=z\prime] - \int_{p(z\prime)}^{p(z)} 2b|v-v^*|dv}{p(z)-p(z\prime)} \leq MTE(v^*) \leq \frac{E[Y|Z=z] - E[Y|Z=z\prime] + \int_{p(z\prime)}^{p(z)} 2b|v-v^*|dv}{p(z)-p(z\prime)}.$$

The bounds depend on the true unobserved propensity score and the difference of the propensity score for different values of z. Then, combining the previous logic with bounding the propensity score will identify the MTE with both discrete instruments and a misreported treatment.

2.4.2 Identification of the MTE with misreporting

In order to identify the MTE first we need to identify $p(z)$. In appendix 2.12 we discuss identification of $p(z)$ as in Acerenza et al. (2021). Subsequent subsections, builds upon the results from Acerenza et al. (2021) restated in the appendix.

For clarity, let
$\Delta_p \equiv p(z) - p(z'), \Delta_{D^*Z}(z', z) \equiv P(D^* = 1|Z = z) - P(D^* = 1|Z = z'), \alpha \equiv P(\varepsilon = 1)$, then let the bounds derived in appendix 2.12 be:

$$
\begin{aligned}
\Delta_{pl} &\equiv \Delta_{D^*Z}(z', z), \\
\Delta_{pu} &\equiv \min\left\{1, 2\alpha + \Delta_{D^*Z}(z', z), 2(1-\alpha) - \Delta_{D^*Z}(z', z)\right\}, \\
p_l(z) &\equiv \max\left\{P(D^* = 1|Z = z) - \alpha, \alpha - P(D^* = 1|Z = z)\right\}, \\
p_u(z) &\equiv \min\left\{P(D^* = 1|Z = z) + \alpha, (1-\alpha) + P(D^* = 0|Z = z)\right\}.
\end{aligned}
$$

The bounds on the propensity score rely on any given level of unconditional misreporting (α). For more details on the bounding strategy, see appendix 2.12. Misreporting enters in two ways. First of all, if the probabilities of misreporting are unknown, $p(z)$ is no longer identified (see appendix 2.12 for more details), which then creates a problem since such quantity appears systematically in the bounding strategies. To solve this problem, we use the previously defined bounds on the misreporting probabilities.

The lack of point identification of $p(z)$ affects the strategy using smoothness restrictions. See for example that:

$$
E[Y|Z = z] - E[Y|Z = z\prime] \leq \int_{p_l(z\prime)}^{p_u(z)} 2b|v - v^*|dv + MTE(v^*)(p(z) - p(z\prime))
$$

The bound of MTE depends on the sign of it since it is not always true that $MTE(v^*)(p(z) - p(z\prime)) \leq MTE(v^*)(p_u(z) - p_l(z\prime))$. These considerations are taken into account in theorem 2 and theorem 3.

In the case of monotonicity and bounded support the issue is similar. Note that for example for a v^* between $p(z\prime))$ and $p(z)$

$$E[Y|Z = z] - E[Y|Z = z\prime] \leq MTE(v^*)[v^* - p(z\prime)] + (Y_{1u} - Y_{0l})[p(z) - v^*]$$

But now to get a bound on the MTE the sign of it matter since it is not always true that $MTE(v^*)(v^* - p(z)) \leq MTE(v^*)(v^* - p_l(z\prime))$. These considerations are taken into account in theorem 1.

2.4.2.1 Monotonicity and Bounded Support

This approach to partially identify the MTE imposes monotonicity on $E[Y_1 - Y_0|V = v^*]$ and combines these with bounded support assumptions and the partially identified propensity score to get bounds on the MTE. Without any misreporting, in section 2.4.1 a procedure on how to bound the MTE for v^* such that $p(z\prime) < v^* < p(z)$ was exposed. Now, following the same logic but for $p_l(z\prime) < v^* < p_u(z)$ we get:

$$E[Y|Z = z] - E[Y|Z = z\prime] \leq MTE(v^*)[v^* - p(z\prime)] + (Y_{1u} - Y_{0l})[p(z) - v^*]$$

The sign of $(Y_{1u} - Y_{0l})$ is known by assumption. Say that is positive (the symmetric argument applies if we assume it is negative). Then:

$$\begin{aligned} E[Y|Z = z] - E[Y|Z = z\prime] &\leq MTE(v^*)[v^* - p(z\prime)] \\ &\quad + (Y_{1u} - Y_{0l})[p(z) - v^*] \\ &\leq MTE(v^*)[v^* - p(z\prime)] + (Y_{1u} - Y_{0l})[p_u(z) - v^*] \end{aligned}$$

Then we have

$$E[Y|Z=z]-E[Y|Z=z\prime]-(Y_{1u}-Y_{0l})[p_u(z)-v^*]\leq MTE(v^*)[v^*-p(z\prime)]$$

If $E[Y|Z=z]-E[Y|Z=z\prime]-(Y_{1u}-Y_{0l})[p_u(z)-v^*]\geq 0$ then $MTE(v^*)\geq 0$ and thus:

$$\begin{aligned}E[Y|Z=z]-E[Y|Z=z\prime]&\leq MTE(v^*)[v^*-p(z\prime)]\\&\quad+(Y_{1u}-Y_{0l})[p(z)-v^*]\\&\leq MTE(v^*)[v^*-p(z\prime)]+(Y_{1u}-Y_{0l})[p_u(z)-v^*]\\&\leq MTE(v^*)[v^*-p_l(z\prime)]+(Y_{1u}-Y_{0l})[p_u(z)-v^*]\end{aligned}$$

Which then provides that:

$$MTE^{+}(v^*)_{lb}\equiv\frac{E[Y|Z=z]-E[Y|Z=z\prime]-(Y_{1u}-Y_{0l})[p_u(z)-v^*]}{v^*-p_l(z\prime)}$$

Note that the bound is divided by v^* implying that for certain v^* and certain values, the bounds might go to infinity[5] so then we need to incorporate the worst-case bounds also for those situations since by construction the variable Y is bounded and so is the treatment effect. Which then provides that:

$$MTE^{+}(v^*)_{lb}\equiv\max\left\{Y_{1l}-Y_{0u},\frac{E[Y|Z=z]-E[Y|Z=z\prime]-(Y_{1u}-Y_{0l})[p_u(z)-v^*]}{v^*-p_l(z\prime)}\right\}$$

Remark 2. *Note that the previous bound applies since*
$E[Y|Z=z]-E[Y|Z=z\prime]-(Y_{1u}-Y_{0l})[p_u(z)-v^*]\leq 0$. *If*
$E[Y|Z=z]-E[Y|Z=z\prime]-(Y_{1u}-Y_{0l})[p_u(z)-v^*]>0$ *then the bound is*
$MTE^{+}(v^*)_{lb}\equiv\frac{E[Y|Z=z]-E[Y|Z=z\prime]-(Y_{1u}-Y_{0l})[p_u(z)-v^*]}{v^*-p_l(z\prime)}$, *but this is logically consistent as long as*
$\frac{E[Y|Z=z]-E[Y|Z=z\prime]-(Y_{1u}-Y_{0l})[p_u(z)-v^*]}{v^*-p_l(z\prime)}\leq Y_{1u}-Y_{0l}$, *if this inequality is violated it means that the model and the assumptions are rejected by the data. A similar consideration applies for the upper bounds using* $E[Y|Z=z]-E[Y|Z=z\prime]-(Y_{1l}-Y_{0u})[p_l(z)-v^*]$ *and the reversed sign.*

[5]The intuition behind the bounds going to infinity is that when one is interested in recovering the MTE of a particular v^* given the monotonicity assumption one is using information of all individuals between v^* and $p(z)$, when v^* is too close to $p(z)$, there is not enough information to be used to recover the bounds of the MTE.

Given the sign of the MTE is known now since we used above the fact that $E[Y|Z=z]-E[Y|Z=z\prime]-(Y_{1u}-Y_{0l})[p_u(z)-v^*]\geq 0$ provides the sign of the MTE, we can also get an upper bound from:

$$E[Y|Z=z]-E[Y|Z=z\prime]\geq (Y_{1l}-Y_{0u})[v^*-p(z\prime)] \\ +(E[Y_1|V=v^*]-E[Y_0|V=v^*])[p(z)-v^*]$$

Using the fact that $MTE(v^*)$ is positive and also assuming that $(Y_{1l}-Y_{0u})\leq 0$ (the symmetric argument applies if we assume is positive):

$$MTE^{+}(v^*)_{ub}\equiv \min\left\{Y_{1u}-Y_{0l},\frac{E[Y|Z=z]-E[Y|Z=z\prime]-(Y_{1l}-Y_{0u})[v^*-p_l(z\prime)]}{p_l(z)-v^*}\right\}$$

Which then gives a lower bound that is valid for every v^* such that $p_l(z\prime)<v^*<p_u(z)$. Additionally we get an upper bound that is valid for every v^* such that $p_l(z\prime)<v^*<p_l(z)$ and between $p_l(z)<v^*<p_u(z)$ we use the worst case upper bounds. This is since the upper bound we found is divided by $p_l(z)-v^*$ and would make no sense when $p_l(z)<v^*$.

Note that if the sign was not identified from $E[Y|Z=z]-E[Y|Z=z\prime]-(Y_{1u}-Y_{0l})[p_u(z)-v^*]$ one could try to identify it from $E[Y|Z=z]-E[Y|Z=z\prime]-(Y_{1l}-Y_{0u})[p_l(z)-v^*]\leq 0$ and then the MTE would be negative and the bounds would change to:

$$MTE^{-}(v^*)_{lb}\equiv \max\left\{Y_{1l}-Y_{0u}\frac{E[Y|Z=z]-E[Y|Z=z\prime]-(Y_{1u}-Y_{0l})[p_u(z)-v^*]}{v^*-p_u(z\prime)}\right\}$$

$$MTE^{-}(v^*)_{ub}\equiv \min\left\{Y_{1u}-Y_{0l},\frac{E[Y|Z=z]-E[Y|Z=z\prime]-(Y_{1l}-Y_{0u})[v^*-p_l(z\prime)]}{p_u(z)-v^*}\right\}$$

Which gives an upper bound that is valid between $p_l(z\prime)<v^*<p_u(z)$. The lower bound is valid between $p_u(z\prime)<v^*<p_u(z)$. For $p_l(z\prime)<v^*<p_u(z\prime)$ the worst case lower bound applies following a similar logic as the one with the opposite sign bounds.

If we cannot use

$E[Y|Z=z]-E[Y|Z=z\prime]-(Y_{1l}-Y_{0u})[p_l(z)-v^*], E[Y|Z=z]-E[Y|Z=z\prime]-(Y_{1u}-Y_{0l})[p_u(z)-v^*]$

to identify the sign of the $MTE(v^*)$ then the bounds are:

$$MTE(v^*)_{lb} = \min\{MTE^-(v^*)_{lb}, MTE^+(v^*)_{lb}\}$$
$$MTE(v^*)_{ub} = \max\{MTE^-(v^*)_{ub}, MTE^+(v^*)_{ub}\}$$

The following theorem summarizes this discussion for a binary z and follows the similar logic for different relative positions of v^* with respect to $p(z)$. The same notion can be extended to more values of the instrument and combining information from different $E[Y|Z=z] - E[Y|Z=z\prime]$. Say, for example, one is interested in bounding the MTE for an instrument that takes three values. One cares about $p(z\prime) < p(z\prime\prime) < v^* < p(z)$ then it can be used both the information from $E[Y|Z=z] - E[Y|Z=z\prime]$ and $E[Y|Z=z] - E[Y|Z=z\prime\prime]$.

Theorem 1. *Under assumptions 1-4 and assuming that $Y_{1u} - Y_{0l} > 0$, $Y_{1l} - Y_{0u} < 0$ the following bounds for the $MTE(v^*)$ are valid for any $p(z\prime) < p(z)$:*

1. *For $v^* < p_l(z\prime)$ If $E[Y|Z=z] - E[Y|Z=z\prime] \leq 0$ ($MTE(v^*) \leq 0$):*

$$MTE(v^*)_{lb1} = Y_{1l} - Y_{0u}$$
$$MTE(v^*)_{ub1} = \min\left\{Y_{1u} - Y_{0l}, \frac{E[Y|Z=z] - E[Y|Z=z\prime]}{\Delta_{pu}}\right\}$$

2. *For $v^* < p_l(z\prime)$ If $E[Y|Z=z] - E[Y|Z=z\prime] \geq 0$ ($MTE(v^*)$ cannot be signed):*

$$MTE(v^*)_{lb2} = Y_{1l} - Y_{0u}$$
$$MTE(v^*)_{ub2} = \min\left\{Y_{1u} - Y_{0l}, \max\left\{\frac{E[Y|Z=z] - E[Y|Z=z\prime]}{\Delta_{pu}}, \frac{E[Y|Z=z] - E[Y|Z=z\prime]}{\Delta_{pl}}\right\}\right\}$$

3. *For $p_l(z\prime) < v^* < p_l(z)$ If $E[Y|Z=z] - E[Y|Z=z\prime] - (Y_{1u} - Y_{0l})[p_u(z) - v^*] \geq 0$ ($MTE(v^*) \geq 0$):*

$$MTE^+(v^*)_{lb3} = \max\left\{Y_{1l} - Y_{0u}, \frac{E[Y|Z=z] - E[Y|Z=z\prime] - (Y_{1u} - Y_{0l})[p_u(z) - v^*]}{v^* - p_l(z\prime)}\right\}$$
$$MTE^+(v^*)_{ub3} = \min\left\{Y_{1u} - Y_{0l}, \frac{E[Y|Z=z] - E[Y|Z=z\prime] - (Y_{1l} - Y_{0u})[v^* - p_l(z\prime)]}{p_l(z) - v^*}\right\}$$

4. *For $p_l(z) < v^* < p_u(z)$ If $E[Y|Z=z] - E[Y|Z=z\prime] - (Y_{1u} - Y_{0l})[p_u(z) - v^*] \geq 0$ $(MTE(v^*) \geq 0)$:*

$$MTE^+(v^*)_{lb4} = \max\left\{Y_{1l} - Y_{0u}, \frac{E[Y|Z=z] - E[Y|Z=z\prime] - (Y_{1u} - Y_{0l})[p_u(z) - v^*]}{v^* - p_l(z\prime)}\right\}$$

$$MTE^+(v^*)_{ub4} = Y_{1u} - Y_{0l}$$

5. *For $p_l(z\prime) < v^* < p_u(z\prime)$ If $E[Y|Z=z] - E[Y|Z=z\prime] - (Y_{1l} - Y_{0u})[p_l(z) - v^*] \leq 0$ $(MTE(v^*) \leq 0)$:*

$$MTE^-(v^*)_{lb5} = Y_{1l} - Y_{0u}$$

$$MTE^-(v^*)_{ub5} = \min\left\{Y_{1u} - Y_{0l}, \frac{E[Y|Z=z] - E[Y|Z=z\prime] - (Y_{1l} - Y_{0u})[v^* - p_l(z\prime)]}{p_u(z) - v^*}\right\}$$

6. *For $p_u(z\prime) < v^* < p_u(z)$ If $E[Y|Z=z] - E[Y|Z=z\prime] - (Y_{1l} - Y_{0u})[p_l(z) - v^*] \leq 0$ $(MTE(v^*) \leq 0)$:*

$$MTE^-(v^*)_{lb6} = \max\left\{Y_{1l} - Y_{0u}\frac{E[Y|Z=z] - E[Y|Z=z\prime] - (Y_{1u} - Y_{0l})[p_u(z) - v^*]}{v^* - p_u(z\prime)}\right\}$$

$$MTE^-(v^*)_{ub6} = \min\left\{Y_{1u} - Y_{0l}, \frac{E[Y|Z=z] - E[Y|Z=z\prime] - (Y_{1l} - Y_{0u})[v^* - p_l(z\prime)]}{p_u(z) - v^*}\right\}$$

7. *Otherwise for $p_l(z\prime) < v^* < p_u(z)$:*

$$MTE(v^*)_{lb7} = \max\{Y_{1l} - Y_{0u}, \min\{MTE^-(v^*)_{lb5}, MTE^+(v^*)_{lb6}\}\}$$

$$MTE(v^*)_{ub7} = \min\{Y_{1u} - Y_{0l}, \max\{MTE^-(v^*)_{ub5}, MTE^+(v^*)_{ub6}\}\}$$

8. *For $v^* > p_u(z)$ if $E[Y|Z=z] - E[Y|Z=z\prime] \geq 0$ $(MTE(v^*) \geq 0)$:*

$$MTE(v^*)_{lb8} = \max\left\{Y_{1l} - Y_{0u}, \frac{E[Y|Z=z] - E[Y|Z=z\prime]}{\Delta_{pu}}\right\}$$

$$MTE(v^*)_{ub8} = Y_{1u} - Y_{0l}$$

9. *For $v^* > p_u(z)$ if $E[Y|Z=z] - E[Y|Z=z\prime] \leq 0$ $(MTE(v^*)$ cannot be signed):*

$$MTE(v^*)_{lb9} = \max\left\{Y_{1l} - Y_{0u}, \min\left\{\frac{E[Y|Z=z] - E[Y|Z=z\prime]}{\Delta_{pu}}, \frac{E[Y|Z=z] - E[Y|Z=z\prime]}{\Delta_{pl}}\right\}\right\}$$

$$MTE(v^*)_{ub9} = Y_{1u} - Y_{0l}$$

Remark 3. *The previous bounds are not sharp. Take for example for any two v_1, v_2. Let v_1 be smaller than v_2 in a way that v_1 is infinitesimally smaller than $p_u(z)$ and v_2 is infinitesimally bigger than $p_u(z)$. Let in this case the MTE to be positive at both values and let positive regression dependence hold. Let $p_l(z) < v_1 < p_u(z)$ and $p_u(z) < v_2 < 1$. In this setting nothing assures that $MTE^+(v_1)_{lb} < MTE^+(v_2)_{lb}$. Nevertheless, since positive regression dependence is assumed $MTE^+(v_1)_{lb}$ is a lower bound for every v^* bigger than v_1 since $MTE^+(v_1)_{lb}$ is the minimum value the true MTE can take at v_1. Thus the lower bound for the MTE at v_2 has a tighter bound that is the maximum between $MTE^+(v_2)_{lb}, MTE^+(v_1)_{lb}$. In practice this might be hard to compute so what can be done to get tighter bounds is instead of taking the v_1 infinitesimal smaller than v_2 take a v_3 that is the previous grid point used to evaluate the MTE and use as a lower bound for the MTE at v_2 the maximum between $MTE^+(v_2)_{lb}, MTE^+(v_3)_{lb}$. So when the researcher is using the results from theorem 1 we recommend that if he cares about the MTE at three different values $v_a < v_b < v_c$ to bound them respectively as $MTE^+(v_a)_{lb}$, $\max\{MTE^+(v_a)_{lb}, MTE^+(v_b)_{lb}\}$,$\max\{MTE^+(v_a)_{lb}, MTE^+(v_b)_{lb}, MTE^+(v_c)_{lb}\}$.*

A similar logic applies with negative regression dependence but for the upper bounds instead of lower bounds.

2.4.2.2 Smoothness assumptions for marginal treatment responses

The results from the previous subsection rely on regression dependence and bounded support of Y. We can introduce smoothness conditions alternatively.

The bounds from theorem 2 builds on the following inequalities due to assumption 5

$$E[Y|Z=z] - E[Y|Z=z\prime] \leq \int_{p_l(z\prime)}^{p_u(z)} 2b|v-v^*|dv + MTE(v^*)(p(z)-p(z\prime)) \tag{2.11}$$

$$E[Y|Z=z] - E[Y|Z=z\prime] \geq -\int_{p_l(z\prime)}^{p_u(z)} 2b|v-v^*|dv + MTE(v^*)(p(z)-p(z\prime)) \tag{2.12}$$

Where the inequalities use the smoothness assumption as in the previous section without misreporting and fact that $\int_{p(z\prime)}^{p(z)} 2b|v-v^*|dv \leq \int_{p_l(z\prime)}^{p_u(z)} 2b|v-v^*|dv$. Note then that similarly as in

the monotonicity case we have a sufficient condition for identifying the sign of the $MTE(v^*)$. If $E[Y|Z=z] - E[Y|Z=z\prime] - \int_{p_l(z\prime)}^{p_u(z)} 2b|v-v^*|dv \geq 0$ then the $MTE(v^*)$ is positive. If $E[Y|Z=z] - E[Y|Z=z\prime] + \int_{p_l(z\prime)}^{p_u(z)} 2b|v-v^*|dv \leq 0$ then the $MTE(v^*)$ is negative.

If $E[Y|Z=z] - E[Y|Z=z\prime] - \int_{p_l(z\prime)}^{p_u(z)} 2b|v-v^*|dv \geq 0$ then from equations 2.11 and 2.12 combined with the bounds on $p(z) - p(z\prime)$ we get:

$$MTE^+(v^*)_{lb} = \frac{E[Y|Z=z] - E[Y|Z=z\prime] - \int_{p_l(z\prime)}^{p_u(z)} 2b|v-v^*|dv}{\Delta_{pu}}$$

$$MTE^+(v^*)_{ub} = \frac{E[Y|Z=z] + E[Y|Z=z\prime] + \int_{p_l(z\prime)}^{p_u(z)} 2b|v-v^*|dv}{\Delta_{pl}}$$

Note that the form of the bounds depend on $|v-v^*|$. In the integral of the absolute value is where the relative position of v^* with respect of $p_l(z\prime), p_u(z)$ will matter. Note that if the v^* of interest is such that $v^* \leq p_l(z\prime)$, the integral involving $|v-v^*|$ is $[\frac{p_u(z)^2 - p_l(z\prime)^2}{2} + v^*(p_l(z\prime) - p_u(z))]$. If the v^* of interest is such that $v^* \geq p_u(z)$ then $[\frac{-p_u(z)^2 + p_l(z\prime)^2}{2} + v^*(-p_l(z\prime) + p_u(z))]$. If the v^* of interest is such that $p_l(z\prime) \leq v^* \leq p_u(z)$ then $[v^{*2} - v^*(p_u(z) + p_l(z\prime)) + (\frac{p_u(z)^2 + p_l(z\prime)^2}{2})]$

The following theorem uses assumption 5 to get nonparametric bounds on the MTE with a binary instrument with $p(z) > p(z\prime)$ and summarizes the previous discussion.

Theorem 2. *If assumptions 1-2 and 5 holds. Then the following bounds are valid:*

1. *If* $E[Y|Z=z] - E[Y|Z=z\prime] - \int_{p_l(z\prime)}^{p_u(z)} 2b|v-v^*|dv \geq 0$*:*

$$MTE^+(v^*)_{lb} = \frac{E[Y|Z=z] - E[Y|Z=z\prime] - \int_{p_l(z\prime)}^{p_u(z)} 2b|v-v^*|dv}{\Delta_{pu}}$$

$$MTE^+(v^*)_{ub} = \frac{E[Y|Z=z] + E[Y|Z=z\prime] + \int_{p_l(z\prime)}^{p_u(z)} 2b|v-v^*|dv}{\Delta_{pl}}$$

2. *If* $E[Y|Z=z] - E[Y|Z=z\prime] + \int_{p_l(z\prime)}^{p_u(z)} 2b|v-v^*|dv \leq 0$*:*

$$MTE^-(v^*)_{lb} = \frac{E[Y|Z=z] - E[Y|Z=z\prime] - \int_{p_l(z\prime)}^{p_u(z)} 2b|v-v^*|dv}{\Delta_{pl}}$$

$$MTE^-(v^*)_{ub} = \frac{E[Y|Z=z] + E[Y|Z=z\prime] + \int_{p_l(z\prime)}^{p_u(z)} 2b|v-v^*|dv}{\Delta_{pu}}$$

3. Otherwise:

$$MTE(v^*)_{lb} = \max\{MTE^-(v^*)_{lb}, MTE^+(v^*)_{lb}\}$$
$$MTE(v^*)_{ub} = \min\{MTE^-(v^*)_{ub}, MTE^+(v^*)_{ub}\}$$

Remark 4. *In some situations like in the case of SNAP, one could be willing to assume that for every level of heterogeneity v, $P(Y_1 < Y_0|V = v) = 1$ holds which means that receiving SNAP is not making anyone more food insecure. This is the "treatment cannot hurt" assumption. In such a case, we would be imposing the sign of the MTE even if we cannot extract it from*
$E[Y|Z = z] - E[Y|Z = z\prime] - \int_{p_l(z\prime)}^{p_u(z)} 2b|v - v^*|dv \geq 0$ *or*
$E[Y|Z = z] - E[Y|Z = z\prime] + \int_{p_l(z\prime)}^{p_u(z)} 2b|v - v^*|dv \leq 0$

Remark 5. *The previous bounds are not assuming there is a known support for Y if the nature of Y is bounded then the previous bounds change in the following way:*

$$\tilde{MTE}(v^*)_{lb} = \max\{MTE(v^*)_{lb}, Y_{1l} - Y_{0u}\}$$
$$\tilde{MTE}(v^*)_{ub} = \min\{MTE(v^*)_{ub}, Y_{1u} - Y_{0u}\}$$

A researcher might be interested in combining the monotonicity assumption on the treatment responses and the smoothness assumption (for monotonicity of the MTE directly and smoothness, see section 3.8). So instead of using assumption 5, the researcher might be willing to use 6. This result is collected in appendix 2.13 under theorem 3.

The choice of b is not arbitrary. The fact that it operates as a tuning parameter might lead to a discretionary use of it to get the desired result; different ways of choosing this parameter are discussed in appendix 2.14.

2.5 Illustration of the results

To illustrate the results from the previous section, we build the following DGP motivated by the application to SNAP in section 3.8. As stated by Kreider et al. (2012) the case of SNAP is sensitive to misreporting. It is more likely to observe people receiving SNAP and erroneously say

they are not receiving it, than people not receiving it saying that they do. This is consistent with setting $(1-D)\varepsilon = 0$, which means no one misreports receiving when they do not receive it. Consider the following DGP:

$$\begin{cases} Y &= DV + (1-D)\frac{V}{4} \\ D &= \mathbb{1}\{Z - V \geq 0\} \\ D^* &= D(1-\varepsilon) \\ \varepsilon &= \mathbb{1}\{V \leq 0.15\} \end{cases} \tag{2.13}$$

$V \sim U(0,1)$, Z takes values be 0.7 or 0.1 with probability 1/2, Z is independent of V. Note $|E[Y_1|V=v_1] - E[Y_1|V=v_2]| = |v_1 - v_2|$, $|E[Y_0|V=v_1] - E[Y_0|V=v_2]| = \frac{1}{4}|v_1 - v_2|$ so b from assumption 5 can be set to 1. Note $MTE(v) = \frac{3}{4}v$ and is thus monotonic. Similarly the marginal treatment responses $E[Y_1|V=v] = v, E[Y_0|V=v] = \frac{1}{4}v$ are monotonic in v. $\alpha = P(\varepsilon = 1) = 0.15$.

The following figure illustrates the different bounds for this particular DGP.

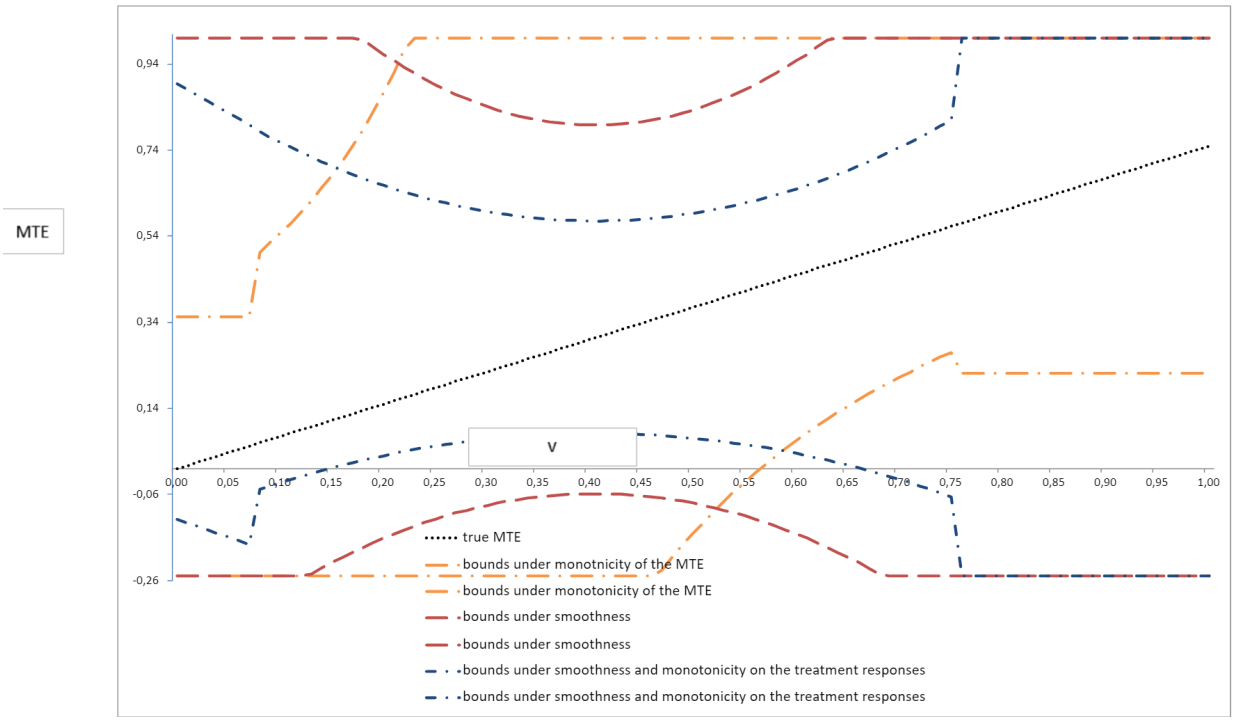

Figure 2.1 Results from the DGP

The limit of the Y axis is the worst case upper bounds (1) and lower bounds (-0.25). The black dotted line is the true MTE curve. In orange, we show the bounds from the monotonicity assumptions on the MTE. Such bounds differ in some zones of the potential values of v^* (the X-axis) from the worst-case bounds. Additionally, they can recover the sign in certain regions. The red lines represent the bounds from using the smoothness assumption alone. We can see that even though it improves from the worst-case bounds in this particular DGP, it cannot recover the sign of the MTE. Finally, the blue lines represent the combination of monotonicity of the treatment responses and smoothness. In such cases improvements over only smoothness are achieved, and the sign of the MTE is recovered in certain regions. We can see that there is no clear dominance of the monotonicity bounds versus the smoothness bounds. In different regions, they develop tighter bounds.

In general, we can see that the bounds improve over the worst-case bounds and have identifying power on the MTE.

Note that the point from Remark 3 is illustrated at $v^* = 0.75$.

Before moving to the following section is in place, it is worth noticing that in this particular DGP, there is no clear dominance from monotonicity over smoothness, but both are valid. Nevertheless, the researcher should carefully think about why either of these holds to choose the route to identification. There might be situations where smoothness is more credible than monotonicity and vice-versa.

2.6 Application: MTE of SNAP on child health when participation is endogenous and misreported

In this section, the developed methods are applied to get bounds on the MTE and ATE of receiving SNAP on the outcome of being food insecure. As stated by Kreider et al. (2012) the Supplemental Nutrition Assistance Program (SNAP), formerly known as the Food Stamp Program, is by far the largest food assistance program in the United States and, as such, constitutes a crucial component of the social safety net in the United States. In any given month

during 2009, SNAP assisted more than 15 million children, and it is estimated that nearly one in two American children will receive assistance during their childhood. Concluding about the program's impact is complex due to two of the fundamental problems studied in this paper. First, a selection problem arises because the decision to participate in SNAP is unlikely to be exogenous. On the contrary, unobserved factors such as expected future health status, parents' human capital characteristics, financial stability, and attitudes towards work and family are all thought to be jointly related to participation in the program and health outcomes such as food security. Families may decide to participate precisely because they expect to be food insecure or in poor health. Second, a nonrandom measurement error problem arises because a large fraction of food stamp recipients fails to correctly report their program participation in household surveys. Using administrative data matched with data from the Survey of Income and Program Participation (SIPP), for example, Bollinger and David (1997) find that errors in self-reported receipt of food stamps exceed 12 percent and are related to respondents' characteristics, including their true participation status, health outcomes, and demographic attributes. Meyer et al. (2009) provide evidence of extensive underreporting of food stamps in the SIPP, the Current Population Survey (CPS), and the Panel Study of Income Dynamics (PSID).

In this context Kreider et al. (2012) studies the average effects using the December Supplement of the 2003 Current Population Survey (CPS). They claim that the CPS has been widely applied to evaluate the association between SNAP and food insecurity and is used by the U.S. Department of Agriculture (USDA) to establish the official food insecurity rates for the United States. In the data, we can observe a self-reported measure of food stamp receipt over the past year, food insecurity over the past year, and the ratio of income to the poverty line[6].

As Kreider et al. (2012) states, the data is rich enough to allow the construction of instrumental variables for SNAP participation used in previous literature. In particular, state identifiers in the CPS apply a more traditional instrumental variable (IV) assumption based on

[6]For further details about the data see Gundersen and Kreider (2008). In there, they state that just over 40 percent of the households report receiving food stamps, and the food insecurity rate among self-reported recipients is 17.9 percentage points higher than among eligible non-recipients (52.3 percent vs. 34.4).

cross-state variation in program eligibility rules. In general terms, program eligibility rules are income requirements (most households must meet both gross and net income limits to qualify for SNAP benefits), resource requirements (households must also meet a resource limit in their bank accounts), work requirements (If you are an able-bodied adult without dependents, between the ages of 18 and 49, and able to work but currently unemployed, you may only be eligible for SNAP benefits for three months within a three-year period) and other eligibility requirements (to be eligible for SNAP benefits, households must also, meet other conditions in addition to the income and resource requirements, such as everyone in your household having, or have applied for, a social security number). To establish the income and resource requirements, each state computes asset tests, but there is variation in how these states evaluate the assets of individuals. More specifically, they may or may not include certain assets that will affect the individuals' eligibility. Merging the Urban Institute's database of state program rules with the CPS data Kreider et al. (2012) create two instrumental variables: an indicator for whether the state uses a simplified semi-annual reporting requirement for earnings and an indicator for whether cars are exempt from the asset test.[7] These instrumental variables are assumed to be valid and independent allows using the current methods to bound the MTE. For example, if one is willing to assume that the state variation in the asset test is exogenous, then instrument independence is satisfied. It is worth noticing that only $LATE$ can be identified from an instrumental variable regression under individual heterogeneity. In this sense, the methods developed here can be used to recover bounds on the average treatment effect, and complement Kreider et al. (2012) results. In this context, the MTE at a particular level of v^* represents the treatment effect receiving SNAP has for a particular level of stigma. Stigma mat is connected with the potential outcomes since, as stated by Earnshaw and Karpyn (2020) stigma manifestations affect health outcomes (and food security as such).

[7]For more details on the construction see Kreider et al. (2012).

More precisely, in this context to illustrate our methods, Y is a food insecurity indicator over the past year, D^* is the self-reported SNAP participation (subject to potential measurement error as stated before), Z is a binary variable for cars exempt from the asset test.

Concerning choosing b for the sake of exposition, we present the results here for several potential values of b. The particular choice of b for the case of SNAP is an open question that is not addressed in this work. In appendix 2.14 we discuss different methods on how to choose b which could be used in this context. Intuitively choosing b is restricting the degree of smoothness the MTE would have. One can draw a parallelism between choosing a linear form (a low b) for the MTE versus choosing a high order polynomial (high b). The linear form is rather restrictive on the behavior of higher-order derivatives (and thus smoothness) compared with the polynomial. A researcher choosing b for SNAP should consider what he thinks the underlying decision-maker optimization problem looks like. If the expected utility, for example, has a quadratic form, or the optimal expected demand of food security is linear under both receiving and not receiving SNAP, then the expected benefit on the optimal choice of food security both under getting SNAP and not getting it could be considered linear.

Consistently with Kreider et al. (2012) a treatment cannot hurt assumption ($P(Y_1 < Y_0|V = v) = 1$) will be introduced. Making then the conservative upper bounds of the MTE to be 0.

The following table summarizes the data and shows the average and median characteristics in the sample.

Table 2.1 Summary statistics

	Mean	Standard Deviation	Median
Food insecure (Y)	0.42	0.49	0
Reporting being on SNAP (D)	0.41	0.49	0
Cars exempt from the asset test (Z)	0.30	0.46	0
Income to poverty line ratio	0.75	0.36	0.75
N	2707	-	-

We can see that around forty-two percent of the people are food insecure, while forty-one report receiving SNAP. Thirty percent have their cars excluded from the asset test, which implies that thirty percent of the individuals in the sample live in states where cars are excluded from the asset test. On average (and in the median), individuals in this sample are below the poverty line.

Each of the graphs in the following figure is computed in the following way. For any given level of b and α, point estimates of the bounds are constructed for the MTE at different values v^* using their sample analogs. To decide which type of bound to use, the relative position of the sample analog estimates of the bounds for the propensity score is calculated. The maximum level of α is twenty percent which is chosen as an arbitrary big upper bound above the existing results from Kreider et al. (2012).

The following figures summarize the results.

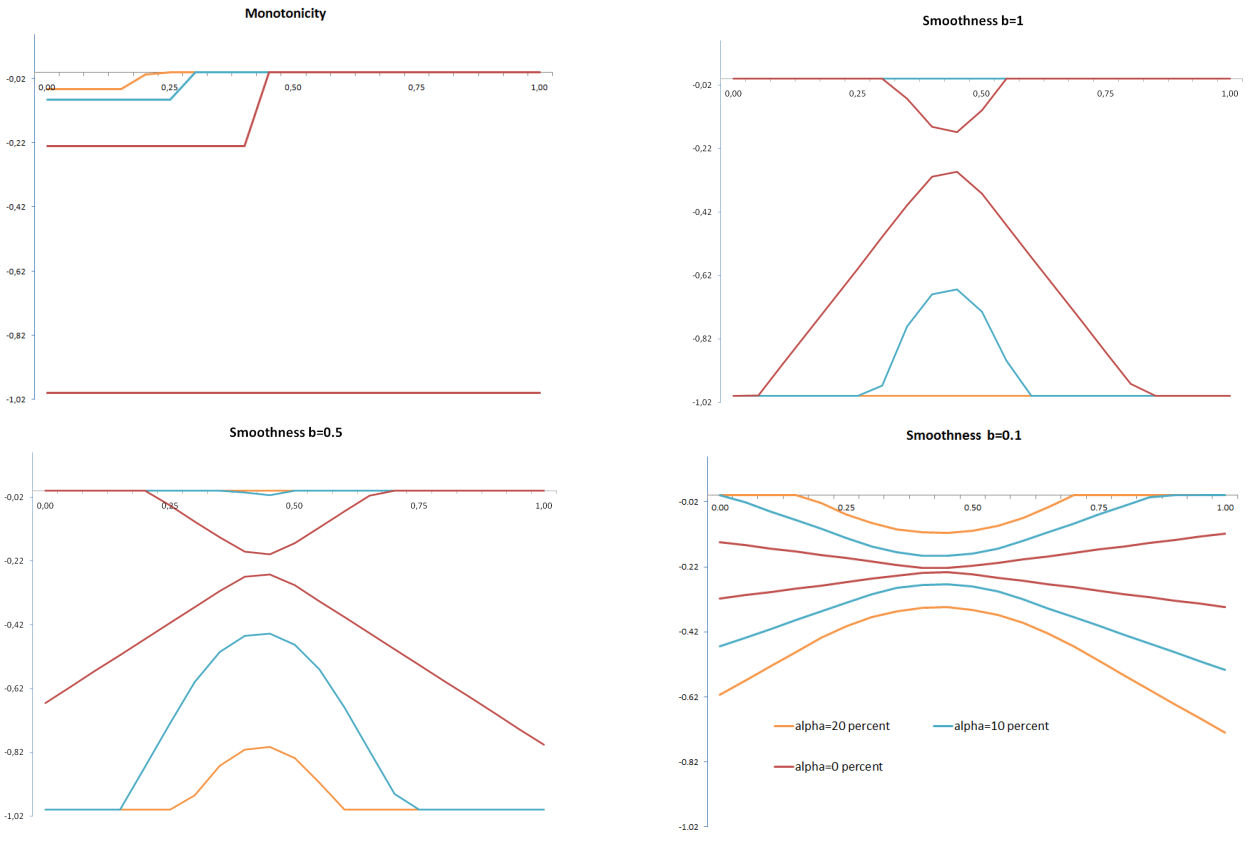

Figure 2.2 Bounds for the MTE (Y axis is the MTE and X axis is the value of v^*)

In orange, we have the upper and lower bounds assuming $\alpha = 0.2$. In blue, we have the upper and lower bounds assuming $\alpha = 0.1$, and in red, we report the upper and lower bounds assuming $\alpha = 0$ (no misreporting). The limits of the Y-axis are the worst-case upper bound (0 under treatment cannot hurt) and the worst-case lower bound (-1). The X-axis goes from 0 to 1, the different potential values of v^*. We can see that the bounds have identification power over different regions of v^* support. We can see that when the level of misreporting decreases, the bounds become tighter since there is less lack of identification due to misreporting. Similarly, when b decreases, we also get tighter bounds consistent with reducing the potential functional forms of the marginal treatment responses.

These bounds are computed by fixing a level of α. If, for example, in the case of $b = 0.1$, the researcher is not interested in $\alpha = 0.2$ rather in $\alpha \leq 0.2$ then all the region between the upper orange curve and the lower orange curve is the identified set consistent with $b = 0.1$ and all the α's less or equal to twenty percent.

We can combine the monotonicity of the MTE with the smoothness of the marginal treatment responses the researcher can take as the upper bound, the minimum of both types of bounds, and as the lower bound, the maximum of both. In addition, the consideration of Remark 3 can be taken into account. The following figure summarizes this consideration for the case of no misreporting and $b = 0.5$.

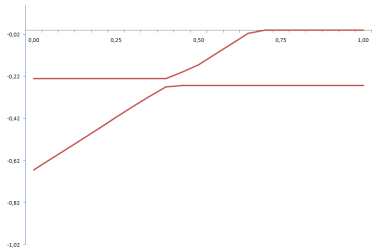

Figure 2.3 Bounds for the MTE combining monotonicity of the MTE with smoothness of the marginal treatment responses (Y axis is the MTE and X axis is the value of v^*)

The figure shows the identified region of the MTE. We can see that the marginal returns to receiving SNAP on food insecurity are increasing for higher levels of stigma. Individuals with higher psychological costs of receiving SNAP would be the ones that benefit the most from taking it given that this higher stigma cost is correlated with poorer health outcomes consistent with Earnshaw and Karpyn (2020), thus SNAP would highly benefit these individuals. The length of the identified set becomes tighter in the region between the estimated observed propensity scores ($\widehat{P}[D^* = 1|Z = 1] = 0.49, \widehat{P}[D^* = 1|Z = 0] = 0.38$) since there is more information being used to bound the MTE.

The previous display also implies a simple way of computing bounds of a parameter of interest such as the ATE. It is known that $ATE = \int_0^1 MTE(v^*)dv^*$. Then we know that $\int_0^1 MTE_{lb}(v^*)dv^* \leq ATE \leq \int_0^1 MTE_{ub}(v^*)dv^*$. So then we can approximate bounds for the ATE as:

$$\frac{1}{N_{v^*}} \sum_{v^*} \widehat{MTE}_{lb}(v^*) \equiv \widehat{ATE}_{lb}$$
$$\frac{1}{N_{v^*}} \sum_{v^*} \widehat{MTE}_{ub}(v^*) \equiv \widehat{ATE}_{ub}$$

Where N_{v^*} is the number of grid points where $MTE(v^*)$ was evaluated and where $\widehat{MTE}$ are the bounds from the previous graphs. Table 2.2 on appendix 2.15 computes the estimates of the bounds on the ATE for the sake of illustration. In the appendix, there is also an alternative way of estimating (and doing inference) on the ATE for the case of smoothness restrictions. In section 2.10 a method for asymptotic normality for an outer-set of the ATE in the case of no misreporting and with smoothness conditions is developed. In section 2.11 a method for asymptotic normality for an outer-set of the ATE in the case of misreporting, treatment cannot hurt assumption, and smoothness conditions can be found.

2.7 Conclusions

In this paper, we provided partial identification results for the Marginal Treatment Effect in the presence of measurement error and a discrete instrument. To do so, given the discrete nature

of the instruments, we introduced different ways of recovering the MTE via nonparametric restrictions building over Mogstad et al. (2018), Brinch et al. (2017) and Acerenza et al. (2021).

Results are illustrated via a numerical example and quantifying the marginal treatment effect of SNAP on food insecurity, a case in which measurement error and endogeneity of treatment are known to be an issue. Results suggest that for most levels of heterogeneity, the treatment reduces the chances of being food insecure. Implying that for different levels of stigma, receiving SNAP reduces food insecurity. Suppose the researcher is willing to assume monotonicity between the stigma cost and the MTE. In that case, evidence suggests that individuals with higher stigma costs (and thus less likely to both take SNAP and report receiving it) are the ones that will benefit the most.

In a more general way, our results can serve as a sensitivity analysis tool for when researchers suspect measurement error and are interested in recovering the MTE.

If no measurement error exists, the results from this paper provided analytical partial identification results of the MTE in the presence of discrete (binary) instruments that can serve as a complement for the already existing results.

2.8 References

Acerenza, S., K. Ban and D. Kédagni. 2021. "Marginal Treatment effects with misclassified treatment." *Working paper*.

Aigner, D. J.. 1973. "Regression with a binary independent variable subject to errors of observation." *Journal of Econometrics* 1:49-59.

Armstrong, T and M. Kolesár. 2020. "Simple and honest confidence intervals in nonparametric regression." *Quantitative Economics*:1-39.

Balakrishnan, N. and C. D. Lai. 2009. "Continuous Bivariate Distributions." *Springer-Verlag New York*.

Bollinger, C., and M. David. 1997. "Modeling Discrete Choice with Response Error: Food Stamp Participation." *Journal of the American Statistical Association* 92(439): 827-835.

Brinch, C.N. , M. Mogstad and M. Wiswall. 2017. "Beyond $LATE$ with a Discrete Instrument." *Journal of Political Economy* 125(4): 985-1039.

Carneiro, P., J. J. Heckman and E. Vytlacil. 2011. "Estimating marginal returns to education." *American Economic Review* 101(6):2754–2781.

Chalak, K. 2017. "Instrumental Variables methods with heterogeneity and mismeasured instruments." *Econometric Theory* 33:69–104.

Chesher, A. 2005. "Nonparametric Identification under Discrete Variation." *Econometrica* 73(5):1525–1550.

Contini, D. and A. M. Richiardi. 2012. "Reconsidering the effect of welfare stigma on unemployment." *Journal of Economic Behavior and Organization* 84(2):224-244.

DiTraglia, F. J. and C. García-Jimeno. 2019. "Identifying the effect of a mis-classified, binary, endogenous regressor." *Journal of Econometrics* 209: 376-390.

Earnshaw, V. and A. Karpyn. 2020. "Understanding stigma and food inequity: a conceptual framework to inform research, intervention, and policy." *Translational Behavioral Medicine* 10(6): 1350–1357.

Gundersen, C., and. B. Kreider. 2008. "Food Stamps and Food Insecurity: What Can Be Learned in the Presence of Nonclassical Measurement Error?" *Journal of Human Resources* 43(2): 352-382.

Haider, S. and Stephens M. 2020. "Correcting for Misclassified Binary Regressors Using Instrumental Variables." *NBER Working paper series* Working Paper 27797.

Hausman, J.A., Abrevaya, J. and Scott-Morton, F.M. 1998. "Misclassification of the dependent variable in a discrete-response setting." *Journal of Econometrics* 87:239–269.

Heckman, J.J. and E. Vytlacil. 1999. "Local Instrumental variables and latent variable models for identifying and bounding treatment effects." *Proceedings of the National Academy of Sciences* 96:4730–4734.

Heckman, J.J. and E. Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica* 73(3):669–738.

Heckman, J.J., S. Urzua and E. Vytlacil. 2006. "Understanding Instrumental Variables in models with essential heterogeneity." *The Review of Economics and Statistics* 88(3):389–432.

Henry, M., Y. Kitamura and B. Salnié. 2014. "Partial Identification of finite mixtures in econometric models." *Quantitative Economics* 5:123–144.

Hernandez, M. and S. Pudney . 2007. "Measurement error in models of welfare participation." *Journal of Public Economics* 91:327–341.

Hu, Y. 2008. "Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution." *Journal of Econometrics* 144:27–61.

Imai, K. and T. Yamamoto. 2010. "Causal inference with differential measurement error: Nonparametric identification and sensitivity analysis." *American Journal of Political Science* 54:543–560.

Jun, S. J., J. Pinkse and H. Xu. 2011. "Tighter bounds in triangular systems." *Journal of Econometrics* 161(2):122–128.

Kédagni, D. and I. Mourifié . 2014. "Tightening bounds in triangular systems." *Economics Letters* 125(3):455–458.

Kédagni, D. 2019. "Identification of Treatment Effects with Mismeasured Imperfect Instruments." *Working paper* Number 19009, Iowa State University.

Kim, W., K. Kwon, S. Kwon and S. Lee. 2018. "The identification power of smoothness assumptions in models with counterfactual outcomes." *Quantitative Economics* 9, 617–642.

Kreider, B., J.V. Pepper, C. Gundersen and D. Jolliffe. 2012. "Identifying the Effects of SNAP(Food Stamps) on Child Health Outcomes When Participation is Endogenous and Misreported." *Journal of the American Statistical Association* 107:432–441.

Kreider, B. and J.V. Pepper. 2007. "Disability and employment: Reevaluating the evidence in light of reporting errors." *Journal of the American Statistical Association* 102:432–441.

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

Lewbel, A. 2007. "Estimation of Average treatment effects with misclasification." *Econometrica* 75(2):537–551.

Machado, C., A.M. Shaikh and E.J. Vytlacil. 2019. "Instrumental variables and the sign of the average treatment effect." *Journal of Econometrics* 212(2):522–555.

Mahajan, A. 2006. "Identification and Estimation of regression models with misclassification." *Econometrica* 74(3):631–665.

Meyer, B., D.W. Mok and J.X. Sullivan. 2006. "The Under-Reporting of Transfers in Household Surveys: Its Nature and Consequences." *Working Paper* University of Chicago, Harris School of Public Policy Studies.

Moffitt, R. 1983. "An Economic Model of Welfare Stigma." *The American Economic Review* 73(5):1023–1035.

Mogstad, M., A. Santos and A. Torgovitsky. 2018. "Using Instrumental Variables For Inference About Policy Relevant Treatment Parameters." *Econometrica* 86(5):1589–1619.

Molinari, F. 2008. "Partial identification of probability distributions with misclassified data." *Journal of Econometrics* 144:81–117.

Molinari, F. 2008. "Partial identification of probability distributions with misclassified data." *Journal of Econometrics* 144:81–117.

Palar, K., Frongillo, E. A., Escobar, J., Sheira, L. A., Wilson, T. E., Adedimeji, A., Merenstein, D., Cohen, M. H., Wentz, E. L., Adimora, A. A., Ofotokun, I., Metsch, L., Tien, P. C., Turan, J. M., and Weiser, S. D. 2018. "Food Insecurity, Internalized Stigma, and Depressive Symptoms Among Women Living with HIV in the United States." *AIDS and behavior* 22(12):3869–3878.

Possebom, V. 2021. "Crime and Mismeasured Punishment: Marginal Treatment Effect with Misclassification." *Working Paper.*

Tommasi, D and L. Zhang. 2020. "Bounding Program Benefits When Participation Is Misreported." *IZA IZA DP No. 13430*

Ura, T. 2018. "Heterogeneous treatment effects with mismeasured endogenous treatment." *Quantitative Economics* 9(3):1335–1370.

2.9 Appendix A: Identification without shape restrictions

The object of interest as noted is $E[Y_1 - Y_0|V = v^*]$. Following Mogstad et al. (2018) this can be expressed as:

$$E[Y_1|V = v^*] - E[Y_0|V = v^*] = \int_0^1 E[Y_1|V = v]w_1 dv + \int_0^1 E[Y_0|V = v]w_0 dv \tag{2.14}$$

Where $w_1 = \delta(v^*)$, $w_0 = -\delta(v^*)$ and $\delta(v^*)$ is Dirac delta measure assigning all the mass at $V = v^*$.

As noted in section 2.3 we can express $E[Y|Z = z] - E[Y|Z = z\prime]$ our IV-like estimand as:

$$\begin{aligned} E[Y|Z = z] - E[Y|Z = z\prime] &= \int_{p(z\prime)}^{p(z)} E[Y_1|V = v]dv - \int_{p(z\prime)}^{p(z)} E[Y_0|V = v]dv \\ &= \int_0^1 E[Y_1|V = v]1\{v \in (p(z\prime), p(z))\}dv \\ &+ \int_0^1 -E[Y_0|V = v]1\{v \in (p(z\prime), p(z))\}dv \\ &\equiv \int_0^1 E[Y_1|V = v]\omega_1 dv + \int_0^1 E[Y_0|V = v]\omega_0 dv \end{aligned} \tag{2.15}$$

Where $\omega_1 = 1\{v \in (p(z\prime), p(z))\}$ and $\omega_0 = -1\{v \in (p(z\prime), p(z))\}$.

As stated in section 2.3, $1\{v \in (p(z\prime), p(z))\}$ is not known since $p(z)$ is not known. Nevertheless, for any fixed $p(z), p(z\prime)$, for $E[Y_d|V = v] \in \mathcal{M}$ for $d = 1, 0$ and $\mathcal{M}$ being a convex space we known from Mogstad et al. (2018) that since the equations 2.14 and 2.15 define linear operators then convexity is carried onto the space of solutions of equation 2.14 subject to 2.15. Then this allows to define a linear programming as in Mogstad et al. (2018) and take into account the implementation considerations they make to get upper bounds and lower bound for the MTE by solving respectively:

$$\max(\min)_{E[Y_1|V=v], E[Y_0|V=v] \in \mathcal{M}} \quad \int_0^1 E[Y_1|V = v]w_1 dv + \int_0^1 E[Y_0|V = v]w_0 dv \tag{2.16}$$

Subject to

$$E[Y|Z = z] - E[Y|Z = z\prime] = \int_0^1 E[Y_1|V = v]\omega_1(p)dv + \int_0^1 E[Y_0|V = v]\omega_0(p)dv$$

Which then give as a solution an interval defined as $MTE_{lb}(v^*, p), MTE_{ub}(v^*, p)$. Where $\omega_d(p), MTE_{lb}(., p), MTE_{ub}(., p)$ is stating the dependence of the program to a particular $p(z), p(z\prime)$.

The following procedure can be repeated for every $p(z), p(z\prime) \in \mathcal{P}$, the identification region of the propensity score ($\mathcal{P}$ defined in section 2.12) and then the set of possible values for the MTE is $\bigcup_{p\in\mathcal{P}} \left(MTE_{lb}(v^*, p), MTE_{ub}(v^*, p)\right)$. In practice the calculation cannot be made for every p since it is infinite-dimensional, but the solution can be approximated taking several grid points in the space $\mathcal{P}$.

2.10 Inference for the ATE with no misreporting and smoothness conditions

Let $\Delta_Y \equiv E[Y|Z = z] - E[Y|Z = z\prime]$. Let $p_1 \equiv p(z)$ and $p_0 \equiv p(z\prime)$. We can use the upper bounds on the MTE to construct upper bounds on the ATE (a similar display would apply for the lower bounds).

Note that from the bounds developed under smoothness conditions it is true that the following are upper bounds for the $MTE(v^*)$ (namely $MTE(v^*)_{ub}$):

$$MTE(v^*)_{ub1} \equiv \frac{\Delta_y + 2b[\frac{p_1^2 - p_0^2}{2} + v^*(p_0 - p_1)]}{p_1 - p_0} \quad \text{if} \quad v^* < p_0 \tag{2.17}$$

$$MTE(v^*)_{ub2} \equiv \frac{\Delta_y + 2b[v^{*2} + \frac{p_1^2 + p_0^2}{2} - v^*(p_0 + p_1)]}{p_1 - p_0} \quad \text{if} \quad p_0 < v^* < p_1 \tag{2.18}$$

$$MTE(v^*)_{ub3} \equiv \frac{\Delta_y + 2b[\frac{-p_1^2 + p_0^2}{2} + v^*(-p_0 + p_1)]}{p_1 - p_0} \quad \text{if} \quad p_1 < v^* \tag{2.19}$$

Then from the fact that $ATE = \int_0^1 MTE(v^*)dv^*$ we can see that:

$$\begin{aligned} ATE_{ub} &= \int_0^1 MTE(v^*)_{ub} \\ &= \int_0^{p_0} MTE(v^*)_{ub1} + \int_{p_0}^{p_1} MTE(v^*)_{ub2} + \int_{p_1}^1 MTE(v^*)_{ub3} \end{aligned} \tag{2.20}$$

Which then combining equations 2.17-2.20 and after some calculus we get:

$$ATE_{ub} = \frac{\Delta_y}{p_1 - p_0} + \frac{2b}{3}\frac{p_1^3 - p_0^3}{p_1 - p_0} - b\frac{p_1^2 - p_0^2}{p_1 - p_0} + b \tag{2.21}$$

Which is a smooth continuous function of p_1, p_0, Δ_y (except at $p_1 = p_0$ which is ruled out by assumption) then if the estimators of p_1, p_0, Δ_y are asymptotically normal (which is the case under standard conditions since they are sample analogs) we get by the continuous mapping theorem that the estimator of ATE_{ub} is also asymptotically normal. Then we can perform valid asymptotic inference on the bounds on the ATE. The bound is an outer set because, as pointed out in the main document, if the variables Y_1, Y_0 are naturally bounded then, so it is the ATE, the bounds here do not incorporate that aspect.

2.11 Appendix B: Inference for the ATE with smoothness conditions and treatment cannot hurt assumption

As in section 2.10, let $\Delta_Y \equiv E[Y|Z=z] - E[Y|Z=z\prime]$. Now as there is misreporting let $p_{1u} \equiv p_u(z), p_{1l} \equiv p_l(z), p_{0u} \equiv p_u(z\prime)$ and $p_{0l} \equiv p_l(z\prime)$.

Also let the lower bound of the difference of the probabilities as in the main text to be Δ_{pl}. In this case adding the assumption that $Y_1 - Y_0 \leq 0$ we are imposing an upper bound on the MTE and ATE to be 0. We are also imposing information on the sign of it which leads to the following lower bounds for the $MTE(v^*)$

$$MTE(v^*)_{lb1} \equiv \frac{\Delta_y - 2b[\frac{p_{1u}^2 - p_{0l}^2}{2} + v^*(p_{0l} - p_{1u})]}{\Delta_{pl}} \quad \text{if} \quad v^* < p_{0l} \tag{2.22}$$

$$MTE(v^*)_{lb2} \equiv \frac{\Delta_y - 2b[v^{*2} + \frac{p_{1u}^2 + p_{0l}^2}{2} - v^*(p_{0l} + p_{1u})]}{\Delta_{pl}} \quad \text{if} \quad p_{0l} < v^* < p_{1u} \tag{2.23}$$

$$MTE(v^*)_{lb3} \equiv \frac{\Delta_y - 2b[\frac{-p_{1u}^2 + p_{0l}^2}{2} + v^*(-p_{0l} + p_{1u})]}{\Delta_{pl}} \quad \text{if} \quad p_{1u} < v^* \tag{2.24}$$

Then from a similar display as in section 2.10 we get:

$$ATE_{lb} = \frac{\Delta_y - \frac{2b}{3}(p_{1u}^3 - p_{0l}^3) + b(p_{1u}^2 - p_{0l}^2) - b(p_{1u} - p_{0l})}{\Delta_{pl}} \tag{2.25}$$

We know that Δ_y, Δ_{pl} can be estimated with the sample analogs, and they are well-behaved estimators that, under standard central limit theory, are asymptotically normal. Note that from the main text $p_{1u} \equiv \min\{P(D^*=1|Z=z) + \alpha, (1-\alpha) + P(D^*=0|Z=z)\}, p_{0l} \equiv \max\{P(D^*=1|Z=z\prime) - \alpha, \alpha - P(D^*=1|Z=z\prime)\}$ where the max, min operators make the asymptotic normality of their sample analogs not possible. But note that $p_{1u} \leq P(D^*=1|Z=z) + \alpha$ and $p_{0l} \geq P(D^*=1|Z=z\prime) - \alpha$. So if the researcher is willing to assume he is using levels of α that are such that $P(D^*=1|Z=z) + \alpha \leq 1$ and $P(D^*=1|Z=z\prime) - \alpha \geq 0$ and also theirs sample analogs, then, he can use $P(D^*=1|Z=z) + \alpha, P(D^*=1|Z=z\prime) - \alpha$ instead of p_{1u}, p_{0l} as the bounds on the probabilities.

In such a case, the sample analogs of these outer bounds are asymptotically normal by the usual central limit theory.[8] In this case then, since the bound on the ATE is a smooth continuous function of $p_{1u}, p_{0l}, \Delta_y, \Delta_{pl}$ and since the sample analogs of $P(D^* = 1|Z = z) + \alpha, P(D^* = 1|Z = z\prime) - \alpha, \Delta_y, \Delta_{pl}$ are asymptotically normal, by standard results we get by the continuous mapping theorem that the estimator of ATE_{lb} is also asymptotically normal. Then we can perform valid asymptotic inference on the bounds on the ATE. This bound is an outer set because, as pointed out in Section 2.10 and also because we are not using the tightest possible bounds on $p(z)$.

2.12 Appendix C: Identification of $p(z)$

As in Acerenza et al. (2021) from the use of assumption 1 we know that the true propensity score is $P(D = 1|Z = z) = p(z)$ where the fact that V is uniform has been used. It is also true by assumption 1 and the structural form of D that an index sufficiency condition holds, thus $P(D = 1|Z = z) = P(D = 1|p(Z) = p(z))$. Note then that from the definitions of probability and the fact that V is uniform

$$p(z) = F_V(p(z)) = P(\varepsilon = 1)F_{V|\varepsilon=1}(p(z)) + P(\varepsilon = 0)F_{V|\varepsilon=0}(p(z)) \tag{2.26}$$

Where $F_{V|\varepsilon}$ is the conditional distribution of V given ε and $P(\varepsilon = 1)$ is the unconditional probability of misreporting which we will call α. The identification strategy will be developed conditional on α. In that sense, as stated earlier, if the researcher has an educated guess of α, he can "plug it in" to get the identification region we develop here. Alternatively, if the researcher believes in a range of , say, for example, all such α's that $\alpha < 0.2$, he can take the union of our bounds for all the α's he considers plausible.

Note now that from the definition of the model, the observed propensity score is such that:

$$P(D^* = 1|Z = z) = P(\varepsilon = 0, V \leq p(z)) + P(\varepsilon = 1, V > p(z)). \tag{2.27}$$

[8] Note that in the development of the bounds on the MTE with misreporting the particular form of the bounds for $p(z)$ was never used. In that sense, the previous results still hold just that now we change tighter bounds of $p(z)$ for wider ones.

Notice that the index sufficiency again implies that
$P\left(Y \in A, D^* = d^* | p(Z) = p(z)\right) = P\left(Y \in A, D^* = d^* | Z = z\right)$ for all z and d^*. This in turn implies that by the definition of joint probability that:

$$P(D^* = 1 | p(Z) = p(z)) = (1-\alpha) F_{V|\varepsilon=0}(p(z)) + \alpha(1 - F_{V|\varepsilon=1}(p(z))). \tag{2.28}$$

The following display on how to identify the difference between any two $p(z), p(z\prime)$ comes from Acerenza et al. (2021). For now, we assume $\alpha \in (0,1)$, since the cases where $\alpha \in \{0,1\}$ can be dealt with separately since no misreporting or complete misreporting do not provide an identification challenge. Combining equations 2.26 with 2.28, and solving for $F_{V|\varepsilon=0}(p)$ and $F_{V|\varepsilon=1}(p)$ in the system of equations, we obtain:

$$F_{V|\varepsilon=1}(p(z)) = \frac{p(z) + \alpha - P(D^* = 1 | p(Z) = p(z))}{2\alpha}, \tag{2.29}$$

$$F_{V|\varepsilon=0}(p(z)) = \frac{p(z) - \alpha + P(D^* = 1 | p(Z) = p(z))}{2(1-\alpha)}. \tag{2.30}$$

Therefore, the above functions need to satisfy all required conditions for a cumulative distribution on $[0,1]$: monotonicity, right-continuity, $F_{V|\varepsilon=1}(0) = F_{V|\varepsilon=0}(0) = 0$, and $F_{V|\varepsilon=1}(1) = F_{V|\varepsilon=0}(1) = 1$. Using the monotonicity condition of CDF's and the fact that the probabilities $F_{V|\varepsilon=1}(p(z))$ and $F_{V|\varepsilon=0}(P(z))$ lie between 0 and 1. For any z and z' such that $p(z') < p(z)$, we have:

$$0 \leq F_{V|\varepsilon=0}(p(z)) - F_{V|\varepsilon=0}(p(z')) \leq 1,$$

$$0 \leq F_{V|\varepsilon=1}(p(z)) - F_{V|\varepsilon=1}(p(z')) \leq 1.$$

which implies

$$0 \leq \frac{P(D^* = 1 | Z = z) - P(D = 1 | Z = z') + p(z) - p(z')}{2(1-\alpha)} \leq 1,$$

$$0 \leq \frac{p(z) - p(z') - P(D^* = 1 | Z = z) + P(D^* = 1 | Z = z')}{2\alpha} \leq 1.$$

This latter inequalities respectively imply

$$-P(D^*=1|Z=z)+P(D^*=1|Z=z') \leq p(z)-p(z') \tag{2.31}$$
$$\leq 2(1-\alpha)-P(D^*=1|Z=z)+P(D^*=1|Z=z'),$$
$$P(D^*=1|Z=z)-P(D^*=1|Z=z') \leq p(z)-p(z') \tag{2.32}$$
$$\leq 2\alpha+P(D^*=1|Z=z)-P(D^*=1|Z=z').$$

Defining $\Delta_{D^*Z}(z,z') \equiv P(D^*=1|Z=z)-P(D^*=1|Z=z')$ the previous bounds are for $0 \leq p(z') < p(z) \leq 1$.

$$\left|\Delta_{D^*Z}(z',z)\right| \leq p(z)-p(z') \leq \min\left\{1, 2\alpha+\Delta_{D^*Z}(z',z), 2(1-\alpha)-\Delta_{D^*Z}(z',z)\right\}.$$

Define for simplicity $\Delta_p \equiv p(z)-p(z')$, $\Delta_{pl} \equiv \Delta_{D^*Z}(z',z)$ and

$\Delta_{pu} \equiv \min\left\{1, 2\alpha+\Delta_{D^*Z}(z',z), 2(1-\alpha)-\Delta_{D^*Z}(z',z)\right\}$

From equations 2.31 and 2.32 and noticing that the properties of the model in the special case when $P(z\prime)=0$, we have[9]

$$-P(D^*=1|Z=z)+P(D^*=1|p(Z)=0) \leq p(z)$$
$$\leq 2(1-\alpha)-P(D^*=1|Z=z)+P(D^*=1|p(Z)=0),$$
$$P(D^*=1|Z=z)-P(D^*=1|p(Z)=0) \leq p(z)$$
$$\leq 2\alpha+P(D^*=1|Z=z)-P(D^*=1|p(Z)=0).$$

Using the condition that $F_{V|\varepsilon=1}(0)=0$, we identify $P(D^*=1|p(Z)=0)=\alpha$. Therefore, the above constraints on $p(z)$ become

$$\alpha-P(D^*=1|Z=z) \leq p(z) \leq 1-\alpha+P(D^*=0|Z=z),$$

$$P(D^*=1|Z=z)-\alpha \leq p(z) \leq \alpha+P(D^*=1|Z=z).$$

[9]Note that the display is using the fact that given the latent index structure of D when $p(z)=0$ given V is normalized to be uniform $P(D=1|p(z)=0)=0$. This result is relying on the structure of the model and it can be applied although there might not exist any observed $p(z^0)$ such that it exists in the data $P(D=1|p(z)=0)=0$ due to the discrete nature of the instruments.

A similar argument holds for the special case where $p(z) = 1$, but this yields the same constraints on $p(Z)$.

Given this display the following bounds which are point-wise sharp (see Acerenza et al. (2021)) hold:

$$p_l(z) \equiv \max\left\{P(D^* = 1|Z = z) - \alpha, \alpha - P(D^* = 1|Z = z)\right\},$$

$$p_u(z) \equiv \min\left\{P(D^* = 1|Z = z) + \alpha, (1 - \alpha) + P(D^* = 0|Z = z)\right\}.$$

Define the set of $p(z)$ satisfying these two constraints as $\mathcal{P}$.

Note that since $p(z)$ must lie between $(0, 1)$, the previous bounds are imposing restrictions on the level of misreporting the data supports. If there is an α such that $p_u(z) > 1$ or $p_l(z) < 0$ then it is a signal that that level of α is not sustained by the data.

To illustrate the previously mentioned notion about instead of having a guess for α having a guess of a set of values, imagine the researcher suspects two potential levels of misreporting α_1 and α_2. In this sense he could use the union of the bounds of both levels of misreporting to get the set for $p(z)$. This would look like:

$$\begin{aligned} p_l(z) = \min\Big(& \max\{P(D^* = 1|Z = z) - \alpha_1, \alpha_1 - P(D^* = 1|Z = z)\}, \\ & \max\{P(D^* = 1|Z = z) - \alpha_2, \alpha_2 - P(D^* = 1|Z = z)\}\Big) \\ p_u(z) = \max\Big(& \min\{\alpha_1 + P(D^* = 1|Z = z), (1 - \alpha) + P(D^* = 0|Z = z)\}, \\ & \min\{\alpha_2 + P(D^* = 1|Z = z), (1 - \alpha) + P(D^* = 0|Z = z)\}\Big) \end{aligned}$$

Instead of using assumption 5 the researcher might be willing to use 6.The next appendix focus on this aspect.

2.13 Appendix D: Identification with monotonicity assumption on the treatment responses and smoothness

Instead of using assumption 5 the researcher might be willing to use 6. In such a case note then that for some v^* between $p(z), p(z\prime)$:

$$\begin{aligned}
E[Y|Z=z] - E[Y|Z=z\prime] &= \int_{p(z\prime)}^{p(z)} E[Y_1|V=v] - E[Y_0|V=v]dv \\
&= \int_{p(z\prime)}^{p(z)} \Big(E[Y_1|V=v] - E[Y_1|V=v^*] - E[Y_0|V=v] \\
&+ E[Y_0|V=v^*] + E[Y_1|V=v^*] - E[Y_0|V=v^*] \Big) dv \\
&= [p(z) - p(z\prime)]MTE(v^*) \\
&+ \int_{p(z\prime)}^{p(z)} E[Y_1|V=v] - E[Y_1|V=v^*]dv + \int_{p(z\prime)}^{p(z)} E[Y_0|V=v^*] - E[Y_0|V=v]dv \\
&= [p(z) - p(z\prime)]MTE(v^*) \\
&+ \int_{p(z\prime)}^{v^*} E[Y_1|V=v] - E[Y_1|V=v^*]dv + \int_{v^*}^{p(z)} E[Y_1|V=v] - E[Y_1|V=v^*]dv \\
&+ \int_{p(z\prime)}^{v^*} E[Y_0|V=v^*] - E[Y_0|V=v]dv + \int_{v^*}^{p(z)} E[Y_0|V=v^*] - E[Y_0|V=v]dv
\end{aligned}$$

Note that between $p(z\prime), v^*$ every v is smaller than v^* such that then by assumption 6 for v^* bigger than v we have $0 \leq E[Y_1|V=v^*] - E[Y_1|V=v] \leq b(v^* - v)$, then $0 \geq -E[Y_1|V=v^*] + E[Y_1|V=v] \geq b(v^* - v)$ thus between $p(z\prime)$ and v^*, $\int_{p(z\prime)}^{v^*} E[Y_1|V=v] - E[Y_1|V=v^*]dv \leq 0$. Similarly between v^* and $p(z)$ we get $\int_{v^*}^{p(z)} E[Y_1|V=v] - E[Y_1|V=v^*]dv \leq \int_{v^*}^{p(z)} b(v - v^*)dv$. Also $\int_{p(z\prime)}^{v^*} E[Y_0|V=v^*] - E[Y_0|V=v]dv \leq \int_{p(z\prime)}^{v^*} b(v^* - v)dv$, $\int_{v^*}^{p(z)} E[Y_0|V=v^*] - E[Y_0|V=v]dv \leq 0$.

Then:

$$\begin{aligned}
E[Y|Z=z]-E[Y|Z=z\prime] &\leq [p(z)-p(z\prime)]MTE(v^*) \\
&+ \int_{v^*}^{p(z)} b(v-v^*)dv \\
&+ \int_{p(z\prime)}^{v^*} b(v^*-v)dv \\
&= [p(z)-p(z\prime)]MTE(v^*) + \int_{p(z\prime)}^{p(z)} b|v-v^*|dv \\
&\leq [p(z)-p(z\prime)]MTE(v^*) + \int_{p_l(z\prime)}^{p_u(z)} b|v-v^*|dv
\end{aligned}$$

Symmetrically,

$$E[Y|Z=z]-E[Y|Z=z\prime] \geq [p(z)-p(z\prime)]MTE(v^*) - \int_{p_l(z\prime)}^{p_u(z)} b|v-v^*|dv$$

Then by a similar display as in the discussion before theorem 2 we can get bounds based on no information about the sign, or based on the information contained in $E[Y|Z=z]-E[Y|Z=z\prime] \pm \int_{p_l(z\prime)}^{p_u(z)} b|v-v^*|dv$. A similar logic applies for $p(z) < v^*$ and $v^* < p(z\prime)$. The following theorem summarizes this result.

Theorem 3. *If assumptions 1-2 and 6 holds. Then the following bounds are valid:*

1. *If* $E[Y|Z=z]-E[Y|Z=z\prime] - \int_{p_l(z\prime)}^{p_u(z)} b|v-v^*|dv \geq 0$*:*

$$MTE^+(v^*)_{lb} = \frac{E[Y|Z=z]-E[Y|Z=z\prime] - \int_{p_l(z\prime)}^{p_u(z)} b|v-v^*|dv}{\Delta_{pu}}$$

$$MTE^+(v^*)_{ub} = \frac{E[Y|Z=z]+E[Y|Z=z\prime] + \int_{p_l(z\prime)}^{p_u(z)} b|v-v^*|dv}{\Delta_{pl}}$$

2. *If* $E[Y|Z=z]-E[Y|Z=z\prime] + \int_{p_l(z\prime)}^{p_u(z)} b|v-v^*|dv \leq 0$*:*

$$MTE^-(v^*)_{lb} = \frac{E[Y|Z=z]-E[Y|Z=z\prime] - \int_{p_l(z\prime)}^{p_u(z)} b|v-v^*|dv}{\Delta_{pl}}$$

$$MTE^-(v^*)_{ub} = \frac{E[Y|Z=z]+E[Y|Z=z\prime] + \int_{p_l(z\prime)}^{p_u(z)} b|v-v^*|dv}{\Delta_{pu}}$$

3. Otherwise:

$$MTE(v^*)_{lb} = \max\{MTE^-(v^*)_{lb}, MTE^+(v^*)_{lb}\}$$
$$MTE(v^*)_{ub} = \min\{MTE^-(v^*)_{ub}, MTE^+(v^*)_{ub}\}$$

2.14 Appendix E: The choice of b

In some applications, choosing b involves some subjective belief about the maximum size of treatment effects or, as above, on the underlying behavior of unobservable taste parameters. Identification results are obtained conditional on those beliefs. One possible route to choose b as proposed by Kim et al. (2018) formally is to rely on Bayesian inference using pre-samples or information from prior elicitation. Using existing experimental results or previous research, one may obtain a posterior distribution regarding b and use a high quantile of the posterior distribution as a possible value of b.

An alternative way of choosing b in the current paper and the application to SNAP is if there are previous results on $LATE$ or ATE for SNAP, we could choose b to be such that is consistent with previous studies on the topic.

Yet another way would be in the same spirit of Armstrong and Kolesár (2020) and a-priory decide on the bigger (or worst case) class of functions the researcher is willing to accept as potential marginal treatment responses. In that sense, if the researcher is willing, for example, to accept the idea that all functions between 0 and 1 with $b = 2$ or less are candidates, then he should present the report for all the values of b consistent with this notion.

The previous ideas all rely on the researcher either having a belief ex-ante or auxiliary data. An alternative way of choosing the b from inside the given data itself is the following. Suppose the support of the instrument has at least three values, z_0, z_1, z_2 and respective propensity scores p_0, p_1, p_2. If the researcher was willing to use only information on z_0, z_2 to get bounds on the

MTE at different values of v^* between 0 and 1 then notice the following equality:

$$E[Y|Z = z_2] - E[Y|Z = z_1] = \int_{p_1}^{p_2} MTE(v)dv$$

The obtained bounds on the MTE could be plugged in to get:

$$E[Y|Z = z_2] - E[Y|Z = z_1] \leq \int_{p_1}^{p_2} MTE_{ub}(v)dv$$
$$E[Y|Z = z_2] - E[Y|Z = z_1] \geq \int_{p_1}^{p_2} MTE_{lb}(v)dv$$

Similarly

$$E[Y|Z = z_1] - E[Y|Z = z_0] \leq \int_{p_0}^{p_1} MTE_{ub}(v)dv$$
$$E[Y|Z = z_1] - E[Y|Z = z_0] \geq \int_{p_0}^{p_1} MTE_{lb}(v)dv$$

Then the range of b could be chosen as consistent with the previous set of inequalities supported by the data. This strategy would require the instrument to take at least three values and would also imply not using all the information available to get the tightest possible bounds on the MTE conditional on b. Nevertheless, this would bring a way to discipline the value of b to be consistent with the observed data.

The best way to choose the tuning parameter b is still an open question that future research can work on. A similar comment can be made about the maximum level of α that sustain a particular marginal treatment effect of interest.

2.15 Appendix F: Bounds on the ATE

Table 2.2 Bounds on the ATE

$\alpha = 0.2$, monotonicity		$\alpha = 0.1$, monotonicity		$\alpha = 0$, monotonicity	
LB	UB	LB	UB	LB	UB
-1.00	-0.01	-1.00	-0.02	-1.00	-0.10
$\alpha = 0.2$, $b = 1$		$\alpha = 0.2$, $b = 0.5$		$\alpha = 0.2$, $b = 0.1$	
LB	UB	LB	UB	LB	UB
-1.00	0.00	-0.96	0.00	-0.48	-0.04
$\alpha = 0.1$, $b = 1$		$\alpha = 0.1$, $b = 0.5$		$\alpha = 0.1$, $b = 0.1$	
LB	UB	LB	UB	LB	UB
-0.94	0.00	-0.81	0.00	-0.38	-0.09
$\alpha = 0$, $b = 1$		$\alpha = 0$, $b = 0.5$		$\alpha = 0$, $b = 0.1$	
LB	UB	LB	UB	LB	UB
-0.72	-0.02	-0.50	-0.05	-0.28	-0.18

CHAPTER 3. TESTING IDENTIFYING ASSUMPTIONS IN BIVARIATE PROBIT MODELS

Santiago Acerenza, Otavio Bartalotti and Désiré Kédagni

Department of Economics, Iowa State University, Ames, IA, 50011, USA

Modified from a manuscript under review in *Journal of Applied Econometrics*

3.1 Abstract

In this paper, we develop a test for the identifying restrictions in bivariate probit models: the joint normality of errors, the instrument exogeneity, the linear index structure, and the relevance condition. We show that the procedure can easily be implemented using existing inference methods for intersection bounds. We discuss ways to relax the assumptions when they are rejected. Empirical examples illustrate the methodology.

3.2 Introduction

Since the seminal work of Heckman (1978), bivariate probit models have earned a lot of attention in social sciences. The bivariate probit model provides enough structure to point-identify traditional parameters of interest such as the average treatment effect (ATE), and its counterparts for the treated (ATT) and untreated (ATU) groups. While researchers recognize the restrictive nature of this model, it has been shown to be a powerful and popular approach in the literature. Influential examples include Evans and Schwab (1995), Neal (1997), and Altonji, Elder, and Taber (2005a,b) who use a bivariate probit model to estimate the effectiveness of Catholic schools. Goldman et al. (2001) developed a bivariate probit model of insurance and mortality to explain the correlation between unobserved health and insurance status, leading to the counter-intuitive results that HIV-infected persons receiving regular medical care using

insurance have a higher probability of death. Finally, Rhine, Greene, and Toussaint-Comeau (2006) model the consumer's decision to patronize check-cashing businesses jointly with the decision to be "unbanked." However, the literature related to testing the validity of these models and its assumptions remains underdeveloped.

This paper derives testable implications for the identifying assumptions in bivariate probit models, and proposes a testing procedure that can be used to check the falsifiability of such models. In the standard bivariate probit model which assumes joint normality of the errors, identification of the usual parameter of interest - the coefficient on the endogenous binary regressor - comes from three main sources: (i) instrument exogeneity, that is, its exclusion from the outcome of interest, (ii) the joint normality of the two latent variables in the triangular system of equations, and (iii) the relevance condition for the instrument. The exogeneity condition alone is not sufficient to obtain point-identification of the coefficient of interest. However, it allows partial identification of the ATE, the joint distribution of the potential outcomes, and other parameters of interest. Sometimes, the bounds on can be uninformative about the key feature of parameter of interest. In such a case, researchers often add restrictions to the model in order to draw conclusions about these features. This paper focuses on testing these identifying restrictions.

There is a growing literature on the testability of the identifying assumptions in various econometric models. Pearl (1994) derived testable implication for instrumental variables when the endogenous regressor is discrete. Balke and Pearl (1997) provided testable inequalities for the local average treatment effect (LATE) assumptions when the outcome, treatment, and instrument are all binary. Heckman and Vytlacil (2005) generalized those results to the case where the outcome has no support restrictions and can be discrete, continuous or mixed, but still with binary treatment and instrument. Kitagawa (2015) and Mourifié and Wan (2017) developed two different statistical tests for those inequalities. Huber and Mellace (2015) developed an alternative test for a version of the LATE assumptions using mean independence instead of full independence. Recently, Kédagni and Mourifié (2020) have complemented and generalized Pearl's (1994) testable inequalities to the case where the outcome and instruments are unrestricted, but

the treatment is discrete. Building on Pearl's (1995) conjecture, Gunsilius (2018) showed that there is no testable restriction in the continuous treatment case. Arai et al. (2018) developed a test for the identifying assumptions in the regression discontinuity design framework.

This paper's first contribution is to provide novel sharp testable equalities that can detect all possible violations of the bivariate probit model. Second, we propose a test for the validity of the identifying assumptions in the bivariate probit model using the previously mentioned sharp testable equalities and other feasible testable implications that are implied by the sharp equalities. All the testable implications (sharp and non-sharp) take the form of conditional moment equalities and inequalities, which can be implemented using existing inferential methods such as Chernozhukov, Lee, and Rosen (2013) or Andrews and Shi (2013). Monte Carlo simulations suggest that the proposed sharp test adequately controls size in large samples, though it tends to over-reject in finite samples. Furthermore, the test has power to detect violations of either the exclusion restriction or the joint normality, separately. Third, we discuss how one could relax the identifying assumptions when they are rejected in the data. Finally, we provide some empirical examples to illustrate our methodology and its practical relevance.

The remainder of the paper is organized as follows. Section 3.3 presents the model and the identifying assumptions. Section 3.4 discusses identification of the model parameters, and introduces the testable implications. Section 3.5 discusses the testing procedure. Section 3.6 includes simulation results about the size and power of the tests derived from 3.4. Section 3.7 discusses how to relax the assumptions when they are rejected. Section 3.8 provides two empirical illustrations. Section **??** extends the results to a bivariate switching regression framework. Finally, Section 3.9 concludes.[1]

[1] Additional results can be found in the appendix. Section 3.11 has the proof of our main propositions. 3.12 contains some additional remarks. Section 3.13 contains the proof of the consistency and asymptotic behavior of the test. Section 3.14 generalizes our framework to a more general copula theory. Section 3.15 has additional results for the applications.

3.3 The baseline model

Consider the following model

$$\begin{cases} Y &= \mathbb{1}\{\beta + \alpha D - U \geq 0\} \\ D &= \mathbb{1}\{\gamma + \delta Z - V \geq 0\} \end{cases} \tag{3.1}$$

where the vector (Y, D, Z) is the observed data, Y is a binary outcome, D is a binary treatment, $Z \in \mathcal{Z}$ is a potential instrument, (U, V) is a vector of latent variables, β, α, γ and δ are the model parameters, while α is of interest. For simplicity, we drop exogenous covariates from the model. All results derived in the paper hold conditional on covariates.

Interesting applications from the papers mentioned above provide examples of variables fitting this setup. Y could be mortality, a measure of college success or labor market success such as employment status, decision to patronize a check-cashing business, D could be health insurance, catholic schooling attendance, having a college degree, the decision to be "unbanked," and Z could be eligibility thresholds of an insurance policy, being Catholic, geographic proximity to Catholic schools, owning a house or college tuition.

Under this framework, the classical bivariate probit model identifying assumptions are the following.

Assumption 7 (Random Assignment)**.** *Z is statistically independent of (U, V).*

Assumption 8 (Normality)**.** *The vector $(U, V)'$ follows the standard bivariate normal distribution with covariance ρ, i.e., $\begin{pmatrix} U \\ V \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma)$, where $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, and $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.*

Assumption 9 (Relevance)**.** $\delta \neq 0$.

Assumption 7 states that the instrument is independent of all the unobservables in the model. In the Catholic school attendance example, it means that being Catholic is unrelated with unobserved factors that influence the decision to attend a Catholic school and student performance. Assumption 8 assumes that the vector of the unobservables in the model is jointly normally distributed. This assumption makes the model fully parametric, and eases the

identification of the model parameters. Assumption 9 states that the instrument is relevant in explaining variation in the treatment variable, e.g., that being Catholic has a direct effect on attending a Catholic school.

Under Assumptions 7, 8 and 9, the parameters β, α, γ, δ and ρ are identified.[2] We can therefore identify the average treatment effect ATE, defined as $\mathbb{E}[Y_1 - Y_0]$, where $Y_1 = \mathbb{1}\{\beta + \alpha - U \geq 0\}$ and $Y_0 = \mathbb{1}\{\beta - U \geq 0\}$. Indeed, under Assumption 8 we have $ATE = \Phi(\beta + \alpha) - \Phi(\beta)$, where $\Phi(.)$ denotes the standard normal cumulative distribution function (cdf).

It is worth noticing that the linear index structure of the model is assumed and will also be tested jointly with the assumptions previously mentioned.

3.4 Testable implications

In this section, we derive testable implications implied by the bivariate probit model described above. First, we heuristically discuss identification of β, α, γ, δ and ρ.

3.4.1 Identification

Assumption 7 implies

$$\mathbb{P}(Y = 1, D = 1|Z = z) = \mathbb{P}(U \leq \beta + \alpha, V \leq \gamma + \delta z), \tag{3.2}$$

$$\mathbb{P}(Y = 1, D = 0|Z = z) = \mathbb{P}(U \leq \beta, V > \gamma + \delta z), \tag{3.3}$$

$$\mathbb{P}(Y = 0, D = 1|Z = z) = \mathbb{P}(U > \beta + \alpha, V \leq \gamma + \delta z), \tag{3.4}$$

$$\mathbb{P}(Y = 0, D = 0|Z = z) = \mathbb{P}(U > \beta, V > \gamma + \delta z). \tag{3.5}$$

Combining equations (3.2) and (3.4), we have $\mathbb{P}(D = 1|Z = z) = \mathbb{P}(V \leq \gamma + \delta z)$, which is equal to $\Phi(\gamma + \delta z)$ under Assumption 8. Then, $\Phi^{-1}(\mathbb{P}(D = 1|Z)) = \gamma + \delta Z$. Hence, γ and δ are identified as follows: $\delta = \frac{Cov\left(\Phi^{-1}(\mathbb{P}(D=1|Z)),Z\right)}{Var(Z)}$, and $\gamma = \mathbb{E}\left[\Phi^{-1}(\mathbb{P}(D = 1|Z))\right] - \delta\mathbb{E}[Z]$.

[2] We briefly discuss these parameters' identification in Section 3.4. See Han and Vytlacil (2017), Han and Lee (2019), Mourifié and Méango (2014) for detailed identification results. Note that Li, Poskitt, and Zhao (2019) provide conditions for identification by functional form in the absence of an instrument. We focus on the testability of the standard bivariate probit model with an excluded variable, as it is the most commonly used.

Under Assumption 8, the latent variables U and V are jointly normal, and we can write the linear projection $U = \rho V + e$, where eV and $e \sim N(0, 1-\rho^2)$. Equation (3.2) implies

$$\begin{aligned}\mathbb{P}(Y=1|D=1,Z=z) &= \mathbb{P}(U \leq \beta + \alpha | V \leq \gamma + \delta z) = \\ &= \mathbb{P}\left(\frac{e}{\sqrt{1-\rho^2}} \leq a - \frac{\rho}{\sqrt{1-\rho^2}} V | V \leq \Phi^{-1}\left(\mathbb{P}(D=1|Z=z)\right) \right),\end{aligned}$$

where $a = \frac{\beta+\alpha}{\sqrt{1-\rho^2}}$. Since this function is strictly increasing in a, we can invert it to identify a for a particular value of ρ. Similarly, using Equation (3.3), we can identify $b = \frac{\beta}{\sqrt{1-\rho^2}}$. Then, $\beta = b\sqrt{1-\rho^2}$, and $\alpha = (a-b)\sqrt{1-\rho^2}$ can be recovered.

Let $a(\rho, z)$ and $b(\rho, z)$ describe a and b as a function of ρ and z. Under Assumption 7, ρ must satisfy $a(\rho, z) = a(\rho, z')$ and $b(\rho, z) = b(\rho, z')$ for all z, $z' \in \mathcal{Z}$. Han and Vytlacil (2017) show that ρ is uniquely determined if $\delta \neq 0$ (i.e., under Assumption 9).[3]

3.4.2 Sharp testable implications

We have

$$\begin{aligned}\mathbb{E}\left[YD|Z=z\right] &= \mathbb{P}(Y=1, D=1|Z=z) \\ &= \mathbb{P}\left(U \leq \beta+\alpha, V \leq \gamma + \delta Z | Z=z\right) \\ &= \mathbb{P}\left(U \leq \beta+\alpha, V \leq \gamma + \delta z\right) = \Phi_\rho(\beta+\alpha, \gamma+\delta z),\end{aligned}$$

where the first equality holds because Y, and D are binary, the second holds from the definition of the model, and third one holds under Assumption 7, and the last one by Assumption 8. We can do the same for $\mathbb{E}\left[Y(1-D)|Z=z\right]$ and $\mathbb{E}\left[(1-Y)D|Z=z\right]$ and get the following sharp equalities.

$$\mathbb{E}\left[YD|Z=z\right] = \mathbb{P}\left(U \leq \beta+\alpha, V \leq \gamma+\delta z\right) = \Phi_\rho(\beta+\alpha, \gamma+\delta z), \tag{3.6}$$

$$\mathbb{E}\left[Y(1-D)|Z=z\right] = \mathbb{P}(U \leq \beta, V > \gamma+\delta z) = \Phi(\beta) - \Phi_\rho(\beta, \gamma+\delta z), \tag{3.7}$$

$$\mathbb{E}\left[(1-Y)D|Z=z\right] = \mathbb{P}(U > \beta+\alpha, V \leq \gamma+\delta z) = \Phi(\gamma+\delta z) - \Phi_\rho(\beta+\alpha, \gamma+\delta z). \tag{3.8}$$

Where $\Phi_\rho(), \Phi()$ represent the joint standard bivariate normal with correlation coefficient equal to ρ and the univariate standard normal respectively.

[3] See Section 4 in Han and Vytlacil (2017) for a more complete discussion.

These equalities are testable since the model parameters are identified and must hold for every value of Z.

The next proposition summarizes all these testable equalities for the bivariate probit model.

Proposition 1. *Assume that Z is continuous. Under Assumptions 7-9 in the bivariate probit model (3.1), the parameters α, β, δ and ρ are identified, and equalities (3.6), (3.7), and (3.8) must hold. Moreover, these equalities are sharp.*

Remark 6 (Sharpness)**.** *In the context of model (3.1), equalities (3.6), (3.7), and (3.8) are sharp in the sense that whenever they hold, it is possible to construct a vector of $(\tilde{Y}, \tilde{D}, \tilde{U}, \tilde{V}, Z)$ that satisfies model (3.1), Assumptions 7-9, and induces the observed distribution on the data (Y, D, Z) where (Y, D, Z) has the same distribution as $(\tilde{Y}, \tilde{D}, Z)$.*

The previous equalities are sharp. We now provide a set of inequalities implied by equations (3.6), (3.7), and (3.8).

3.4.3 Non-sharp testable implications

In this section, we derive non sharp testable implications.

Equation (3.2) implies $\mathbb{P}(Y = 1, D = 1|Z = z) \leq \mathbb{P}(U \leq \beta + \alpha)$ for all z. Note that the right-hand side does not depend on Z. Under Assumption 8, we have $\mathbb{P}(U \leq \beta + \alpha) = \Phi(\beta + \alpha)$. Thus, the following testable implication must hold under model (3.1) and Assumptions 7-8:

$$\sup_z \mathbb{P}(Y = 1, D = 1|Z = z) \leq \Phi(\beta + \alpha).$$

Similarly, using equations (3.3), (3.4), and (3.5), we obtain the following testable implications

$$\begin{aligned} \sup_z \mathbb{P}(Y = 1, D = 0|Z = z) &\leq \Phi(\beta), \\ \sup_z \mathbb{P}(Y = 0, D = 1|Z = z) &\leq 1 - \Phi(\beta + \alpha), \\ \sup_z \mathbb{P}(Y = 0, D = 0|Z = z) &\leq 1 - \Phi(\beta), \end{aligned}$$

respectively. These four inequalities impose upper bounds on the joint distribution of (Y, D) conditional on the instrument Z. Note that only information about the marginal distribution of the unobserved heterogeneity Y_1 and Y_0 is used to derive these inequalities. They imply the Pearl (1994) instrumental inequalities, which are obtained by adequately taking the summation of two of the previous inequalities, respectively:

$$\sup_z \mathbb{P}(Y = 1, D = 1|Z = z) + \sup_z \mathbb{P}(Y = 0, D = 1|Z = z) \leq 1,$$
$$\sup_z \mathbb{P}(Y = 1, D = 0|Z = z) + \sup_z \mathbb{P}(Y = 0, D = 0|Z = z) \leq 1.$$

It is worth pointing out that the Pearl (1994) inequalities were derived in a more general potential outcome model, with no restriction on the distribution of the unobservables. In that context, the ATE is only partially identified, while it is point-identified in the current model. We now follow Kédagni and Mourifié (2020) to use restrictions that this model imposes on the joint distribution of (Y_0, Y_1) to derive further testable implications. We have

$$\begin{aligned}\mathbb{P}(Y_0, Y_1) = \mathbb{P}(U \leq \beta + \alpha, U \leq \beta) &= \mathbb{P}(U \leq \beta + \alpha, U \leq \beta, D = 1|Z = z) \\ &\quad + \mathbb{P}(U \leq \beta + \alpha, U \leq \beta, D = 0|Z = z) \\ &\leq \mathbb{P}(U \leq \beta + \alpha, D = 1|Z = z) + \mathbb{P}(U \leq \beta, D = 0|Z = z) \\ &= \mathbb{P}(Y = 1, D = 1|Z = z) + \mathbb{P}(Y = 1, D = 0|Z = z),\end{aligned}$$

where the first equality holds from Assumption 7 and the law of total probability, the second inequality holds from the monotonicity of a probability measure, and the last equality follows from the model definition. Under Assumption 8, we have $\mathbb{P}(U \leq \beta + \alpha, U \leq \beta) = \Phi(\min(\beta + \alpha, \beta))$. Therefore, by taking the infimum over z, the following must hold under the assumptions:

$$\Phi(\min(\beta + \alpha, \beta)) \leq \inf_z \mathbb{P}(Y = 1|Z = z).$$

Using a similar reasoning, we derive an additional testable implication of this form. The new inequalities impose lower bounds on the pairwise sum of the joint distribution of (Y, D) given the instrument Z. We summarize the discussion in the following proposition.

Proposition 2. *Under Assumptions 7, 8 and 9, the parameters α, β, δ and ρ are identified, and the following inequalities hold:*

$$\sup_z \mathbb{E}[YD|Z=z] \leq \Phi(\beta+\alpha), \tag{3.9}$$

$$\sup_z \mathbb{E}[Y(1-D)|Z=z] \leq \Phi(\beta), \tag{3.10}$$

$$\sup_z \mathbb{E}[(1-Y)D|Z=z] \leq 1-\Phi(\beta+\alpha), \tag{3.11}$$

$$\sup_z \mathbb{E}[(1-Y)(1-D)|Z=z] \leq 1-\Phi(\beta), \tag{3.12}$$

$$\Phi(\min(\beta+\alpha,\beta)) \leq \inf_z \mathbb{E}[Y|Z=z], \tag{3.13}$$

$$1-\Phi(\max(\beta+\alpha,\beta)) \leq \inf_z \mathbb{E}[1-Y|Z=z]. \tag{3.14}$$

Inequalities (3.9)-(3.14) are implied by the sharp equalities derived in the previous section, and imply the generalized instrumental inequalities of Kédagni and Mourifié (2020), since we are testing a stronger set of assumptions than the ones considered in their paper. We can estimate the vector of parameters $(\alpha, \beta, \delta, \gamma, \rho)$ by maximum likelihood, and then use the estimates $\hat{\alpha}, \hat{\beta}$ to test these inequalities using the intersection bounds framework of Chernozhukov, Lee, and Rosen (2013) or Andrews and Shi (2013).[4]

3.5 Testing procedure

To test equalities 3.6-3.8 we convert them into inequalities and then we write them in the in the Chernozhukov, Lee, and Rosen (2013) intersection bounds framework. Then we can use the Chernozhukov et al. (2015) Stata package for direct implementation. We then can also write the inequalities (3.9) to (3.14) in this same framework.

[4]There exist two extra inequalities but they are redundant and, as such, reduce the power (and possibly the size) of the test, which selects only binding constraints (Andrews and Soares, 2010).These two inequalities are collected in Remark 7 in Section 3.12.

Testing equalities 3.6-3.8 that must hold for any value of Z is equivalent to testing the following:

$$\begin{aligned}
\sup_z \mathbb{E}\left[YD - \Phi_\rho(\beta+\alpha, \gamma+\delta z)|Z=z\right] &\leq 0, \\
\sup_z \mathbb{E}\left[-YD + \Phi_\rho(\beta+\alpha, \gamma+\delta z)|Z=z\right] &\leq 0, \\
\sup_z \mathbb{E}\left[Y(1-D) - \Phi(\beta) + \Phi_\rho(\beta, \gamma+\delta z)|Z=z\right] &\leq 0, \\
\sup_z \mathbb{E}\left[-Y(1-D) + \Phi(\beta) - \Phi_\rho(\beta, \gamma+\delta z)|Z=z\right] &\leq 0, \\
\sup_z \mathbb{E}\left[(1-Y)D - \Phi(\gamma+\delta z) + \Phi_\rho(\beta+\alpha, \gamma+\delta z)|Z=z\right] &\leq 0, \\
\sup_z \mathbb{E}\left[-(1-Y)D + \Phi(\gamma+\delta z) - \Phi_\rho(\beta+\alpha, \gamma+\delta z)|Z=z\right] &\leq 0,
\end{aligned}$$

This is such because we can always write for example $\mathbb{E}\left[YD|Z=z\right] = \Phi_\rho(\beta+\alpha, \gamma+\delta z)$ as jointly $(\mathbb{E}\left[YD|Z=z\right] \leq \Phi_\rho(\beta+\alpha, \gamma+\delta z); \mathbb{E}\left[YD|Z=z\right] \geq \Phi_\rho(\beta+\alpha, \gamma+\delta z))$ which can also in turn be written as $(\mathbb{E}\left[YD|Z=z\right] \leq \Phi_\rho(\beta+\alpha, \gamma+\delta z); \mathbb{E}\left[-YD|Z=z\right] \leq -\Phi_\rho(\beta+\alpha, \gamma+\delta z))$.

Which can in turn also be written as
$(\mathbb{E}\left[YD - \Phi_\rho(\beta+\alpha, \gamma+\delta z)|Z=z\right] \leq 0; \mathbb{E}\left[-YD + \Phi_\rho(\beta+\alpha, \gamma+\delta z)|Z=z\right] \leq 0)$. These two inequalities must hold for every Z so we take the supremum over z.

Testing inequalities (3.9) to (3.14) would be equivalent to testing the following:

$$\begin{aligned}
\sup_z \mathbb{E}\left[YD - \Phi(\beta+\alpha)|Z=z\right] &\leq 0,\\
\sup_z \mathbb{E}[Y(1-D) - \Phi(\beta)|Z=z] &\leq 0,\\
\sup_z \mathbb{E}\left[(1-Y)D - 1 + \Phi(\beta+\alpha)|Z=z\right] &\leq 0,\\
\sup_z \mathbb{E}[(1-Y)(1-D) - 1 + \Phi(\beta)|Z=z] &\leq 0,\\
\sup_z \mathbb{E}\left[\Phi(\min(\beta+\alpha,\beta)) - Y|Z=z\right] &\leq 0,\\
\sup_z \mathbb{E}[1 - \Phi(\max(\beta+\alpha,\beta)) - (1-Y)|Z=z] &\leq 0.
\end{aligned}$$

To implement the test, we replace α and β by their maximum likelihood estimators (MLE), denoted $\hat{\alpha}$ and $\hat{\beta}$, respectively. In Section 3.13 we show the asymptotic properties of the test are unaffected by using $\hat{\alpha}$ and $\hat{\beta}$ when nonparametric estimators for the conditional expectations are used.

We now briefly describe the method. First, write the inequalities above as the null hypothesis.

$$H_0 : \theta_0 \equiv \max_{j\in\{1,\ldots,6\}} \sup_{z\in\mathcal{Z}} \theta(z,j) \leq 0,$$

where $\theta(z,j) \equiv \mathbb{E}[W_j|Z=z]$, and W_j represents the expression in the conditional expectation for inequality j. For example, $W_1 = YD - \Phi(\hat{\alpha}+\hat{\beta})$. The decision rule for the test is given by Chernozhukov, Lee, and Rosen (2013), we reject H_0 if

$$\hat{\theta}_{1-\alpha} \equiv \max_{j\in\{1,\ldots,6\}} \sup_{z\in\mathcal{Z}} \left\{\hat{\theta}(z,j) - k_{1-\alpha}\hat{s}(z,j)\right\} > 0,$$

where $\hat{\theta}(z,j)$ is the local linear estimator for $\theta(z,j)$, $\hat{s}(z,j)$ its standard error, and $k_{1-\alpha}$ is a critical value at the significance level α. Details about the implementation can be found in Section 3.15.

3.6 Monte Carlo simulations

This section presents simulation results for the size and power of the test for the validity of the bivariate probit model and assumptions based on the intersection bounds framework (Chernozhukov, Lee, and Rosen, 2013) exploiting the equalities derived in Section 3.4.2.[5]

3.6.0.1 Size

Consider the following data generating process (DGP) where Assumptions 7, 8 and 9 hold for the bivariate probit model:

$$\begin{cases} Y &= \mathbb{1}\{D - U \geq 0\} \\ \\ D &= \mathbb{1}\{2Z - V \geq 0\} \end{cases} \tag{3.15}$$

where $\begin{pmatrix} U \\ V \end{pmatrix} \sim N(0, \Sigma)$ with $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$, $Z \sim \mathcal{U}_{[-3,3]}$, and $Z \perp\!\!\!\perp (U, V)$.

Table 3.1 shows the false rejection rates for different nominal sizes and sample sizes. Each simulation relies on 500 replications. We can see that for smaller sample sizes the test tends to over reject. While for bigger samples the tendency tends to revert but approaching proper cover for a nominal size of 1 percent.

Table 3.1 Rejection Frequency (clrbound)

	Series method		
Nominal Size	10%	5%	1%
$n = 200$	41%	39%	33%
$n = 1000$	16%	13%	8%
$n = 10000$	5%	3%	1%

Based on 500 replications.

[5] For results on the nonsharp equalities see the working paper version of the document.

3.6.0.2 Power

We consider two DGPs, with different violations of Assumptions 7-9 to provide evidence of the test's power.

In the first DGP, we consider violations of random assignment (Assumption 7), while normality and relevance hold. We run 500 replications for samples of size 10000. In this DGP, the coefficient ρ captures simultaneously the degree of endogeneity of the treatment D and the extent of violation of the random assignment assumption[6]. The results are reported in Table 3.2.

$$\begin{cases} Y &= \mathbb{1}\{D - U \geq 0\} \\ \\ D &= \mathbb{1}\{2Z - V \geq 0\} \end{cases} \tag{3.16}$$

where $\begin{pmatrix} U \\ V \\ Z \end{pmatrix} \sim N(0, \Sigma)$ with $\Sigma = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$.

Table 3.2 Rejection Frequency (clrbound)

	Series method		
Nominal Size	10%	5%	1%
$\rho = 0.0$	5%	3%	1%
$\rho = 0.1$	100%	100%	100%
$\rho = 0.5$	100%	100%	100%

Based on 500 replications with sample size 10000.

We can see in table 3.2 that the test is very power full detecting violations from independence.

In the second DGP, we consider the case when the normality assumption is violated while random assignment and relevance hold. The joint distribution of the errors is a combination of a normal and a log-normal.

[6]For results where the either U or V are not independent of Z see the working paper version of this document.

$$\begin{cases} Y &= \mathbb{1}\{D - U \geq 0\} \\ \\ D &= \mathbb{1}\{2Z - V \geq 0\} \end{cases} \tag{3.17}$$

where $U = (1-\lambda)\ln U^* + (\lambda)^*$, $V = (1-\lambda)\ln V^* + (\lambda)^*$, λ between 0 and 1, $\begin{pmatrix} U^* \\ V^* \end{pmatrix} \sim N(0, \Sigma)$ with $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$, and $Z \perp\!\!\!\perp (U^*, V^*)$. Table 3.3 presents the results based on 500 replications and 10000 observations for different values of λ.[7]

Table 3.3 Rejection Frequency (clrbound)

	Series method		
Nominal Size	10%	5%	1%
$\lambda = 0.1$	100%	100%	100%
$\lambda = 0.5$	100%	100%	100%
$\lambda = 0.8$	100%	100%	100%
$\lambda = 1$	100%	100%	100%

Based on 500 replications.

We can see that the test is very powerful even for mild violations of the assumptions.

3.7 What to do when the testable implications are rejected

When the test rejects the researcher could relax some of the identifying assumptions in order to study identification of the treatment effect.

For example, one can relax the normality assumption and bound the ATE under the random assignment assumption following Kédagni and Mourifié (2020).

Suppose that only Assumption 7 holds. Then, by the identification results of Kédagni and Mourifié (2020, Proposition 1) can be used to obtain the following bounds on the potential

[7] In the working paper version, we consider a similar violation of the normality assumption, replacing the log-normal with a uniform distribution.

outcome means $\mathbb{E}[Y_0]$ and $\mathbb{E}[Y_1]$, respectively:

$$\max\left\{\sup_z \mathbb{E}[Y(1-D)|Z=z], 1-\inf_z \mathbb{E}[1-Y|Z=z] - \inf_z \mathbb{E}[(1-Y)(1-D)+YD|Z=z]\right\}$$
$$\leq \mathbb{E}[Y_0] \leq$$
$$\min\left\{1-\sup_z \mathbb{E}[(1-Y)(1-D)|Z=z], \inf_z \mathbb{E}[Y|Z=z] + \inf_z \mathbb{E}[Y(1-D)+(1-Y)D|Z=z]\right\},$$

and

$$\max\left\{\sup_z \mathbb{E}[YD|Z=z], 1-\inf_z \mathbb{E}[1-Y|Z=z] - \inf_z \mathbb{E}[Y(1-D)+(1-Y)D|Z=z]\right\}$$
$$\leq \mathbb{E}[Y_1] \leq$$
$$\min\left\{1-\sup_z \mathbb{E}[(1-Y)D|Z=z], \inf_z \mathbb{E}[Y|Z=z] + \inf_z \mathbb{E}[(1-Y)(1-D)+YD|Z=z]\right\}.$$

Constructing confidence bounds for these potential outcome means is challenging as the identified sets involve the summation of extrema over the support of the instrument. We follow Kédagni and Mourifié (2020) to combine a sample splitting approach with the intersection bounds framework.

Suppose that we have two independent copies $(Y^{(1)}, D^{(1)}, Z^{(1)})$ and $(Y^{(2)}, D^{(2)}, Z^{(2)})$ of the data (Y, D, Z). In practice, two independent copies of the data can be obtained by randomly splitting the sample into two subsamples if the original data are independent and identically distributed.

Then, we can write these identified sets in the intersection bounds framework to perform inference. We have:[8]

$$\max\left\{\sup_z \mathbb{E}\left[Y^{(1)}\left(1-D^{(1)}\right)|Z^{(1)}=z\right],\right.$$
$$\left.\sup_{z,z'} \mathbb{E}\left[Y^{(1)} - \left(1-Y^{(2)}\right)\left(1-D^{(2)}\right) - Y^{(2)}D^{(2)}|Z^{(1)}=z, Z^{(2)}=z'\right]\right\}$$
$$\leq \mathbb{E}[Y_0] \leq$$
$$\min\left\{\inf_z \mathbb{E}\left[1-\left(1-Y^{(1)}\right)\left(1-D^{(1)}\right)|Z^{(1)}=z\right],\right.$$
$$\left.\inf_{z,z'} \mathbb{E}\left[Y^{(1)} + Y^{(2)}\left(1-D^{(2)}\right) + \left(1-Y^{(2)}\right)D^{(2)}|Z^{(1)}=z, Z^{(2)}=z'\right]\right\},$$

[8] See Kédagni and Mourifié (2020) for more details on the procedure.

and

$$\max\left\{\sup_z \mathbb{E}\left[Y^{(1)}D^{(1)}|Z^{(1)}=z\right],\right.$$
$$\left.\sup_{z,z'}\mathbb{E}\left[Y^{(1)}-Y^{(2)}\left(1-D^{(2)}\right)-\left(1-Y^{(2)}\right)D^{(2)}|Z^{(1)}=z,Z^{(2)}=z'\right]\right\}$$
$$\leq \mathbb{E}[Y_1] \leq$$
$$\min\left\{\inf_z \mathbb{E}\left[1-\left(1-Y^{(1)}\right)D^{(1)}|Z^{(1)}=z\right],\right.$$
$$\left.\inf_{z,z'}\mathbb{E}\left[Y^{(1)}+\left(1-Y^{(2)}\right)\left(1-D^{(2)}\right)+Y^{(2)}D^{(2)}|Z^{(1)}=z,Z^{(2)}=z'\right]\right\}.$$

We can therefore obtain bounds on the ATE by taking the difference of the bounds on $\mathbb{E}[Y_1]$ and $\mathbb{E}[Y_0]$. We again can use the Chernozhukov et al. (2015) Stata package to implement these bounds. If one of those bounds is empty, then Assumption 7 is rejected. In such a case, the researcher could resort to other identification strategies such as the monotone instrumental variable approach ($\mathbb{E}[Y_d|Z=z]$ is monotone in z for each d) developed by Manski and Pepper (2000), or some sensitivity analysis like the one developed by Altonji, Elder, and Taber (2005b).

The above bounds can be tightened by, in addition to Assumption 7, imposing monotonicity of the outcome Y in the treatment D, as proposed by Machado, Shaikh, and Vytlacil (2019). Note that this assumption is implicit in the bivariate probit specification.

Assumption 10 (Monotonicity of Y in D)**.** *Either* $Y_1 \geq Y_0$ *a.s. or* $Y_1 \leq Y_0$ *a.s.*.

As pointed out by Machado, Shaikh, and Vytlacil (2019), Assumption 10 is weaker than the "monotone treatment response" considered in Manski (1997), and Manski and Pepper (2000), which assumes that the direction of the monotonicity is known *a priori*. Then, the following proposition holds.[9]

[9]The proof of the proposition is in Appendix 3.11. The identified set in Proposition 3 takes the form of intersection bounds, and can be implemented using Chernozhukov et al.'s (2015) Stata package.

Proposition 3. *Under Assumptions 7 and 10, sharp bounds for* $\mathbb{E}[Y_1]$ *and* $\mathbb{E}[Y_0]$ *are:*

$$\begin{cases} \sup_z \mathbb{E}[Y|Z=z] \leq \mathbb{E}[Y_1] \leq \inf_z \mathbb{E}[1-(1-Y)D|Z=z] \\ \\ \sup_z \mathbb{E}[Y(1-D)|Z=z] \leq \mathbb{E}[Y_0] \leq \inf_z \mathbb{E}[Y|Z=z] \end{cases} \tag{3.18}$$

or

$$\begin{cases} \sup_z \mathbb{E}[YD|Z=z] \leq \mathbb{E}[Y_1] \leq \inf_z \mathbb{E}[Y|Z=z] \\ \\ \sup_z \mathbb{E}[Y|Z=z] \leq \mathbb{E}[Y_0] \leq \inf_z \mathbb{E}[1-(1-Y)(1-D)|Z=z] \end{cases} \tag{3.19}$$

Sharp bounds on the ATE are obtained by taking the difference of the bounds on $\mathbb{E}[Y_1]$ and $\mathbb{E}[Y_0]$. The sign of the ATE is identified only if one of the identified sets in Equations (3.18) and (3.19) is empty. Note that Balke and Pearl (1997) showed that adding monotonicity of D in Z does not improve the bounds for the ATE under Assumption 7 when the outcome, treatment and instrument are all binary. Hence, we conjecture and prove in Appendix **??** that imposing monotonicity of the treatment in the instrument (which is also implicit in the bivariate probit model) will not further tighten the bounds if the model is not misspecified.

If we consider the case in which only Assumption 8 holds, Mourifié and Méango (2014) showed that when $\delta = 0$ the model is generally underidentified. In such scenario α and β are identified up to the coefficient of correlation ρ, the degree of endogeneity in the model.[10] Therefore, knowing the trivial bounds $(-1, 1)$ on ρ, we can partially identify α and β. In general, the bounds on the parameters α and β may be wide (and possibly uninformative) without additional restrictions. Alternatively, ρ and the selection on unobservables could be bounded by the degree of selection on observables as proposed by Altonji, Elder, and Taber (2005b). Finally, one can restrict ρ and ask what values are plausible using some economic argument, as suggested by Rosenbaum and Rubin (1983), and Rosenbaum (1995).

[10] See discussion in Subsection 3.4.1.

3.8 Empirical illustrations

To illustrate the usefulness of the tests developed in Section 3.5, we apply the methodology to the data sets of two policy relevant recent papers. Our first empirical example uses the data set from Zimmer (2017) and Han and Lee (2019) to analyze the effect of access to health insurance on individuals' decision to visit a doctor. The second application revisits Gao et al. (2018) to analyze how land tenure arrangements affect Chinese farmers' adoption of straw retention.

3.8.1 The effect of insurance on doctor visits

Han and Lee (2019) analyze how health insurance coverage affects an individual's decision to visit a doctor. In this example, Y and D are indicators for whether an individual has a doctor visit, and is covered by private health insurance, respectively. The instrument Z is the number of employees in the firm at which the individual works. The reasoning for instrument's validity holds that larger firms are more likely to provide health insurance to their workforce.

The data comes from the 2010 wave of the Medical Expenditure Panel Survey (MEPS). As in Han and Lee (2019), we focus on all the visits that occurred in January 2010, restrict the sample to contain individuals with age between 25 and 64, and exclude individuals who have retained any federal or state insurance in 2010. Furthermore, individuals who are self-employed or unemployed are excluded from the analysis.

Table 3.4 presents the estimates and standard errors for the parameters in the model, obtained by a bivariate probit framework. The first column presents estimates for the selection into treatment, and indicates that the number of employees in a firm increases the likelihood that an individual has private health insurance coverage. The second column reports a positive effect of having private health insurance on doctor visits, as economic theory predicts. Furthermore, the coefficient of correlation between the unobservables in the selection and outcome equation (ρ) is negative and significant. This seems consistent with the idea of adverse selection, since sicker employees are more likely to obtain private insurance and to visit a doctor's office.

Table 3.4 Bivariate probit specification

	MLE	
	Private insurance	Doctor visit
Nb employees	0.3661^{***}	
	(0.0170)	
Private insurance		0.6388^{***}
		(0.1189)
Constant	0.4374^{***}	-1.3166^{***}
	(0.0154)	(0.0707)
ρ		-0.2605^{***}
		(0.0777)
n	$7,555$	$7,555$

Standard errors (in parentheses); ***: significant at 1% level.

However, when we test for the validity of the bivariate probit model as described above, the model is rejected at all three conventional levels 1%, 5% and 10%.

Confidence Bounds on ATE under Assumption 7

Following the discussion in Section 3.7, we construct confidence bounds on the potential outcome means under Assumption 7 only, as described in Kédagni and Mourifié (2020).

The first column of Table 3.5 reports estimates of the bivariate probit model for the potential outcome means ($\mathbb{E}[Y_1] = \Phi(\hat{\beta} + \hat{\alpha})$, $\mathbb{E}[Y_0] = \Phi(\hat{\beta})$), and the ATE ($\mathbb{E}[Y_1 - Y_0]$), while the second and third columns report the 95% confidence lower and upper bounds, respectively, under the random assignment assumption only. We can see that the point estimate for $\mathbb{E}[Y_0]$ and $\mathbb{E}[Y_1]$ lies within its confidence region, while the point estimate for lies outside its confidence bounds. These results suggest that the rejection of the bivariate model is due to the joint normality and the relevance assumptions. However, we cannot reject the null hypothesis that the ATE is equal to the value 0.155 provided by the bivariate probit model, as it lies within the confidence region for the ATE: $[-0.496, 0.292]$. This suggests that there may exist a data generating process compatible with Assumption 7 that can yield this value, but it is not the standard bivariate probit model. The

confidence bounds on ATE do not allow us to draw a conclusion about the direction of the effect of private insurance on doctor visits. The effect can be positive, zero or negative.

Table 3.5 Confidence sets for parameters

Parameters	Biprobit estimates	95% conf. LB	95% conf. UB
$\mathbb{E}[Y_0]$	0.0940	0.0495	0.6467
$\mathbb{E}[Y_1]$	0.2490	0.1507	0.3414
ATE	0.1550	-0.4960	0.2920

conf.: confidence; LB: lower bound; UB: upper bound.

3.8.2 Do farmers adopt fewer conservation practices on rented land?

Since conservation practices are costly, some farmers adopt them and some do not. The economic theory suggests that land tenants will adopt less conservation practices. Because of land tenure insecurity, they invest less in their lands, especially those with higher initial costs or long payoff horizon. Gao et al. (2018) investigate this theory in the case of Chinese farmers' adoption of straw retention, a key conservation practice to curb air pollution from burning crop residues. The standard bivariate probit model has been one of their model specifications. The data is drawn from a rural household survey conducted by Henan Agricultural University in 2016 in Henan Province, a major grain production province in central China.[11] The data consists of 1659 plot-level observations from 670 farmer households for analysis. The dependent variable Y is an indicator for the adoption of straw retention, and the treatment variable D is a dummy for rented plot. The authors use a ratio of annual income for family's migrant workers to annual agricultural profits for the farmer household as an instrument.[12] The first column of Table 3.6 suggests that families with higher ratio of migrants' income to agricultural profits are less likely to rent land to augment their income. The second column seems to support the idea that farmers who rent land are less likely to adopt straw retention. However, as in the previous empirical example, our test

[11] We thank Wendong Zhang for sharing the data with us.
[12] See Gao et al. (2018) for more details on the construction of the instrument.

for the bivariate probit model is rejected at all three significance levels. We then proceed again to construct confidence regions for the potential outcome means and the ATE under the random assignment assumption only.

Table 3.6 Bivariate probit specification

	MLE	
	Rent	Adoption
Mig-ag-ratio	-0.0109^{**}	
	(0.0049)	
Rent		-2.0984^{***}
		(0.2216)
Constant	-1.2556^{***}	0.6403^{***}
	(0.0505)	(0.0332)
ρ		0.9559^{*}
		(0.0987)
n	1,659	1,659

Standard errors in parentheses; ***: significant at 1% level, **: 5%, *: 10%.

Confidence Bounds on ATE under Assumption 7

The first column of Table 3.7 reports the bivariate probit estimates for the potential outcome means and the ATE, and the second and third columns display the 95% confidence lower and upper bounds for these parameters, respectively.

We can see that both bivariate probit estimates for $\mathbb{E}[Y_0]$ and $\mathbb{E}[Y_1]$ lie outside their confidence sets. The non-emptiness of the confidence regions suggests that Assumption 7 alone is not rejected by the data. The fact that the point estimate for the ATE lies within its confidence set suggests that the value -0.67 cannot be rejected as the effect of land leasing or tenure on straw retention, but the data generating process cannot be the bivariate probit model. Assumption 7 alone does not allow us to draw a conclusion on the direction of the effect, since the confidence set for the ATE contains zero.

Table 3.7 Confidence sets for parameters

Parameters	Biprobit estimates	95% conf. LB	95% conf. UB
$\mathbb{E}[Y_0]$	0.7390	0.6026	0.7724
$\mathbb{E}[Y_1]$	0.0724	0.0393	0.9813
ATE	-0.6666	-0.7331	0.3787

conf.: confidence; LB: lower bound; UB: upper bound.

3.9 Conclusion

This paper develops a falsification test for the identifying assumptions in bivariate probit models. We derive sharp testable equalities for the model assumptions. We then propose a testing procedure, which we express in the form of conditional moment inequalities. We implement those inequalities using existing inferential methods. The test can easily be implemented using the Stata package developed by Chernozhukov et al. (2015). We provide some simulation results on the performance of the test. We find that the test tends to over-reject in finite samples. However in large samples, the test controls size. We discuss ways to relax the assumptions when they are rejected. Finally, we provide some empirical examples, illustrating that the bivariate model, despite its nice feature that leads to point-identication of model parameters, could be restrictive in some cases. Our proposed procedure should serve as a screening test for the validity of the bivariate probit specification.

While the test we develop in this paper can easily handle discrete covariates, it is difficult to include continuous covariates in the implementation of the procedure. We believe this question could be further explored in future research.

3.10 References

Altonji, J. G., T. E. Elder, and C. R. Taber. 2005a. "An evaluation of instrumental variable strategies for estimating the effects of catholic schooling." *Journal of Human Resources* 40 (4):791–821.

Altonji, Joseph, Todd Elder, and Christopher Taber. 2005b. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy* 113 (1):151–184.

Andrews, D. W. K. and X. Shi. 2013. "Inference Based on Conditional Moment Inequalities." *Econometrica* 81:609–666.

Andrews, D. W. K. and G. Soares. 2010. "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection." *Econometrica* 78 (1):119–157.

Arai, Y., Y-C. Hsu, T. Kitagawa, I. Mourifé, and Y. Wan. 2018. "Testing identifying assumptions in fuzzy regression discontinuity designs." *Cemmap Working Paper* CWP50/18.

Balke, Alexander and Judea Pearl. 1997. "Bounds on Treatment Effects from Studies with Imperfect Compliance." *Journal of the American Statistical Association* 92 (439):1171–1176.

Chernozhukov, Victor, Wooyoung Kim, Sokbae Lee, and Adam M. Rosen. 2015. "Implementing Intersection Bounds in Stata." *Stata Journal* 15 (1):21–44.

Chernozhukov, Victor, Sokbae Lee, and Adam M. Rosen. 2013. "Intersection Bounds: Estimation and Inference." *Econometrica* 81 (2):667–737. URL http://dx.doi.org/10.3982/ECTA8718.

Evans, W. N. and R. M. Schwab. 1995. "Finishing high school and starting college: Do Catholic schools make a difference?" *Quarterly Journal of Economics* 110 (4):941–974.

Galichon, Alfred and Marc Henry. 2011. "Set identification in models with multiple equilibria." *The Review of Economic Studies* 78 (4):1264–1298.

Gao, Li, Wendong Zhang, Yingdan Mei, Abdoul G. Sam, Yu Song, and Shuqin Jin. 2018. "Do farmers adopt fewer conservation practices on rented land? Evidence from straw retention in China." *Land Use Policy* 79:609–621.

Goldman, D., j. Bhattacharya, D. Mccaffrey, N. Duan, A. Leibowitz, G. Joyce, and S. Morton. 2001. "Effect of Insurance on Mortality in an HIV-Positive Population in Care." *Journal of American Statistical Association* 96 (455):883–894.

Gunsilius, F. 2018. "Testability of the exclusion restriction in continuous instrumental variable models." *Working Paper* .

Han, S. and S. Lee. 2019. "Estimation in a generalization of bivariate probit models with dummy endogenous regressors." *Journal of Applied Econometrics* :1–22.

Han, S. and E. J. Vytlacil. 2017. "Identification in a generalization of bivariate probit models with dummy endogenous regressors." *Journal of Econometrics* 199:63–73.

Heckman, James J. 1978. "Dummy Endogenous Variables in a Simultaneous Equation System." *Econometrica* 46 (4):931–959.

Heckman, James J and Edward Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica* 73 (3):669–738.

Huber, Martin and Giovanni Mellace. 2015. "Testing Instrument Validity for LATE Identification Based on Inequality Moment Constraints." *The Review of Economics and Statistics* 97 (2):398–411.

Kédagni, D. and I. Mourifié. 2020. "Generalized Instrumental Inequalities: Testing the Instrumental Variable Independence Assumption." *Biometrika* 107 (3):661–675.

Kitagawa, Toru. 2015. "A Test for Instrument Validity." *Econometrica* 83:2043–2063.

Li, Chuhui, D.S. Poskitt, and Xueyan Zhao. 2019. "The bivariate probit model, maximum likelihood estimation, pseudo true parameters and partial identification." *Journal of Econometrics* 209:94–113.

Machado, C., A. Shaikh, and E. Vytlacil. 2019. "Instrumental Variables and the Sign of the Average Treatment Effect." *Journal of Econometrics* 212:522–555.

Manski, C. F. and J. Pepper. 2000. "Monotone Instrumental Variables: With an Application to the Returns to Schooling." *Econometrica* 68:997–1010.

Manski, Charles F. 1997. "Monotone Treatment Response." *Econometrica* 65 (6):1311–1334.

Mourifié, I. and R. Méango. 2014. "A note on the identification in two equations probit model with dummy endogenous regressor." *Economics Letters* 125:360–363.

Mourifié, I. and Y. Wan. 2017. "Testing Local Average Treatment Effect Assumptions." *The Review of Economics and Statistics* 99 (2):305–313.

Neal, D. A. 1997. "The effects of catholic secondary schooling on educational achievement." *Journal of Labor Economics* 15 (1, Part 1):98–123.

Pearl, J. 1995. "Causal inference from indirect experiments." *Artificial Intelligence in Medicine* 7:561–582.

Pearl, Judea. 1994. "On the Testability of Causal Models with Latent and Instrumental Variables." *Uncertainty in Artificial Intelligence* 11:435–443.

Rhine, S. L., W. H. Greene, and M. Toussaint-Comeau. 2006. "The importance of check-cashing businesses to the unbanked: Racial/ethnic differences." *Review of Economics and Statistics* 88 (1):146–157.

Rosenbaum, P. R. 1995. *Observational Studies*. New York: Springer-Verlag.

Rosenbaum, P. R. and D. B. Rubin. 1983. "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome." *Journal of the Royal Statistical Society. Series B (Methodological)* 45 (2):212–218.

Zimmer, David. 2017. "Using copulas to estimate the coefficient of a binary endogenous regressor in a Poisson regression: Application to the effect of insurance on doctor visits." *Health Economics* 27:545–456.

3.11 Appendix A: Proof of Propositions 1, 2 and 3

3.11.1 Proof of Proposition 1

Proof. Assume that $Cov\left(\Phi^{-1}\left(\mathbb{P}(D=1|Z)\right), Z\right) \neq 0$. Let the parameters α, β, ρ, γ, δ be defined as previously identified in Subsection 3.4.1. Then $\delta \neq 0$ by definition. Suppose now that equalities (3.6), (3.7) and (3.8) hold. We need to show that there exists a vector $\left(\tilde{U}, \tilde{V}, \tilde{Y}, \tilde{D}, Z\right)$ such that Z is independent of $\left(\tilde{U}, \tilde{V}\right)$, $\begin{pmatrix} \tilde{U} \\ \tilde{V} \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma)$, where $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, and $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, and the joint distribution of $\left(\tilde{Y}, \tilde{D}, Z\right)$ is the same as that of (Y, D, Z). Define the joint density of $\left(\tilde{U}, \tilde{V}\right)$ conditional on Z as

$$f_{(\tilde{U},\tilde{V}|Z)}(u, v|z) = \frac{1}{\sqrt{1-\rho^2}} \phi\left(\frac{u-\rho v}{\sqrt{1-\rho^2}}\right) \phi(v),$$

where $\phi(t) = \exp(-t^2/2)$, and define

$$\begin{cases} \tilde{Y} &= \mathbb{1}\{\beta + \alpha\tilde{D} - \tilde{U} \geq 0\} \\ \\ \tilde{D} &= \mathbb{1}\{\gamma + \delta Z - \tilde{V} \geq 0\} \end{cases}$$

It is clear that Z is independent from $(\tilde{U}, \tilde{V})$ since the conditional density of $\left(\tilde{U}, \tilde{V}\right)|Z = z$ does not depend on z. We can easily show that the above joint density of $\left(\tilde{U}, \tilde{V}\right)$ is equal to

$$f_{(\tilde{U},\tilde{V})}(u, v) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(u\ v)\Sigma^{-1}(u\ v)'\right\}.$$

Therefore, $\begin{pmatrix} \tilde{U} \\ \tilde{V} \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma)$ and thus has the same distribution as $\begin{pmatrix} U \\ V \end{pmatrix}$ and also their marginals. It remains to show that

$$\mathbb{P}\left(\tilde{Y} = y, \tilde{D} = d | Z = z\right) = \mathbb{P}\left(Y = y, D = d | Z = z\right)$$

for all y, d, and z. We have

$$\begin{aligned}
\mathbb{P}\left(\tilde{Y} = 1, \tilde{D} = 1 | Z = z\right) &= \mathbb{P}\left(\tilde{U} \leq \beta + \alpha, \tilde{V} \leq \gamma + \delta z\right), \text{ since } Z \text{ indep. of } (U, \tilde{V}) \\
&= \Phi_\rho(\beta + \alpha, \gamma + \delta z), \text{ from Joint Normality of } (\tilde{U}, \tilde{V}) \\
&= \mathbb{E}\left[YD | Z = z\right], \text{ from Equality (3.6)} \\
&= \mathbb{P}\left(Y = 1, D = 1 | Z = z\right).
\end{aligned}$$

Similarly we have $\mathbb{P}\left(\tilde{Y} = 1, \tilde{D} = 0 | Z = z\right) = \mathbb{P}\left(Y = 1, D = 0 | Z = z\right)$ from joint normality and equality (3.7). Finally, we have $\mathbb{P}\left(\tilde{Y} = 0, \tilde{D} = 1 | Z = z\right) = \mathbb{P}\left(Y = 0, D = 1 | Z = z\right)$ from joint normality and(3.8). □

3.11.2 Proof of Proposition 2

Proof. We begin with **inequality (3.9)**. We have

$$\begin{aligned}
\mathbb{E}[YD | Z = z] &= \mathbb{P}(Y = 1, D = 1 | Z = z), \\
&= \mathbb{P}(U \leq \alpha + \beta, V \leq \delta z | Z = z), \\
&\leq \mathbb{P}(U \leq \alpha + \beta | Z = z), \\
&= \mathbb{P}(U \leq \alpha + \beta) = \Phi(\alpha + \beta),
\end{aligned}$$

where the last two equalities follow from Assumptions 7 and 8, respectively. Thus, by taking the supremum of the left-hand side over z, we have $\sup_z \mathbb{E}[YD | Z = z] \leq \Phi(\alpha + \beta)$.

Inequality (3.10)

Similar to the previous reasoning, we have

$$\begin{aligned}\mathbb{E}[Y(1-D)|Z=z] &= \mathbb{P}(Y=1, D=0|Z=z),\\ &= \mathbb{P}(U \leq \beta, V > \delta z|Z=z),\\ &\leq \mathbb{P}(U \leq \beta) = \Phi(\beta).\end{aligned}$$

Thus $\sup_z \mathbb{E}[Y(1-D)|Z=z] \leq \Phi(\beta)$.

Inequality (3.11)

$$\begin{aligned}\mathbb{E}[(1-Y)D|Z=z] &= \mathbb{P}(Y=0, D=1|Z=z),\\ &= \mathbb{P}(U > \alpha+\beta, V \leq \delta z),\\ &\leq P(U > \alpha+\beta) = 1-\Phi(\alpha+\beta).\end{aligned}$$

Thus $\sup_z \mathbb{E}[(1-Y)D|Z=z] \leq 1-\Phi(\alpha+\beta)$.

Inequality (3.12)

$$\begin{aligned}\mathbb{E}[(1-Y)(1-D)|Z=z] &= \mathbb{P}(Y=0, D=0|Z=z),\\ &= \mathbb{P}(U > \beta, V > \delta z),\\ &\leq \mathbb{P}(U > \beta) = 1-\Phi(\beta).\end{aligned}$$

Thus $\sup_z \mathbb{E}[(1-Y)(1-D)|Z=z] \leq 1-\Phi(\beta)$.

Inequality (3.13)

$$\begin{aligned}\mathbb{P}(U \leq \alpha+\beta, U \leq \beta) &= \mathbb{P}(U \leq \alpha+\beta, U \leq \beta, D=1|Z=z) + \mathbb{P}(U \leq \alpha+\beta, U \leq \beta, D=0|Z=z),\\ &\leq \mathbb{P}(U \leq \alpha+\beta, D=1|Z=z) + \mathbb{P}(U \leq \beta, D=0|Z=z),\\ &= \mathbb{P}(Y=1, D=1|Z=z) + \mathbb{P}(Y=1, D=0|Z=z),\end{aligned}$$

where the first equality holds from Assumption 7 and the law of total probability. Therefore

$$\Phi(\min(\alpha+\beta, \beta)) \leq \inf_z \mathbb{E}[YD + Y(1-D)|Z=z]$$

since $\mathbb{P}(U \leq \alpha+\beta, U \leq \beta) = \Phi(\min(\alpha+\beta, \beta))$ under Assumption 8.

Inequality (3.14)

$$\begin{aligned}\mathbb{P}(U > \alpha+\beta, U > \beta) &= \mathbb{P}(U > \alpha+\beta, U > \beta, D = 1|Z = z) + \mathbb{P}(U > \alpha+\beta, U > \beta, D = 0|Z = z),\\ &\leq \mathbb{P}(U > \alpha+\beta, D = 1|Z = z) + \mathbb{P}(U > \beta, D = 0|Z = z),\\ &= \mathbb{P}(Y = 0, D = 1|Z = z) + \mathbb{P}(Y = 0, D = 0|Z = z).\end{aligned}$$

Thus $1 - \Phi(\max(\alpha+\beta, \beta)) \leq \inf_z \mathbb{E}[(1-Y)D + (1-Y)(1-D)|Z = z]$. □

3.11.3 Proof of Proposition 3

Proof. We show the validity of the bounds in Proposition 3, and propose a joint distribution on (Y_0, Y_1) that achieves the lower bound on Y_1 and the upper bound on Y_0, and vice versa. First, we combine the condition $Y_1 \geq Y_0$ with Assumption 7. Results for condition $Y_1 \leq Y_0$ are obtained symmetrically by defining $Y = 1 - Y$, $Y_1 = 1 - Y_1$, and $Y_0 = 1 - Y_0$. Define the correspondence G between the unobservables (Y_0, Y_1) and the observables (Y, D):

$$\begin{aligned}G\{(0,0)\} &= \{(0,0),(0,1)\},\\ G\{(0,1)\} &= \{(0,0),(1,1)\},\\ G\{(1,1)\} &= \{(1,0),(1,1)\}.\end{aligned}$$

Let $p_{ij} \equiv \mathbb{P}(Y_0 = i, Y_1 = j)$. By Galichon and Henry (2011, Theorem 1), we have that all restrictions on the unconditional joint distribution of (Y_0, Y_1) and the marginals of Y_0 and Y_1 are given by: for all $A \subset \{(0,0),(0,1),(1,0),(1,1)\}$,

$$\begin{aligned}\mathbb{P}((Y, D) \in A|Z = z) &\leq \mathbb{P}(G(Y_0, Y_1) \cap A \neq \emptyset|Z = z)\\ &= \mathbb{P}(G(Y_0, Y_1) \cap A \neq \emptyset),\ \forall\ z \in \mathcal{Z},\end{aligned}$$

that is:

for singletons,

$$\mathbb{P}(Y=1,D=1|Z=z)\leq p_{01}+p_{11}, \tag{3.20}$$

$$\mathbb{P}(Y=0,D=1|Z=z)\leq p_{00}, \tag{3.21}$$

$$\mathbb{P}(Y=1,D=0|Z=z)\leq p_{11}, \tag{3.22}$$

$$\mathbb{P}(Y=0,D=0|Z=z)\leq p_{00}+p_{01}; \tag{3.23}$$

for pairs,

$$\mathbb{P}(Y=1,D=1|Z=z)+\mathbb{P}(Y=1,D=0|Z=z)\leq p_{11}+p_{01}, \tag{3.24}$$

$$\mathbb{P}(Y=1,D=1|Z=z)+\mathbb{P}(Y=0,D=1|Z=z)\leq 1, \tag{3.25}$$

$$\mathbb{P}(Y=1,D=1|Z=z)+\mathbb{P}(Y=0,D=0|Z=z)\leq 1, \tag{3.26}$$

$$\mathbb{P}(Y=1,D=0|Z=z)+\mathbb{P}(Y=0,D=1|Z=z)\leq p_{00}+p_{11}, \tag{3.27}$$

$$\mathbb{P}(Y=1,D=0|Z=z)+\mathbb{P}(Y=0,D=0|Z=z)\leq 1, \tag{3.28}$$

$$\mathbb{P}(Y=0,D=1|Z=z)+\mathbb{P}(Y=0,D=0|Z=z)\leq p_{00}+p_{01}; \tag{3.29}$$

for triplets,

$$\mathbb{P}(Y=1,D=1|Z=z)+\mathbb{P}(Y=1,D=0|Z=z)+\mathbb{P}(Y=0,D=1|Z=z)\leq 1,$$

$$\mathbb{P}(Y=1,D=1|Z=z)+\mathbb{P}(Y=1,D=0|Z=z)+\mathbb{P}(Y=0,D=0|Z=z)\leq 1,$$

$$\mathbb{P}(Y=1,D=1|Z=z)+\mathbb{P}(Y=0,D=1|Z=z)+\mathbb{P}(Y=0,D=0|Z=z)\leq 1,$$

$$\mathbb{P}(Y=1,D=0|Z=z)+\mathbb{P}(Y=0,D=1|Z=z)+\mathbb{P}(Y=0,D=0|Z=z)\leq 1.$$

Inequalities (3.25)-(3.26), (3.28), and all inequalities for the triplets are redundant. Inequality (3.24) implies (3.20), inequality (3.29) implies (3.23), and inequalities (3.21)-(3.22) imply (3.27). Hence, the set of non-redundant inequalities are inequalities (3.21), (3.22), (3.24) and (3.29). Since $p_{01}+p_{11}=\mathbb{E}[Y_1]$, (3.24) implies

$$\sup_z \mathbb{E}[Y|Z=z]\leq \mathbb{E}[Y_1].$$

Since $p_{00} = 1 - (p_{01} + p_{11}) = 1 - \mathbb{E}[Y_1]$, inequality (3.21) imples

$$\mathbb{E}[Y_1] \leq \inf_z \mathbb{E}[1 - (1 - Y)D|Z = z].$$

Since $p_{00} + p_{01} = \mathbb{P}(Y_0 = 0) = 1 - \mathbb{E}[Y_0]$, inequality (3.29) implies

$$\mathbb{E}[Y_0] \leq \inf_z \mathbb{E}[Y|Z = z].$$

Finally, since $p_{11} = 1 - (p_{00} + p_{01}) = \mathbb{E}[Y_0]$, inequality (3.22) implies

$$\sup_z \mathbb{E}[Y(1 - D)|Z = z] \leq \mathbb{E}[Y_0].$$

Therefore, we obtain the bounds in Equation (3.18).

To show sharpness, suppose that $\sup_z \mathbb{E}[Y(1 - D)|Z = z] \leq \inf_z \mathbb{E}[Y|Z = z]$, and $\sup_z \mathbb{E}[Y|Z = z] \leq \inf_z \mathbb{E}[1 - (1 - Y)D|Z = z]$. Define $\tilde{p}_{11} = \inf_z \mathbb{E}[Y|Z = z]$, $\tilde{p}_{00} = \inf_z \mathbb{E}[1 - Y|Z = z]$, and $\tilde{p}_{01} = 1 - \inf_z \mathbb{E}[1 - Y|Z = z] - \inf_z \mathbb{E}[Y|Z = z]$. Those quantities are well-defined probabilities. Indeed, $\tilde{p}_{11} + \tilde{p}_{01} + \tilde{p}_{00} = 1$, $\tilde{p}_{11} \geq 0$, $\tilde{p}_{00} \geq 0$, and $\tilde{p}_{01} = \sup_z \mathbb{E}[Y|Z = z] - \inf_z \mathbb{E}[Y|Z = z] \geq 0$. Since $\tilde{p}_{10} = 0$ by definition, we have $Y_1 \geq Y_0$ a.s.. Define $\mathbb{P}(\tilde{Y}_1 = i, \tilde{Y}_0 = j|Z = z) = \tilde{p}_{ij}$. Then, Assumption 7 holds. Moreover, we have $\mathbb{E}[\tilde{Y}_1] = \tilde{p}_{11} + \tilde{p}_{01} = \sup_z \mathbb{E}[Y|Z = z]$, and $\mathbb{E}[\tilde{Y}_0] = \tilde{p}_{11} + \tilde{p}_{10} = \inf_z \mathbb{E}[Y|Z = z]$. Therefore, the lower bound for $\mathbb{E}[Y_1]$ in Equation (3.18) is achieved, while the upper bound for $\mathbb{E}[Y_0]$ is achieved by the same joint distribution.

On the other hand, define $\tilde{p}_{11} = \sup_z \mathbb{E}[Y(1 - D)|Z = z]$, $\tilde{p}_{00} = \sup_z \mathbb{E}[(1 - Y)D|Z = z]$, and $\tilde{p}_{01} = 1 - \sup_z \mathbb{E}[Y(1 - D)|Z = z] - \sup_z \mathbb{E}[(1 - Y)D|Z = z]$. As before, we can check that those quantities are well-defined probabilities. Indeed, $\tilde{p}_{11} + \tilde{p}_{01} + \tilde{p}_{00} = 1$, $\tilde{p}_{11} \geq 0$, $\tilde{p}_{00} \geq 0$, and

$$\tilde{p}_{01} = \inf_z \mathbb{E}[1 - (1 - Y)D|Z = z] - \sup_z \mathbb{E}[Y(1 - D)|Z = z] \geq \sup_z \mathbb{E}[Y|Z = z] - \inf_z \mathbb{E}[Y|Z = z] \geq 0,$$

where the second inequality follows from the conditions

$$\sup_z \mathbb{E}[Y(1 - D)|Z = z] \leq \inf_z \mathbb{E}[Y|Z = z], \text{ and } \sup_z \mathbb{E}[Y|Z = z] \leq \inf_z \mathbb{E}[1 - (1 - Y)D|Z = z].$$

Since $\tilde{p}_{10} = 0$ by definition, we have $Y_1 \geq Y_0$ a.s.. Define $\mathbb{P}(\tilde{Y}_1 = i, \tilde{Y}_0 = j|Z = z) = \tilde{p}_{ij}$. Then, Assumption 7 holds. Moreover, we have $\mathbb{E}[\tilde{Y}_1] = \tilde{p}_{11} + \tilde{p}_{01} = \inf_z \mathbb{E}[1 - (1 - Y)D|Z = z]$, and

$\mathbb{E}[\tilde{Y}_0] = \tilde{p}_{11} + \tilde{p}_{10} = \sup_z \mathbb{E}[Y(1-D)|Z=z]$. Therefore, the upper bound for $\mathbb{E}[Y_1]$ in Equation (3.18) is achieved, while the lower bound for $\mathbb{E}[Y_0]$ is achieved by the same joint distribution.

Now, suppose $Y_1 \leq Y_0$ a.s., and define $Y = 1 - Y$, $Y_1 = 1 - Y_1$, and $Y_0 = 1 - Y_0$. Then, $Y_1 \geq Y_0$, and $Y = Y_1 D + Y_0(1-D)$. Using the previous results, we have the following sharp bounds for Y_1 and Y_0:

$$\sup_z \mathbb{E}[Y|Z=z] \leq \mathbb{E}[Y_1] \leq \inf_z \mathbb{E}[1-(1-Y)D|Z=z],$$

$$\sup_z \mathbb{E}[Y(1-D)|Z=z] \leq \mathbb{E}[Y_0] \leq \inf_z \mathbb{E}[Y|Z=z].$$

After rewriting these inequalities in terms of Y, Y_1, and Y_0, we obtain the bounds in (3.19).

Finally, we propose the following joint distribution on $(\tilde{Y}_0, \tilde{Y}_1, D)$ given Z, which is compatible with the data (Y, D, Z), and satisfies Assumptions 7 and 10:

$$\mathbb{P}(\tilde{Y}_0 = 0, \tilde{Y}_1 = 0, D = 1|Z=z) = \mathbb{P}(Y = 0, D = 1|Z=z),$$

$$\mathbb{P}(\tilde{Y}_0 = 0, \tilde{Y}_1 = 0, D = 0|Z=z) = \tilde{p}_{00} - \mathbb{P}(Y = 0, D = 1|Z=z),$$

$$\mathbb{P}(\tilde{Y}_0 = 0, \tilde{Y}_1 = 1, D = 0|Z=z) = \mathbb{P}(Y = 0, D = 0|Z=z) - \tilde{p}_{00} + \mathbb{P}(Y = 0, D = 1|Z=z),$$

$$\mathbb{P}(\tilde{Y}_0 = 1, \tilde{Y}_1 = 1, D = 0|Z=z) = \mathbb{P}(Y = 1, D = 0|Z=z),$$

$$\mathbb{P}(\tilde{Y}_0 = 0, \tilde{Y}_1 = 1, D = 1|Z=z) = \tilde{p}_{01} + \tilde{p}_{00} - \mathbb{P}(Y = 0|Z=z),$$

$$\mathbb{P}(\tilde{Y}_0 = 1, \tilde{Y}_1 = 1, D = 1|Z=z) = \mathbb{P}(Y = 1, D = 1|Z=z) - \tilde{p}_{01} - \tilde{p}_{00} + \mathbb{P}(Y = 0|Z=z),$$

$$\mathbb{P}(\tilde{Y}_0 = 1, \tilde{Y}_1 = 0, D = 0|Z=z) = 0,$$

$$\mathbb{P}(\tilde{Y}_0 = 1, \tilde{Y}_1 = 0, D = 1|Z=z) = 0.$$

□

3.12 Appendix B: Additional remarks

Remark 7. *There exist two extra inequalities that are redundant. They are:*

$$\Phi(\beta) - \Phi(\min(\beta + \alpha, \beta)) \leq \inf_z \mathbb{E}[(1-Y)D + Y(1-D)|Z=z], \tag{3.30}$$

$$\Phi(\beta + \alpha) - \Phi(\min(\beta + \alpha, \beta)) \leq \inf_z \mathbb{E}[YD + (1-Y)(1-D)|Z=z]. \tag{3.31}$$

Inequality (3.30)

$$\begin{aligned}\mathbb{P}(U > \alpha+\beta, U \le \beta) &= \mathbb{P}U > \alpha+\beta, U \le \beta, D=1|Z=z) + \mathbb{P}(U > \alpha+\beta, U \le \beta, D=0|Z=z),\\ &\le \mathbb{P}(U > \alpha+\beta, D=1|Z=z) + \mathbb{P}(U \le \beta, D=0|Z=z),\\ &= \mathbb{P}(Y=0, D=1|Z=z) + \mathbb{P}(Y=1, D=0|Z=z).\end{aligned}$$

Thus $\Phi(\beta) - \Phi(\min(\beta+\alpha, \beta)) \le \inf_z \mathbb{E}[(1-Y)D + Y(1-D)|Z=z]$.

Inequality (3.31)

$$\begin{aligned}\mathbb{P}(U \le \alpha+\beta, U > \beta) &= \mathbb{P}(U \le \alpha+\beta, U > \beta, D=1|Z=z) + \mathbb{P}(U \le \alpha+\beta, U > \beta, D=0|Z=z),\\ &\le \mathbb{P}(U \le \alpha+\beta, D=1|Z=z) + \mathbb{P}(U > \beta, D=0|Z=z),\\ &= \mathbb{P}(Y=1, D=1|Z=z) + \mathbb{P}(Y=0, D=0|Z=z).\end{aligned}$$

Thus $\Phi(\alpha+\beta) - \Phi(\min(\alpha+\beta, \beta)) \le \inf_z \mathbb{E}[YD + (1-Y)(1-D)|Z=z]$.

We can get(3.30) by combining (3.9) and (3.12).

From (3.12) we have:

$\sup_z \mathbb{E}[1-D-Y+YD|Z=z] = \sup_z \mathbb{E}[(1-Y)(1-D)|Z=z] \le 1-\Phi(\beta)$.

Thus, $\sup_z \mathbb{E}[-D-Y+YD|Z=z] \le -\Phi(\beta)$ *or* $\inf_z \mathbb{E}[D+Y-YD|Z=z] \ge \Phi(\beta)$.

From (3.9) we know that $\sup_z \mathbb{E}[YD|Z=z] \le \Phi(\beta+\alpha)$.

We have

$$\begin{aligned}\inf_z \mathbb{E}[(1-Y)D + Y(1-D)|Z=z] &= \inf_z \mathbb{E}[(1-Y)D + Y - YD|Z=z]\\ &\ge \inf_z \mathbb{E}[(1-Y)D + Y|Z=z] + \inf_z \mathbb{E}[-YD|Z=z]\\ &= \inf_z \mathbb{E}[D - YD + Y|Z=z] - \sup_z \mathbb{E}[YD|Z=z]\\ &\ge \Phi(\beta) - \Phi(\beta+\alpha)\end{aligned}$$

where in the first equality we just expand the interior of the expectation, in the first inequality and second equality we use the properties of inf *and* sup. *In the third inequality we use (3.12) and (3.9).*

Furthermore, we have $\inf_z \mathbb{E}[(1-Y)D + Y(1-D)|Z=z] \ge 0 = \Phi(\beta) - \Phi(\beta)$. *Hence (3.30) holds whenever (3.12) and (3.9) hold.*

Similarly (3.31) can be obtained by combining (3.10) and (3.11).

3.13 Appendix C: Validity of the plug-in approach

The proofs for the validity of the plug-in approach for (3.9)-(3.12) and of (3.6-3.8) once they are converted into inequalities are similar. So, we only show the proof for (3.9).

Proof. Define $\mathbb{E}[YD|Z=z]=g(z)$. Let $\hat{g}(z)$ be a nonparametric estimator for $g(z)$ with convergence rate $\sqrt{nh}$ where $h \longrightarrow 0$ as $n \longrightarrow \infty$, such that:

$$\sqrt{nh}(\hat{g}(z)-g(z)) \xrightarrow{d} N(0,V(z)).$$

If $\alpha,\ \beta$ are known, we will have:

$$\sqrt{nh}[(\hat{g}(z)-\Phi(\alpha+\beta))-(g(z)-\Phi(\alpha+\beta))]=\sqrt{nh}[\hat{g}(z)-g(z)] \xrightarrow{d} N(0,V(z))$$

But, actually we do not know $\alpha,\ \beta$, and we replace them by their MLE estimators $\hat{\alpha},\ \hat{\beta}$. We need to show that the above asymptotic distribution is not affected by this plug-in approach. We have

$$\sqrt{nh}[(\hat{g}(z)-\Phi(\hat{\alpha}+\hat{\beta}))-(g(z)-\Phi(\alpha+\beta))]=\sqrt{nh}[\hat{g}(z)-g(z)]-\sqrt{h}[\sqrt{n}(\Phi(\hat{\alpha}+\hat{\beta})-\Phi(\alpha+\beta))]$$

We know that the first part $\sqrt{nh}[\hat{g}(z)-g(z)] \xrightarrow{d} N(0,V(z))$. Since $\alpha,\ \beta$ are estimated by MLE, we know that: $\Phi(\hat{\alpha}+\hat{\beta})-\Phi(\alpha+\beta)=O_p(n^{-\frac{1}{2}})$. Then, $\sqrt{n}(\Phi(\hat{\alpha}+\hat{\beta})-\Phi(\alpha+\beta))]=O_p(1)$, and since $h=o(1)$, it is clear that $\sqrt{h}=o_p(1)$. Hence,
$\sqrt{h}[\sqrt{n}(\Phi(\hat{\alpha}+\hat{\beta})-\Phi(\alpha+\beta))]=o_p(1)\times O_p(1)=o_p(1)$ (product rule). Therefore, by the asymptotic equivalence lemma, $\sqrt{nh}[(\hat{g}(z)-\Phi(\hat{\alpha}+\hat{\beta}))-(g(z)-\Phi(\alpha+\beta))] \xrightarrow{d} N(0,V(z))$. □

We are now going to show that the plug-in approach works for inequality (3.13). Similar reasoning holds for inequality (3.14).

Proof. Define $\theta(z)=\mathbb{E}[Y|Z=z]$ and $\hat{\theta}(z)$ is a nonparametric estimator for $\theta(z)$ such that

$$\sqrt{nh}\left\{\hat{\theta}(z)-\theta(z)\right\} \longrightarrow N(0,\Omega_1(z))$$

where $h \longrightarrow 0$ as $n \longrightarrow \infty$. We want to show that

$$\sqrt{nh}\left\{\left(\Phi(\min(\hat{\beta}+\hat{\alpha},\hat{\beta}))-\hat{\theta}(z)\right)-(\Phi(\min(\beta+\alpha,\beta))-\theta(z))\right\} \longrightarrow N(0,\Omega_1(z)).$$

It suffices to show that $\sqrt{nh}\left\{\Phi(\min(\hat{\beta}+\hat{\alpha},\hat{\beta}))-\Phi(\min(\beta+\alpha,\beta))\right\}=o_p(1)$, from the asymptotic equivalence lemma. For any $\delta>0$, we have:

$$\begin{aligned}
&\mathbb{P}\left(\sqrt{nh}\left\{\Phi(\min(\hat{\beta}+\hat{\alpha},\hat{\beta}))-\Phi(\min(\beta+\alpha,\beta))\right\}>\delta\right)\\
&\quad=\mathbb{P}\left(\sqrt{nh}\left\{\Phi(\min(\hat{\beta}+\hat{\alpha},\hat{\beta}))-\Phi(\min(\beta+\alpha,\beta))\right\}>\delta,\hat{\alpha}\geq 0,\alpha\geq 0\right)\\
&\qquad+\mathbb{P}\left(\sqrt{nh}\left\{\Phi(\min(\hat{\beta}+\hat{\alpha},\hat{\beta}))-\Phi(\min(\beta+\alpha,\beta))\right\}>\delta,\hat{\alpha}\leq 0,\alpha\leq 0\right)\\
&\qquad+\mathbb{P}\left(\sqrt{nh}\left\{\Phi(\min(\hat{\beta}+\hat{\alpha},\hat{\beta}))-\Phi(\min(\beta+\alpha,\beta))\right\}>\delta,\hat{\alpha}>0,\alpha<0\right)\\
&\qquad+\mathbb{P}\left(\sqrt{nh}\left\{\Phi(\min(\hat{\beta}+\hat{\alpha},\hat{\beta}))-\Phi(\min(\beta+\alpha,\beta))\right\}>\delta,\hat{\alpha}<0,\alpha>0\right),\\
&\quad\leq\mathbb{P}\left(\sqrt{nh}\left\{\Phi(\hat{\beta})-\Phi(\beta)\right\}>\delta\right)\\
&\qquad+\mathbb{P}\left(\sqrt{nh}\left\{\Phi(\hat{\beta}+\hat{\alpha})-\Phi(\beta+\alpha)\right\}>\delta\right)\\
&\qquad+\mathbb{P}\left(\hat{\alpha}>0,\alpha<0\right)\\
&\qquad+\mathbb{P}\left(\hat{\alpha}<0,\alpha>0\right)
\end{aligned}$$

By the delta method $\sqrt{n}\left\{\Phi(\hat{\beta})-\Phi(\beta)\right\}=O_p(1)$ and $\sqrt{n}\left\{\Phi(\hat{\beta}+\hat{\alpha})-\Phi(\beta+\alpha)\right\}=O_p(1)$. Since $\sqrt{h}=o(1)$, we conclude using the product rule that $\sqrt{nh}\left\{\Phi(\hat{\beta})-\Phi(\beta)\right\}=o_p(1)$ and $\sqrt{nh}\left\{\Phi(\hat{\beta}+\hat{\alpha})-\Phi(\beta+\alpha)\right\}=o_p(1)$. Therefore, $\mathbb{P}\left(\sqrt{nh}\left\{\Phi(\hat{\beta})-\Phi(\beta)\right\}>\delta\right)\longrightarrow 0$ as $n\longrightarrow\infty$, and $\mathbb{P}\left(\sqrt{nh}\left\{\Phi(\hat{\beta}+\hat{\alpha})-\Phi(\beta+\alpha)\right\}>\delta\right)\longrightarrow 0$ as $n\longrightarrow\infty$. It remains to show that $\mathbb{P}\left(\hat{\alpha}>0,\alpha<0\right)$ and $\mathbb{P}\left(\hat{\alpha}<0,\alpha>0\right)$ go to zero as n goes to zero. We have

$$\begin{aligned}
\hat{\alpha}>0,\alpha<0&\Longrightarrow\exists\ \epsilon_1,\ \epsilon_2:\ \hat{\alpha}\geq\epsilon_1>0,\ \alpha\leq\epsilon_2<0,\\
&\Longrightarrow\hat{\alpha}-\alpha\geq\epsilon_1-\epsilon_2>0.
\end{aligned}$$

Then $\mathbb{P}\left(\hat{\alpha}>0,\alpha<0\right)\leq P\left(\hat{\alpha}-\alpha\geq\epsilon_1-\epsilon_2\right)\longrightarrow 0$ as $n\longrightarrow\infty$, because $\hat{\alpha}$ is a consistent estimator for α. By a similar argument, $\mathbb{P}\left(\hat{\alpha}<0,\alpha>0\right)\longrightarrow 0$ as $n\longrightarrow\infty$. □

3.14 Appendix D: Further Extensions

3.14.1 Adding exogenous covariates

Suppose we have the following specification:

$$\begin{cases} Y &= \mathbb{1}\{\alpha D + \beta' X - U \geq 0\} \\ \\ D &= \mathbb{1}\{\delta Z + \lambda' X - V \geq 0\} \end{cases} \tag{3.32}$$

The testable implications in Proposition 1 become

$$\mathbb{E}\left[YD|Z=z, X=x\right] = \Phi_\rho(\beta' x + \alpha, \lambda' x + \delta z), \tag{3.33}$$

$$\mathbb{E}\left[Y(1-D)|Z=z, X=x\right] = \Phi(\beta' x) - \Phi_\rho(\beta' x, \lambda' x + \delta z), \tag{3.34}$$

$$\mathbb{E}\left[(1-Y)D|Z=z, X=x\right] = \Phi(\lambda' x + \delta z) - \Phi_\rho(\beta' x + \alpha, \lambda' x + \delta z). \tag{3.35}$$

The testable implications in Proposition 2 become

$$\begin{array}{lcl} \sup_z \mathbb{E}[YD|Z=z, X=x] & \leq & \Phi(\alpha + \beta' x), \\ \\ \sup_z \mathbb{E}[Y(1-D)|Z=z, X=x] & \leq & \Phi(\beta' x), \\ \\ \sup_z \mathbb{E}[(1-Y)D|Z=z, X=x] & \leq & 1 - \Phi(\alpha + \beta' x), \\ \\ \sup_z \mathbb{E}[(1-Y)(1-D)|Z=z, X=x] & \leq & 1 - \Phi(\beta' x), \\ \\ \Phi(\min(\alpha + \beta' x, \beta' x)) & \leq & \inf_z \mathbb{E}[YD + Y(1-D)|Z=z, X=x], \\ \\ 1 - \Phi(\max(\alpha + \beta' x, \beta' x)) & \leq & \inf_z \mathbb{E}[(1-Y)D + (1-Y)(1-D)|Z=z, X=x]. \end{array} \tag{3.36}$$

3.14.2 Extension to generalized bivariate models

Suppose we still have the model (3.1) but instead of Assumption 7 and 8 we have:

Assumption 11. *Z is independent of (U, V).*

Assumption 12. *F_U and F_V are known marginal distributions of U and V, respectively, that are strictly increasing, are absolutely continuous with respect to Lebesgue measure, and such that $E[U] = E[V] = 0$ and $Var(U) = Var(V) = 1$.*

Assumption 13. *$(U, V) \sim F_{UV}(U, V) = C(F_U(U), F_V(V); \rho)$ where $C(,,;\rho)$ is a copula known up to scalar parameter $\rho \in \Omega$ such that $C : (0,1)^2 \to (0,1)$ is twice differentiable in its arguments and ρ*

We can repeat the procedure in section 3.4.2 to get the next equalities relying on the fact that the copula is known and that the marginals are absolutely continuous and stritcly increasing.

$$\mathbb{E}[YD|Z = z] = \mathbb{P}(U \leq \beta + \alpha, V \leq \gamma + \delta z) \tag{3.37}$$

$$= P(F_U(U) \leq F_U(\beta + \alpha), F_V(V) \leq F_V(\gamma + \delta z)) \tag{3.38}$$

$$= C(F_U(\beta + \alpha), F_V(\gamma + \delta z), \rho), \tag{3.39}$$

$$\mathbb{E}[Y(1 - D)|Z = z] = \mathbb{P}(U \leq \beta, V > \gamma + \delta z) \tag{3.40}$$

$$= \mathbb{P}(F_U(U) \leq F_U(\beta), F_V(V) > F_V(\gamma + \delta z)) \tag{3.41}$$

$$= F_U(\beta) - C(F_U(\beta), F_V(\gamma + \delta z), \rho), \tag{3.42}$$

$$\mathbb{E}[(1 - Y)D|Z = z] = \mathbb{P}(U > \beta + \alpha, V \leq \gamma + \delta z) \tag{3.43}$$

$$= \mathbb{P}(F_U(U) > F_U(\beta + \alpha), F_V(V) \leq F_V(\gamma + \delta z)) \tag{3.44}$$

$$= F_V(\gamma + \delta z) - C(F_U(\beta + \alpha), F_V(\gamma + \delta z), \rho). \tag{3.45}$$

We can repeat the procedure in section 3.11 to get the next inequalities relying on the fact that the copula is known. Here we assume w.l.o.g. that $\gamma = \beta = 0$, but as shown in the previous

displays the results carry when this simplification is not made. Then, we get:

$$\begin{aligned}\mathbb{E}[YD|Z=z] &= \mathbb{P}(Y=1, D=1|Z=z),\\ &= \mathbb{P}(U\leq\alpha, V\leq\delta z|Z=z),\\ &\leq \mathbb{P}(U\leq\alpha|Z=z),\\ &= \mathbb{P}(U\leq\alpha) = F_U(\alpha)\end{aligned}$$

Thus $\sup_z \mathbb{E}[YD|Z=z] \leq F_U(\alpha)$.

$$\begin{aligned}\mathbb{E}[Y(1-D)|Z=z] &= \mathbb{P}(Y=1, D=0|Z=z),\\ &= \mathbb{P}(U\leq 0, V>\delta z|Z=z),\\ &\leq \mathbb{P}(U\leq 0) = F_U(0).\end{aligned}$$

Thus $\sup_z \mathbb{E}[Y(1-D)|Z=z] \leq F_U(0)$

$$\begin{aligned}\mathbb{E}[(1-Y)D|Z=z] &= \mathbb{P}(Y=0, D=1|Z=z),\\ &= \mathbb{P}(U>\alpha, V\leq\delta z),\\ &\leq P(U>\alpha) = 1-F_U(\alpha).\end{aligned}$$

Thus $\sup_z \mathbb{E}[(1-Y)D|Z=z] \leq 1-F_U(\alpha)$.

$$\begin{aligned}\mathbb{E}[(1-Y)(1-D)|Z=z] &= \mathbb{P}(Y=0, D=0|Z=z),\\ &= \mathbb{P}(U>0, V>\delta z),\\ &\leq \mathbb{P}(U>0) = 1-F_U(0).\end{aligned}$$

Thus $\sup_z \mathbb{E}[(1-Y)(1-D)|Z=z] \leq 1-F_U(0)$.

$$\begin{aligned}\mathbb{P}(U\leq\alpha, U\leq 0) &= \mathbb{P}(U\leq\alpha, U\leq 0, D=1|Z=z) + \mathbb{P}(U\leq\alpha, U\leq 0, D=0|Z=z),\\ &\leq \mathbb{P}(U\leq\alpha, D=1|Z=z) + \mathbb{P}(U\leq 0, D=0|Z=z),\\ &= \mathbb{P}(Y=1, D=1|Z=z) + \mathbb{P}(Y=1, D=0|Z=z).\end{aligned}$$

Thus $F_U(\min(\alpha, 0)) \leq \inf_z \mathbb{E}[YD + Y(1-D)|Z=z]$.

$$\begin{aligned}\mathbb{P}(U > \alpha, U > 0) &= \mathbb{P}(U > \alpha, U > 0, D = 1|Z = z) + \mathbb{P}(U > \alpha, U > 0, D = 0|Z = z),\\ &\leq \mathbb{P}(U > \alpha, D = 1|Z = z) + \mathbb{P}(U > 0, D = 0|Z = z),\\ &= \mathbb{P}(Y = 0, D = 1|Z = z) + \mathbb{P}(Y = 0, D = 0|Z = z).\end{aligned}$$

Thus $1 - F_U(\max(\alpha, 0)) \leq \inf_z \mathbb{E}[(1 - Y)D + (1 - Y)(1 - D)|Z = z]$

Now that we have these more general equalities and inequalities, building upon Han and Lee (2019) we can extend our test to situations where the marginal distributions are known, and the copula structure has certain conditions. Furthermore, if there are exogenous covariates X (thus combining section 3.14.1 with section 3.14.2) and if the distributions of X and Z are absolutely continuous with respect to Lebesgue measure, the assumption of known marginals can be dropped as discussed in Han and Vytlacil (2017) section 6.

3.15 Appendix E: Additional results for the application

3.15.1 Summary Statistics

This section includes extra summary statistics for the empirical examples.

Table 3.8 Summary Statistics for empirical example 1

	Total	
Observations	7,555	
Doctor visit	0.1820	(0.3859)
Private insurance	0.6567	(0.4749)
Nb employees (standardized)	-0.0054	(0.9970)

Average and standard deviation (in parentheses)

Table 3.9 Summary Statistics for empirical example 2

	Total	
Observations	1,659	
Adoption	0.7077	(0.4550)
Rent	0.0934	(0.2911)
Mig-ag-ratio	6.7027	(11.5441)

Average and standard deviation (in parentheses)

3.15.2 Implementation: Chernozhukov, Lee, and Rosen (2013) conditions and commands

3.15.2.1 Algorithm

Let β_n denotes the coefficients of the series terms if series estimation is used or of the local polynomial. K denotes the dimension of β_n and I_k denotes the K dimensional identity matrix. Let $p_n(z) = \frac{\partial \theta_n(z,\hat{\beta}_n)}{\partial \beta_n}$. Inference is conducted in the following way:

1. Set $\tilde{\gamma}_n \equiv 1 - 0.1/log(n)$. Recall in our proccedure $\mathcal{J} : 1..j..6$. Simulate a large number R of draws denoted $Z_1, ...Z_R$ from the K-variate standard normal distribution $N(0, I_K)$.

2. Compute $\hat{\Omega}_n$ a consistent estimator for the large sample variance of $\sqrt{n}(\hat{\beta}_n - \beta_n)$.

3. For each $z, j \in \mathcal{Z}, \mathcal{J}$, compute $\hat{g}(z,j) = p_n(z,j); \hat{\Omega}_n^{1/2}$ and set $\hat{s}(z,j) = ||\hat{g}(z,j)||/\sqrt{n}$.

4. Compute $k(\tilde{\gamma}_n) = \tilde{\gamma}_n$-quantile of $\{\sup_{z,j\in\mathcal{Z},\mathcal{J}}(\hat{g}(z,j)'Z_r/||\hat{g}(z,j)||)$ for $r = 1, ...R\}$ and $\widehat{Z,J}_n = \{z, j \in \mathcal{Z}, \mathcal{J} : \hat{\theta}_n(z,j) \geq \min_{z,j\in\mathcal{Z},\mathcal{J}}[\hat{\theta}_n(z,j) - k(\tilde{\gamma}_n)\hat{s}(z,j)] - 2k(\tilde{\gamma}_n)\}$

5. Compute $k_{1-\alpha} = 1 - \alpha$ quantile of $\{\sup_{z,j\in\widehat{Z,J}_n}(\hat{g}(z,j); Z_r/||\hat{g}(z,j))||$ for $r = 1, ..., R\}$ Then set: $\hat{\theta}_{1-\alpha} = \sup_{z,j\in\mathcal{Z},\mathcal{J}}[\hat{\theta}(z,j) - k_{1-\alpha}\hat{s}(z,j)]$.

3.15.2.2 Commands

Find below the code for the implementation of the test in Stata for the sharp implications.

use " C:/Users/acerenza/application/HLData.DTA", clear

gen $Y = Doctorvisit$

gen $D = Privateins$

gen $Z = Stdnbemp$

biprobit $(Y = D)$ $(D = Z)$

matrix define $C = e(b)$

scalar $rho = e(rho)$

gen $ldepen1 = Y * D - binormal(C[1,1] + C[1,2], C[1,3] * Z + C[1,4], rho)$

gen $ldepen2 = Y * (1 - D) + normal(C[1,2]) - binormal(C[1,2], C[1,3] * Z + C[1,4], rho)$

gen $ldepen3 =$

$(1 - Y) * D + normal(C[1,3] * Z + C[1,4]) - binormal(C[1,1] + C[1,2], C[1,3] * Z + C[1,4], rho)$

gen $udepen1 = -(Y * D - binormal(C[1,1] + C[1,2], C[1,3] * Z + C[1,4], rho))$

gen $udepen2 = -(Y * (1 - D) + normal(C[1,2]) - binormal(C[1,2], C[1,3] * Z + C[1,4], rho))$

gen $udepen3 =$

$-((1-Y) * D + normal(C[1,3] * Z + C[1,4]) - binormal(C[1,1] + C[1,2], C[1,3] * Z + C[1,4], rho))$

*** Define the range for the instruments

$centile(Z)$, $centile(199)$

scalar $LBZ = r(c_1)$

scalar $UBZ = r(c_2)$

sum Z

gen $RZ = LBZ + n * (UBZ - LBZ)/200$

replace $RZ =$. if $n > 200$

clrbound (ldepen1 Z RZ) (ldepen2 Z RZ) (ldepen3 Z RZ) (udepen1 Z RZ) (udepen2 Z RZ) (udepen3 Z RZ), low met("series") level(0.5 0.9 0.95 0.99) norseed rnd(20000)

Find below the code for the implementation of the test in Stata for the nonsharp implications.

use " C:/Users/acerenza/application/HLData.DTA", clear

biprobit (Doctorvisit=Privateins) (Privateins= Stdnbemp)

matrix define $C = e(b)$

gen $Y = Doctorvisit$

gen $D = Privateins$

gen $ldepen1 = Y * D - normal(C[1,1] + C[1,2])$

gen $ldepen2 = Y * (1 - D) - normal(C[1,2])$

gen $ldepen3 = (1 - Y) * D - 1 - normal(C[1,1] + C[1,2])$

gen $ldepen4 = (1 - Y) * (1 - D) - 1 - normal(C[1,2])$

gen $ldepen5 = normal(\min(C[1,1] + C[1,2], C[1,2])) - Y$

gen $ldepen6 = 1 - normal(\max(C[1,1] + C[1,2], C[1,2])) - (1 - Y)$

*** Define the range for the instruments gen $Z = Stdnbemp$

$centile(Z)$, $centile(199)$

scalar $LBZ = r(c_1)$

scalar $UBZ = r(c_2)$

sum Z

gen $RZ = LBZ + n * (UBZ - LBZ)/200$

replace $RZ = .$ if $n > 200$

clrbound (ldepen1 Z RZ) (ldepen2 Z RZ) (ldepen3 Z RZ) (ldepen4 Z RZ) (ldepen5 Z RZ) (ldepen6 Z RZ) , low met("local") level(0.5 0.9 0.95 0.99) norseed rnd(20000)

CHAPTER 4. ASYMPTOTIC THEORY FOR M-ESTIMATORS UNDER CLUSTERING AND WITH MISSING DATA

Santiago Acerenza

Department of Economics, Iowa State University, Ames, IA, 50011, USA

Modified from a manuscript to be submitted to *Oxford Bulletin of Economics and Statistics*

4.1 Abstract

This paper provides identification and inference results for a general M-estimator in the presence of missing data and cluster data dependence in the selection variables. The clustered sample selection mechanism is a consequence of the fact that the sample's selection depends on individuals on the same cluster. First, we achieve identification of the parameter of interest defined as the solution to a population maximization problem extending results from Wooldridge (2002). Secondly, we formalize conditions for an M-estimator with missing data to be consistent when the number of clusters increases and the size of the clusters remain fixed. We then characterize the inverse probability weighting estimator's asymptotic properties that address the clustered sample selection mechanism. An additional result is providing the form of the inconsistency induced in the parameters of interest by ignoring the dependence in the inverse probability weighting estimation. We provide evidence of the finite sample properties of the bias using simulations.

4.2 Introduction

This paper develops identification, consistency, and asymptotic inference results for M-estimators in the context of clustering and sample selection happening at the same time. The inference results are achieved using weighting procedures. We identify the parameter of interest

using Inverse Probability weighting (IPW). Then we show the consistency of two stage M-estimators with clustering in this context and the correct asymptotic variance form. Thirdly, we show the existence and specify the form of bias induced on the estimate of the parameter of interest when cluster-specific heterogeneity in the selection is ignored. Previous literature focused on situations in which data dependence in the selection mechanism in a very restricted form. This paper contributes to that literature by considering a more general structure and a more general cluster dependence structure. The fundamental contribution of this work is looking at dependence at the selection stage for a general setting such as the one provided by M-estimation.

Sample selection and clustering can have effects over the correct way of doing inference and on the identification of population parameters of interest, also if they are occurring together, which has not been studied in the literature, but it is a common phenomenon in real-life data. For example, in randomized field experiments, individuals are clustered in villages, and some treatment effect is of interest (for example, the effect of information on protection against sexually transmitted diseases), but the estimates would be biased if sample selection is not taken into account. Also, in those settings, the probability of observing an individual is correlated with the other village members, which also affects the proper way of correcting the selection process.

Economic and social agents do not make random choices about consumption, health, and work decisions. As a consequence, observations in empirical studies are in many cases selected so that they are not independent of the outcome variables. In this context, the outcome variable (and the decision to participate in the survey) are related across individuals. Sample selection leads to invalid inference and/or lack of identification of parameters of interest. Non-random selection is a source of bias in empirical research and a fundamental aspect of many economic, social, and medical processes. Researchers of different areas are concerned about drawing proper conclusions from the data, but it is often the case that random samples are not available and results are biased because of sample selection. Solon et al. (2015) provides relevant examples where bias arises due to selection in the context of commuting and returns to schooling. Suppose one studies commuters' choice of transit mode (such as the choice between driving to work or

taking the bus). In that case, one might be interested in how certain explanatory variables (such as bus fare and walking distance from bus stops) affect the probability of choosing one model versus the other. Researchers would estimate a probit or logit model for the binary choice between transit modes given random sampling. But in many cases, the sample is not drawn as a random sample of commuters but as a choice-based sample. Suppose the survey was conducted studying the mode choice for work trips at the station or parking lots instead of their homes. In that case, this might over-represent one mode and under-represents the other relative to the population distribution of commuting choices. Thus, conventional estimation generally results in inconsistent parameter estimation. If one cares about returns to schooling, one will frame their analysis within a linear regression of log earnings on years of schooling with controls for other variables such as years of work experience. If the regression were estimated with the full Panel Study of Income Dynamics (PSID) without any correction for the oversampling of the low-income population, this would lead to inconsistent estimation of the regression parameters. The sampling would be endogenous because the sampling criterion, family income, is related to the error term in the regression for log earnings. Another relevant example is when the equation of interest is a wage offer equation for the population of all adults of working age. Nevertheless, the wage offer is observed only for working adults; using a sample of working people to estimate the wage offer equation may result in the inconsistent estimation of the population wage offer function if the observations are not properly weighted. See Wooldridge (2002).

Following the seminal work of Little and Rubin (2002), and Vella (1998), the ways of solving missing data problems are imputation methods, likelihood-based approaches, and weighting approaches. Among the weighting approaches, there is weighting by the inverse of the probability of being observed. This will be the approach adopted here, which is widely used in medical and social sciences. As stated in Li et al. (2013) IPW relies on the intuitive idea of creating a pseudo-population of weighted copies of the complete sample to remove the selection.

In addition to selection, researchers might be dealing with data that was not collected at an individual level. Maybe the institution conducting the survey did not make a random draw of

individuals working on companies, but they did a random draw of companies and interviewed everyone inside that company. This is what is known as sampling at a group or cluster level. One can also think about this in a different way, which is the usual concept of clustering. This consists of individuals with some correlated outcomes within a group, but individuals of different groups are independent. In this second case, the clustering is not by the sampling but occurs because of some underlying characteristics of individuals or natural grouping (such as villages in randomized control trials). For example, one can do a random sample of individuals in neighborhoods and be interested in the perception of security of the neighborhoods. Their characteristics might share a form of dependence that is strongly driven by the underlying population relationship (decide to live in the same neighborhood) rather than how the sample was collected. Clustering is a well-known feature of data sets, and recently, many methods have been developed for the correct use of these clusters to extract information from samples. For a detailed discussion on clustering and more examples of scenarios with clustering, see Abadie et al. (2017), Asri et al. (2016) and references therein.

4.2.1 Literature review

4.2.1.1 Literature related to IPW and missing data

Following Wooldridge (1999), Wooldridge (2001), Wooldridge (2002) and Wooldridge (2007) we will deal with the setup in which the outcome of interest is observed only for a selected part of the sample, while the covariates are fully observed for the whole sample. Similarly, we use IPW as the solution to the selection. Differently from Wooldridge (1999) and Wooldridge (2001) we deal with sample selection and clustering instead of stratification or variable probability samples. Differently from Wooldridge (2002) and Wooldridge (2007) we focus on a sample selection correction that accounts for the possibility of clustering, which was not accounted for in the paper mentioned above. Our paper is also closely related to Negi (2019) that develops a new class of IPW M-estimators that deal with non-random assignment and missing outcomes by combining propensity score weighting with weighting for missing data. Differently from this work Negi

(2019) is interested in recovering the causal impact of a non-random treatment, so she introduces a weighting scheme that accounts for the endogeneity of the treatment variable and the bias due to the missing data patterns. This paper assumes the treatment is not suffering from endogeneity (and thus without missing data problems, the parameter of interest is identified in the population) and focuses on the missing data when there is the possibility of clustering, which is not considered in Negi (2019). Missing data has also been treated in a general method of moments(GMM) estimation as in Prokhorov and Schmidt (2009). In contrast with Prokhorov and Schmidt (2009) we focus on an M-estimation framework and allow for the possibility of cluster dependence in the selection stage.

4.2.1.2 Literature related to clustering

The literature on clustering is well established. Some of the closest related papers to ours are Rahmani (2019) that deals with a multistage sampling design including stratification and clustering under an M-estimators framework but not estimating the probability weights and without dealing with clustering like the current paper. Rahmani (2019) does not focus on clustering at the first stage and also does not model the clustering in the sample selection stage as an endogenous choice individuals make based on their group membership as this paper does. Also, we are not dealing with stratification as covered in Rahmani´s paper. Clustering without missing data in an M-estimation context has been studied by Asri et al. (2016) and Xu (2019). In a GMM framework, there is, for example, Bhattacharya (2005). The current paper complements this literature by using a general framework such as the M-estimation in a context of clustering and differs from it by including a sample selection stage that can be explicitly modeled where the clustering is happening.

4.2.1.3 Literature related to clustering and sample selection jointly

Skinner and D 'Arrigo (2011) recover the estimate of the unconditional population average of a variable of interest in a linear index model, assuming normality of the errors in the probability

of being observed, and modeling the clustering by assuming it is an unobserved heterogeneity that behaves like a random effect. The setup used in the current work extends theirs by allowing for a more general estimation structure. This is obtained using M-estimation. Additionally, we are not using any distributional assumption of the error term and the unobserved heterogeneity.

Balan and Jankovic (2018) analyze non-response in longitudinal data using the Generalized Estimating Equations (GEE) setup to recover probability weights and derive the asymptotic theory of the estimators of the probability of the data missing. Nevertheless, they do not derive asymptotic results in a second stage like in this setting, where we give a formula for an asymptotic bias and the asymptotic variance of the parameters of interest. In their setting, the IPWs is the subject of interest itself and not inference about another parameter. Besides this, the first stage treatment is equivalent to ours with the difference that they are dealing with time dependence, and we are dealing with cluster dependence.

Semykina and Wooldridge (2018) consider estimating binary response models on an unbalanced panel, where the dependent variable is missing due to non-random selection. In that paper, they consider estimating sample selection models and treatment effects using a fully parametric approach, where the error distribution is assumed to be normal in both the outcome and selection equations. Then, they consider a semiparametric estimator of binary response panel data models with sample selection robust to a variety of error distributions. The estimator employs a control function approach similar to ours to account for endogenous selection and permits consistent estimation of scaled coefficients and relative effects. Finally, Semykina and Wooldridge (2010) consider estimation of panel data models with sample selection when the equation of interest contains endogenous explanatory variables and unobserved heterogeneity. Assuming that appropriate instruments are available, they propose two estimation procedures that correct for selection in the presence of endogenous regressors. The first correction procedure is parametric and is valid under the assumption that the errors in the selection equation are normally distributed. The second procedure estimates the model parameters semiparametrically using series estimators. The error terms may be heterogeneously distributed and serially

dependent in both selection and primary equations in the proposed correction procedures. In this paper, we deal with M-estimation, and our results hold without imposing distributional assumptions on the unobserved components. Also, we are using parametric estimation procedures (M-estimators) that are more general than the approaches taken in Semykina and Wooldridge (2010) parametric analysis. Yet, Semykina and Wooldridge (2018) deal with binary models, which can only be fitted into our framework by imposing extra distributional assumptions on the errors, but their results are only valid for binary models. Last but not least, our procedures are not accounting for the possibility of endogenous regressors.

4.2.2 Outline of the paper

The rest of the paper organized as follows: Section 4.3 presents the main setup, section 4.4 introduces, identification, consistency and asymptotic variance results, section 4.5 the bias results, section 4.6 some simulation results regarding the bias due to inconsistent estimation of the first stage and Section 4.7 concludes. Section 4.9 contains the proofs to the propositions in the text. Section 4.10 contains the tables from the simulation exercises.

4.3 Setup

We are interested in recovering a feature of the population, like the mean of income conditional on educational level. This can be represented by a parametric model, indexed by a vector of parameters. This will be a vector $\theta \subset \Theta \subset R^P$. This vector solves the population problem:

$$\min_{\theta \subset \Theta} \quad E[q(w, \theta)] \tag{4.1}$$

Where $q(.)$ is a known criterion that is minimized. Let w_{ci} be the realization of the random variable w for observation i in cluster c. It is worth mentioning that cluster dependence arises in the selection stage not in the outcome equation. Define
$Q(\mathbf{w_c}, .) = \begin{bmatrix} q(w_{c1}, .) & q(w_{c2},).... & q(w_{cN}, .) \end{bmatrix}^T$ as a vector that contains the realization of the

criterion function for the N individuals in cluster[1] c. Define 1_N as a vector of ones of dimension N. The population element related to the sample analog $Q(\mathbf{w_c}, .)$ is: $Q(\mathbf{w}, \theta)$, where $\mathbf{w}$ is a vector of size N of w. More precisely: $Q(\mathbf{w}, .) = \begin{bmatrix} q(w_1, .) & q(w_2, .) .... & q(w_N, .) \end{bmatrix}^T$, where w is an $M \times 1$ random vector taking values in a subset of R^M and c is omitted here because it is the element of the population not an element from the random sampled clusters c. Here we are introducing the clustering with clusters of size N. We will observe a random sample of clusters, each of one containing a fixed number of individuals that are allowed to have a certain dependence structure.[2] In this context, $\boldsymbol{w}$ is a random vector which elements are $w_1, ..., w_i, ...w_N$ and $\boldsymbol{w}_c$ is a sampled cluster which elements are $w_{c1}, ..., w_{ci}, ...w_{cN}$.

We focus on recovering the solution to 4.1 when there is sample selection/missing data. We first make a standard identification assumption in the M-estimators context.

Assumption 14. θ_0 *uniquely solves 4.1*

Following Wooldridge (2002), we define indicator variables that take the value 1 when the unit is observed and 0 otherwise. In this sample selection scenario, covariates will always be observed, but not the outcome variable. For example, we might have a question on test scores in a survey of students in different classrooms, and some of them may choose not to answer this question. Yet, we do observe their demographics. Thus, these observations are missing in the relevant question but not in other covariates that might be of interest themselves. In this context, answering or not might be more likely depending on the classroom (thus allowing for clustering in the sample selection stage).

Extending the indicator variable idea to this setup, we can say that if we have a random sample of C independent clusters each of size N (where N is the total number of individuals in that cluster or a random sample of size N of the individuals inside that cluster) the observed

[1] This paper focuses in clusters with the same amount of observations N but results can be extended relying on Hansen and Lee (2019) to different amount of observations in all the clusters

[2] We will stack the sample analogs of q in this way to explicitly take into account for the potential correlation structure inside a cluster/group, that's why we introduce $Q(.)$ instead of minimizing just $q(.)$ We will use the law of large numbers and central limit theorems at a cluster level. Along the rest of this paper bold notation denotes random vectors where each element has the same mean and any element will be indexed by subscript "i", when we introduce the sampling at a cluster level we will add subscript "c" in front of "i".

sample moment that is related to our object of interest 4.1 taking into account the fact that there is sample selection is:

$$\min_{\theta \subset \Theta} \quad \frac{\sum_{c=1}^{C} \sum_{i=1}^{N} s_{ci} q(w_{ci}, \theta)}{NC} = \min_{\theta \subset \Theta} \quad \frac{\sum_{c=1}^{C} \mathbf{s_c} Q(\mathbf{w_c}, \theta) \frac{1}{N}}{C} \tag{4.2}$$

where s_{ci} is an indicator function that takes value 1 if the outcome for sampled individual i inside cluster c is observed, and $\mathbf{s_c}$ is a vector of N sampled individuals in cluster c and $\mathbf{s_c} = \begin{bmatrix} s_{c1} & s_{c2} .... & s_{cN} \end{bmatrix}$. The random vector version of $\mathbf{s_c}$ is $\mathbf{s} = \begin{bmatrix} s_1 & s_2 .... & s_N \end{bmatrix}$ Where $\mathbf{s}$ and $\mathbf{s_c}$ are multiplying element by element of $Q(.)$. By the analogy principle and certain regularity conditions (that will be treated below), as C grows, this problem converges to [3]:

$$\min_{\theta \subset \Theta} \quad E[sq(w, \theta)] \tag{4.3}$$

Without further assumptions the solution to 4.3 is not identifying θ_0. We are interested in recovering the solution to 4.1, θ_0. To do this, first, we will impose conditions specified in assumption 15 below related to the behavior of the selection vector. Note here the selection is playing a role combined with clustering. In the presence of clustering, the selection of any given individual in a cluster is correlated with selecting the rest of the individuals in that cluster. This leads to the necessity of considering the dependence structure to get a corrected estimator that will consistently estimate the parameter of interest.

Remark 8. *Note that the usual sample analog of the population object of interest is* $\min_{\theta \subset \Theta} \quad \frac{\sum_{c=1}^{C} \sum_{i=1}^{N_1} q(w_{ci}, \theta)}{N_1 C}$. *With* $N_1 \leq N$ *due to the fact that some people are not being observed in the sample.*

Example-OLS *: In the standard OLS case,* $q(w, \theta) = \frac{(y - x\theta)^2}{2}$. *Where* $w = (y, x) \subset R^{p+1}$. *Then the population minimization problem is:*

$$\min_{\theta \subset \Theta} \quad E[\frac{(y - x\theta)^2}{2}] \tag{4.4}$$

[3] Note that the cluster size N is a constant, and we are not using it for asymptotics, so it is not affecting the optimization.

Suppose $N = 2$, that is, all clusters have size 2. Then, $Q(\mathbf{w}, .) = \left[\frac{(y_1 - x_1\theta)^2}{2}, \quad \frac{(y_2 - x_2\theta)^2}{2}\right]^T$. For a sampled cluster c, Q would be: $Q(\mathbf{w_c}, .) = \left[\frac{(y_{c1} - x_{c1}\theta)^2}{2}, \quad \frac{(y_{c2} - x_{c2}\theta)^2}{2}\right]^T$. The sample analog version of the population minimization problem:

$$\min_{\theta \subset \Theta} \quad \frac{\sum_{c=1}^{C}\sum_{i=1}^{2} s_{ci}\frac{(y_{ci} - x_{ci}\theta)^2}{2}}{2C} = \min_{\theta \subset \Theta} \quad \frac{\sum_{c=1}^{C} \mathbf{s_c}\left[\frac{(y_{c1} - x_{c1}\theta)^2}{2}, \quad \frac{(y_{c2} - x_{c2}\theta)^2}{2}\right]^T \frac{1}{2}}{C} \tag{4.5}$$

This is the estimator when you have two observations per cluster, but you do not observe all the individuals,so for some clusters there is only one person observed, others are not observed at all.

Next we will specify an assumption for for the selection mechanism that generates the vector s that allows us to evaluate dependence, efficiency concerns and use a two step M-estimation methods to recover the parameter of interest.

Assumption 15.

$$E(\mathbf{s}|\mathbf{z}) = P(\mathbf{s}|\mathbf{z}) = \mu(\mathbf{z}, \beta), \tag{4.6}$$

In the previous assumption $\beta \subset R^{M+1}$ is a vector of unknown parameters and $\mu(\mathbf{z}, .) = \left[\mu(z_1, h(\mathbf{z}), .) \quad \mu(z_2, h(\mathbf{z}), .) .... \quad \mu(z_N, h(\mathbf{z}), .)\right]$. Let z be an $M \times 1$ random vector taking values in a subset of R^M and the bold version be an array of size N of this z. Let $\mu(.)$ be some known function that models the relationship between each element of $\mathbf{s}$ and the explanatory variables $\mathbf{z}$. Here $h(\mathbf{z})$ is explicitly allowing that each probability is potentially affected by the values of the covariates for the individuals inside the cluster. Under assumption 15, each element of a cluster has the probability of being observed depending on its characteristics and on the observed characteristics of other people in the cluster. This assumption is on the same spirit as the work of Skinner and D 'Arrigo (2011) and Vazquez-Bare (2018). The idea is to find a sufficient statistic for the network effect. In those two works, the sufficient statistic is usually the number of members on a particular network (or cluster); here, we assume the sufficient statistic is knowledge about all members' covariates. Assumption 15 is a form of strict exogeneity applied to the clustering context expanding these two studies. Similar to strict exogeneity for the SUR or

panel data case, it means that the expectation of the error term on the probabilities (not the outcomes) conditional in all the covariates of individuals inside the cluster is zero. Intuitively, this means that once we control for the characteristics of all individuals in that cluster, the selection mechanism is properly accounted for. This is similar to the idea introduced by Mundlak (1978) in a panel data context. Instead of modeling the common unobserved heterogeneity that would be part of the error, we directly model it as a function of covariates of everyone in the cluster. Conditional to this function, an individual's selection behavior is independent of each other.

Assumption 15 stated that the vector $\mathbf{s}$ of random variables s, has error components that are mean independent of $\mathbf{z}$. We also know the functional form of the probability of observing $\mathbf{s}$, this is $\mu(.)$. Misspecification of $\mu(.)$ would lead to inconsistent estimation of parameters on both the selection stage and the structural model. This will be discussed in section 4.5.

It is worth noting that inconsistency of the parameters θ, β would arise only if the dependence on observables is ignored, but inference is affected by cluster dependence in both stages of the estimation so it should be taken into account accordingly.

***Example continued**: $\mu(z_i, h(\mathbf{z}))$ can be for example the logistic link. $h(\mathbf{z})$ can be the sample average of the observations in any cluster so $\frac{z_1+z_2}{2}$. $\beta = (\beta_1, \gamma) \subset R^{p+2}$ Then, for any member of the cluster we have: $\mu(z_i, h(\mathbf{z})) = \frac{1}{1+e^{-(z_i\beta_1 + \frac{z_i+z_{-i}}{2}\gamma)}}$. Then*
$\mu(\mathbf{z}, \beta) = \left[\frac{1}{1+e^{-(z_1\beta_1 + \frac{z_1+z_2}{2}\gamma)}}, \quad \frac{1}{1+e^{-(z_2\beta_1 + \frac{z_2+z_1}{2}\gamma)}} \right]$.

4.4 Main results

4.4.1 Identification

To recover the parameters of interest that would be obtained by solving 4.1 if there was no sample selection, we extend the IPW approach in Wooldridge (2002) to take into account the cluster dependence in both the selection mechanism and outcomes. Instead of minimizing $E(\boldsymbol{s}Q(\boldsymbol{w}, .))$, we minimize a re-weighted version of this object, specifically: $E(\boldsymbol{s}\boldsymbol{p}^T Q(\boldsymbol{w}, .))$, which identifies the population parameter of interest. Here T stands for transpose.

Lemma 1 extends Wooldridge (2002) Lemma 3.1 for random vectors instead of scalar random variables allowing for cluster dependence.

Define

$$\mu(\mathbf{z},\beta) \equiv \left[\mu(z_1,h(\mathbf{z}),\beta) \quad \mu(z_2,h(\mathbf{z}),\beta).... \quad \mu(z_N,h(\mathbf{z}),\beta)\right] \equiv \left[p(s_1,\beta), \quad p(s_2,\beta).... \quad p(s_N,\beta)\right]$$

Lemma 1. *Under assumptions 14-15, and assuming $P(\mathbf{s}|\mathbf{z})$ is positive definite and for any real value function such that $E[\mathbf{sp^T}|Q(\mathbf{w},.)|]$ is finite, where:*

- $\mathbf{sp^T} = \left[s_1/p(s_1,\beta), \quad s_2/p(s_2,\beta).... \quad s_N/p(s_N,\beta)\right]$
- $|Q(.)|$ *is the euclidean norm of* $Q(.)$

Then:

$$E[\mathbf{sp^T}Q(\mathbf{w},.)] = E[q(w_1,.) + ... + q(w_N,.)]$$

Note that if inside each cluster observations do not present cluster dependence outside the selection stage then, $E[\frac{1}{N}\mathbf{sp^T}Q(\mathbf{w},.)] = \frac{1}{N}E[q(w_1,.) + ... + q(w_N,.)] = E[q(w_i,.)]$ which is the element we want to recover. Thus optimizing $E[\frac{1}{N}\mathbf{sp^T}Q(\mathbf{w},.)]$ is the same as optimizing $E[q(w_i,.)]$.

The proof of this lemma and remaining propositions can be found in Section 4.9.

This result extends the usual IPW identification of the population expectation to allow for clustering. This result combined with assumption 14 allows us to identify θ_0.

Example continued:

$$\mathbf{sp^T}Q(.) = \left[\frac{s_1}{1+e^{-(z_1\beta_1+\frac{z_1+z_2}{2}\gamma)}}, \quad \frac{s_2}{1+e^{-(z_2\beta_1+\frac{z_2+z_1}{2}\gamma)}}\right] \cdot \left[\frac{(y_1-x_1\theta)^2}{2}, \quad \frac{(y_2-x_2\theta)^2}{2}\right]^T$$

Lemma 1 provides the basis for identification in this setup.

4.4.2 Estimation

Following the analogy principle, the proposed estimator is given by the sample analogue of $\min_\theta E[\mathbf{sp^T}Q(\mathbf{w},.)]$ which under the conditions in proposition 4 will allow us to recover θ_0 asymptotically. To implement the estimator, the propensity score needs to be estimated; we assume that we have a root-C consistent parametric estimator of the propensity score correctly,

taking into account the dependence as stated in assumption 15. In practical terms, we assume we know the functional form of the dependence among individuals inside a given group. More precisely, we know which function of the Z of other individuals inside my cluster is the one we need to include as a control.

With the results from lemma 1, we can now propose the cluster version of the IPW estimator. Let $\tilde{\beta}$ be a parametric estimate of β in $\mu(.,\beta)$. The IPW estimator is:

$$\min_{\theta\in\Theta} \quad \frac{\sum_{c=1}^{C}\sum_{i=1}^{N}\frac{s_{ci}}{\mu(z_{ci},\tilde{\beta})}q(w_{ci},\theta)}{NC} \tag{4.7}$$

The main technical contribution from the usual case is that both the probabilities and the optimization criterion are allowed to be correlated within a cluster.

We state two results that extend Wooldridge (2002) Theorem 3.1 and Theorem 4.1 to this setup. Proposition 4 shows consistency for the estimator proposed that deals with clustering and missing data. Proposition 5 is about the form of the asymptotic variance necessary to perform inference.

Proposition 4. *Let*

1. $\left\{\mathbf{s_c}=\begin{bmatrix}s_{c1} & s_{c2}.... & s_{cN}\end{bmatrix},\mathbf{w_c}=\begin{bmatrix}w_{c1} & w_{c2}.... & w_{cN}\end{bmatrix},\mathbf{z_c}=\begin{bmatrix}z_{c1} & z_{c2}.... & z_{cN}\end{bmatrix}\right\}_1^C$ *Be an independent random sample of C clusters each of size N.*

2. *Let all the w_i be such that $E(q(w_i)) = E(q(w_{i'}))\quad \forall i,i' \in N$.*

3. *Θ be a compact set.*

4. *For each w_i, $q(.)$ be continuous.*

5. *For each $\theta \in \Theta$, $q(.)$ be Lebesgue measurable.*

6. *$|g(.)| \leq b(w)$ where $E(b(w)) < \infty$. Also directly implied by this using the triangle inequality $|Q(.)| - B(\mathbf{w}) \leq 0$ where B is another function of w.*

7. *Conditions for Lemma 1 holds.*

8. $\tilde{\beta}$ *is a root-C consistent estimator of* β*.*

9. For some $\delta > 0, \quad p(s_{ci}|\mathbf{z}_c) \geq \delta \quad \forall \quad z_c$ *in a subset of* R^{2M}

Then, $\hat{\theta}$*, the solution to* $\min_{\theta \subset \Theta} \frac{\sum_{c=1}^{C}\sum_{i=1}^{N}\frac{s_{ci}}{\mu(z_{ci},\tilde{\beta})}q(w_{ci},\theta)}{NC}$ *converges in probability to* θ_0*, the solution of Equation 4.1.*

Condition 1 states the nature of the sampling process and is standard with the difference that the condition is stated at a cluster level instead of at an individual level. This is saying that instead of randomly sampling individuals, we take the random sample to the cluster level.

Condition 2 is needed in the cluster setting to recover from the clustered estimator $Q(.)$ the relevant parameter θ_0. The intuition behind this condition is that even though the marginal distribution of the random variables inside the cluster may be different, they must have the same mean for the researcher to recover and identify from any given sample the parameter of interest. [4] One can think of this intuitively as imposing a sort of mean stationary condition but for clustered data instead of a time series.

Conditions 3, 4, and 5 are common in the literature and serve as regularity conditions of the stochastic process. Condition 6 assures the existence of a uniform weak law of large numbers (UWLLN) for q and Q; the UWLLN is central for the consistency argument while continuity of q and compactness of the parameter space can be relaxed but with significant added technical effort. Condition 7 is needed for identification. Condition 8 states the existence of a consistent estimator of the propensity score. An empirical researcher intending to use this procedure and that wants to get a consistent estimator of the propensity score can use particular examples of the procedure proposed here. This type of procedures can be as the ones proposed in Liang and Zeger (1986a), Liang and Zeger (1986b), Zeger et al. (1988), Lee et al. (1993) or more recently Pereda-Fernández (2019). Condition 9 is needed for a UWLLN to hold for $\frac{\sum_{c=1}^{C}\sum_{i=1}^{N}\frac{s_{ci}}{\mu(z_{ci},\tilde{\beta})}q(w_{ci},\theta)}{NC}$.

The next is the proposition related to the asymptotic variance needed for inference.

[4] Condition 2 relaxes the usual stronger condition that inside a cluster, all individuals have the same distribution, allowing more generality. Assuming the same distribution would rule out heteroscedasticity of individuals inside the cluster.

Proposition 5. *Let*

1. *Conditions from Proposition 4 hold.*
2. θ_0 *be in the interior of* Θ.
3. $\nabla_\theta Q(\mathbf{w},\theta)1_n$ *be differentiable on the interior of* Θ. *Let* $s(\mathbf{w},\theta) \equiv \nabla_\theta Q(\mathbf{w},\theta)1_n$.
4. $|H(\mathbf{w},\theta)|$, *the absolute value of the Hessian of the objective function* $Q(.)\frac{1}{N}$ *be bounded by* $B(\mathbf{w})$ *with* $E(B(\mathbf{w})) < \infty$.
5. $E[H(\mathbf{w},\theta_0)]$ *and* $E(\mathbf{sp^T}H(\mathbf{w},\theta_0))$ *be non-singular.*
6. $E[s(\mathbf{w},\theta_0)] = 0$ *and each element of* $s(\mathbf{w},\theta)$ *has a finite second moment for every* $\theta \in \Theta$.
7. $E[\mathbf{sp^T}|H(\mathbf{w},\theta_0)|]$ *is finite.*
8. *Let the asymptotic variance of* $C^{1/2}(\tilde{\beta} - \beta)$ *exist and be finite.*
9. $\mu(\mathbf{z},\beta)$ *be continuous on the compact set* $\mathcal{B}$ *and twice continuously differentiable on the interior of* $\mathcal{B}$.
10. *Call the gradient with respect to* β *of* $\mathbf{sp}$, $\nabla\mathbf{sp} = \begin{bmatrix} -s_1\mu(z_1,h(\mathbf{z}),\beta)^{-2}\nabla_\beta\mu(z_1,h(\mathbf{z}),\beta), \\ -s_2\mu(z_2,h(\mathbf{z}),\beta)^{-2}\nabla_\beta\mu(z_2,h(\mathbf{z}),\beta), \\ ..... \\ -s_N\mu(z_N,h(\mathbf{z}),\beta)^{-2}\nabla_\beta\mu(z_N,h(\mathbf{z}),\beta) \end{bmatrix}$.

    and $\widehat{\nabla\mathbf{sp}}$ *the same but replacing* β *with the consistent estimator.*

Then, we can represent

$$C^{\frac{1}{2}}(\hat{\theta} - \theta_0) = -(E[H(\mathbf{w},\theta_0)])^{-1}(C^{\frac{-1}{2}}\sum_{c=1}^{C}\mathbf{sp^T_c}s(\mathbf{w_c},\theta_0) + (E[\nabla\mathbf{sp^T}s(\mathbf{w},\theta_0)])C^{\frac{1}{2}}(\tilde{\beta} - \beta)) + o_p(1)$$

This representation emphasizes the asymptotic relationship of the distribution of the parameter estimator on the second stage with the first stage. Also, that allows us to get our estimator's correct standard asymptotic representation when clustering is present.

Condition 2 and 3 are standard. They are needed if the outcome expectation is a nonlinear function; mean value expansions are used to determine the asymptotic behavior of the estimators that might not have a closed-form. Condition 4 and 5 are also standard and assure consistency of the sample analog of the hessian to the expected value of the hessian. Condition 6 is saying that the first-order condition of the population maximization problem is only zero at the true parameter. Is it some sense a restatement of assumption 14 in the context of a differentiable estimator. Condition 7 is needed, so Lemma 1 holds for the Hessian. Condition 8 states the existence of an estimator of the propensity score that is consistent and asymptotically normal. Condition 9 allows using a linear expansion around the propensity score. Notice $\mu(.,.)$ is a vector-valued function taking into account the clustering nature of the stochastic process. Condition 10 is just a definition that is useful and simplifies the notation of the proof.

Proposition 5 extends the results in Wooldridge (2002) and allows researchers to consider cluster data with sample selection when the clustering occurs. It also generalizes Skinner and D 'Arrigo (2011) as noted above.

Remark 9. ***The representation shows the connection between the first stage estimation and the second stage****, which will in-turn affect the variance. If one is interested in computing the variance, and the first stage is estimated with an M-estimator (where the Jacobian with respect to β for individual i in cluster c is denoted as $S_{ci}(\beta)$), we can see that the problem we are solving jointly is:*

$$\nabla_\theta \frac{\sum_{c=1}^{C}\sum_{i=1}^{N} \frac{s_{ci}}{\mu(z_{ci},\tilde{\beta})} q(w_{ci},\theta)}{NC} = 0,$$
$$\sum_{c=1}^{C}\sum_{i=1}^{N} S_{ci}(\beta) = 0,$$

The previous system determinate $\hat{\theta}, \hat{\beta}$, then under standard regularity conditions as the ones in the previous two propositions, doing a mean value expansion, using central limit theory and weak law of large numbers at a cluster level, assuming w_{ci}, z_{ci} are i.i.d. inside a given cluster once selection is accounted for, and applying standard results from Newey and McFadden (1994) we get that $C^{\frac{1}{2}}(\hat{\theta}-\theta, \hat{\beta}-\beta)$ is asymptotically normal with mean 0 and variance-covariance matrix $A^{-1}BA^{-1}$.

Where:

$$A = \begin{pmatrix} -E\Big(\nabla_\theta \nabla_\theta \frac{s}{\mu(z,\beta)} q(w,\theta)\Big) & -E\Big(\nabla_\beta \nabla_\theta \frac{s}{\mu(z,\beta)} q(w,\theta)\Big) \\ -E\Big(\nabla_\theta S(\beta)\Big) & -E\Big(\nabla_\beta S(\beta)\Big) \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}$$

$$B = \begin{pmatrix} Var\Big(\nabla_\theta \frac{s}{\mu(z,\hat{\beta})} q(w,\theta)\Big) & Cov\Big(\nabla_\theta \frac{s}{\mu(z,\hat{\beta})} q(w,\theta), S(\beta)\Big) \\ Cov\Big(\nabla_\theta \frac{s}{\mu(z,\hat{\beta})} q(w,\theta), S(\beta)\Big) & Var\Big(S(\beta)\Big) \end{pmatrix} = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

The following example illustrates an estimation procedure that fulfills the conditions of the proposition.

***Example**: Let $q(w,\theta) = \frac{(y - x\theta)^2}{2}$. Let s for any given individual in some cluster be $s_{ci} = 1(\beta_1 z_{ci} + \beta_2 \mathbf{z}_{-ci} - u \geq 0)$, where β_1 is a vector of coefficients associated to the individuals z and β_2 to the z of all others individuals in the cluster. Letting u have a logistic distribution we know: $P(s_{ci}|\mathbf{z_c}) = \frac{exp(\beta_1 z_{ci} + \beta_2 \mathbf{z}_{-ci})}{exp(\beta_1 z_{ci} + \beta_2 \mathbf{z}_{-ci}) + 1}$ which can be estimated consistently at root-C by standard maximum likelihood results. Based on this, one can propose the following estimator for θ:*

$$\min_{\theta \in \Theta} \quad \frac{\sum_{c=1}^{C} \sum_{i=1}^{N} \frac{s_{ci}}{\left(\frac{exp(\hat{\beta}_1 z_{ci} + \hat{\beta}_2 \mathbf{z}_{-ci})}{exp(\hat{\beta}_1 z_{ci} + \hat{\beta}_2 \mathbf{z}_{-ci}) + 1}\right)} \frac{(y_{ci} - x_{ci}\theta)^2}{2}}{NC}$$

Remark 10. *Results in this section show how inference needs to be adjusted to reflect clustering. In the next section, we show that even in some common cases, ignoring clustering could lead to improper inference and inconsistency of the estimator of the parameter of interest.*

4.5 Bias and Bias correction

This section considers the situation when the clustering in the first stage is not properly corrected. In this context, this is equivalent to omitting a relevant variable from the first stage. Such a situation can make assumption 15 to fail, leading to inconsistent estimates of the first stage parameters (and given the parametric first stage, inconsistent probability weights), and thus to an inconsistent second stage. In a sample selection context, omitting relevant variables from the other members in the cluster directly affect the probability of any given individual being in the sample.[5] This omission in the first stage affects the estimation of the parameter of interest.

[5] A relevant example that model the cluster dependence in the probability weighting stage is Zeger et al. (1988) that specifically models the unobserved heterogeneity as a random effect.

Once we formally state this, we will propose a method for correcting this asymptotic bias in our context. Here, we focus on the following situation:

Assumption 16.

$$\begin{aligned} \mathbf{s} &= \mu(\mathbf{z}, \beta) + \epsilon, \\ E(\epsilon|\mathbf{z}) &= \mathbf{0}, \end{aligned} \tag{4.8}$$

Nevertheless, we will have incorrectly specified $\mu(.)$ and have:

$$\begin{aligned} \mathbf{s} &= M(\mathbf{z}, \beta) + \xi, \\ E(\xi|\mathbf{z}) &\neq \mathbf{0}, \end{aligned} \tag{4.9}$$

Where

$\xi = \epsilon + [\mu(\mathbf{z}, \beta) - M(\mathbf{z}, \beta)]$.

In the spirit of White (1981), White (1982) one can think of this M as the misspecified function the researcher believes the underlying probability of selection takes. In this clustering context, we happen to omit the Z of other members of the cluster that specify the dependence of the probability of being observed of a given individual. This will be consistent with assumption 16 and violate 15. This would lead to an inconsistent estimation. An example would be specifying $s_i = 1[cz_i > \xi_i]$ when the correct model is $s_i = 1[cz_i + f(\mathbf{z}) > \epsilon_i]$. In this case, the exogeneity is violated because z_i is correlated with $\mathbf{z}$.

Inconsistency in the first stage generates inconsistency in the second stage.

Proposition 6. *1. The weak law of large numbers holds for the cluster averages of $\mathbf{sp_c^T H}(\mathbf{w_c}, \theta)$ and $\nabla_\beta \mathbf{sp_c^T s}(\mathbf{w_c}, \theta)$.*

2. Conditions for Lemma 1 hold.

3. $E[s(\mathbf{w}, \theta_0)] = 0$ and each element of $s(\mathbf{w}, \theta)$ has a finite second moment for every $\theta \in \Theta$.

4. $\hat{\beta}$ an estimator for the true β will be inconsistent of the form $plim(\hat{\beta} - \beta) = A$, A is $O_p(1)$ and different from zero.

Then: $\hat{\theta}$ is inconsistent and the asymptotic bias is:

$plim(\hat{\theta} - \theta_0) = [E[\widehat{\mathbf{sp^T}} H(\mathbf{w}, \bar{\theta})]^{-1}][E[\nabla_\beta \hat{\mathbf{sp^T}} s(\mathbf{w}, \theta_0) A] + o_p(1)$

Condition 1 is needed so we can characterize the asymptotic behavior of the estimator. If no law of large number holds then, we cannot get asymptotic properties. Condition 1 here is expressed; higher-order conditions could be stated and would be analogous of the Proposition 4 conditions 1, 3, 4, 5, and 6. Here we are not using a UWLLN, but the results would not change given we are just focusing on these particular elements of the parameter space Ω. Condition 2 is stated to show that the inconsistency arises even if the true parameter of interest is identified as long as the first stage is improperly dealt with. Condition 3 is needed to follow a usual M-estimation procedure of mean value expansion. Condition 4 is stated to get a finite form of the asymptotic bias. The inconsistency could potentially lead to infinity, but this would not let us get a finite form of the asymptotic bias. Under assumption 16 $\hat{\beta}$ an estimator for the true β will be inconsistent. Even though we can state conditions similar to the ones in Proposition 4 for an estimator of the β associated to $\mathbf{s} = M(\mathbf{z}, \beta) + \xi$ to be consistent, given the fact that $\mathbf{M} \neq \mu$, both would, in general, converge to different places and the form of the difference between the convergence points will in general not know.

Under these conditions in proposition 6, we see how inconsistency on the first stage affects the second stage. If $E[\nabla_\beta \mathbf{sp_c^T} s(.)] = 0$ then there is no inconsistency "pass-through". The intuition behind this condition is that if the nuisance parameters do not affect the determination of the parameter of interest, then inconsistencies in the estimation of these nuisance parameters will not affect the estimation of θ. See Newey and McFadden (1994) for details.

We are concerned in inverse probability weighting and as we show in Lemma 1 and Propositions 4-5 the nuisance parameters affect the parameter of interest estimation and identification, so $E[\nabla_\beta sp^T s(.)] \neq 0$. In a clustering context then, it will be relevant to take into account the way the dependence exists and interacts with the sample selection mechanism and outcomes. Dependence of individuals in a cluster can operate as an omitted variable that induces

inconsistency of the whole procedure. In our context, we chose to correct by using a Mundlak Device.

To provide some intuition of the relevance of this problem, imagine the following situation. A researcher is interested in measuring the effect of education on labor force participation of married individuals. But the sample has some selection since not all the individuals answer to the survey precisely because of the way they decide to spend their time. To correct this, the researcher uses IPW for each individual, including as controls age, gender, political preferences, and income. Each individual is married, so their choices to answer the survey may correlate with their partners' age, gender, and political preferences. In this sense, controlling for the other member of the marriage (the cluster) is relevant. Missing such an effect may lead to improperly solving the selection problem and thus not recovering the effect education has on labor force participation.

4.6 Simulations

This section presents simulation evidence related to the finite sample behavior of the bias of the previous section.

The data generating process (DGP) for the simulations will be the following one. For every individual i in a cluster c:

$y_{ci} = 1 + x_{1ci} + x_{2ci} + u_{ci}$

For each individual inside any cluster, y_{ci} is observed or not depending on the behaviour of s_{ci}, where:

$s_{ci} = 1[x_{2ci} + \bar{x}_{3c} + \bar{x}_{3c}^2 + \max_{i \in c} x_{3ci} + \epsilon_{ci} \geq 0]$

Where $x_{1ci} = b_{ci} \times x_{2ci} + v_{ci}, b_{ci}$ is Bernoulli(0.3); x_{2ci}, x_{3ci} are jointly Normal with correlation coefficient equal to -0.9; v_{ci}, u_{ci} are draws from a $N(0,1)$; $\bar{x}_{3c}$ is the average of variable x_3 for a given cluster; similarly $\bar{x}_{3c}^2$, but for x_{3ci}^2; $max_{i \in c} x_{3ci}$ is the maximum value of x_{3ci} across all individuals i in cluster c and ϵ_{ci} is drawn from a standard logistic distribution. $1[.]$ is the indicator function that is 1 if we observe y for that individual and 0 otherwise. The relationship between

$\bar{x}_{3c}$ $\bar{x}_{3c}^2$, $max_{i \in c} x_{3ci}$ and x_{2ci}, x_{3ci} is the key aspect of this simulation that allows a non-linear relationship between observations inside a cluster and the cluster membership itself.

Note that dependence is modeled via

$\bar{x}_{3c} + \bar{x}_{3c}^2 + max_{i \in c} x_{3ci}$.

An alternative way of doing this would be write:

$s_{ci} = 1[x_{2ci} + d_c + \epsilon_{ci} \geq 0]$

Where d_c is a common unobserved heterogeneity. In this simulation we are assuming:

$d_c = \bar{x}_{3c} + \bar{x}_{3c}^2 + max_{i \in c} x_{3ci}$

The researcher will estimate the parameters of the following equation[6], he is interest in the slope of x_{1ci}:

$y_{ci} = 1 + x_{1ci} + e_{ci}$

Where $e_{ci} = x_{2ci} + u_{ci}$.

As we care about the bias in the first stage that might be inducing bias in the second stage, we will propose two estimations of the propensity score used for weighting. The first one consists in doing a logit for:

$s_{ci} = 1[x_{2ci} + \omega_{ci} \geq 0]$

Where the true ω_{ci} is $\omega_{ci} = \bar{x}_{3c} + \bar{x}_{3c}^2 + max_{i \in c} x_{3ci} + \epsilon_{ci}$, but we will incorrectly assume that ω_{ci} is drawn from a standard logistic distribution.

In this case, omitting d_c would lead to inconsistency because of the correlation between x_{2ci} and x_{3ci}. Additionally, a logit only makes sense when the error is logistic. In this case the errors is the sum of $\bar{x}_{3c} + \bar{x}_{3c}^2 + max_{i \in c} x_{3ci} + \epsilon_{ci}$ which has a logistic component but the sum of this elements is not logistic.

After doing this logistic regression called the "naive" correction, we do another logit, the "Full" correction using the right equation. We compare the estimates of $y_{ci} = 1 + x_{1ci} + e_{ci}$ when we have no correction, the naive correction and the full correction.

[6] If the second stage is correctly specified, the selection does not affect the estimates for the slopes since no selection is not being controlled for.

Table 4.2 in section 4.10 shows mean squared errors of the estimated logistic regressions. We can see that for a small number of clusters, the full correction is imprecise. This is because we have fewer observations to estimate cluster-specific coefficients. Once we increase the sample size (which in this case consists of increasing the number of clusters), the full correction performs better than the naive one.[7]

Table 4.1 in section 4.10 shows the results for the estimates of the parameters of interest, the slope of the coefficient of x_1. We can see that the full correction always performs at least as well as the naive correction. In that sense, the correction is either reporting the same MSE as the naive correction or is improving it only slightly. The simulations also show that it is worse to correct improperly than to not correct at all.

4.7 Conclusions

In this paper, we provide results combining clustering, M-estimation, and missing data. We show the form of the inconsistency induced in the parameters of interest by not taking into account the dependence on the propensity score of individuals inside a cluster.

We extend an identification result from Wooldridge (2002) to a situation with dependence in the selection stage.

Thirdly, we formalize conditions for an M-estimator under clustering and missing data to be consistent when the number of clusters increases and the size of the clusters remain fixed. Results rely on random sampling of clusters and usual regularity conditions on the propensity score and the objective function. An extension of these results would be doing asymptotics using variability in size and growth of the clusters. This could be achieved by combining Xu (2019) and Hansen and Lee (2019) results with missing data. Another extension could be related to dependence among groups instead of independent clusters. We then characterize the form and behavior of the asymptotic variance of this estimator that involves IPW.

[7] Results are invariant to different choices of N.

Finally, we present simulations exploring the finite sample properties of the correction mechanism and see that for our DGP, the full correction always performs at least as good as the naive correction. In that sense, the correction either reports the same MSE as the naive correction or improves it, although only slightly.

4.8 References

Abadie, A.; Athey, S.; Imbens, G. and Wooldridge, J. When Should You Adjust Standard Errors for Clustering? *NBER working paper series*. 2017.

Asri, M.; Blanke, D. and Gabriel, E. Weighted M-estimators for multivariate clustered data. *Statistics and Probability Letters*,Vol 112. pages 26–34. 2016.

Balan, M. and Jankovic, D. Asymptotic theory for longitudinal data with missing responses adjusted by inverse probability weights. *Working paper*. 2018.

Bhattacharya, D. Asymptotic inference from Multi-stage samples. *Journal of Econometrics*,Vol 126.No 1. pages 145–171. 2005.

Hansen, B. and Lee, S. Asymptotic theory for clustered samples. *Journal of Econometrics*,Vol 210. No 2. pages 268–290. 2019.

Lee, A.; Scott, A. and Soo, S. Comparing liang-zeger estimates with maximum likelihood in bivariate logistic regression. *Journal of Statistical Computation and Simulation*,Vol 44. No 3-4. pages 133–148. 1993.

Liang, K. and Zeger, S. Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*,Vol 73. No 1. pages 13–22. 1986.

Liang, K. and Zeger, S. Data Analysis for Discrete and Continuous Outcomes. *Biometrika*,Vol 42. No 1. pages 121–130. 1986.

Li, L.; Shen, C.; Li, X. and Robins, J. On weighting approaches for missing data. *Statistical Methods in Medical Research*,Vol 22. No 1. pages 14–30. 2013.

Little, R. and Rubin, D. Statistical Analysis with missing data. *Wiley Series in Probability and Statistics*. 2002.

Mundlak, Y. On the pooling of time series and cross section data. *Econometrica*.Vol 46. pages 69–85. 1978.

Newey, W. and McFadden, D. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*.Vol 4. pages 2111–2245. 1994.

Negi, A. Doubly weighted M-estimation for nonrandom assignment and missing outcomes. *Unpublished*. 2019.

Pereda-Fernández, S. Copula-Based Random Effects Models for Clustered Data. *Journal of Business and Economic Statistics*, DOI: 10.1080/07350015.2019.1688665 . 2019.

Prokhorov , A. and Schmidt, P. GMM redundancy results for general missing data problems. *Journal of Econometrics*, Vol 142.No 1. pages 47–55. 2009.

Rahmani, I. Asymptotic Inference of M-Estimator from Multistage Samples with Variable Probability in the Final Stage. *Unpublished manuscript*.

Vazquez-Bare, G. Analysis of Spillover Effects in Randomized Experiments. *A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy (Economics) in the University of Michigan* . 2018

Semykina, A. and Wooldridge, J. Binary Response Panel Data Models with Sample Selection and Self Selection. *Journal of Applied Econometrics*, Vol 33. No 2. pages 179–197. 2018.

Semykina, A. and Wooldridge, J. Estimating panel data models in the presence of endogeneity and selection. *Journal of Econometrics*, Vol 157. No 2. pages 375–380. 2010.

Skinner, C. J. and D 'Arrigo, J. Inverse probability weighting for clustered non response. *Biometrika*, Vol 98. No 4. pages 953–966. 2011.

Solon, G. ; Haider,S. and Wooldridge, J. What Are We Weighting For? *The Journal of Human Resources*, Vol 50. No 2. pages 301–316. 2015.

Vella, F. Estimating Models with Sample Selection Bias: A Survey . *The Journal of Human Resources*, Vol 33.No 2. pages 127–169. 1998.

White, H. Consequences and Detection of Misspecified Nonlinear Regression Models. *Journal of the American Statistical Association*, Vol 76. pages 419–433. 1981.

White, H. Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, Vol 50. No 1. pages 1–25. 1982.

Wooldridge, J. Asymptotic properties of weighted M-estimators for variable probability samples. *Econometrica*, Vol 67. No 6. pages 1385–1406. 1999.

Wooldridge, J. Asymptotic properties of weighted M-estimators for standard stratified samples. *Econometric theory*, Vol 17. pages 451–470. 2001.

Wooldridge, J. Inverse probability weighted M-estimators for sample selection, attrition, and stratification. *Portuguese Economic Journal*, Vol 1.No 2. pages 117–139. 2002.

Wooldridge, J. Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, Vol 142.No 2. pages 1281–1301. 2007.

Xu, R. Asymptotic Properties of M-estimators with Finite Populations under Cluster Sampling and Cluster Assignment. *Unpublished*. 2019.

Zeger, S; Liang, K and Albert, P. Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics*, Vol 44.No 4. pages 1049–1060. 1988.

4.9 Appendix A: Proofs

Proof of Lemma 1: Given $E[\mathbf{sp^T}|Q(\mathbf{w},.)|]$ finite, we can apply law of iterated expectations $E[E[\mathbf{sp^T}Q(\mathbf{w},.)|\mathbf{z},\mathbf{w}]] = E[E[\mathbf{sp^T}|\mathbf{z},\mathbf{w}]E[Q(\mathbf{w},.)|\mathbf{w},\mathbf{z}]] = E[E[\mathbf{sp^T}|\mathbf{z}]E[Q(\mathbf{w},.)|\mathbf{w},\mathbf{z}]] = E[E[1_n][Q(\mathbf{w},.)|\mathbf{w},\mathbf{z}]] = E[1_nQ(\mathbf{w},.)] = E[q(w_1,.) + ... + q(w_N,.)]$

□

Where the first equality comes from the fact that **sp** only depends on **z** and Q is independent of **sp** conditional on the value of **z**. The second equality comes by the definition of **sp** that is a vector of each s_i divided by a constant that is their expectation. Then, applying expectation conditional on **z** gives that same constant for each element. Thus we have $p(s_1|\mathbf{z})/p(s_1|\mathbf{z})$. The last equality comes from the definition of conditional expectation and the definition of Q.

Proof of Proposition 4: Under previous conditions:
$\frac{\sum_{c=1}^{C}\sum_{i=1}^{N}\frac{s_{ci}}{\mu(z_{ci},\beta)}q(w_{ci},\theta)}{NC} = \frac{\sum_{c=1}^{C}\widehat{\mathbf{sp^T}}_{\mathbf{c}}Q(\mathbf{w_c},\theta)\frac{1}{N}}{C}$

Where $\widehat{\mathbf{sp}}$ replaces **sp** with the consistent estimator of the propensity score instead of the true propensity score. The subscript c denotes the sampled cluster version of the random vector **sp**
$\frac{\sum_{c=1}^{C}\widehat{\mathbf{sp^T}}_{\mathbf{c}}Q(\mathbf{w_c},\theta)\frac{1}{N}}{C} \leq \frac{\sum_{c=1}^{C}\frac{1}{\delta}b(w_c)}{C}$

Now recall conditions 5, 6 and 9 of Proposition 4, those imply: $E[\frac{b(w)}{\delta}] < \infty$

Also, $|Q(\mathbf{w},\theta)| - B(\mathbf{w}) \leq 0$

Then, making use of condition 7 of the Proposition 4 and Lemma 2.4 of Newey and McFadden (1994) we get that

$\frac{\sum_{c=1}^{C} \widehat{\mathbf{sp}}_{\mathbf{c}}^{\mathbf{T}} Q(\mathbf{w_c},\theta)\frac{1}{N}}{C}$ converges uniformly in probability to $E[\mathbf{sp}^{\mathbf{T}} Q(\mathbf{w},\theta)\frac{1}{N}]$. Then, by Lemma 1 this is equal to $E[1_n Q(\mathbf{w},\theta)\frac{1}{N}] = E[(q(w_1,.) + ... + q(w_N))\frac{1}{N}]$ then using the fact that the w_is have the same expected value $= E[(q(w,.) + ... + q(w))\frac{1}{N}] = E[q(w,.)]$. Then by Theorem 2.1 of Newey and McFadden (1994) we get that $\widehat{\theta}$ converges in probability to θ_0.

□

Proof of Proposition 5: The proof goes by usual arguments of M-estimators. First make a mean value expansion around θ_0 of $\frac{\sum_{c=1}^{C} \widehat{\mathbf{sp}^{\mathbf{T}}}_{\mathbf{c}} s(\mathbf{w_c},\hat{\theta})}{C} = \mathbf{0}$ and pre multiply by $C^{\frac{-1}{2}}$ to get:

$\mathbf{0} = C^{\frac{-1}{2}} \sum_{c=1}^{C} \widehat{\mathbf{sp}^{\mathbf{T}}}_{\mathbf{c}} s(\mathbf{w_c},\theta_0) + C^{-1} \sum_{c=1}^{C} \widehat{\mathbf{sp}^{\mathbf{T}}}_{\mathbf{c}} H(\mathbf{w_c},\bar{\theta}) C^{\frac{1}{2}}(\hat{\theta} - \theta_0)$

Where $\bar{\theta}$ is in the parameter space Θ and lies between θ and $\hat{\theta}$. Thus:

$C^{\frac{1}{2}}(\hat{\theta} - \theta_0) = (C^{-1} \sum_{c=1}^{C} \widehat{\mathbf{sp}^{\mathbf{T}}}_{\mathbf{c}} H(\mathbf{w_c},\bar{\theta}))^{-1}(-C^{\frac{-1}{2}} \sum_{c=1}^{C} \widehat{\mathbf{sp}^{\mathbf{T}}}_{\mathbf{c}} s(\mathbf{w_c},\theta_0))$

After using the fact that a) $\bar{\theta}$ is between the true and the estimated θ, b) under the conditions of Lemma 2.4 of Newey and McFadden (1994) holds and c) Lemma 1 holds:

$C^{-1} \sum_{c=1}^{C} \widehat{\mathbf{sp}^{\mathbf{T}}}_{\mathbf{c}} H(\mathbf{w_c},\bar{\theta})$ converges uniformly in probability to: $E[H(\mathbf{w},\theta_0)]$. Then, we can express:

$C^{\frac{1}{2}}(\hat{\theta} - \theta_0) = (E[H(\mathbf{w},\theta_0)])^{-1}(-C^{\frac{-1}{2}} \sum_{c=1}^{C} \widehat{\mathbf{sp}^{\mathbf{T}}}_{\mathbf{c}} s(\mathbf{w_c},\theta_0)) + o_p(1)$

Now, doing another mean value expansion of $\sum_{c=1}^{C} \widehat{\mathbf{sp}^{\mathbf{T}}}_{\mathbf{c}} s(\mathbf{w_c},\theta_0)$ around β we get:

$C^{\frac{1}{2}}(\hat{\theta} - \theta_0) = -(E[H(\mathbf{w},\theta_0)])^{-1}(C^{\frac{-1}{2}} \sum_{c=1}^{C} \mathbf{sp}^{\mathbf{T}}{}_{\mathbf{c}} s(\mathbf{w_c},\theta_0) + (\sum_{c=1}^{C} \nabla \hat{\mathbf{sp}}^{\mathbf{T}} s(\mathbf{w_c},\theta_0)) C^{\frac{-1}{2}}(\tilde{\beta} - \beta)) + o_p(1)$

Where $\nabla \hat{sp}^T$ is ∇sp^T evaluated at a $\hat{\beta}$ between the true β and the estimated $\tilde{\beta}$.

After applying a similar idea as we did with $H(.)$ we get:

$C^{\frac{1}{2}}(\hat{\theta} - \theta_0) = -(E[H(\mathbf{w},\theta_0)])^{-1}(C^{\frac{-1}{2}} \sum_{c=1}^{C} \mathbf{sp}^{\mathbf{T}}{}_{\mathbf{c}} s(\mathbf{w_c},\theta_0) + (E[\nabla \mathbf{sp}^{\mathbf{T}} s(\mathbf{w},\theta_0)]) C^{\frac{1}{2}}(\tilde{\beta} - \beta)) + o_p(1)$

□

Proof of Proposition 6:

By construction of M-estimation for a given sample we have:

$\frac{\sum_{c=1}^{C} \widehat{\mathbf{sp}^{\mathbf{T}}}_{\mathbf{c}} s(\mathbf{w_c},\hat{\theta})}{C} = \mathbf{0}$

Doing a mean value expansion around the true theta, we get:

$\mathbf{0} = C^{-1} \sum_{c=1}^{C} \widehat{\mathbf{sp}^{\mathbf{T}}}_{\mathbf{c}} s(\mathbf{w_c},\theta_0) + C^{-1} \sum_{c=1}^{C} \widehat{\mathbf{sp}^{\mathbf{T}}}_{\mathbf{c}} H(\mathbf{w_c},\bar{\theta})(\hat{\theta} - \theta_0)$ or:

$(\hat{\theta} - \theta_0) = -[C^{-1} \sum_{c=1}^{C} \widehat{\mathbf{sp^T}}_{\mathbf{c}} H(\mathbf{w_c}, \bar{\theta})]^{-1} [C^{-1} \sum_{c=1}^{C} \widehat{\mathbf{sp^T}}_{\mathbf{c}} s(\mathbf{w_c}, \theta_0)]$

Now doing another mean value expansion but around β we get:

$(\hat{\theta} - \theta_0) = -[C^{-1} \sum_{c=1}^{C} \widehat{\mathbf{sp^T}}_{\mathbf{c}} H(\mathbf{w_c}, \bar{\theta})]^{-1} C^{-1} [\sum_{c=1}^{C} \mathbf{sp_c^T} s(\mathbf{w_c}, \theta_0) + \sum_{c=1}^{C} \nabla_{\beta} \hat{\mathbf{sp}}_{\mathbf{c}}^{\mathbf{T}} s(\mathbf{w_c}, \theta_0)(\hat{\beta} - \beta)]$

where $\nabla_{\beta} \hat{sp}^T$ is $\nabla_{\beta} sp^T$ evaluated at a β between the true one and the estimated one.

Now, applying probability limits, we can see that by the WLLN and Lemma 1:

$C^{-1}[\sum_{c=1}^{C} \widehat{\mathbf{sp^T}}_{\mathbf{c}} H(\mathbf{w_c}, \bar{\theta})]^{-1}$ converges to $E[\widehat{\mathbf{sp^T}} H(\mathbf{w}, \bar{\theta})]^{-1}$ and

$C^{-1}[\sum_{c=1}^{C} \mathbf{sp_c^T} s(\mathbf{w_c}, \theta_0) + \sum_{c=1}^{C} \nabla_{\beta} \hat{\mathbf{sp}}_{\mathbf{c}}^{\mathbf{T}} s(\mathbf{w_c}, \theta_0)(\hat{\beta} - \beta)]$ to

$E[s(\mathbf{w}, \theta_o) + \nabla_{\beta} \hat{\mathbf{sp}}^{\mathbf{T}} s(\mathbf{w}, \theta_0) A]$ where by conditions of the proposition $E[s(\mathbf{w}, \theta_o)] = \mathbf{0}$.

Then, the estimator of θ converges asymptotically to:

$p\lim(\hat{\theta} - \theta_0) = [E[\widehat{\mathbf{sp^T}} H(\mathbf{w}, \bar{\theta})]^{-1}][E[\nabla_{\beta} \hat{\mathbf{sp}}^{\mathbf{T}} s(\mathbf{w}, \theta_0) A] + o_p(1)$

4.10 Appendix B: Tables

In this appendix the tables from the simulation results are presented.

The following table summarises the second stage mean squared error from the simulations.

Results are shown for different cluster size and number of clusters sampled.

Results systematically show that a wrong correction can be worst than no correction.

Results seem to suggest that once the number of clusters is big enough, the size of the cluster is not relevant. The second table summarises the second first mean squared error from the

Table 4.1 Mean squared errors with respect to the true parameter of interest of the second stage. 10000 replications.

Sample Size		No correction	Naïve Correction	Full Correction
N=2, C=20	slope of X1	0,130	0,147	0,136
N=2, C=100	slope of X1	0,076	0,085	0,074
N=2, C=1000	slope of X1	0,059	0,063	0,056
N=2, C=10000	slope of X1	0,056	0,063	0,054
N=2, C=100000	slope of X1	0,056	0,063	0,053
N=4, C=20	slope of X1	0,075	0,093	0,087
N=4, C=100	slope of X1	0,060	0,072	0,063
N=4, C=1000	slope of X1	0,052	0,063	0,054
N=20, C=1000	slope of X1	0,053	0,054	0,053

simulations.

Results are shown for different cluster size and number of clusters sampled.

Results systematically show that the full correction performs better, as expected.

Results also seem to suggest that the size of the cluster is not relevant.

Table 4.2 Mean squared errors with respect to the true parameters of the sample selection stage. 10000 replications.

Sample Size		Naïve Correction	Full Correction
N=2	constant	0,36	0,65
C=20	slope of X2	0,39	0,79
	Slope of mean X3	-	20,67
	Slope of max X3	-	16,06
	Slope of mean (X3 squared)	-	20,38
N=2	constant	0,21	0,07
C=100	slope of X2	0,30	0,07
	Slope of mean X3	-	1,64
	Slope of max X3	-	1,32
	Slope of mean (X3 squared)	-	1,20
N=2	constant	0,21	0,01
C=1000	slope of X2	0,27	0,01
	Slope of mean X3	-	0,12
	Slope of max X3	-	0,10
	Slope of mean (X3 squared)	-	0,08
N=2	constant	0,20	0,00
C=10000	slope of X2	0,29	0,00
	Slope of mean X3	-	0,01
	Slope of max X3	-	0,01
	Slope of mean (X3 squared)	-	0,01
N=2	constant	0,21	0,00
C=100000	slope of X2	0,29	0,00
	Slope of mean X3	-	0,00
	Slope of max X3	-	0,00
	Slope of mean (X3 squared)	-	0,00
N=4	constant	0,60	0,53
C=20	slope of X2	0,16	0,17
	Slope of mean X3	-	7,29
	Slope of max X3	-	4,80
	Slope of mean (X3 squared)	-	11,74
N=4	constant	0,49	0,07
C=100	slope of X2	0,11	0,02
	Slope of mean X3	-	0,83
	Slope of max X3	-	0,57
	Slope of mean (X3 squared)	-	1,21
N=4	constant	0,47	0,01
C=1000	slope of X2	0,10	0,00
	Slope of mean X3	-	0,08
	Slope of max X3	-	0,06
	Slope of mean (X3 squared)	-	0,11
N=20	constant	0,44	0,00
C=1000	slope of X2	0,09	0,00
	Slope of mean X3	-	0,05
	Slope of max X3	-	0,04
	Slope of mean (X3 squared)	-	0,10

CHAPTER 5. GENERAL CONCLUSION

In this dissertation, I have explored the challenges heterogeneity, miss-measurement, and missingness have on inference and identification of parameters of interest (or treatment effects).

The first paper provided partial identification results for the Marginal Treatment Effect with measurement error and a discrete instrument. Given the discrete nature of the instruments, we introduced different ways of recovering the MTE via nonparametric restrictions. Results were illustrated via a numerical example and quantifying the marginal treatment effect of SNAP on food insecurity, a case in which measurement error and endogeneity of treatment are known to be an issue. Results suggest that for most levels of heterogeneity, the treatment reduces the chances of being food insecure. Results from this paper can also serve as a sensitivity analysis tool for when researchers suspect measurement error and are interested in recovering a parameter such as the MTE that allows for the presence of heterogeneous effects of treatment.

In the second paper, we develop a falsification test for identifying assumptions in bivariate probit models. We derived sharp testable equalities for the model assumptions and non-sharp inequalities. Both of them which we express in the form of conditional moment inequalities. We implement those inequalities using existing inferential methods. We also provide some empirical examples, illustrating that the bivariate model, despite its nice feature that leads to point-identification of model parameters, could be restrictive in some cases. Our proposed procedure then serves as a screening test for the validity of the bivariate probit specification.

In the third paper, we provide results combining clustering, M-estimation, and missing data. Then, we extend existing identification results to a situation with dependence in the selection stage. Thirdly, we formalize conditions for an M-estimator under clustering and missing data to be consistent when the number of clusters increases and the size of the clusters remain fixed.