MARGINAL TREATMENT EFFECTS WITH MISCLASSIFIED TREATMENT

SANTIAGO ACERENZA, KYUNGHOON BAN, AND DÉSIRÉ KÉDAGNI

Iowa State University - Department of Economics

Abstract. This paper studies identification of the marginal treatment effect (MTE) when a binary treatment variable is misclassified. We show under standard assumptions that the MTE is identified as the derivative of the conditional expectation of the observed outcome given the true propensity score, which is partially identified. We characterize the identified set for this propensity score, and then for the MTE. We use our MTE bounds to derive bounds on other commonly used parameters in the literature. We show that our bounds are tighter than the existing bounds for the local average treatment effect. We illustrate the practical relevance of our derived bounds through some numerical and empirical results.

Keywords: Heterogeneous treatment effects, misclassification, instrumental variable, set identification.

JEL subject classification: C14, C31, C35, C36.

1. Introduction

The existence of measurement error in a treatment variable makes the identification of many parameters used in the causal inference literature challenging. When the treatment variable is binary, it is well-understood in the literature that the measurement error is non-classical, that is, it depends on the true treatment. Even in the homogeneous treatment effect framework, measurement errors in a binary regressor can result in severe identification deterioration of regression coefficients (Kreider, 2010). Ura (2018) appears to be the first to investigate the identifying power of an instrumental variable (IV) in the (unobserved)

Date: The first draft was of May 1, 2021. The present version is as of September 3, 2021. We thank Otávio Bartalotti, Helle Bunzel, Sukjin Han, Guido Imbens, Shakeeb Khan, Brent Kreider, Ismael Mourifié, Vitor Possebom, Denni Tommasi, Takuya Ura, Lina Zhang, seminar participants at Iowa State, UMass Amherst (Statistics), Syracuse, Tilburg, AfES 2021, NASMES 2021, and IAAE 2021 for helpful discussions and comments. All errors are ours. Corresponding address: 518 Farm House Lane, 260 Heady Hall, Ames, IA, 50011, USA. Email address: dkedagni@iastate.edu.

heterogeneous treatment effect model when the treatment is endogenous and mismeasured. He derives bounds on the local average treatment effect (LATE) when a binary instrument is available.[1] Calvi, Lewbel, and Tommasi (2018) propose a new estimand for the LATE, called the measurement robust LATE (MR-LATE), and obtain point-identification in an alternative framework. Yanagi (2019) shows that with the help of exogenous covariates, point-identification of the LATE can be obtained under some conditions when there is misclassification in the treatment. Tommasi and Zhang (2020) extend the results in Ura (2018), and Calvi, Lewbel, and Tommasi (2018) to the case with multivalued discrete instruments.

This paper investigates the identification of the MTE in settings where a binary treatment variable is misclassified, and a valid instrument is available. We show under standard assumptions that the MTE is identified as the derivative of the conditional expectation of the observed outcome given the true propensity score, which is partially identified. We provide a tractable characterization (which is an outer set) of the identified set for this propensity score, and then for the MTE. We also derive functional sharp bounds for this propensity score when the misclassification is non-differential. We use our MTE bounds to derive bounds on other commonly used parameters in the literature. In particular, we show that our bounds for the LATE are tighter than the existing Ura's (2018) and Tommasi and Zhang's (2020) bounds when the instrument is discrete. We illustrate the practical relevance of our derived bounds through some numerical and empirical results. More precisely, we apply our methodology on data from the third wave of the Indonesia Family Life Survey to measure marginal returns to upper secondary or higher schooling while allowing for the possibility that education be mismeasured.

Related literature. Several papers have extensively studied issues related to misclassification in treatment variables. See, for example, Aigner (1973), Bollinger (1996), Hausman, Abrevaya, and Scott-Morton (1998), Molinari (2008), Hu (2008), Hu and Schennach (2008), etc. Recently, Mahajan (2006) uses an additional instrument, which he called "instrument-like variable," to nonparametrically identify the regression function in models with a misclassified binary regressor. Considering the same model as Mahajan (2006) under different assumptions, Lewbel (2007) also uses a "second" instrument to nonparametrically identify the average treatment effect (ATE) when the treatment is misclassified. However, their results hold when the true treatment is exogenous. DiTraglia and Garcia-Jimeno (2019) show

[1]Mahajan (2006) and Lewbel (2007) allow for heterogeneous treatment effect through observed covariates.

that the identification result in Mahajan (2006) does not extend to the case of an endogenous treatment. In that context, they derive bounds on the average treatment effect under standard assumptions. Nguimkeu, Denteh, and Tchernis (2019) study a homogenous treatment effect linear regression model in which a binary regressor is potentially misclassified and endogenous. They use exclusion restrictions for both the participation equation and measurement error equation to identify the regression coefficient with endogenous participation and one-sided endogenous misreporting. Millimet (2011) studies the performance of several commonly used estimators in the causal inference literature when there is measurement error in a binary treatment, and warns researchers about the consequences of ignoring the presence of measurement errors. Kreider et al. (2012) partially identify the average effects of food stamps on health outcomes of children when participation is endogenous and misreported by using relatively weak nonparametric assumptions and information from auxiliary data. Our paper studies potential heterogeneity in the treatment effect through the marginal treatment effect when the treatment is endogenous and mismeasured. A more related work to ours is Ura (2020), who investigates heterogeneous treatment effects in the presence of a misclassified endogenous binary treatment variable through the instrumental variable quantile regression model.

Our paper also complements the work of Battistin and Sianesi (2011) and Battistin, De Nadai, and Sianesi (2014), who investigate the identification of average returns to education in the United Kingdom when attainment is potentially measured with error. While these authors focus on average returns, we investigate marginal returns, which may reveal (unobserved) heterogeneity in the treatment effects that would otherwise be hidden when looking at only average effects. Our work is also related to Chalak's (2017) who discusses the interpretation of various estimands (Wald, local IV) when the instrument is mismeasured, but the treatment variable is correctly measured. Differently from his framework, the instrument is observed with no error, while the treatment variable is potentially misreported. Recently, Jiang and Ding (2020) study identification in the binary IV model when allowing for simultaneous measurement errors in the instrument, treatment and/or outcome. They derive sharp bounds on the LATE assuming non-differential measurement errors and a valid IV. Our framework encompasses multivalued discrete/continuous outcomes and instruments, while the treatment variable is maintained binary. But, we only allow for misclassification in the treatment variable. Using a framework similar to ours, Possebom (2021) derives complementary identification results for the MTE, allowing for dependence between the instrument and the misclassification variable. While we focus on

the case where the instrument is completely randomly assigned, we show in Appendix E how our approach can be extended to the situation where there is dependence between the instrument and the misclassification variable. This extension is similar to Ura's (2018) in the LATE framework. Note that Kasahara and Shimotsu (2021) study identification in regression models when an endogenous binary regressor is misclassified, allowing for correlation between the instrument and the misclassification error. They show identification of the regression coefficient when a "special" covariate in the outcome equation is excluded from the misclassification probability. While they allow for heterogeneity in the average effect through observed covariates, we focus on unobserved heterogeneity using marginal treatment effects.

Outline. The remainder of the paper is organized as follows. Section 2 introduces the model and discusses the assumptions. Section 3 presents the main identification results, Section 4 discusses how information about the false positive/negative misclassification rates can be helpful for our identification approach. Section 5 illustrates the empirical relevance of the MTE bounds, Section 6 presents some extensions, Section 7 provides a real world empirical example, and Section 8 concludes. Proofs of the main results are relegated to the appendix.

2. Analytical Framework

Consider the following model:[2]

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

where the vector (Y, D, Z) represents the observed data, while the vector $(Y_1, Y_0, D^*, \varepsilon)$ is latent. In this model, $Y \in \mathcal{Y}$ is the observed outcome, $D^* \in \{0,1\}$ is the unobserved true treatment variable, while $D \in \{0,1\}$ is the observed mismeasured treatment, $\varepsilon \in \{0,1\}$ represents the decision to misreport, Y_0 and Y_1 are the potential outcomes that would have been observed if the true treatment D^* had been externally set to 0 and 1, respectively. The variable $Z \in \mathcal{Z}$ is an instrument. In this paper, we are interested in identifying the marginal treatment effect defined as

$$MTE(p) \equiv \mathbb{E}[Y_1 - Y_0 | V = p].$$

[2]We show in Appendix A that the specification $D = D^*(1-\varepsilon) + (1-D^*)\varepsilon$ is without loss of generality.

Example 1 (Marginal returns to schooling (leading example))**.** *In this example, we assume that the researcher is interested in measuring marginal returns to college education. It is well-documented that education is usually mismeasured. For example, Black, Sanders, and Taylor (2003) find that more than a third of respondents to the U.S. Census claiming to hold a professional degree have no such degree. In this case, the variable Y is earnings/wage, and D is the indicator for college degree. The variable Z could be distance to college. The latent variable V could be interpreted as the cost of going to college, while Y_1 is the potential earnings for someone with a college degree, and Y_0 is the potential earnings for someone without a college degree. The variable D^* is the individual's true indicator for college degree.* □

Example 2 (Marginal effects of masks)**.** *The variable Y could be the indicator that an individual tests positive to Covid-19, D^* the indicator that the individual actually wears masks, D the indicator that the individual reports wearing masks, V the disutility (cost) of wearing masks, and Z could be a shifter for the benefit of wearing masks (number of children).*[3] *People could report wearing masks while they actually do not (for example, because of social pressure). They could also pretend to wear mask while they do not wear it properly. On the other hand, someone could report not wearing regularly a mask (because of political reasons for example), while she actually does (because of her underlying health conditions). These facts could lead to misclassification in the report of mask wearing, and therefore induce some bias in the measurement of the marginal effects of mask wearing on the positivity rate. Y_1 could be the indicator that the individual tests positive to Covid-19 while she is wearing masks, and Y_0 could be the indicator that the individual tests positive to Covid-19 while she is not wearing masks. The effect of wearing masks on the positivity rate could be heterogeneous. Healthy individuals tend to think that they are immune or they will survive if they are infected. For them, the disutility of wearing masks may be higher, and they may be less likely to wear masks. We are interested in measuring the effect of wearing masks on the positivity rate for different levels of the disutility, i.e., $\mathbb{E}[Y_1 - Y_0|V = p]$.* □

We will use the following assumptions for identification:

[3]Children are less likely to get the virus, and therefore less likely to contaminate others. Hence, the variable number of children in the households is likely to satisfy the exclusion restriction assumption. We may add the number of adults in the households as a control variable to the model. However, the validity of number of children as instrument remains questionable, as is the distance to college instrument in Example 1.

Assumption 1 (Random assignment)**.** *The instrument Z is independent of (Y_d, V, ε), i.e., $Z \perp\!\!\!\perp (Y_d, V, \varepsilon)$, for each $d \in \{0,1\}$.* □

Assumption 1 requires that Z be a valid instrument, in the sense that it is statistically independent of all the unobservables in the model. This is a commonly used assumption in the literature. In our framework, the measurement error is nonclassical by definition of the model. Indeed, we can rewrite $D = D^* + (1-2D^*)\varepsilon$. So, the measurement error $(1-2D^*)\varepsilon$ is dependent on the true unobserved treatment D^*. This fact is well-documented and understood in the literature. See Aigner (1973) Mahajan (2006), Lewbel (2007), Kreider et al. (2012), Ura (2018), Yanagi (2019), etc.

Assumption 2 (Absolute continuity of V)**.** *The latent variable V is absolutely continuous. Without loss of generality, the unconditional distribution of V is uniform over $[0,1]$, and the support of the function $P(z)$ is included in $[0,1]$.* □

This assumption is standard in the literature and has been considered in Heckman and Vytlacil (1999, 2001, 2005), Carneiro and Lee (2009), Carneiro, Heckman, and Vytlacil (2010, 2011), etc. It does not require that the conditional density of V given ε exists, as will be apparent in the different specifications we consider in the paper. This assumption implies the following:

$$\mathbb{P}(\varepsilon = 1)F_{V|\varepsilon=1}(p) + \mathbb{P}(\varepsilon = 0)F_{V|\varepsilon=0}(p) = p \quad \text{for all } p \in [0,1], \tag{2.2}$$

where $F_{V|\varepsilon}$ denotes the conditional distribution of V given ε.

Assumption 3 (Continuous instrument)**.** *The instrument Z is continuous such that the support of the random variable $P(Z)$ is an interval.* □

Assumption 3 is also standard in the literature and is crucial for our identification methodology for the MTE. In Section 6, we show how our methodology can be used to identify multiple LATEs when the instrument is discrete. Identification results for the MTE with discrete instruments have been developed in Acerenza (2021), which built on insights from the current paper and Mogstad, Santos, and Torgovitsky (2018).

Assumption 4 (Upper bound on misclassification rate)**.** *The (unconditional) misclassification rate $\alpha \equiv \mathbb{P}(\varepsilon = 1)$ has a known upper bound $\bar{\alpha}$, that is, $\alpha \in [0, \bar{\alpha}]$.* □

A similar assumption to Assumption 4 has been considered in Horowitz and Manski (1995), Kreider and Pepper (2007), Molinari (2008), Kreider et al. (2012), etc. For example,

one can combine information from different sources (e.g., government, universities, etc.) to bound the extent of the misclassification. The case $\bar{\alpha} = 1$ corresponds to the scenario where the researcher is agnostic about the range of the misclassification rate. All our derived results still hold in this case.

3. Identification Results

3.1. **Identification of the MTE.** We have

$$\begin{aligned}
\mathbb{E}\left[Y|P(Z)=p\right] &= \mathbb{E}\left[Y_1 D^* + Y_0(1-D^*)|P(Z)=p\right], \\
&= \mathbb{E}\left[Y_1 \mathbb{1}\left\{V \leqslant p\right\} + Y_0 \mathbb{1}\left\{V > p\right\}\right], \\
&= \int_0^p \mathbb{E}\left[Y_1|V=v\right] dv + \int_p^1 \mathbb{E}\left[Y_0|V=v\right] dv, \qquad (3.1)
\end{aligned}$$

where the first equality holds from the definition of the model, the second holds from Assumption 1, and the third equality holds from Assumption 2. Under Assumption 3, we can differentiate each side of the equation with respect to p. Hence, we obtain

$$\frac{\partial \mathbb{E}[Y|P(Z)=p]}{\partial p} = \mathbb{E}\left[Y_1 - Y_0|V=p\right].$$

Below, we summarize this result in Lemma 1.

Lemma 1. *Suppose that model (2.1) along with Assumptions 1–3 hold. Then, the marginal treatment effect is identified as*

$$MTE(p) = \frac{\partial \mathbb{E}[Y|P(Z)=p]}{\partial p},$$

where the function $P(z)$ is partially identified as explained below. □

The result in Lemma 1 shows that the marginal treatment still has the local instrumental variable interpretation as in Heckman and Vytlacil (1999, 2001, 2005), except that the true propensity score $P(z)$ is now set-identified, because of the presence of misclassification. The intuition is that the conditional expectation $\mathbb{E}[Y|P(Z)=p]$ can be decomposed between the treatment and control groups as usual, even if these groups are unobserved due to the presence of misclassification. The above result holds whether there is misclassification or not, since the model assumes that misreporting does not have a direct effect on the outcome variable. Now, the fact that the treatment and control groups are observed with some noise leaves the propensity score $P(z)$ partially identified.

3.2. **Identification of $P(z)$.** For any Borel set A, we have

$$\begin{aligned}\mathbb{P}(Y \in A, D = 1|Z = z) &= \mathbb{P}(Y \in A, D = 1, D^* = 1|Z = z) + \mathbb{P}(Y \in A, D = 1, D^* = 0|Z = z), \\ &= \mathbb{P}(Y_1 \in A, \varepsilon = 0, V \leqslant P(z)) + \mathbb{P}(Y_0 \in A, \varepsilon = 1, V > P(z)), \qquad (3.2)\end{aligned}$$

where the first equality holds from the law of total probability, and the second equality follows from the definition of the model and Assumption 1. In the special case where $A = \mathcal{Y}$, we have

$$\mathbb{P}(D = 1|Z = z) = \mathbb{P}(\varepsilon = 0, V \leqslant P(z)) + \mathbb{P}(\varepsilon = 1, V > P(z)). \qquad (3.3)$$

When there is no misclassification in the treatment, i.e., $\varepsilon = 0$ a.s., then $P(z)$ is identified as the propensity score $\mathbb{P}(D = 1|Z = z)$, since the distribution of V is normalized to be uniform over $[0, 1]$. When the treatment is *completely* misclassified, i.e., $\varepsilon = 1$ a.s., $P(z)$ is identified under the previous normalization as $\mathbb{P}(D = 0|Z = z)$. We can rewrite the last equality as follows:

$$\mathbb{P}(D = 1|Z = z) = (1 - \alpha)F_{V|\varepsilon=0}(P(z)) + \alpha(1 - F_{V|\varepsilon=1}(P(z))). \qquad (3.4)$$

We have $\mathbb{P}(D^* = 1|Z = z) = \mathbb{P}(V \leqslant P(z)) = P(z)$. Thus, $P(z)$ is the true (unidentified) propensity score. We show that the propensity score $P(z)$ is partially identified using Equations (2.2) and (3.4). Equation (3.4) implies

$$\mathbb{P}(D = 1|P(Z) = p) = (1 - \alpha)F_{V|\varepsilon=0}(p) + \alpha(1 - F_{V|\varepsilon=1}(p)).$$

For now, we assume $\alpha \in (0, 1)$, since the cases where $\alpha \in \{0, 1\}$ can be dealt with separately. Combining this with Equation (2.2), and solving for $F_{V|\varepsilon=0}(p)$ and $F_{V|\varepsilon=1}(p)$ in the system of equations, we obtain:

$$\begin{aligned}F_{V|\varepsilon=1}(p) &= \frac{p + \alpha - \mathbb{P}(D = 1|P(Z) = p)}{2\alpha}, \\ F_{V|\varepsilon=0}(p) &= \frac{p - \alpha + \mathbb{P}(D = 1|P(Z) = p)}{2(1 - \alpha)}.\end{aligned}$$

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 $F_{V|\varepsilon=1}(1) = F_{V|\varepsilon=0}(1) = 1$. In general, it will be difficult to nonparametrically characterize the sharp identification region for the propensity score function $P(z)$ using those conditions. We are going to focus on the monotonicity condition 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:

$$\begin{aligned} 0 &\leqslant F_{V|\varepsilon=0}(P(z)) - F_{V|\varepsilon=0}(P(z')) \leqslant 1, \\ 0 &\leqslant F_{V|\varepsilon=1}(P(z)) - F_{V|\varepsilon=1}(P(z')) \leqslant 1. \end{aligned}$$

which implies

$$\begin{aligned} 0 &\leqslant \frac{\mathbb{P}(D=1|Z=z) - \mathbb{P}(D=1|Z=z') + P(z) - P(z')}{2(1-\alpha)} \leqslant 1, \\ 0 &\leqslant \frac{P(z) - P(z') - \mathbb{P}(D=1|Z=z) + \mathbb{P}(D=1|Z=z')}{2\alpha} \leqslant 1. \end{aligned}$$

This latter inequalities respectively imply

$$\begin{aligned} -\mathbb{P}(D=1|Z=z) + \mathbb{P}(D=1|Z=z') &\leqslant P(z) - P(z') \qquad (3.5) \\ &\leqslant 2(1-\alpha) - \mathbb{P}(D=1|Z=z) + \mathbb{P}(D=1|Z=z'), \\ \mathbb{P}(D=1|Z=z) - \mathbb{P}(D=1|Z=z') &\leqslant P(z) - P(z') \qquad (3.6) \\ &\leqslant 2\alpha + \mathbb{P}(D=1|Z=z) - \mathbb{P}(D=1|Z=z'). \end{aligned}$$

In the special case where $P(z') = 0$, we have

$$\begin{aligned} -\mathbb{P}(D=1|Z=z) + \mathbb{P}(D=1|P(Z)=0) &\leqslant P(z) \\ &\leqslant 2(1-\alpha) - \mathbb{P}(D=1|Z=z) + \mathbb{P}(D=1|P(Z)=0), \\ \mathbb{P}(D=1|Z=z) - \mathbb{P}(D=1|P(Z)=0) &\leqslant P(z) \\ &\leqslant 2\alpha + \mathbb{P}(D=1|Z=z) - \mathbb{P}(D=1|P(Z)=0). \end{aligned}$$

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

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

Similar argument holds for the special case where $P(z) = 1$, but this yields the above same constraints on $P(z')$. Hence, the following proposition holds.

Proposition 1. *Suppose that model (2.1) along with Assumptions 1, 2, and 4 hold. We have the following bounds on $P(z)$: $LB(z) \leqslant P(z) \leqslant UB(z)$, where*

$$\begin{aligned} LB(z) &\equiv \inf_{\alpha\in[0,\bar{\alpha}]} \max\left\{\mathbb{P}(D=1|Z=z) - \alpha, \alpha - \mathbb{P}(D=1|Z=z)\right\}, \\ UB(z) &\equiv \sup_{\alpha\in[0,\bar{\alpha}]} \min\left\{\mathbb{P}(D=1|Z=z) + \alpha, (1-\alpha) + \mathbb{P}(D=0|Z=z)\right\}. \end{aligned}$$

These bounds are pointwise sharp. □

In Appendix B, we provide two different specifications for the relationship between the decision to misreport ε and the unobserved heterogeneity V that achieve the above bounds on $P(z)$. However, these bounds are not necessarily *functionally sharp* in the language of Mourifié, Henry, and Méango (2020), as taking the difference of the bounds for $P(z)$ and $P(z')$ will not necessarily yield the tightest bounds for the difference $P(z)-P(z')$. We show this in Subsection 3.3 below.

For the rest of the paper, we derive our results for each value of $\alpha \in [0, \bar{\alpha}]$. As in Proposition 1, one can take the infimum of the lower bound on the parameter of interest over the range $[0, \bar{\alpha}]$, and similarly the supremum of the upper bound over $[0, \bar{\alpha}]$.

3.3. **Analytical bounds for the MTE.** In this subsection, we provide analytical bounds on the MTE. Define $\Delta_{YZ}(z', z) \equiv \mathbb{E}[Y|Z = z] - \mathbb{E}[Y|Z = z']$ and $\Delta_{DZ}(z', z) \equiv \mathbb{E}[D|Z = z] - \mathbb{E}[D|Z = z']$. Inequalities (3.5) and (3.6) imply the following bounds on the $P(z)-P(z')$ when $0 \leqslant P(z') < P(z) \leqslant 1$.

$$\left|\Delta_{DZ}(z', z)\right| \leqslant P(z) - P(z') \leqslant \min\left\{1, 2\alpha + \Delta_{DZ}(z', z), 2(1-\alpha) - \Delta_{DZ}(z', z)\right\}.$$

These above bounds on the difference $P(z)-P(z')$ can be tightened using Equations (3.10) and (3.11). Notice that the model implies the following index sufficiency result:

$$\mathbb{P}\left(Y \in A, D = d | P(Z) = P(z)\right) = \mathbb{P}\left(Y \in A, D = d | Z = z\right) \text{ for all } z \text{ and } d. \tag{3.7}$$

Indeed, similar to Equation (3.2), the following holds under Assumption 1:

$$\mathbb{P}(Y \in A, D = 1 | P(Z) = P(z)) \quad = \quad \mathbb{P}(Y_1 \in A, \varepsilon = 0, V \leqslant P(z)) + \mathbb{P}(Y_0 \in A, \varepsilon = 1, V > P(z)).$$

From Equation (3.2), we can show under Assumptions 1 and 2 that

$$\begin{aligned}\mathbb{P}(Y \in A, D = 1 | Z = z) \quad &= \quad \int_0^{P(z)} \mathbb{P}\left(Y_1 \in A, \varepsilon = 0 | V = v\right) dv \\ &\qquad + \int_{P(z)}^1 \mathbb{P}\left(Y_0 \in A, \varepsilon = 1 | V = v\right) dv. \end{aligned} \tag{3.8}$$

A similar result holds for the observed control group:

$$\begin{aligned}\mathbb{P}(Y \in A, D = 0 | Z = z) \quad &= \quad \int_0^{P(z)} \mathbb{P}\left(Y_1 \in A, \varepsilon = 1 | V = v\right) dv \\ &\qquad + \int_{P(z)}^1 \mathbb{P}\left(Y_0 \in A, \varepsilon = 0 | V = v\right) dv. \end{aligned} \tag{3.9}$$

The above derived equalities allow us to characterize the functional identified set for $P(z)$.

Definition 1. *The identified set for the function $P : \mathcal{Z} \to [0,1]$ is the collection*

$$\{P(z) : \ 0 \leqslant P(z) \leqslant 1, \ z \in \mathcal{Z}\}$$

such that there exists a joint distribution on $(Y_0, Y_1, \varepsilon, V, Z)$ that satisfies model (2.1), Assumptions 1, 2, 4, and Equations (3.7)–(3.9). □

This characterization of the identified set for $P(z)$ is broad, but less tractable. We are going to derive analytical expressions for the MTE bounds based on the previous results.

Using equality (3.8), we have for $p > p'$

$$\begin{aligned}
&\mathbb{P}(Y \in A, D = 1|P(Z) = p) - \mathbb{P}(Y \in A, D = 1|P(Z) = p') \\
&\quad = \int_{p'}^{p} \mathbb{P}\left(Y_1 \in A, \varepsilon = 0|V = v\right) dv \\
&\qquad - \int_{p'}^{p} \mathbb{P}\left(Y_0 \in A, \varepsilon = 1|V = v\right) dv.
\end{aligned} \tag{3.10}$$

Similarly, from (3.9) we have

$$\begin{aligned}
&\mathbb{P}(Y \in A, D = 0|P(Z) = p) - \mathbb{P}(Y \in A, D = 0|P(Z) = p') \\
&\quad = \int_{p'}^{p} \mathbb{P}\left(Y_1 \in A, \varepsilon = 1|V = v\right) dv \\
&\qquad - \int_{p'}^{p} \mathbb{P}\left(Y_0 \in A, \varepsilon = 0|V = v\right) dv.
\end{aligned} \tag{3.11}$$

Indeed, the density version of Equations (3.10) and (3.11) holds and implies respectively:

$$\begin{aligned}
f_{Y,D|P(Z)}\left(y, 1|P(z)\right) - f_{Y,D|P(Z)}\left(y, 1|P(z')\right) &= \int_{P(z')}^{P(z)} f_{Y_1,\varepsilon|V}(y, 0|v)dv \\
&\quad - \int_{P(z')}^{P(z)} f_{Y_0,\varepsilon|V}(y, 1|v)dv,
\end{aligned}$$

and

$$\begin{aligned}
f_{Y,D|P(Z)}\left(y, 0|P(z)\right) - f_{Y,D|P(Z)}\left(y, 0|P(z')\right) &= \int_{P(z')}^{P(z)} f_{Y_1,\varepsilon|V}(y, 1|v)dv \\
&\quad - \int_{P(z')}^{P(z)} f_{Y_0,\varepsilon|V}(y, 0|v)dv,
\end{aligned}$$

where $f_{X|W}\left(x|w\right)$ is the conditional density of X given $\{W = w\}$ that is absolutely continuous with respect to a known dominating measure μ_X. The index sufficiency implication

of the model (3.7) implies $f_{Y,D|P(Z)}(y,d|P(z)) = f_{Y,D|Z}(y,d|z)$. Combining these results with the triangle inequality, we have

$$\begin{aligned}\left|f_{Y,D|Z}(y,1|z) - f_{Y,D|Z}\left(y,1|z'\right)\right| &\leqslant \int_{P(z')}^{P(z)} f_{Y_1,\varepsilon|V}(y,0|v)dv \\ &\quad + \int_{P(z')}^{P(z)} f_{Y_0,\varepsilon|V}(y,1|v)dv.\end{aligned}$$

Therefore, by integrating each side of the last inequality over the support $\mathcal{Y}$, and using the Fubini-Tonelli theorem, we have

$$\begin{aligned}\int_{\mathcal{Y}} \left|f_{Y,D|Z}(y,1|z) - f_{Y,D|Z}\left(y,1|z'\right)\right| d\mu_Y(y) &\leqslant \int_{P(z')}^{P(z)} \mathbb{P}(\varepsilon=0|V=v)dv + \int_{P(z')}^{P(z)} \mathbb{P}(\varepsilon=1|V=v)dv, \\ &= \int_{P(z')}^{P(z)} dv = P(z) - P(z').\end{aligned}$$

Hence, we have $TV_{(Y,D=1)}(z',z) \leqslant P(z) - P(z')$, where

$$TV_{(Y,D=d)}(z',z) \equiv \int_{\mathcal{Y}} \left|f_{Y,D|Z}(y,d|z) - f_{Y,D|Z}\left(y,d|z'\right)\right| d\mu_Y(y).$$

Using a similar argument for the second equality implied by (3.11), we have $TV_{(Y,D=0)}(z',z) \leqslant P(z) - P(z')$. Therefore, we obtain the following bounds on the difference $P(z) - P(z')$:

$$\begin{aligned}&\max\left\{\left|\Delta_{DZ}(z',z)\right|, TV_{(Y,D=1)}(z',z), TV_{(Y,D=0)}(z',z)\right\} \\ &\qquad \leqslant P(z) - P(z') \leqslant \min\left\{1, 2\alpha + \Delta_{DZ}(z',z), 2(1-\alpha) - \Delta_{DZ}(z',z)\right\}.\end{aligned}$$

We can show that $\max\left\{TV_{(Y,D=1)}(z',z), TV_{(Y,D=0)}(z',z)\right\} \geqslant |\Delta_{DZ}(z',z)|$.[4] Consequently, we have

$$\begin{aligned}&\max\left\{TV_{(Y,D=1)}(z',z), TV_{(Y,D=0)}(z',z)\right\} \\ &\qquad \leqslant P(z) - P(z') \leqslant \min\left\{1, 2\alpha + \Delta_{DZ}(z',z), 2(1-\alpha) - \Delta_{DZ}(z',z)\right\}. \quad (3.12)\end{aligned}$$

These above bounds on $P(z) - P(z')$ are tighter than the ones one would get by taking the difference of the pointwise bounds derived previously in Proposition 1.

Define

$$\begin{aligned}LB_p(z',z) &\equiv \max\left\{TV_{(Y,D=1)}(z',z), TV_{(Y,D=0)}(z',z)\right\}, \\ UB_p(z',z) &\equiv \min\left\{1, 2\alpha + \Delta_{DZ}(z',z), 2(1-\alpha) - \Delta_{DZ}(z',z)\right\}.\end{aligned}$$

[4]To show this, use the fact that for any μ-integrable function h, $\left|\int_{\mathcal{Y}} h(y)d\mu_Y(y)\right| \leqslant \int_{\mathcal{Y}} |h(y)|\, d\mu_Y(y)$, and $\int_{\mathcal{Y}} f_{Y,D|Z}(y,d|z)\, d\mu_Y(y) = \mathbb{P}(D=d|Z=z)$.

Suppose $LB_p(z',z) \neq 0$ and $UB_p(z',z) \neq 0$. Then, the following holds.

$$\begin{cases} \frac{\Delta_{YZ}(z',z)}{UB_p(z',z)} \leqslant \frac{\Delta_{YZ}(z',z)}{P(z)-P(z')} \leqslant \frac{\Delta_{YZ}(z',z)}{LB_p(z',z)} & \text{if } \Delta_{YZ}(z',z) \geqslant 0, \\ \frac{\Delta_{YZ}(z',z)}{LB_p(z',z)} \leqslant \frac{\Delta_{YZ}(z',z)}{P(z)-P(z')} \leqslant \frac{\Delta_{YZ}(z',z)}{UB_p(z',z)} & \text{if } \Delta_{YZ}(z',z) < 0. \end{cases}$$

Hence, we have

$$\begin{aligned} \min\left\{\frac{\Delta_{YZ}(z',z)}{UB_p(z',z)}, \frac{\Delta_{YZ}(z',z)}{LB_p(z',z)}\right\} \\ \leqslant \frac{\mathbb{E}[Y|P(Z)=P(z)] - \mathbb{E}[Y|P(Z)=P(z')]}{P(z)-P(z')} \leqslant \\ \max\left\{\frac{\Delta_{YZ}(z',z)}{UB_p(z',z)}, \frac{\Delta_{YZ}(z',z)}{LB_p(z',z)}\right\}. \end{aligned}$$

Therefore, we can take the limit of each side when z' goes to z. Suppose that $\lim_{z'\to z} \frac{\Delta_{YZ}(z',z)}{UB_p(z',z)}$, $\lim_{z'\to z} \frac{\Delta_{YZ}(z',z)}{LB_p(z',z)}$, and $\lim_{z'\to z} \frac{\mathbb{E}[Y|P(Z)=P(z)]-\mathbb{E}[Y|P(Z)=P(z')]}{P(z)-P(z')}$ exist.[5] Then, using the fact that the functions min and max are continuous, we obtain

$$\begin{aligned} \min\left\{\lim_{z'\to z}\frac{\Delta_{YZ}(z',z)}{UB_p(z',z)}, \lim_{z'\to z}\frac{\Delta_{YZ}(z',z)}{LB_p(z',z)}\right\} \\ \leqslant MTE(P(z)) \leqslant \\ \max\left\{\lim_{z'\to z}\frac{\Delta_{YZ}(z',z)}{UB_p(z',z)}, \lim_{z'\to z}\frac{\Delta_{YZ}(z',z)}{LB_p(z',z)}\right\}. \end{aligned} \tag{3.13}$$

These bounds may not be sharp, but they provide a *tractable outer set* of the identified set for $MTE(P(z))$. In practice, we may set $P(z)$ to be equal to the midpoint of its bounds derived in the previous subsection.

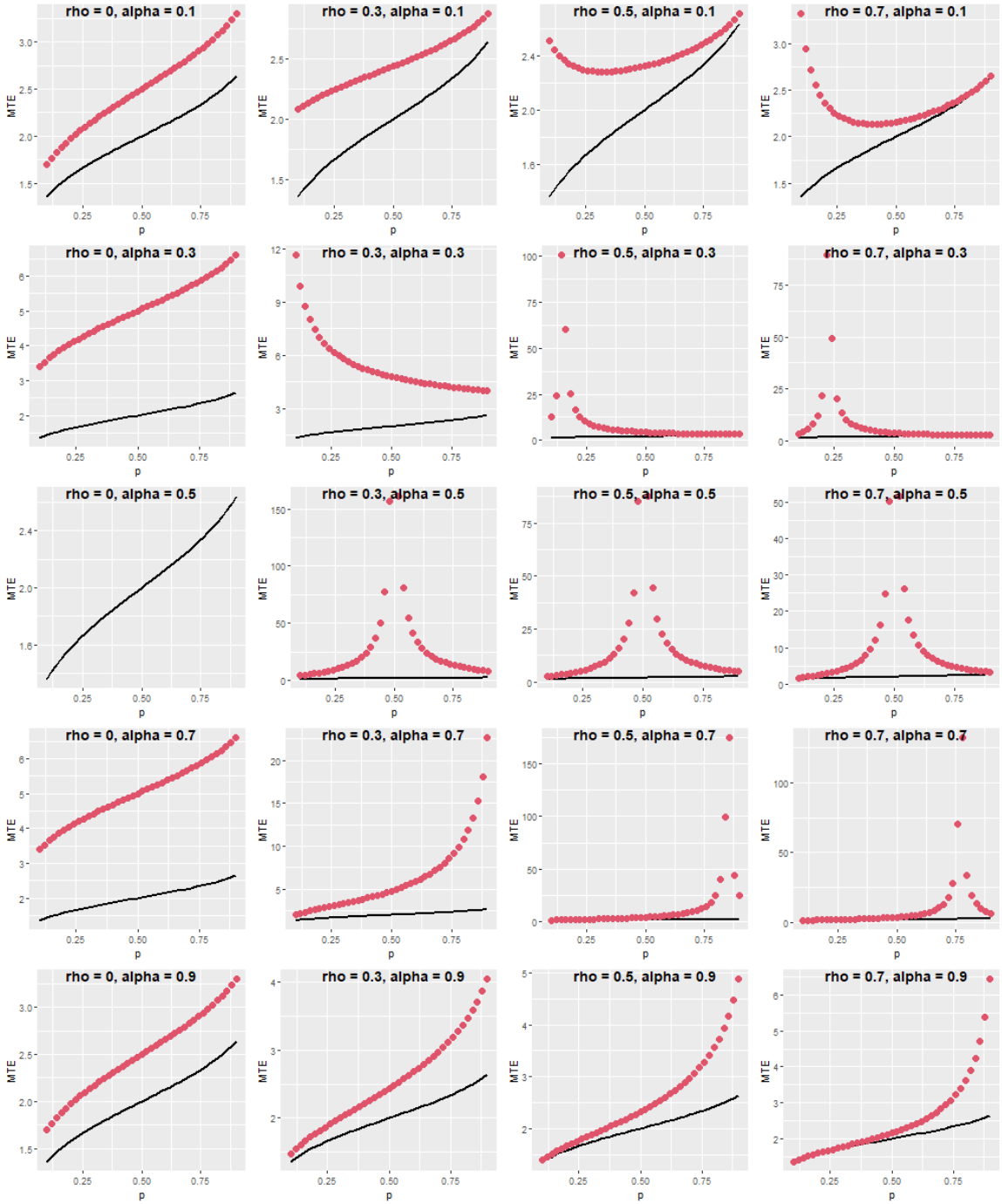
Numerical illustration. We consider the following data generating process (DGP)

$$\begin{cases} Y &= \beta D^* + U \\ D^* &= \mathbb{1}\{V \leqslant \Phi(2Z)\} \\ D &= D^*(1-\varepsilon) + (1-D^*)\varepsilon \\ \varepsilon &= \mathbb{1}\{\xi \leqslant \alpha\} \end{cases} \tag{3.14}$$

[5]Then, we have $\lim_{z'\to z} \frac{\mathbb{E}[Y|P(Z)=P(z)]-\mathbb{E}[Y|P(Z)=P(z')]}{P(z)-P(z')} = \frac{\partial \mathbb{E}[Y|P(Z)=p]}{\partial p}|_{p=P(z)} = MTE(P(z))$. When the first two limits do not exist, we replace them by lim inf and lim sup in the lower and upper bounds of (3.13).

where $V = \Phi(V^*), \xi = \Phi(\xi^*)$, $\begin{pmatrix} \beta \\ U \\ V^* \\ \xi^* \\ Z \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma)$, with $\mu = \begin{pmatrix} 2 \\ 2 \\ 0 \\ 0 \\ 2 \end{pmatrix}$, and $\Sigma = \begin{pmatrix} 1 & 0.5 & 0.5 & 0.5 & 0 \\ 0.5 & 1 & 0.5 & 0.5 & 0 \\ 0.5 & 0.5 & 1 & \rho & 0 \\ 0.5 & 0.5 & \rho & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$.

Details on the DGP are provided in Subsection G.1 in the appendix. As we can see on Figure 1, the upper bound becomes closer to the true MTE as the misclassification rate α approaches 0 or 1. The coefficient ρ captures the degree of dependence between the unobserved heterogeneity V and the misreporting ε. It is unclear how this dependence parameter ρ affects the bounds.



* The black line represents the true MTE from the model.
* The dots are the approximated upper bound at each grid point of p.

FIGURE 1. Numerical Illustration of the MTE Upper Bound

3.4. **When the distribution of V given ε is absolutely continuous.** In this subsection we are going to characterize the (sharp) identified set for the MTE. We add the non-differential measurement error assumption to the set of our identifying assumptions.

Assumption 5 (Non-differential misclassification)**.** *The misclassification variable ε is independent of Y_d conditional on V, i.e., $\varepsilon \perp\!\!\!\perp Y_d|V$, for each $d \in \{0,1\}$.* □

This assumption states that conditional on the unobserved heterogeneity that drives the selection into treatment, misreporting is independent of the potential outcomes. Combined with Assumption 1, it implies that misreporting is independent of the outcome conditional on the true treatment. This assumption could be too restrictive, as there may exist some returns to misreporting. In our leading example, there could exist some "returns to lying" about college completion, as discussed in Hu and Lewbel (2012), and DiTraglia and Garcia-Jimeno (2019).

We have

$$\begin{aligned}\mathbb{P}(Y \in A, D = 1|P(Z) = p) &= \mathbb{P}(Y_1 \in A, \varepsilon = 0, V \leqslant p) + \mathbb{P}(Y_0 \in A, \varepsilon = 1, V > p), \\ &= (1-\alpha)\int_0^p \mathbb{P}\left(Y_1 \in A|V = v, \varepsilon = 0\right) f_{V|\varepsilon=0}(v)dv \\ &\quad +\alpha\int_p^1 \mathbb{P}\left(Y_0 \in A|V = v, \varepsilon = 1\right) f_{V|\varepsilon=1}(v)dv. \qquad (3.15)\end{aligned}$$

where the first equality follows from the results derived in the previous subsection, and the second equality holds from the law of iterated expections. Therefore, by taking the derivatives of both sides of this equality with respect to p, we obtain the following:

$$\begin{aligned}\frac{\partial \mathbb{P}(Y \in A, D = 1|P(Z) = p)}{\partial p} &= (1-\alpha) f_{V|\varepsilon=0}(p)\mathbb{P}\left(Y_1 \in A|V = p, \varepsilon = 0\right) \\ &\quad -\alpha f_{V|\varepsilon=1}(p)\mathbb{P}\left(Y_0 \in A|V = p, \varepsilon = 1\right), \\ &= (1-\alpha) f_{V|\varepsilon=0}(p)\mathbb{P}\left(Y_1 \in A|V = p\right) \\ &\quad -\alpha f_{V|\varepsilon=1}(p)\mathbb{P}\left(Y_0 \in A|V = p\right), \qquad (3.16)\end{aligned}$$

where the second equality holds under Assumption 5. Similarly, we can show that

$$\begin{aligned}\frac{\partial \mathbb{P}(Y \in A, D = 0|P(Z) = p)}{\partial p} &= \alpha f_{V|\varepsilon=1}(p)\mathbb{P}\left(Y_1 \in A|V = p\right) \\ &\quad -(1-\alpha) f_{V|\varepsilon=0}(p)\mathbb{P}\left(Y_0 \in A|V = p\right). \qquad (3.17)\end{aligned}$$

Applying equality (3.16) to the special case where $A = \mathcal{Y}$, and using the fact that $f_V(p) = 1$ (since $V \sim \mathcal{U}_{[0,1]}$), we have

$$\begin{aligned}(1-\alpha)f_{V|\varepsilon=0}(p) - \alpha f_{V|\varepsilon=1}(p) &= \frac{\partial \mathbb{P}(D=1|P(Z)=p)}{\partial p},\\ (1-\alpha)f_{V|\varepsilon=0}(p) + \alpha f_{V|\varepsilon=1}(p) &= 1.\end{aligned}$$

Therefore,

$$\begin{aligned}f_{V|\varepsilon=0}(p) &= \frac{1+\frac{\partial \mathbb{P}(D=1|P(Z)=p)}{\partial p}}{2(1-\alpha)},\\ f_{V|\varepsilon=1}(p) &= \frac{1-\frac{\partial \mathbb{P}(D=1|P(Z)=p)}{\partial p}}{2\alpha}.\end{aligned}$$

Hence, the function P must satisfy the following conditions:

$$\frac{1+\frac{\partial \mathbb{P}(D=1|P(Z)=p)}{\partial p}}{2(1-\alpha)} \geqslant 0, \tag{3.18}$$

$$\frac{1-\frac{\partial \mathbb{P}(D=1|P(Z)=p)}{\partial p}}{2\alpha} \geqslant 0, \tag{3.19}$$

$$\int_0^1 \frac{1+\frac{\partial \mathbb{P}(D=1|P(Z)=p)}{\partial p}}{2(1-\alpha)} dp = 1, \tag{3.20}$$

$$\int_0^1 \frac{1-\frac{\partial \mathbb{P}(D=1|P(Z)=p)}{\partial p}}{2\alpha} dp = 1, \tag{3.21}$$

$$0 \leqslant \frac{\alpha f_{V|\varepsilon=1}(p)\kappa_0(A;p) - (1-\alpha)\, f_{V|\varepsilon=0}(p)\kappa_1(A;p)}{\alpha f_{V|\varepsilon=1}(p) - (1-\alpha)\, f_{V|\varepsilon=0}(p)} \leqslant 1, \tag{3.22}$$

$$0 \leqslant \frac{(1-\alpha)\, f_{V|\varepsilon=0}(p)\kappa_0(A;p) - \alpha f_{V|\varepsilon=0}(p)\kappa_1(A;p)}{\alpha f_{V|\varepsilon=1}(p) - (1-\alpha)\, f_{V|\varepsilon=0}(p)} \leqslant 1, \tag{3.23}$$

$$\begin{aligned}\int_0^1 (1-\alpha)\, f_{V|\varepsilon=0}(p) &\frac{\alpha f_{V|\varepsilon=1}(p)\kappa_0(A;p) - (1-\alpha)\, f_{V|\varepsilon=0}(p)\kappa_1(A;p)}{\alpha f_{V|\varepsilon=1}(p) - (1-\alpha)\, f_{V|\varepsilon=0}(p)} dp\\ &= P(Y \in A, D=1|P(Z)=1),\end{aligned} \tag{3.24}$$

$$\begin{aligned}\int_0^1 \alpha f_{V|\varepsilon=1}(p) &\frac{(1-\alpha)\, f_{V|\varepsilon=0}(p)\kappa_0(A;p) - \alpha f_{V|\varepsilon=0}(p)\kappa_1(A;p)}{\alpha f_{V|\varepsilon=1}(p) - (1-\alpha)\, f_{V|\varepsilon=0}(p)} dp\\ &= P(Y \in A, D=1|P(Z)=0),\end{aligned} \tag{3.25}$$

$$\mathbb{P}\left(Y \in A, D=d|Z=z\right) = \mathbb{P}\left(Y \in A, D=d|P(Z)=P(z)\right), \tag{3.26}$$

for all p such that $\frac{\partial \mathbb{P}(D=1|P(Z)=p)}{\partial p} \neq 0$, all $d \in \{0,1\}$, and all Borel set $A \subset \mathcal{Y}$, where $\kappa_1(A;p) = \frac{\partial \mathbb{P}(Y\in A, D=1|P(Z)=p)}{\partial p}$, and $\kappa_0(A;p) = \frac{\partial \mathbb{P}(Y\in A, D=0|P(Z)=p)}{\partial p}$. The constraints (3.22)

and (3.23) come from the fact that $\mathbb{P}(Y_1 \in A|V=p)$ and $\mathbb{P}(Y_0 \in A|V=p)$ are probabilities. In fact, using equalities (3.16) and (3.17) we can solve for $\mathbb{P}(Y_1 \in A|V=p)$ and $\mathbb{P}(Y_0 \in A|V=p)$ as follows:

$$\begin{aligned}\mathbb{P}(Y_1 \in A|V=p) &= \frac{\alpha f_{V|\varepsilon=1}(p)\kappa_0(A;p) - (1-\alpha) f_{V|\varepsilon=0}(p)\kappa_1(A;p)}{\alpha f_{V|\varepsilon=1}(p) - (1-\alpha) f_{V|\varepsilon=0}(p)},\\ \mathbb{P}(Y_0 \in A|V=p) &= \frac{(1-\alpha) f_{V|\varepsilon=0}(p)\kappa_0(A;p) - \alpha f_{V|\varepsilon=0}(p)\kappa_1(A;p)}{\alpha f_{V|\varepsilon=1}(p) - (1-\alpha) f_{V|\varepsilon=0}(p)}.\end{aligned}$$

Equations (3.24) and (3.25) are like terminal conditions, and come from Equation (3.15). Note that conditions (3.20) and (3.21) are equivalent, and come from the fact that density functions integrate to 1, while Equations (3.18) and (3.19) are the non-negativity conditions for density functions. The following proposition holds.

Proposition 2. *Suppose that model (2.1) along with Assumptions 1–5 hold. In addition, suppose that the distribution of V given ε is absolutely continuous. For a given $\alpha \in [0, \bar{\alpha}]$, the constraints (3.18)–(3.26) yield the (sharp) identified set for the function $P : \mathcal{Z} \to [0,1]$, and therefore for the MTE.* □

The proof of Proposition 2 is given in Appendix D. Although this proposition provides sharp identification region for the propensity score $P(z)$ and the MTE, this identified set is not tractable.

4. How can knowledge about false positive/negative rates help?

4.1. When false positive/negative misclassification probabilities do not depend on the instrument.

Equation (3.3) implies

$$\begin{aligned}\mathbb{P}(D=1|Z=z) &= \mathbb{P}(\varepsilon=0|V \leqslant P(z))\mathbb{P}(V \leqslant P(z)) + \mathbb{P}(\varepsilon=1|V > P(z))\mathbb{P}(V > P(z)),\\ &= (1-\alpha_1(z))P(z) + \alpha_0(z)(1-P(z)),\\ &= (1-\alpha_0(z)-\alpha_1(z))P(z) + \alpha_0(z),\end{aligned}$$

where the second equality holds from Assumption 2, $\alpha_1(z) \equiv \mathbb{P}(\varepsilon = 1|V \leqslant P(z))$ is the false negative misclassification rate, and $\alpha_0(z) \equiv \mathbb{P}(\varepsilon = 1|V > P(z))$ is the false positive misclassification rate. Suppose that $\alpha_0(z) + \alpha_1(z) < 1$.[6] Then,

$$P(z) = \frac{\mathbb{P}(D=1|Z=z) - \alpha_0(z)}{1-\alpha_0(z)-\alpha_1(z)}.$$

[6]This constraint is known as the *monotonicity condition* is the literature on misclassification, and is different from the monotonicity restriction imposed on the treatment selection in model (2.1).

The misclassification rate functions $\alpha_0(z)$ and $\alpha_1(z)$ can be partially identified using the conditions that they lie within the interval $[0,1]$, and that $P(z) \in [0,1]$. We are going to follow the literature (Hausman, Abrevaya, and Scott-Morton, 1998) to first assume that the misclassification rates $\alpha_0(z)$ and $\alpha_1(z)$ are constant across z.[7] This case will occur if for example ε is independent of V. In such a case, $\alpha_0 = \alpha_1 = \alpha$, and the misclassification is symmetric. The true propensity score $P(z)$ is therefore identified up to the misclassification probabilities α_0 and α_1 as follows:

$$P(z) = \frac{\mathbb{P}(D=1|Z=z) - \alpha_0}{1-\alpha_0-\alpha_1},$$

where α_0 and α_1 are partially identified:

$$(\alpha_0,\alpha_1) \in \left\{ [0,1]^2 : \alpha_0+\alpha_1 < 1, \text{ and } 0 \leqslant \frac{\mathbb{P}(D=1|Z=z)-\alpha_0}{1-\alpha_0-\alpha_1} \leqslant 1 \text{ for all } z \right\},$$

which is equivalent to:

$$(\alpha_0,\alpha_1) \quad \in \quad \left\{ [0,1]^2 : \alpha_0+\alpha_1 < 1, \alpha_0 \leqslant \inf_z \mathbb{P}(D=1|Z=z), \alpha_1 \leqslant \inf_z \mathbb{P}(D=0|Z=z) \right\}.$$

Hence, the MTE is partially identified as a function of α_0 and α_1:

$$\begin{aligned} MTE(p;\alpha_0,\alpha_1) &= \frac{\partial \mathbb{E}[Y|P(Z)=p]}{\partial p}, \\ &= \frac{\partial \mathbb{E}\left[Y|\mathbb{P}(D=1|Z)=(1-\alpha_0-\alpha_1)p+\alpha_0\right]}{\partial p}, \\ &= (1-\alpha_0-\alpha_1) LIV\left((1-\alpha_0-\alpha_1)p+\alpha_0\right), \end{aligned}$$

where $LIV(p) \equiv \frac{\partial \mathbb{E}[Y|\mathbb{P}(D=1|Z)=p]}{\partial p}$ is the local instrumental variable (LIV) estimand. When the misclassification is symmetric, the above formula becomes:

$$MTE(p;\alpha) = (1-2\alpha) LIV\left((1-2\alpha)p+\alpha\right),$$

where

$$\alpha \leqslant \min\left\{ \inf_z \mathbb{P}(D=1|Z=z), \inf_z \mathbb{P}(D=0|Z=z) \right\}.$$

Numerical illustration of this special case. We assume in this illustration that the researcher knows that the rate of misclassification α is less than $1/2$. Consider the same example from the previous section (3.14) where $\rho = 0$ (i.e., ε is independent of V). Details on this illustration are given in Subsection G.2 in the appendix.

[7] Note however that in a recent paper, Haider and Stephens Jr. (2020) show that this assumption is invalid in routine empirical settings. In Subsections 4.2 and F.1, we discuss how one can allow the false positive/negative rates to depend on z.

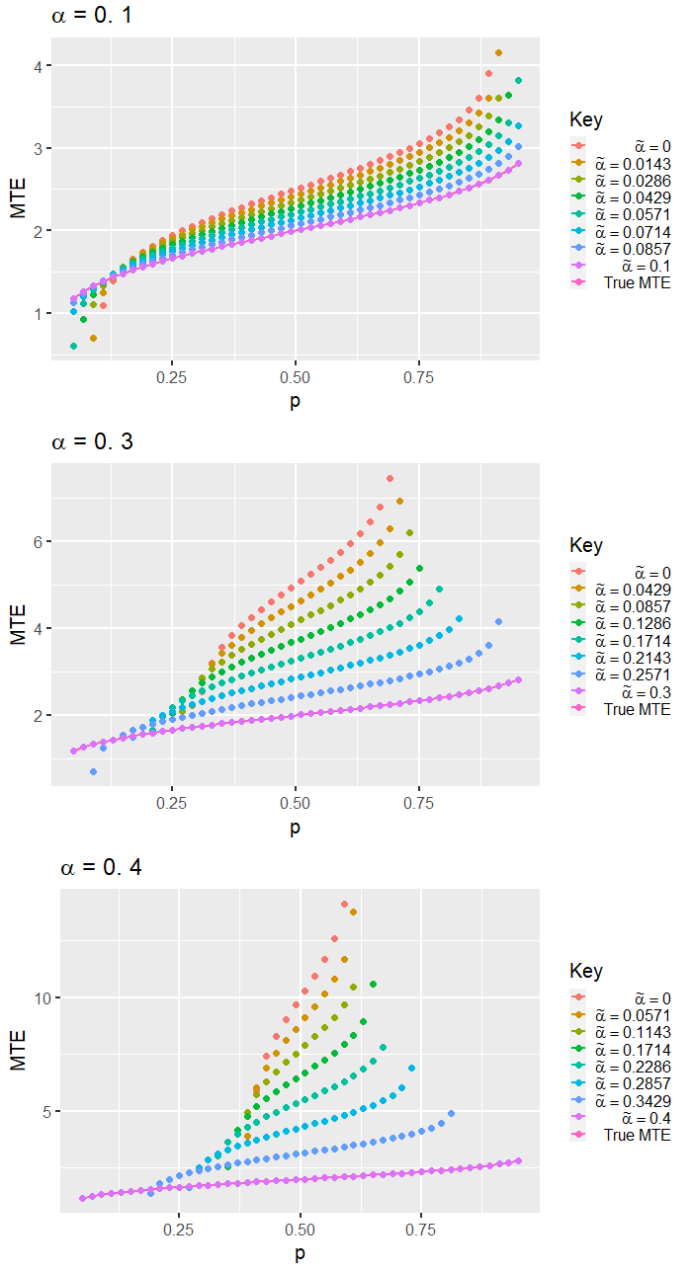


FIGURE 2. Numerical Illustrations of $MTE(p;\tilde{\alpha})$

Figure 2 numerically shows how $MTE(p;\tilde{\alpha})$ varies across different values of $\tilde{\alpha}$ where $\tilde{\alpha} \in [0, \alpha]$. Each panel illustrates the bounds for different values of the true α from 0.1 to 0.4, and 8 different values of $\tilde{\alpha}$ within the identification region of α are used to show the

corresponding plots for the MTE. We note that the identification region for the true MTE grows as the true α approaches 0.5, but the region always contains the true MTE.

4.2. **When false positive/negative rates depend on the instrument: semiparametric approach.** In this subsection, we allow for asymmetric misclassification. Unlike the symmetric case where the instrument Z is required to be independent of the misclassification variable ε, the asymmetric misclassification that we consider allow for dependence between Z and ε.

Under the assumption that Z is independent of V, the following equation still holds

$$\mathbb{P}(D=1|Z=z) \quad = \quad (1-\alpha_0(z)-\alpha_1(z))P(z)+\alpha_0(z), \tag{4.1}$$

where now $\alpha_1(z) \equiv \mathbb{P}(\varepsilon=1|V \leqslant P(z), Z=z)$ is the false negative misclassification rate, and $\alpha_0(z) \equiv \mathbb{P}(\varepsilon=1|V>P(z), Z=z)$ is the false positive misclassification rate. We are going to consider a semiparametric form of these misclassification rates. Before we do that, we show that there exists a semiparametrization for which the observed propensity score $\mathbb{P}(D=1|Z=z)$ is robust to any misclassification.

Proposition 3. *Suppose that Z is independent of V, $0<P(z)<1$ for all z, and $\alpha(z) \equiv \mathbb{P}(\varepsilon=1|Z=z) \leqslant \min\{\mathbb{P}(D=1|Z=z), \mathbb{P}(D=0|Z=z)\}$ for all z. Under the semiparametrization $\alpha_1(z)=\frac{\alpha(z)}{P(z)}$ and $\alpha_0(z)=\frac{\alpha(z)}{1-P(z)}$, we have $P(z)=\mathbb{P}(D=1|Z=z)$.* □

The proof of Proposition 3 follows from Equation (4.1).

Now, consider the following semiparametrization $\alpha_1(z)=\alpha_1 P(z)$ and $\alpha_0(z)=\alpha_0(1-P(z))$, where $\alpha_1, \alpha_0 \in [0, 1/2)$. This imposes that the false negative (positive) rate $\alpha_1(z)$ ($\alpha_0(z)$) is proportional to the probability of being treated (untreated). Then Equation (4.1) becomes

$$\mathbb{P}(D=1|Z=z) \quad = \quad (\alpha_0-\alpha_1)\big[P(z)\big]^2+(1-2\alpha_0)P(z)+\alpha_0.$$

When $\alpha_1=\alpha_0=\alpha$, we obtain $P(z)=\frac{\mathbb{P}(D=1|Z=z)-\alpha}{1-2\alpha}$, which is the same function we obtain under symmetric misclassification. In Subsection F.1, we discuss the cases $\alpha_0>\alpha_1$ and $\alpha_0<\alpha_1$.

5. Empirical Relevance of the MTE bounds

The identification of the MTE can help reveal the presence of heterogeneity in the treatment effect. It can also be useful in the estimation of policy relevant treatment effect

parameters (PRTEs) or conventional parameters such as the ATE, the average treatment effect on the treated (ATT), the average treatment effect on the untreated (ATU), the LATE, etc. Tables 1 and 2, which we borrow from Heckman and Vytlacil (2005), show the link between the MTE and those parameters. Unlike the weights in Heckman and Vytlacil (2005), the weights for the parameters ATT, ATU, and PRTE are not point-identified in our setting. They are only partially identified, as is the true propensity score $P(Z)$. Like the MTE, these policy parameters are also partially identified. In the next section, we explicitly derive analytical bounds for different LATEs when the instrument is multivalued discrete.

TABLE 1. Treatment effects as weighted averages of the MTE

$ATE = \mathbb{E}\left[Y_1 - Y_0\right] = \int_0^1 MTE\left(p\right) \cdot \omega_{ATE}\left(p\right) \, \mathrm{d}p$
$ATT = \mathbb{E}\left[Y_1 - Y_0 \vert D^* = 1\right] = \int_0^1 MTE\left(p\right) \cdot \omega_{ATT}\left(p\right) \, \mathrm{d}p$
$ATU = \mathbb{E}\left[Y_1 - Y_0 \vert D^* = 0\right] = \int_0^1 MTE\left(p\right) \cdot \omega_{ATU}\left(p\right) \, \mathrm{d}p$
$LATE(\underline{p}, \overline{p}) = \mathbb{E}\left[Y_1 - Y_0 \vert V \in \left[\underline{p}, \overline{p}\right]\right] = \int_{\underline{p}}^{\overline{p}} MTE\left(p\right) \cdot \omega_{LATE}\left(p\right) \, \mathrm{d}p$
$PRTE = \dfrac{\mathbb{E}\left[Y_a - Y_{a'}\right]}{\int_0^1 \left(F_{P_{a'}}\left(p\right) - F_{P_a}\left(p\right)\right)\left(p\right) \, \mathrm{d}p} = \int_0^1 MTE\left(p\right) \cdot \omega_{PRTE}\left(p, a, a'\right) \, \mathrm{d}p$ for two policies a and a' that affect only Z

TABLE 2. Weights

$\omega_{ATE}\left(p\right) = 1$
$\omega_{ATT}\left(p\right) = \dfrac{\int_p^1 f_{P(Z)}\left(u\right) \, \mathrm{d}u}{\int_0^1 \left(\int_p^1 f_{P(Z)}\left(u\right) \, \mathrm{d}u\right) \, \mathrm{d}p}$
$\omega_{ATU}\left(p\right) = \dfrac{\int_0^p f_{P(Z)}\left(u\right) \, \mathrm{d}u}{\int_0^1 \left(\int_0^p f_{P(Z)}\left(u\right) \, \mathrm{d}u\right) \, \mathrm{d}p}$
$\omega_{LATE}\left(p\right) = \dfrac{1}{\overline{p} - \underline{p}}$
$\omega_{PRTE}\left(p, a, a'\right) = \dfrac{F_{P_{a'}}\left(p\right) - F_{P_a}\left(p\right)}{\int_0^1 F_{P_{a'}}\left(p\right) - F_{P_a}\left(p\right) \, \mathrm{d}p}$

6. Extension to discrete instruments

Assumption 6 (Discrete instrument)**.** *The instrument Z is discrete with support $\{z_1, z_2, \ldots, z_K\}$ and the propensity score $p_\ell \equiv \mathbb{P}[D^* = 1|Z = z_\ell]$ satisfies $0 \leqslant p_1 < p_2 < \ldots < p_K \leqslant 1$.*

□

This assumption states that the ordering of the true propensity score is known, but the support $\{z_1, z_2, \ldots, z_K\}$ of the instrument does not necessarily have the same ranking. This assumption does require monotonicity in the propensity score. For example, when the false positive and negative rates do not depend on z, we have shown is Subsection 4.1 that the propensity score can be written as:

$$P(z) = \frac{\mathbb{P}(D = 1|Z = z) - \alpha_0}{1 - \alpha_0 - \alpha_1}.$$

As we can see, in such a case, the ordering of the true propensity score p_ℓ is the same as that of the reported propensity score $\mathbb{P}(D = 1|Z = z_\ell)$.

Let us replace p by p_ℓ and p' by $p_{\ell-1}$, and sum up (3.10) and (3.11). We have

$$\begin{aligned}
&\mathbb{P}(Y \in A|P(Z) = p_\ell) - \mathbb{P}(Y \in A|P(Z) = p_{\ell-1}) \\
&\quad = \int_{p_{\ell-1}}^{p_\ell} \mathbb{P}\left(Y_1 \in A|V = v\right) dv - \int_{p_{\ell-1}}^{p_\ell} \mathbb{P}\left(Y_0 \in A|V = v\right) dv, \\
&\quad = (p_\ell - p_{\ell-1})\mathbb{P}\left(Y_1 \in A|p_{\ell-1} < V \leqslant p_\ell\right) - (p_\ell - p_{\ell-1})\mathbb{P}\left(Y_0 \in A|p_{\ell-1} < V \leqslant p_\ell\right).
\end{aligned}$$

Therefore,

$$\mathbb{P}\left(Y_1 \in A|p_{\ell-1} < V \leqslant p_\ell\right) - \mathbb{P}\left(Y_0 \in A|p_{\ell-1} < V \leqslant p_\ell\right) = \frac{\mathbb{P}(Y \in A|P(Z) = p_\ell) - \mathbb{P}(Y \in A|P(Z) = p_{\ell-1})}{p_\ell - p_{\ell-1}}.$$

The analog of the result holds with expectations. Hence, we identify the MTE up to the function $P(z)$ as follows:

$$\begin{aligned}
\mathbb{E}\left[Y_1 - Y_0|p_{\ell-1} < V \leqslant p_\ell\right] &= \frac{\mathbb{E}[Y|P(Z) = p_\ell] - \mathbb{E}[Y|P(Z) = p_{\ell-1}]}{p_\ell - p_{\ell-1}}, \\
&= \frac{\mathbb{E}[Y|Z = z_\ell] - \mathbb{E}[Y|Z = z_{\ell-1}]}{p_\ell - p_{\ell-1}}.
\end{aligned}$$

We have

$$p_\ell - p_{\ell-1} = (p_L - p_1) - \sum_{k \neq \ell} (p_k - p_{k-1}). \tag{6.1}$$

Equations (3.12) and (6.1) imply the following additional bounds on $p_\ell - p_{\ell-1}$:

$$\begin{aligned}
&\max\left\{TV_{(Y,D=1)}(z_1,z_L), TV_{(Y,D=0)}(z_1,z_L)\right\} \\
&\qquad - \sum_{k\neq\ell} \min\left\{1, 2\alpha + \Delta_{DZ}(z_{k-1},z_k), 2(1-\alpha) - \Delta_{DZ}(z_{k-1},z_k)\right\} \\
&\leqslant p_\ell - p_{\ell-1} \leqslant \\
&\qquad \min\left\{1, 2\alpha + \Delta_{DZ}(z_1,z_L), 2(1-\alpha) - \Delta_{DZ}(z_1,z_L)\right\} \\
&\qquad\qquad - \sum_{k\neq\ell} \max\left\{TV_{(Y,D=1)}(z_{k-1},z_k), TV_{(Y,D=0)}(z_{k-1},z_k)\right\}.
\end{aligned}$$

Therefore, the following bounds hold for $p_\ell - p_{\ell-1}$:

$$LB_p(z_{\ell-1},z_\ell) \leqslant p_\ell - p_{\ell-1} \leqslant UB_p(z_{\ell-1},z_\ell),$$

where

$$\begin{aligned}
LB_p(z_{\ell-1},z_\ell) &\equiv \max\left\{LB_p^1(z_{\ell-1},z_\ell), LB_p^2(z_{\ell-1},z_\ell)\right\}, \\
UB_p(z_{\ell-1},z_\ell) &\equiv \min\left\{UB_p^1(z_{\ell-1},z_\ell), UB_p^2(z_{\ell-1},z_\ell)\right\}, \\
LB_p^1(z_{\ell-1},z_\ell) &= \max\left\{TV_{(Y,D=1)}(z_{\ell-1},z_\ell), TV_{(Y,D=0)}(z_{\ell-1},z_\ell)\right\}, \\
LB_p^2(z_{\ell-1},z_\ell) &= \max\left\{TV_{(Y,D=1)}(z_1,z_L), TV_{(Y,D=0)}(z_1,z_L)\right\} \\
&\qquad - \sum_{k\neq\ell} \min\left\{1, 2\alpha + \Delta_{DZ}(z_{k-1},z_k), 2(1-\alpha) - \Delta_{DZ}(z_{k-1},z_k)\right\}, \\
UB_p^1(z_{\ell-1},z_\ell) &= \min\left\{1, 2\alpha + \Delta_{DZ}(z_{\ell-1},z_\ell), 2(1-\alpha) - \Delta_{DZ}(z_{\ell-1},z_\ell)\right\}, \\
UB_p^2(z_{\ell-1},z_\ell) &= \min\left\{1, 2\alpha + \Delta_{DZ}(z_1,z_L), 2(1-\alpha) - \Delta_{DZ}(z_1,z_L)\right\} \\
&\qquad - \sum_{k\neq\ell} \max\left\{TV_{(Y,D=1)}(z_{k-1},z_k), TV_{(Y,D=0)}(z_{k-1},z_k)\right\}.
\end{aligned}$$

The proposition below holds.

Proposition 4. *Suppose that model (2.1) along with Assumptions 1–3, and 6 hold. Then, we have the following bounds for* $LATE(p_{\ell-1},p_\ell) \equiv \mathbb{E}\left[Y_1 - Y_0 | p_{\ell-1} < V \leqslant p_\ell\right]$*:*

$$\begin{aligned}
&\min\left\{\frac{\Delta_{YZ}(z_\ell,z_{\ell-1})}{UB_p(z_\ell,z_{\ell-1})}, \frac{\Delta_{YZ}(z_\ell,z_{\ell-1})}{LB_p(z_\ell,z_{\ell-1})}\right\} \\
&\qquad \leqslant LATE(p_{\ell-1},p_\ell) \leqslant \\
&\qquad\qquad \max\left\{\frac{\Delta_{YZ}(z_\ell,z_{\ell-1})}{UB_p(z_\ell,z_{\ell-1})}, \frac{\Delta_{YZ}(z_\ell,z_{\ell-1})}{LB_p(z_\ell,z_{\ell-1})}\right\}.
\end{aligned} \tag{6.2}$$

□

At this point, we do not have a result on the sharpness of the bounds in Proposition 4. This could be investigated in future work. However, when α is unknown (i.e., $\bar{\alpha} = 1$), these bounds are tighter than the existing bounds in Tommasi and Zhang (2020). Moreover, when α is unknown and the instrument Z is binary, the bounds in Proposition 4 are tighter than the existing Ura (2018) bounds, which appear to be the tightest in the literature before our work. In such a case, we show in Appendix C that the bounds are sharp. Indeed, when Z is binary and α is unknown, our bounds on $p_1 - p_0$ are

$$\max\left\{TV_{(Y,D=1)}(0,1), TV_{(Y,D=0)}(0,1)\right\} \leqslant p_1 - p_0 \leqslant 1,$$

while Ura's (2018) bounds are

$$\frac{1}{2}TV_{(Y,D=1)}(0,1) + \frac{1}{2}TV_{(Y,D=0)}(0,1) \leqslant p_1 - p_0 \leqslant 1.$$

7. Empirical illustration

To illustrate our methodology, we use data from the third wave of the Indonesia Family Life Survey (IFLS) fielded from June through November 2000. We build upon Carneiro, Lokshin, and Umapathi (2017) who estimate average and marginal returns to schooling in Indonesia using a semiparametric selection model. The authors use exogenous geographic variation in access to upper secondary schools to identify their model when ignoring the presence of measurement errors in the treatment variable. In such an analysis, researchers control for several family and village characteristics, namely father's and mother's education, an indicator of whether the community of residence was a village, religion, whether the location of residence is rural, province dummies, and distance from the village of residence to the nearest health post.

The IFLS is a household and community level panel survey that was conducted in 1993, 1997 and 2000. The sample was drawn from 321 randomly selected villages, spread among 13 Indonesian provinces containing 83% of the population, and consists of males aged 25–60 employees in public and private sectors. Females are excluded from the sample because of low labor force participation, self-employed workers are also excluded because it is difficult to measure their earnings. The sample size is 2608.

Following Carneiro, Lokshin, and Umapathi (2017), we define the dependent variable in the analysis as the log of the hourly wage (Y), which is constructed from self-reported monthly wages and hours worked per week. The treatment variable (D) is the indicator that the individual has an upper secondary or higher education (i.e., he completed at least 10 years of education). As we argue in Example 1, people often misreport their education

level. So, we observe their true education level with some measurement errors. The control variables (X) are indicator variables for age, indicators for the level of schooling completed by each of the parents (no education, elementary education, secondary education, and an indicator for unreported parental education), an indicator for whether the individual was living in a village at age 12, indicators for the province of residence, an indicator of rural residence, and distance (in kilometers) from the office of the head of the community of residence to the nearest community health post.

The instrumental variable (Z) for schooling is the distance (in kilometers) from the office of the community head to the nearest secondary school. The main assumption from Carneiro, Lokshin, and Umapathi (2017) is that if we consider two individuals with equally educated parents, with the same religion, living in a village which is located in an area that is equally rural, in the same province, and at the same distance of a health post, then distance to the nearest secondary school is uncorrelated with direct determinants of wages other than schooling. The authors present evidence that this assumption is likely to hold, suggesting that the IV is valid. In particular, they show that, once the previously mentioned variables are controlled for, there is no dependence between the distance to the nearest secondary school and whether the individual ever failed a grade in elementary school, how many times he repeated a grade in elementary school, and whether he had to work while attending elementary school. In addition, they show (using a different sample) that the distance variable is unrelated to test scores (Math, Bahasa, Science, and Social Studies) in elementary school. However, the validity of the distance to the nearest secondary school instrument remains highly questionable. For this reason, this exercise should be seen as illustrative.

Estimation Results. Estimation procedure follows the same steps as in the numerical illustration in Subsection 4.1, with the exception that we do not know the true DGP. We first assume symmetric misclassification, where ε is independent of V. This assumption implies that the conditional distribution of V given $\varepsilon = 1$ is linear. Next, we consider alternative assumptions about the misclassification mechanism in Appendix F, where ε depends on V and the conditional distributions of V given $\varepsilon = 1$ are respectively concave and convex. The results are roughly similar across these three different specifications. However, the MTE curve is fuzzier in the latter specifications. Finally, in Appendix F.1, we present the empirical results when allowing for dependence between the instrument and the misclassification. The pattern is similar to the previous specifications, though the MTE curve is more fuzzy.

By taking the infimum of estimates of $\mathbb{E}[D|Z=z]$ and $\mathbb{E}[1-D|Z=z]$ over z,[8] we partially identify the misclassification rate that can be supported by the data under such an assumption. We find that α must lie between 0 and 0.1574. We first estimate $E[D|Z=z]$ nonparametrically using a local linear method. Afterwards, we estimate the MTE nonparametrically using a local quadratic approximation, as recommended by Fan and Gijbels (1996) for estimating a first-order derivative, for 8 different values of $\tilde{\alpha}$ chosen within the previously identified region. To do so, we use the *R* package *nprobust* developed by Calonico, Cattaneo, and Farrell (2019). Figure 3 shows the results without the control variables X, but using only (Y, D, Z). It is difficult to add covariates nonparametrically. We net out the control variables X from the outcome variable by using residuals from a linear projection of Y on X. We then estimate $\mathbb{E}[D|Z, X]$ using a logit specification. We set X equal to its average in order to evaluate the MTE. The results are shown on Figure 5. Figure 4 displays the MTE when using the logit specification of $\mathbb{E}[D|Z]$ without controls.

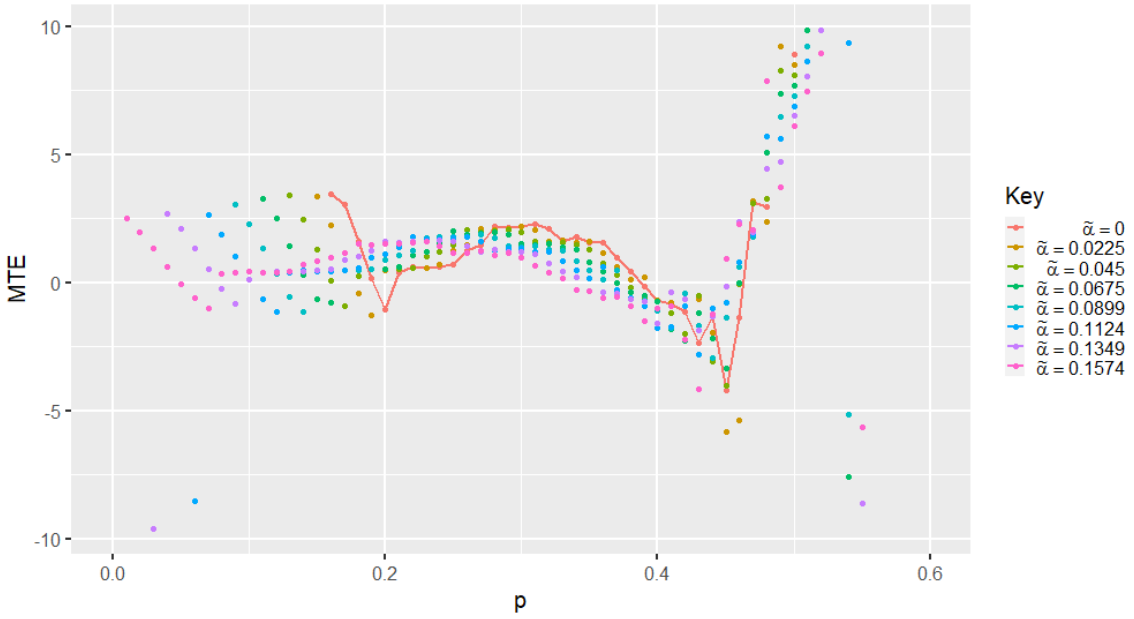


FIGURE 3. Estimated MTE Region

It is interesting that for all results, we observe some regions of p where the MTE is positive, and other regions where the MTE is negative. The estimate of the identified set for the MTE suggests that the return to upper secondary schooling is heterogeneous

[8]We are not conducting inference in this paper. We are treating the estimates are the true quantities. Inference will be explored in future research

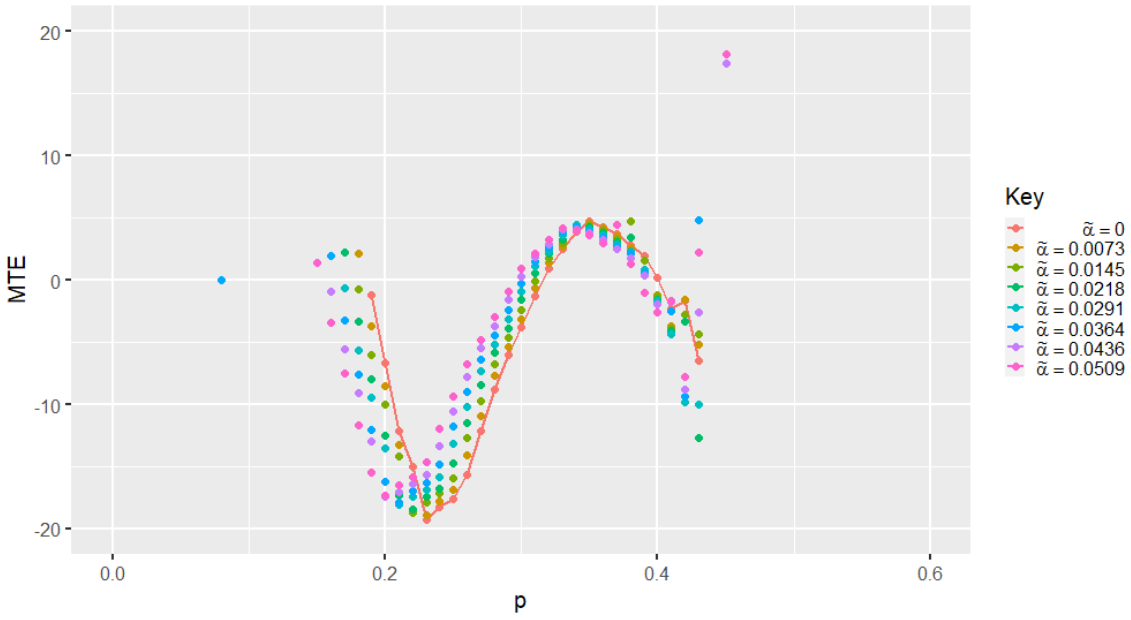


FIGURE 4. Estimated MTE Region (Logit)

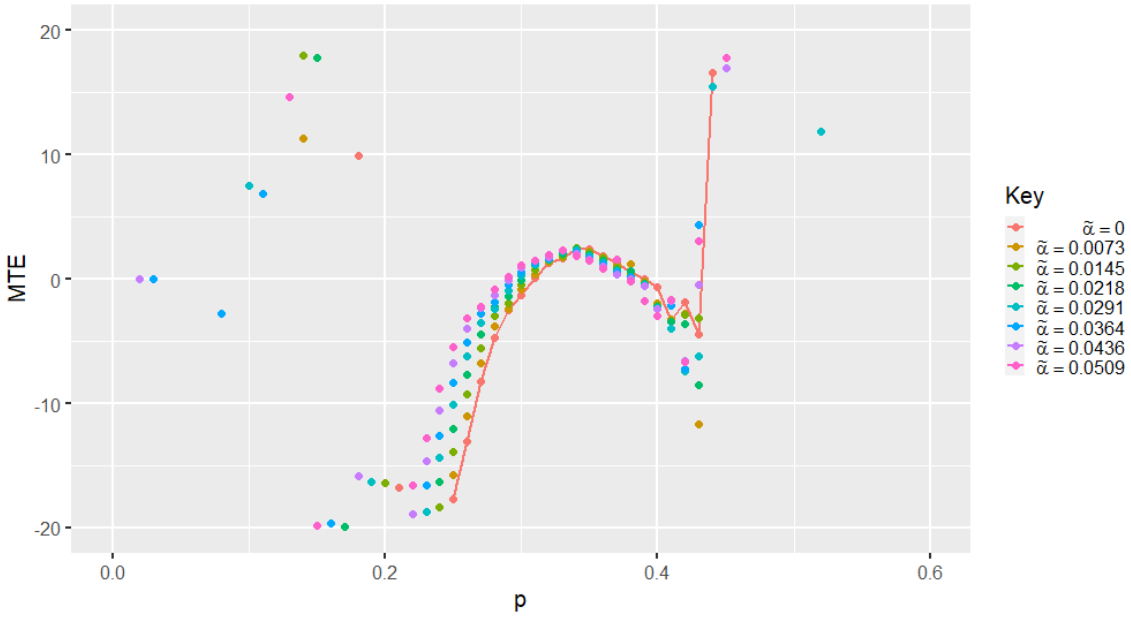


FIGURE 5. Estimated MTE Region (Logit, Residuals)

in Indonesia. Even when allowing for the presence of misreporting, the sign of the MTE is identified for some values of the quantiles of the unobserved cost of attending upper secondary school. However, we also note that MTE estimates are hard to be obtained for some regions of p because the mismeasured propensity score $\mathbb{E}[D|Z=z]$ does not have full support $[0,1]$. Particularly, for $\tilde{\alpha} \leqslant 0.1574$, we have extremely dispersed MTE estimates for p above 0.5 and do not even have any estimates for the higher values of p.

8. Conclusion

In this paper, we show that the MTE is generally partially identified in the presence of misclassification. We show that the MTE is equal to the derivative of the expectation of the observed outcome conditional on the true propensity score, which is partially identified. We provide nonparametric characterization of the identified set for the propensity score and the MTE. We use our MTE bounds to derive bounds on other commonly used parameters in the literature. We show that our bounds are tighter than the existing bounds for the local average treatment effect. We illustrate the methodology numerically and empirically. We investigate the measurement of the return to upper secondary schooling in Indonesia, and find that the return is heterogeneous for people at the cost margin. It is positive for some values of the schooling cost, while it is negative for other values.

We have not developed an inference method for the derived identified set for the MTE in this work. We believe that constructing a confidence set for this identified set could be worth exploring in future research.

Appendix A. Discussion about the model specification

One might think that the specification $D = D^*(1-\varepsilon) + (1-D^*)\varepsilon$ is too restrictive. But, it is general. To see this, we can write $D = D^* + (D - D^*) = D^* + \xi$, where $\xi = D - D^*$. We have $(D, D^*, \xi) \in \{(0,0,0),(1,1,0),(1,0,1),(0,1,-1)\}$. We can see that $D = 1 - D^*$ if $\xi \in \{-1,1\}$ and $D = D^*$ otherwise. Hence, we can write $D = D^*\left(1 - \mathbb{1}\left\{\xi \in \{-1,1\}\right\}\right) + (1-D^*)\mathbb{1}\left\{\xi \in \{-1,1\}\right\}$. By setting $\varepsilon = \mathbb{1}\left\{\xi \in \{-1,1\}\right\}$, we have $D = D^*(1-\varepsilon) + (1-D^*)\varepsilon$.

An easier way to see the specification is the following:

$$\begin{aligned} D &= D^*\left(1 - \mathbb{1}\left\{D \neq D^*\right\}\right) + (1-D^*)\mathbb{1}\left\{D \neq D^*\right\}, \\ &= D^*\left(1-\varepsilon\right) + (1-D^*)\varepsilon, \end{aligned}$$

where $\varepsilon = \mathbb{1}\left\{D \neq D^*\right\}$.

Appendix B. Proofs of Proposition 1

Since the function max and min are continuous, there exist α^* the inf in the lower bound on $P(z)$ is attained. Similar result holds for the upper. We propose two misclassification scenarios that yield the lower or upper bound for each value of α.

B.1. $\varepsilon = \mathbb{1}\{V \leqslant \alpha\}$**.** Here, we assume that the misclassification occurs when the unobserved heterogeneity V is less than or equal to α. Then, we have $\mathbb{P}(\varepsilon = 1) = \alpha$, and $F_{V|\varepsilon=1}(p) = \min\{p/\alpha, 1\}$. Hence, the conditional distribution of V given $\varepsilon = 1$ is concave. Using Equation (2.2), we obtain that $F_{V|\varepsilon=0}(p) = \frac{p - \min\{p,\alpha\}}{1-\alpha}$.

If $\alpha \leqslant P(z)$, then Equation (3.4) implies

$$\mathbb{P}(D=1|Z=z) = (1-\alpha)\left(\frac{P(z)-\alpha}{1-\alpha}\right) + \alpha(1 - \min\{P(z)/\alpha, 1\}) = P(z) - \alpha,$$

which leads to

$$P(z) = \mathbb{P}(D=1|Z=z) + \alpha. \tag{B.1}$$

If $\alpha \geqslant P(z)$, then Equation (3.4) implies

$$\mathbb{P}(D=1|Z=z) = (1-\alpha)\left(\frac{P(z)-P(z)}{1-\alpha}\right) + \alpha(1 - P(z)/\alpha) = \alpha - P(z),$$

which in turn implies

$$P(z) = \alpha - \mathbb{P}(D=1|Z=z). \tag{B.2}$$

B.2. $\varepsilon = \mathbb{1}\{V > 1-\alpha\}$. Given this specification for the misclassification, we have $\mathbb{P}(\varepsilon = 1) = \alpha$, and $F_{V|\varepsilon=0}(p) = \min\{\frac{p}{1-\alpha}, 1\}$. From there, we have $F_{V|\varepsilon=1}(p) = \max\{0, 1 - \frac{1-p}{\alpha}\}$. Hence, the conditional distribution of V given $\varepsilon = 1$ is convex.

From Equation (3.4), we have:

$$(1-\alpha)\min\left\{\frac{P(z)}{1-\alpha}, 1\right\} + \alpha\left(1 - \max\left\{0, 1 - \frac{1-P(z)}{\alpha}\right\}\right) = \mathbb{P}(D=1|Z=z).$$

If $1-\alpha \leqslant P(z)$, the above equation becomes:

$$(1-\alpha) + \alpha\left(1 - 1 + \frac{1-P(z)}{\alpha}\right) = \mathbb{P}(D=1|Z=z),$$

which implies

$$P(z) = \mathbb{P}(D=0|Z=z) + 1 - \alpha. \tag{B.3}$$

If $1-\alpha \geqslant P(z)$, the equation becomes:

$$(1-\alpha)\frac{P(z)}{1-\alpha} + \alpha(1-0) = \mathbb{P}(D=1|Z=z),$$

which implies

$$P(z) = \mathbb{P}(D=1|Z=z) - \alpha. \tag{B.4}$$

Appendix C. Proof of sharpness when the instrument is binary

In this section, we assume that the researcher has no information about the misclassification rate α. Suppose $\Delta_{DZ}(0,1) \neq 0$, and for each d, either $f_{Y,D|Z}(y,d|1) - f_{Y,D|Z}(y,d|0) \geqslant 0$ or $f_{Y,D|Z}(y,d|1) - f_{Y,D|Z}(y,d|0) \leqslant 0$ for all y. Define

$$\begin{aligned}
LB_p^1(0,1) &\equiv \max\left\{TV_{(Y,D=1)}(0,1), TV_{(Y,D=0)}(0,1)\right\} = |\Delta_{DZ}(0,1)|, \\
\tilde{P}(0) &\equiv \frac{1}{2}\left[1 + \frac{LB_p^1(0,1)}{\Delta_{DZ}(0,1)}\left(1 - 2\mathbb{E}[D|Z=1]\right)\right], \\
\tilde{P}(1) &\equiv \tilde{P}(0) + LB_p^1(0,1).
\end{aligned}$$

Notice that $\tilde{P}(0)$ and $\tilde{P}(1)$ are well-defined probabilities. Define a joint distribution on $(\tilde{Y}_0, \tilde{Y}_1, \tilde{V}, \tilde{\varepsilon})$.

$$\begin{aligned}
&\mathbb{P}(\tilde{V} \leqslant v) = v,\\
&\mathbb{P}(\tilde{Y}_1 \leqslant y_1, \tilde{\varepsilon} = 1 | \tilde{V} \leqslant \tilde{P}(z)) = \mathbb{P}(Y \leqslant y_1, D = 0 | Z = z),\\
&\mathbb{P}(\tilde{Y}_1 \leqslant y_1, \tilde{\varepsilon} = 0 | \tilde{V} \leqslant \tilde{P}(z)) = \mathbb{P}(Y \leqslant y_1, D = 1 | Z = z),\\
&\mathbb{P}(\tilde{Y}_0 \leqslant y_0, \tilde{\varepsilon} = 0 | \tilde{V} > \tilde{P}(z)) = \mathbb{P}(Y \leqslant y_0, D = 0 | Z = z),\\
&\mathbb{P}(\tilde{Y}_0 \leqslant y_0, \tilde{\varepsilon} = 1 | \tilde{V} > \tilde{P}(z)) = \mathbb{P}(Y \leqslant y_0, D = 1 | Z = z),
\end{aligned}$$

for each $z \in \{0, 1\}$. Define

$$\begin{aligned}
&\mathbb{P}(\tilde{Y}_1 \leqslant y_1, \tilde{Y}_0 \leqslant y_0, \tilde{\varepsilon} = \ell, \tilde{V} \leqslant \tilde{P}(z) | Z = \check{z}) =\\
&\qquad \mathbb{P}(\tilde{Y}_1 \leqslant y_1 | \tilde{\varepsilon} = \ell, \tilde{V} \leqslant \tilde{P}(z)) \mathbb{P}(\tilde{Y}_0 \leqslant y_0 | \tilde{\varepsilon} = \ell, \tilde{V} \leqslant \tilde{P}(z)) \mathbb{P}(\tilde{\varepsilon} = \ell | \tilde{V} \leqslant \tilde{P}(z)) \tilde{P}(z),\\
&\mathbb{P}(\tilde{Y}_1 \leqslant y_1, \tilde{Y}_0 \leqslant y_0, \tilde{\varepsilon} = \ell, \tilde{V} > \tilde{P}(z) | Z = \check{z}) =\\
&\qquad \mathbb{P}(\tilde{Y}_1 \leqslant y_1 | \tilde{\varepsilon} = \ell, \tilde{V} > \tilde{P}(z)) \mathbb{P}(\tilde{Y}_0 \leqslant y_0 | \tilde{\varepsilon} = \ell, \tilde{V} > \tilde{P}(z)) \mathbb{P}(\tilde{\varepsilon} = \ell | \tilde{V} > \tilde{P}(z)) (1 - \tilde{P}(z)),
\end{aligned}$$

for each $\ell \in \{0, 1\}$, and each $z \in \{0, 1\}$.

We can verify that the above proposed joint distribution is well-defined, $Z \perp\!\!\!\perp (\tilde{Y}_0, \tilde{Y}_1, \tilde{V}, \tilde{\varepsilon})$ (Assumption 1), and $\tilde{V} \sim \mathcal{U}_{[0,1]}$ (Assumption 2), the joint distribution is compatible with the data, that is, it satisfies the following conditions:

$$\begin{aligned}
\mathbb{P}(Y \in A, D = 1 | Z = z) &= \mathbb{P}(\tilde{Y}_1 \in A, \tilde{\varepsilon} = 0, \tilde{V} \leqslant \tilde{P}(z)) + \mathbb{P}(\tilde{Y}_0 \in A, \tilde{\varepsilon} = 1, \tilde{V} > \tilde{P}(z)),\\
\mathbb{P}(Y \in A, D = 0 | Z = z) &= \mathbb{P}(\tilde{Y}_1 \in A, \tilde{\varepsilon} = 1, \tilde{V} \leqslant \tilde{P}(z)) + \mathbb{P}(\tilde{Y}_0 \in A, \tilde{\varepsilon} = 0, \tilde{V} > \tilde{P}(z)),
\end{aligned}$$

for all z, and all Borel set A.

Now, we are going to show that $\mathbb{E}[\tilde{Y}_1 - \tilde{Y}_0 | \tilde{P}(0) < \tilde{V} \leqslant \tilde{P}(1)] = \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{LB_p^1(0,1)}$. From the last equations, we have

$$\mathbb{P}(Y \in A | Z = z) = \mathbb{P}(\tilde{Y}_1 \in A, \tilde{V} \leqslant \tilde{P}(z)) + \mathbb{P}(\tilde{Y}_0 \in A, \tilde{V} > \tilde{P}(z)).$$

Therefore,

$$\begin{aligned}
\mathbb{P}(Y \in A | Z = 1) - \mathbb{P}(Y \in A | Z = 0) &= \mathbb{P}(\tilde{Y}_1 \in A, \tilde{P}(0) < \tilde{V} \leqslant \tilde{P}(1))\\
&\qquad - \mathbb{P}(\tilde{Y}_0 \in A, \tilde{P}(0) < \tilde{V} \leqslant \tilde{P}(z)).
\end{aligned}$$

Using Bayes' rule, we have

$$\begin{aligned}
&\mathbb{P}(Y \in A | Z = 1) - \mathbb{P}(Y \in A | Z = 0)\\
&\qquad \left[\mathbb{P}(\tilde{Y}_1 \in A | \tilde{P}(0) < \tilde{V} \leqslant \tilde{P}(1)) - \mathbb{P}(\tilde{Y}_0 \in A | \tilde{P}(0) < \tilde{V} \leqslant \tilde{P}(1))\right] \left(\tilde{P}(1) - \tilde{P}(0)\right).
\end{aligned}$$

Hence,

$$\mathbb{P}(\tilde{Y}_1 \in A|\tilde{P}(0) < \tilde{V} \leqslant \tilde{P}(1)) - \mathbb{P}(\tilde{Y}_0 \in A|\tilde{P}(0) < \tilde{V} \leqslant \tilde{P}(1)) = \frac{\mathbb{P}(Y \in A|Z=1) - \mathbb{P}(Y \in A|Z=0)}{\tilde{P}(1) - \tilde{P}(0)}$$

Finally, the expectation version of the result also holds, and we obtain

$$\mathbb{E}[\tilde{Y}_1 - \tilde{Y}_0|\tilde{P}(0) < \tilde{V} \leqslant \tilde{P}(1)] = \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{LB_p^1(0,1)}.$$

We have just shown that one of the bounds for the $LATE$ is achieved. Next, we will show that the other bound $\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]$ is also achievable. This case implies that $P(1) - P(0) = 1$, which is possible only when $P(1) = 1$ and $P(0) = 0$, that is, there is full compliance: $D^* = Z$. This is possible if $\mathbb{P}(D=0|Z=1) = \mathbb{P}(D=1|Z=0)$. Indeed, $\tilde{P}(1) = 1$ and $\tilde{P}(0) = 0$ imply that $\mathbb{P}(D=1|Z=0) = \mathbb{P}(\tilde{\varepsilon}=1)$, and $\mathbb{P}(D=0|Z=1) = \mathbb{P}(\tilde{\varepsilon}=1)$.

Appendix D. Proof of sharpness when the instrument is continuous

Proof. For each $\alpha \in (0,1)$, we need to find a joint distribution on the vector $\left(\tilde{Y}_1, \tilde{Y}_0, \tilde{V}, \tilde{\varepsilon}, Z\right)$, such that it satisfies model (2.1) and Assumptions 1, 2, and 5, and induces the joint distribution on (Y, D, Z). For any function $P(z)$ satisfying the constraints in (3.18)-(3.26), define:

$$\begin{aligned}
\mathbb{P}(\tilde{\varepsilon}=1|Z=z) &= \alpha, \\
f_{\tilde{V}|\tilde{\varepsilon}=0,Z=z}(p) &= \frac{1 + \frac{\partial \mathbb{P}(D=1|P(Z)=p)}{\partial p}}{2(1-\alpha)}, \\
f_{\tilde{V}|\tilde{\varepsilon}=1,Z=z}(p) &= \frac{1 - \frac{\partial \mathbb{P}(D=1|P(Z)=p)}{\partial p}}{2\alpha}, \\
\mathbb{P}(\tilde{Y}_1 \leqslant y|\tilde{V}=p, \tilde{\varepsilon}=1, Z=z) &= \frac{\alpha f_{\tilde{V}|\tilde{\varepsilon}=1}(p)\kappa_0(y;p) - (1-\alpha) f_{\tilde{V}|\tilde{\varepsilon}=0}(p)\kappa_1(y;p)}{\alpha f_{\tilde{V}|\tilde{\varepsilon}=1}(p) - (1-\alpha) f_{\tilde{V}|\tilde{\varepsilon}=0}(p)}, \\
\mathbb{P}(\tilde{Y}_1 \leqslant y|\tilde{V}=p, \tilde{\varepsilon}=0, Z=z) &= \frac{\alpha f_{\tilde{V}|\tilde{\varepsilon}=1}(p)\kappa_0(y;p) - (1-\alpha) f_{\tilde{V}|\tilde{\varepsilon}=0}(p)\kappa_1(y;p)}{\alpha f_{\tilde{V}|\tilde{\varepsilon}=1}(p) - (1-\alpha) f_{\tilde{V}|\tilde{\varepsilon}=0}(p)}, \\
\mathbb{P}(\tilde{Y}_0 \leqslant y|\tilde{V}=p, \tilde{\varepsilon}=1, Z=z) &= \frac{(1-\alpha) f_{\tilde{V}|\tilde{\varepsilon}=0}(p)\kappa_0(y;p) - \alpha f_{\tilde{V}|\tilde{\varepsilon}=1}(p)\kappa_1(y;p)}{\alpha f_{\tilde{V}|\tilde{\varepsilon}=1}(p) - (1-\alpha) f_{\tilde{V}|\tilde{\varepsilon}=0}(p)}, \\
\mathbb{P}(\tilde{Y}_0 \leqslant y|\tilde{V}=p, \tilde{\varepsilon}=0, Z=z) &= \frac{(1-\alpha) f_{\tilde{V}|\varepsilon=0}(p)\kappa_0(y;p) - \alpha f_{\tilde{V}|\varepsilon=1}(p)\kappa_1(y;p)}{\alpha f_{\tilde{V}|\tilde{\varepsilon}=1}(p) - (1-\alpha) f_{\tilde{V}|\varepsilon=0}(p)},
\end{aligned}$$

where $\kappa_1(y;p) = \frac{\partial \mathbb{P}(Y \leqslant y, D=1|P(Z)=p)}{\partial p}$, and $\kappa_0(y;p) = \frac{\partial \mathbb{P}(Y \leqslant y, D=0|P(Z)=p)}{\partial p}$. Define

$$\mathbb{P}(\tilde{Y}_0 \leqslant y_0, \tilde{Y}_1 \leqslant y_1 | \tilde{V} = p, \varepsilon = \ell, Z = z) = \mathbb{P}(\tilde{Y}_0 \leqslant y_0 | \tilde{V} = p, \varepsilon = \ell, Z = z)\mathbb{P}(\tilde{Y}_1 \leqslant y_1 | \tilde{V} = p, \varepsilon = \ell, Z = z).$$

It is easy to check that the above quantities are well-defined probabilities/distributions under the constraints (3.18)-(3.25), and the vector $\left(\tilde{Y}_1, \tilde{Y}_0, \tilde{V}, \tilde{\varepsilon}, Z\right)$ satisfies Assumptions 1, 2, and 5. Define

$$\begin{cases} \tilde{Y} &= \tilde{Y}_1 \tilde{D}^* + \tilde{Y}_0(1 - \tilde{D}^*) \\ \tilde{D}^* &= \mathbb{1}\left\{\tilde{V} \leqslant P(Z)\right\} \\ \tilde{D} &= \tilde{D}^*(1 - \tilde{\varepsilon}) + (1 - \tilde{D}^*)\tilde{\varepsilon} \end{cases} \tag{D.1}$$

We will now show that the vector $\left(\tilde{Y}, \tilde{D}, Z\right)$ has the same distribution as the vector (Y, D, Z). We have

$$\begin{aligned} \mathbb{P}(\tilde{Y} \leqslant y, \tilde{D} = 1 | Z = z) &= \mathbb{P}(\tilde{Y} \leqslant y, \tilde{D} = 1 | P(Z) = P(z)), \\ &= (1-\alpha) \int_0^{P(z)} \mathbb{P}\left(\tilde{Y}_1 \leqslant y | \tilde{V} = v, \tilde{\varepsilon} = 0\right) f_{\tilde{V}|\tilde{\varepsilon}=0}(v) dv \\ &\quad + \alpha \int_{P(z)}^1 \mathbb{P}\left(\tilde{Y}_0 \leqslant y | \tilde{V} = v, \tilde{\varepsilon} = 1\right) f_{\tilde{V}|\tilde{\varepsilon}=1}(v) dv, \\ &= (1-\alpha) \int_0^1 \mathbb{P}\left(\tilde{Y}_1 \leqslant y | \tilde{V} = v, \tilde{\varepsilon} = 0\right) f_{\tilde{V}|\tilde{\varepsilon}=0}(v) dv \\ &\quad - (1-\alpha) \int_{P(z)}^1 \mathbb{P}\left(\tilde{Y}_1 \leqslant y | \tilde{V} = v, \tilde{\varepsilon} = 0\right) f_{\tilde{V}|\tilde{\varepsilon}=0}(v) dv \\ &\quad + \alpha \int_{P(z)}^1 \mathbb{P}\left(\tilde{Y}_0 \leqslant y | \tilde{V} = v, \tilde{\varepsilon} = 1\right) f_{\tilde{V}|\tilde{\varepsilon}=1}(v) dv, \\ &= (1-\alpha) \int_0^1 \mathbb{P}\left(\tilde{Y}_1 \leqslant y | \tilde{V} = v, \tilde{\varepsilon} = 0\right) f_{\tilde{V}|\tilde{\varepsilon}=0}(v) dv \\ &\quad + \int_{P(z)}^1 -(1-\alpha) f_{\tilde{V}|\tilde{\varepsilon}=0}(v) \mathbb{P}\left(\tilde{Y}_1 \leqslant y | \tilde{V} = v, \tilde{\varepsilon} = 0\right) \\ &\quad + \alpha f_{\tilde{V}|\tilde{\varepsilon}=1}(v) \mathbb{P}\left(\tilde{Y}_0 \leqslant y | \tilde{V} = v, \tilde{\varepsilon} = 1\right) dv, \end{aligned}$$

where the first equality holds from Equation (3.26). Given the definition of our DGP, we have

$$\begin{aligned} &-(1-\alpha) f_{\tilde{V}|\tilde{\varepsilon}=0}(v) \mathbb{P}\left(\tilde{Y}_1 \leqslant y | \tilde{V} = v, \tilde{\varepsilon} = 0\right) + \alpha f_{\tilde{V}|\tilde{\varepsilon}=1}(v) \mathbb{P}\left(\tilde{Y}_0 \leqslant y | \tilde{V} = v, \tilde{\varepsilon} = 1\right) \\ &\quad = -\kappa_1(y;v) = -\frac{\partial \mathbb{P}(Y \leqslant y, D = 1 | P(Z) = v)}{\partial v}. \end{aligned}$$

Therefore,

$$\begin{aligned}\int_{P(z)}^{1} &-(1-\alpha) f_{\tilde{V}|\tilde{\varepsilon}=0}(v)\mathbb{P}\left(\tilde{Y}_1 \leqslant y|\tilde{V}=v,\tilde{\varepsilon}=0\right)+\alpha f_{\tilde{V}|\tilde{\varepsilon}=1}(v)\mathbb{P}\left(\tilde{Y}_0 \leqslant y|\tilde{V}=v,\tilde{\varepsilon}=1\right)dv \\ &= \mathbb{P}(Y\leqslant y, D=1|P(Z)=P(z)) - \mathbb{P}(Y\leqslant y, D=1|P(Z)=1), \\ &= \mathbb{P}(Y\leqslant y, D=1|Z=z) - \mathbb{P}(Y\leqslant y, D=1|P(Z)=1)\end{aligned}$$

At this point, it remains to show that

$$\mathbb{P}(Y\leqslant y, D=1|P(Z)=1) = (1-\alpha)\int_0^1 \mathbb{P}\left(\tilde{Y}_1\leqslant y|\tilde{V}=v,\tilde{\varepsilon}=0\right) f_{\tilde{V}|\tilde{\varepsilon}=0}(v)dv.$$

This equality holds from condition (3.24).

Similarly, $\mathbb{P}(\tilde{Y}\leqslant y, \tilde{D}=0|Z=z) = \mathbb{P}(Y\leqslant y, D=0|Z=z)$. □

Appendix E. Allowing for dependence between misclassification and IV

Similarly to Equation 3.1, we have under $Z \perp\!\!\!\perp (Y_d, V)$, $d=0,1$:

$$f_{Y|Z}(y|z) = \int_0^{P(z)} f_{Y_1|V}(y|v)\,dv + \int_{P(z)}^1 f_{Y_0|V}(y|v)\,dv.$$

Hence, for any $P(z') < P(z)$ we have

$$\begin{aligned}f_{Y|Z}(y|z) - f_{Y|Z}\left(y|z'\right) &= \int_{P(z')}^{P(z)} f_{Y_1|V}(y|v)dv \\ &\quad - \int_{P(z')}^{P(z)} f_{Y_0|V}(y|v)dv,\end{aligned}$$

Using the triangle inequality, we have

$$\begin{aligned}\left|f_{Y|Z}(y|z) - f_{Y|Z}\left(y|z'\right)\right| &\leqslant \int_{P(z')}^{P(z)} f_{Y_1|V}(y|v)dv \\ &\quad + \int_{P(z')}^{P(z)} f_{Y_0|V}(y|v)dv.\end{aligned}$$

Therefore, by integrating each side over the support $\mathcal{Y}$ and using the Fubini-Tonelli theorem, we have

$$\int_{\mathcal{Y}} \left|f_{Y|Z}(y|z) - f_{Y|Z}\left(y|z'\right)\right| d\mu_Y(y) \leqslant \int_{P(z')}^{P(z)} dv + \int_{P(z')}^{P(z)} dv = 2\left(P(z)-P(z')\right).$$

Hence, we have $TV_Y(z',z) \leqslant P(z) - P(z') \leqslant 1$, where

$$TV_Y(z',z) \equiv \frac{1}{2}\int_{\mathcal{Y}} \left|f_{Y|Z}(y|z) - f_{Y|Z}\left(y|z'\right)\right| d\mu_Y(y).$$

Then, we have

$$\min\left\{\frac{\Delta_{YZ}(z',z)}{TV_Y(z',z)},\Delta_{YZ}(z',z)\right\} \leqslant \frac{\mathbb{E}[Y|P(Z)=P(z)]-\mathbb{E}[Y|P(Z)=P(z')]}{P(z)-P(z')} \leqslant \max\left\{\frac{\Delta_{YZ}(z',z)}{TV_Y(z',z)},\Delta_{YZ}(z',z)\right\}.$$

Therefore, the following bounds hold for the MTE:

$$\min\left\{\lim_{z'\to z}\frac{\Delta_{YZ}(z',z)}{TV_Y(z',z)},0\right\} \leqslant MTE(P(z)) \leqslant \max\left\{\lim_{z'\to z}\frac{\Delta_{YZ}(z',z)}{TV_Y(z',z)},0\right\}. \tag{E.1}$$

These bounds are wider those derived in Subsection 3.3 under Assumptions 1-4.

Appendix F. Additional empirical results

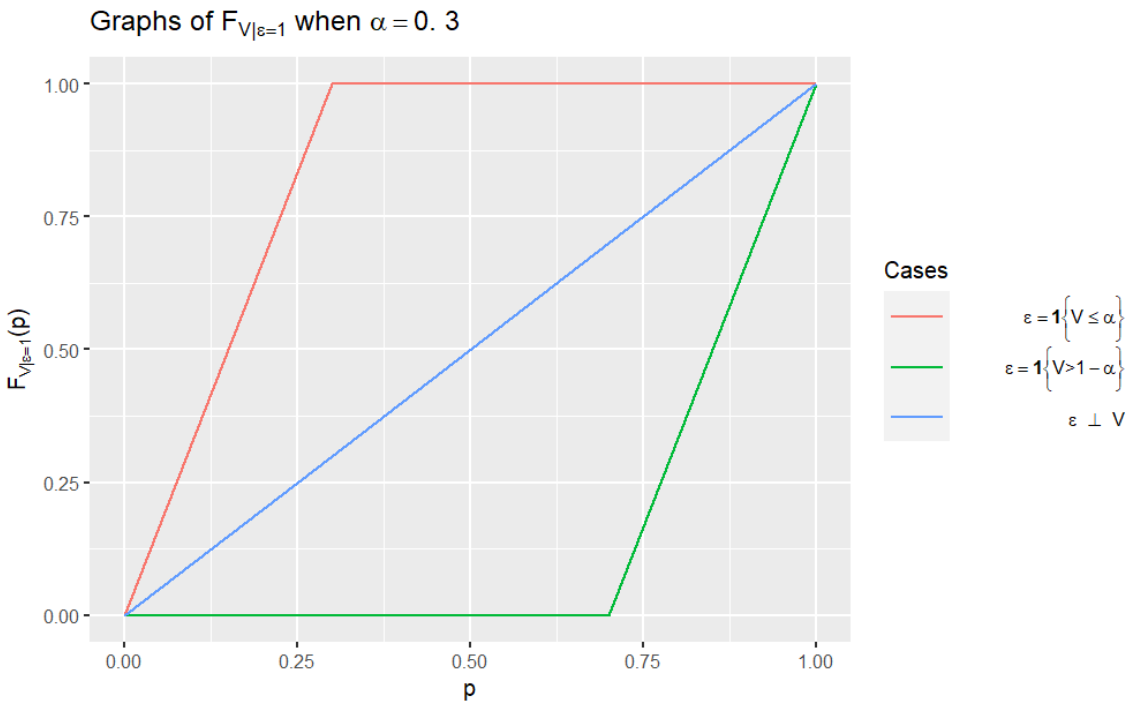


Figure 6. Illustration of $F_{V|\varepsilon=1}$ under Various Specifications

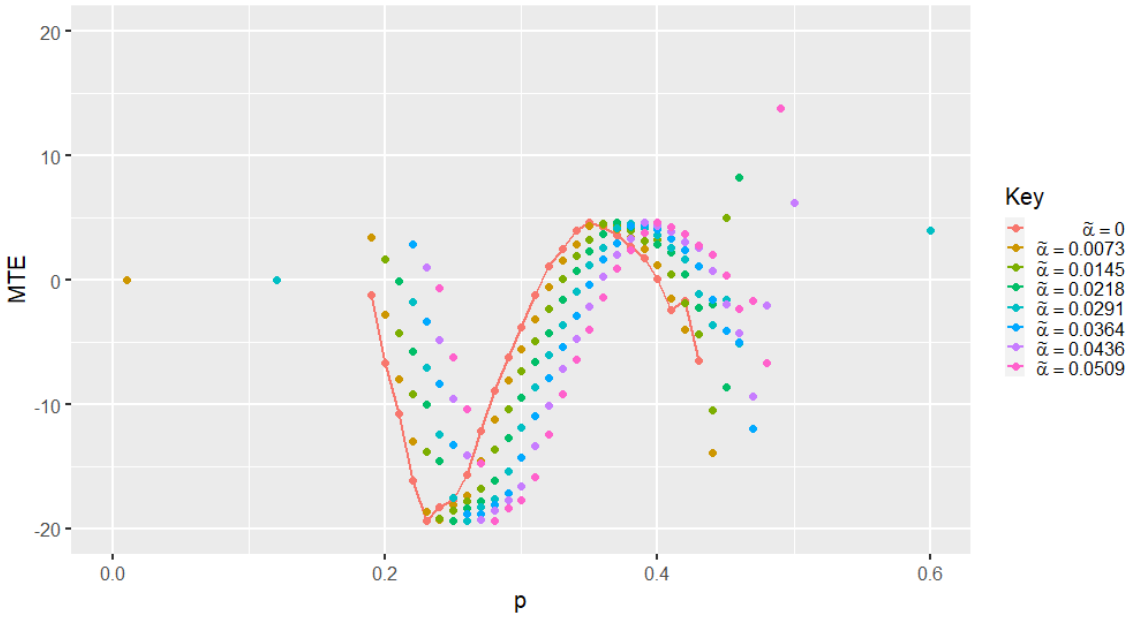


FIGURE 7. Estimated MTE Region ($\varepsilon = \mathbb{1}\{V \leqslant \alpha\}$, Logit)

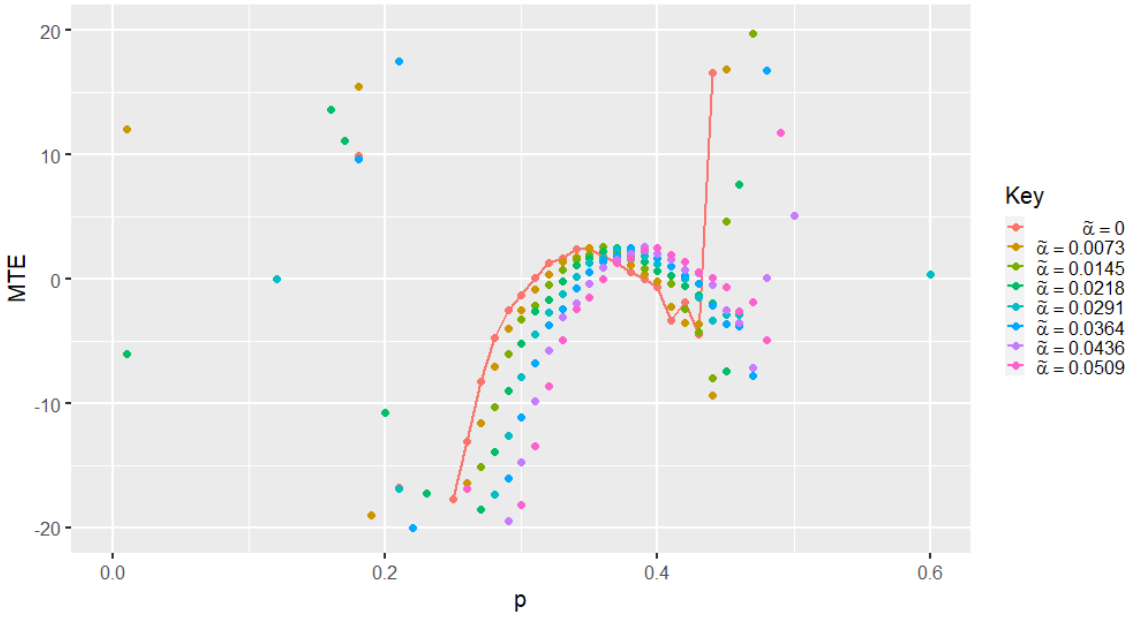


FIGURE 8. Estimated MTE Region ($\varepsilon = \mathbb{1}\{V \leqslant \alpha\}$, Logit, Residuals)

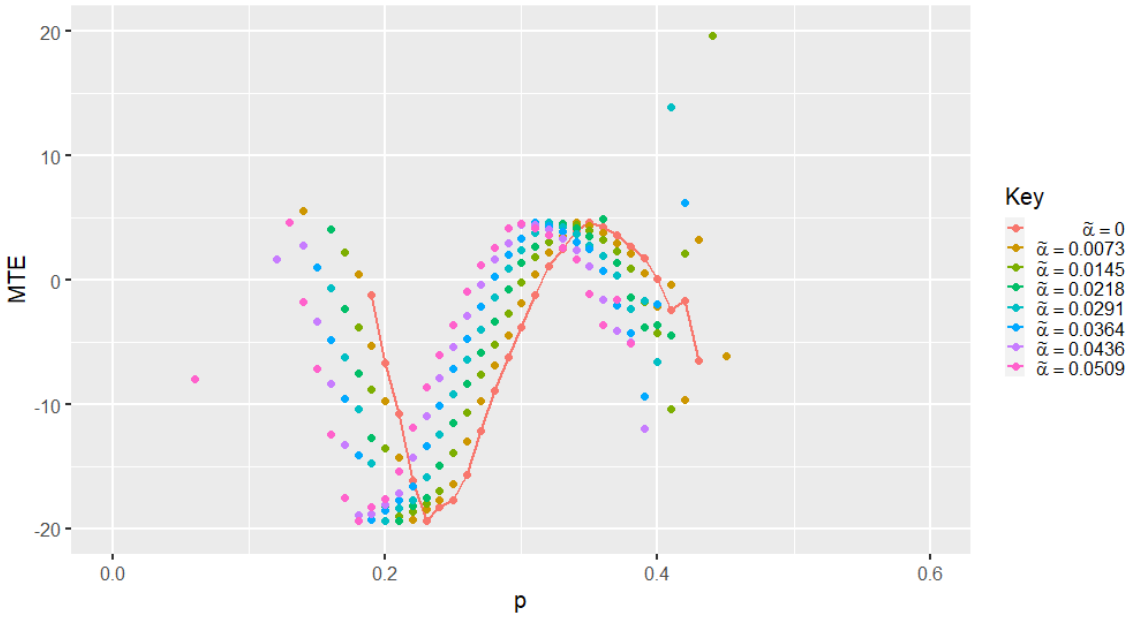


FIGURE 9. Estimated MTE Region ($\varepsilon = \mathbb{1}\{V > 1 - \alpha\}$, Logit)

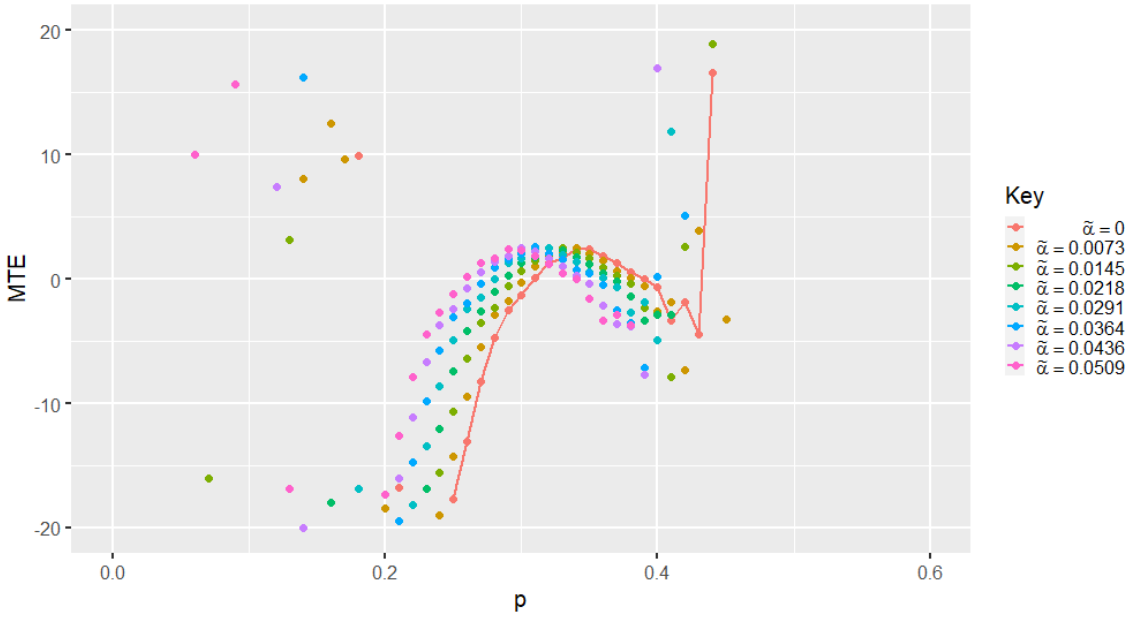


FIGURE 10. Estimated MTE Region ($\varepsilon = \mathbb{1}\{V > 1 - \alpha\}$, Logit, Residuals)

F.1. **Allowing false positive/negative rates to depend on the instrument.** Consider the following semi-parameterization of the misclassification rates $\alpha_0(z)$ and $\alpha_1(z)$:

$$\begin{aligned}\alpha_0(z) &\equiv \alpha_0 \mathbb{P}\big(V > P(z)\big) \\ &= \alpha_0\big(1 - P(z)\big), \\ \alpha_1(z) &\equiv \alpha_1 \mathbb{P}\big(V \leqslant P(z)\big) \\ &= \alpha_1 P(z),\end{aligned}$$

for parameters α_0 and α_1, where

$$(\alpha_0, \alpha_1) \in \Big[0, \frac{1}{2}\Big)^2$$

Hence, Equation (3.3) implies

$$\begin{aligned}\mathbb{P}(D = 1|Z = z) &= \Big(1 - \alpha_0\big(1 - P(z)\big) - \alpha_1 P(z)\Big)P(z) + \alpha_0\big(1 - P(z)\big) \\ &= (\alpha_0 - \alpha_1)\big(P(z)\big)^2 + (1 - 2\alpha_0)P(z) + \alpha_0,\end{aligned}$$

and we can solve for $P(z)$.

F.1.1. *Case 1:* $\alpha_0 = \alpha_1$. $P(z)$ can be solved for:

$$P(z) = \frac{\mathbb{P}(D = 1|Z = z) - \alpha}{1 - 2\alpha},$$

where $\alpha \equiv \alpha_0 = \alpha_1$.

Thus, α is partially identified:

$$\alpha \in \left\{ \Big[0, \frac{1}{2}\Big) : P(z) = \frac{\mathbb{P}(D = 1|Z = z) - \alpha}{1 - 2\alpha} \in [0, 1] \text{ for all } z \right\},$$

Note that we do not assume independence between ε and V, but the same results in Figures 3, 4, or 5 would be obtained.

F.1.2. *Case 2:* $\alpha_0 > \alpha_1$. We define the discriminant as

$$\Delta(z) = (1 - 2\alpha_0)^2 - 4(\alpha_0 - \alpha_1)\big(\alpha_0 - \mathbb{P}(D = 1|Z = z)\big).$$

Note that $\Delta(z)$ should be non-negative to have $P(z)$ real-valued, and

$$\begin{aligned}P_1(z) &\equiv \frac{-(1 - 2\alpha_0) - \sqrt{\Delta(z)}}{2(\alpha_0 - \alpha_1)}, \\ P_2(z) &\equiv \frac{-(1 - 2\alpha_0) + \sqrt{\Delta(z)}}{2(\alpha_0 - \alpha_1)},\end{aligned}$$

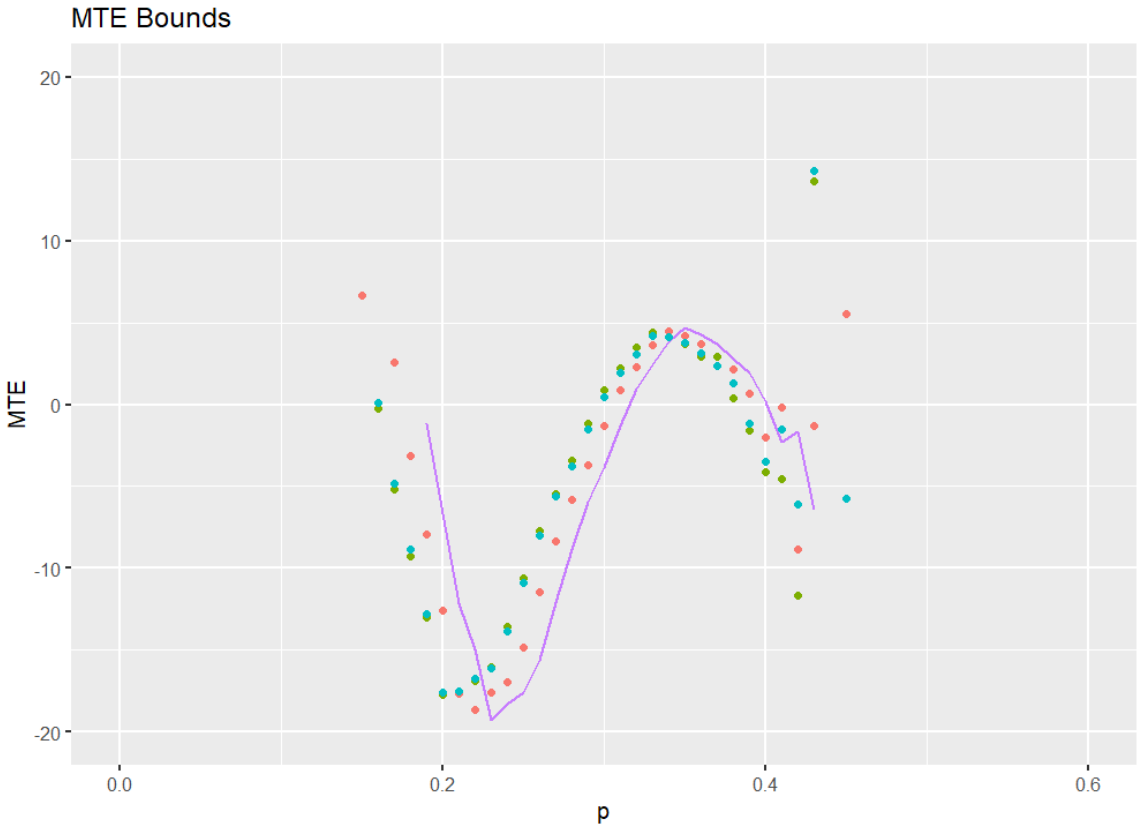
are the two possible candidates for $P(z)$.

Hence, α_0 and α_1 are partially identified:

$$(\alpha_0, \alpha_1) \in \left\{ \left[0, \frac{1}{2}\right)^2 : \alpha_0 > \alpha_1, \Delta(z) \geqslant 0, \text{ and } P_2(z) \in [0,1] \text{ for all } z \right\},$$

and $P(z) = P_2(z)$ as $P_1(z) < 0$ given the other conditions.

Using each point in the identification region for (α_0, α_1), we get $P(z)$ and estimate MTE with the `lprobust` package. The result is shown in Figure 11. If we impose the false negative rate α_1 is zero (i.e., anyone who has a least an upper secondary education reports correctly), then we obtain Figure 12.

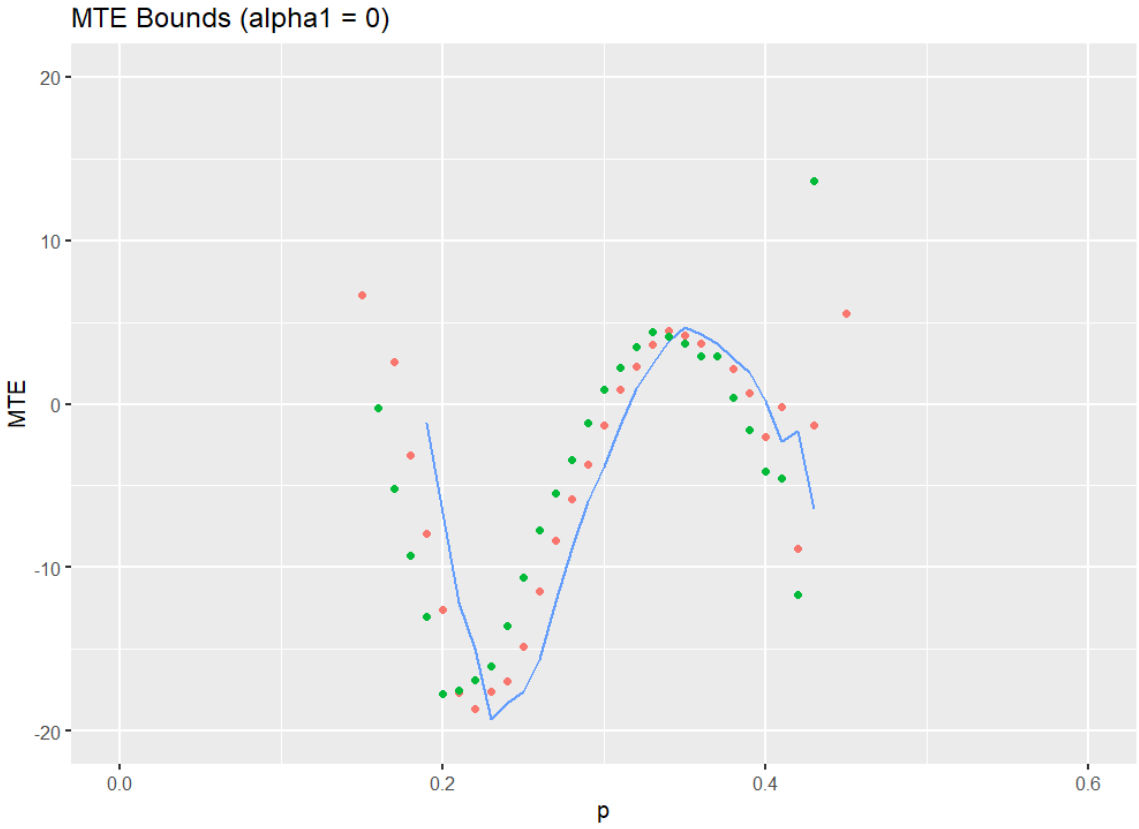


* The solid line represents the naive MTE estimates with $\alpha_0 = \alpha_1 = 0$, and the dots are the estimates with different grid points of (α_0, α_1), which are illustrated with different colors.

FIGURE 11. Estimated MTE Region (Logit)

The following Figures 13 and 14 show the same results using the residuals.[9]

[9]The same identification region for (α_0, α_1) is obtained using the residuals.



* The solid line represents the naive MTE estimates with $\alpha_0 = \alpha_1 = 0$, and the dots are the estimates with different grid points of (α_0, α_1), which are illustrated with different colors.

FIGURE 12. Estimated MTE Region (Logit, $\alpha_1 = 0$)

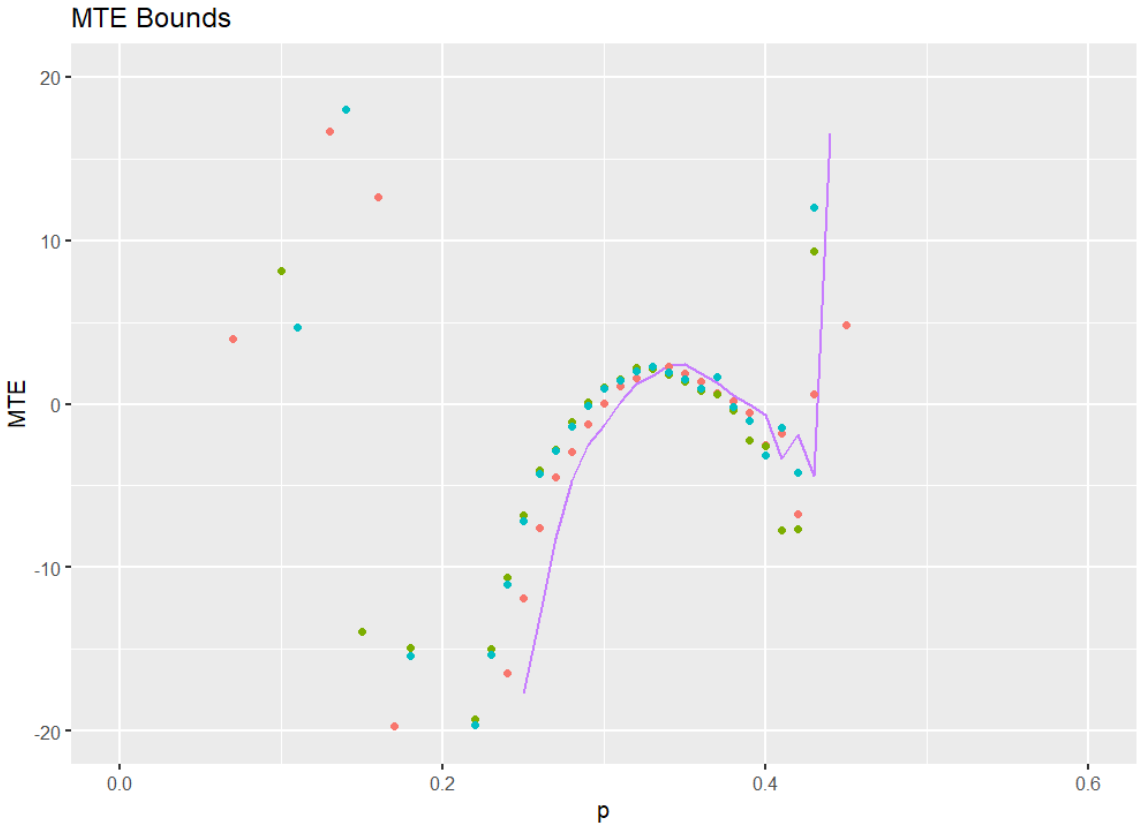
F.1.3. *Case 3:* $\alpha_0 < \alpha_1$. As $P_1(z)$ and $P_2(z)$ can be either positive or negative, now α_0 and α_1 are partially identified as

$$(\alpha_0, \alpha_1) \in \left\{ \left[0, \frac{1}{2}\right)^2 : \alpha_0 < \alpha_1, \Delta(z) \geqslant 0, \text{ and either } P_1(z) \text{ or } P_2(z) \in [0,1] \text{ for all } z \right\},$$

and $P(z)$ can be either $P_1(z)$ or $P_2(z)$ given (α_0, α_1, z).[10]

Likewise, we get $P(z)$ and estimate MTE with the `lprobust` package. The results are shown in Figures 15 and 16, where the latter shows the result with $\alpha_0 = 0$.

[10]If there were some non-empty range of z such that $0 \leqslant P_1(z) \leqslant 1$ and $0 \leqslant P_2(z) \leqslant 1$, we would obtain two different sets of MTE estimates using either $P_1(z)$ or $P_2(z)$ for those z range, and take the union of the two sets of MTE estimates. However, in our illustration, we have only one of $P_1(z)$ and $P_2(z)$ that is in $[0,1]$.

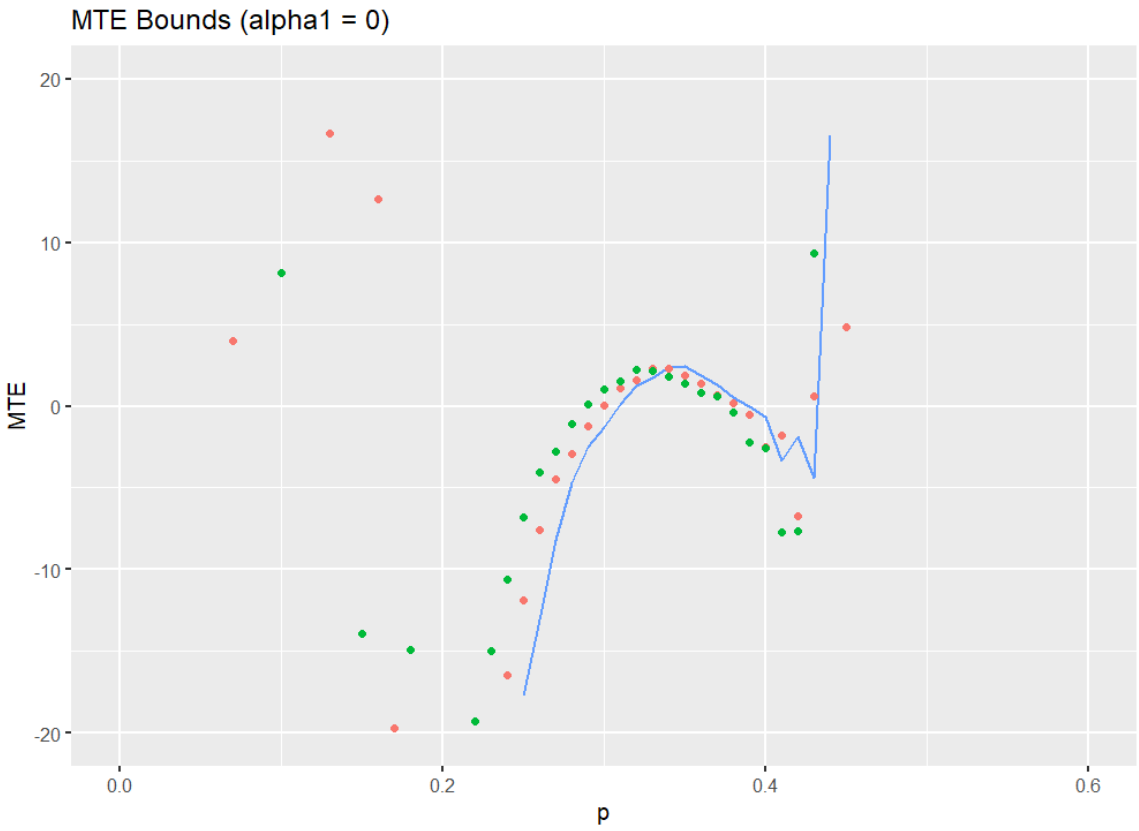


* The solid line represents the naive MTE estimates with $\alpha_0 = \alpha_1 = 0$, and the dots are the estimates with different grid points of (α_0, α_1), which are illustrated with different colors.

FIGURE 13. Estimated MTE Region (Logit, Residuals)

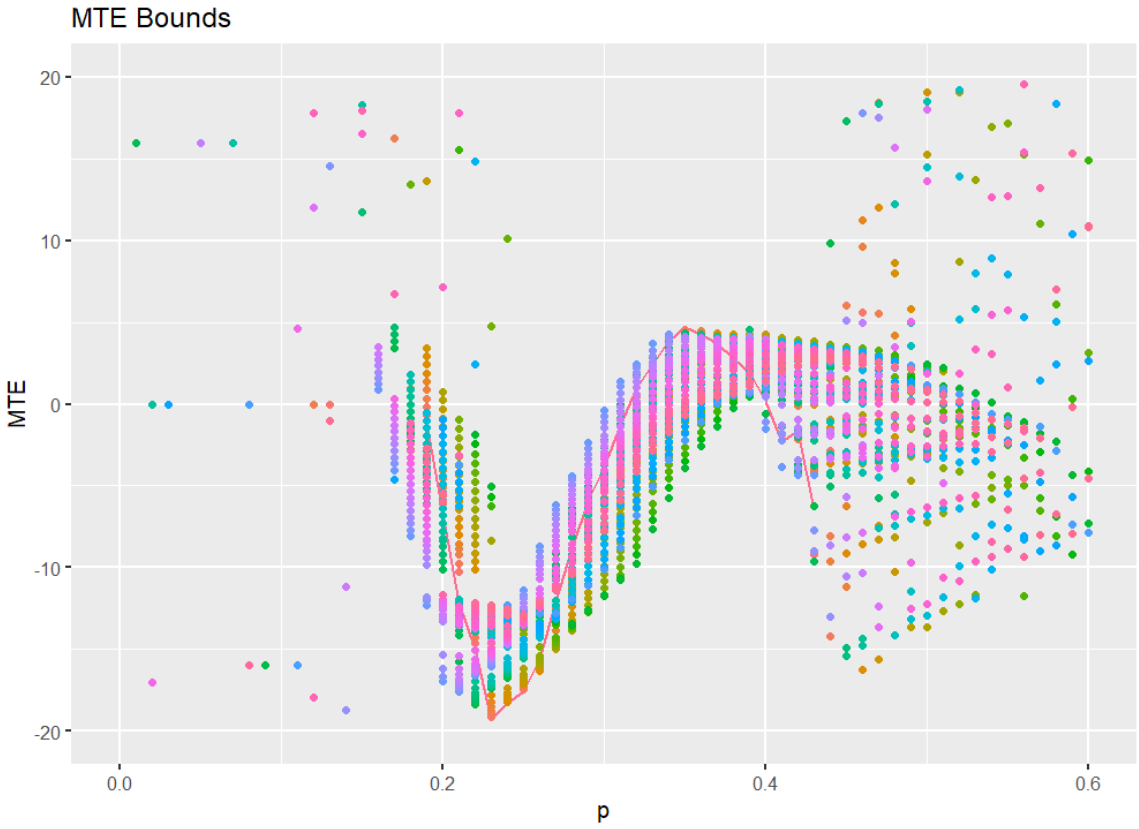
The following Figures 17 and 18 show the same results using the residuals.[11]

[11]The same identification region for (α_0, α_1) is obtained using the residuals.



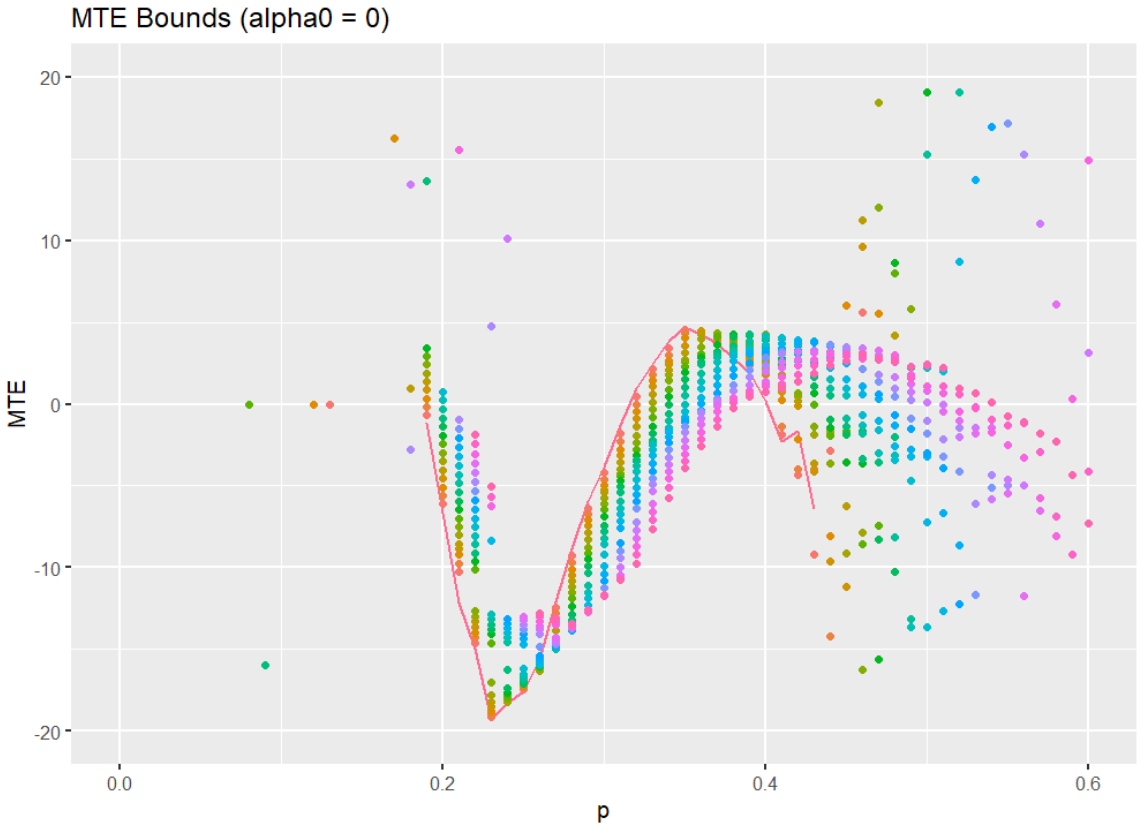
* The solid line represents the naive MTE estimates with $\alpha_0 = \alpha_1 = 0$, and the dots are the estimates with different grid points of (α_0, α_1), which are illustrated with different colors.

FIGURE 14. Estimated MTE Region (Logit, Residuals, $\alpha_1 = 0$)



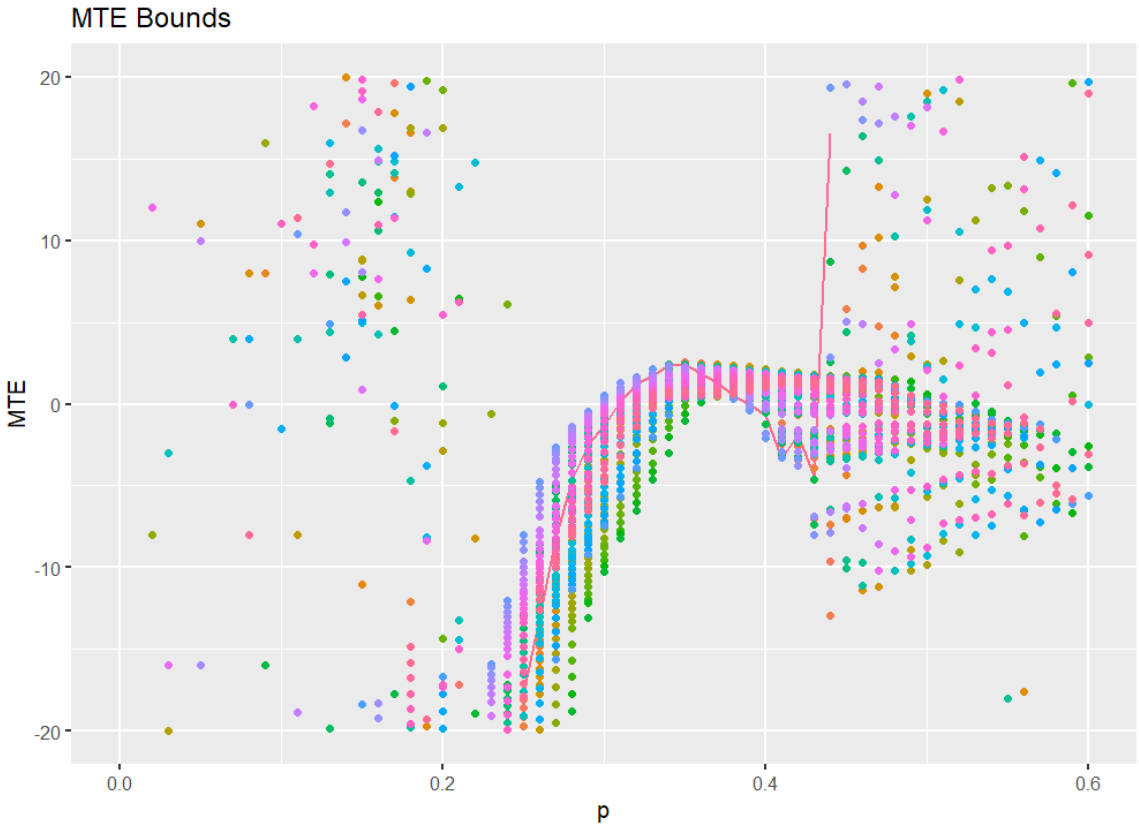
* The solid line represents the naive MTE estimates with $\alpha_0 = \alpha_1 = 0$, and the dots are the estimates with different grid points of (α_0, α_1), which are illustrated with different colors.

FIGURE 15. Estimated MTE Region (Logit)



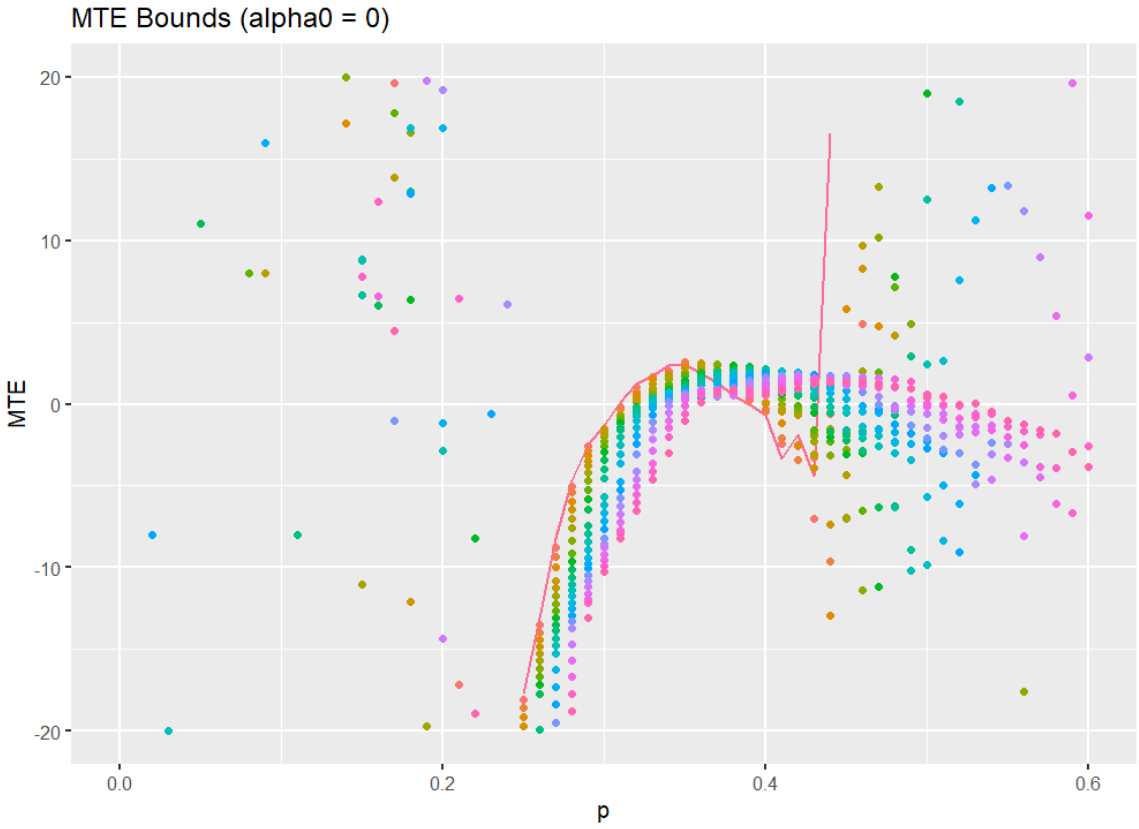
* The solid line represents the naive MTE estimates with $\alpha_0 = \alpha_1 = 0$, and the dots are the estimates with different grid points of (α_0, α_1), which are illustrated with different colors.

FIGURE 16. Estimated MTE Region (Logit, $\alpha_1 = 0$)



* The solid line represents the naive MTE estimates with $\alpha_0 = \alpha_1 = 0$, and the dots are the estimates with different grid points of (α_0, α_1), which are illustrated with different colors.

FIGURE 17. Estimated MTE Region (Logit, Residuals)



* The solid line represents the naive MTE estimates with $\alpha_0 = \alpha_1 = 0$, and the dots are the estimates with different grid points of (α_0, α_1), which are illustrated with different colors.

FIGURE 18. Estimated MTE Region (Logit, Residuals, $\alpha_1 = 0$)

Appendix G. Details on the numerical illustrations

G.1. **Numerical illustration in Section 3.3.** In this example, we have:

$$\begin{aligned}
\mathbb{E}[Y|P(Z)=P(z)] &= \mathbb{E}[\beta D^*|P(Z)=P(z)]+\mathbb{E}[U|P(Z)=P(z)],\\
&= \mathbb{E}[\beta \mathbb{1}\{V\leqslant P(z)\}]+\mathbb{E}[U],\\
&= \mathbb{E}[\beta|V\leqslant P(z)]\mathbb{P}(V\leqslant P(z))+\mathbb{E}[U],\\
&= \int_0^{P(z)}\mathbb{E}[\beta|V=v]dv+\mathbb{E}[U],\\
&= \int_0^{P(z)}\mathbb{E}[\beta|V^*=\Phi^{-1}(v)]dv+\mathbb{E}[U].
\end{aligned}$$

Hence, $MTE(P(z))=\mathbb{E}\left[\beta|V^*=\Phi^{-1}(P(z))\right]=0.5\Phi^{-1}(P(z))+2$.

Moreover, note that

$$\mathbb{P}(\varepsilon=1)=\mathbb{P}(\xi\leqslant\alpha)=\alpha,$$

and thus it can be shown that

$$\begin{aligned}
F_{V|\varepsilon=1}(p) &= \mathbb{P}(V\leqslant p|\varepsilon=1),\\
&= \frac{\mathbb{P}(V\leqslant p,\xi\leqslant\alpha)}{\mathbb{P}(\xi\leqslant\alpha)},\\
&= \alpha^{-1}\Phi_2\big[\Phi^{-1}(p),\Phi^{-1}(\alpha);\rho\big],\\
&\equiv \alpha^{-1}C(p,\alpha;\rho),
\end{aligned}$$

where $\Phi_2[\cdot,\rho]$ is the bivariate normal distribution function with a correlation coefficient of ρ, and $C(p,\alpha;\rho)$ is the bivariate normal copula.

Similarly, we have

$$\begin{aligned}
F_{V|\varepsilon=0}(p) &= \mathbb{P}(V\leqslant p|\varepsilon=0),\\
&= \frac{\mathbb{P}(V\leqslant p,\xi>\alpha)}{\mathbb{P}(\xi>\alpha)},\\
&= (1-\alpha)^{-1}\Big\{\Phi_2\big[\Phi^{-1}(p),\Phi^{-1}(1);\rho\big]-\Phi_2\big[\Phi^{-1}(p),\Phi^{-1}(\alpha);\rho\big]\Big\},\\
&= (1-\alpha)^{-1}\Big\{\Phi\big[\Phi^{-1}(p)\big]-\Phi_2\big[\Phi^{-1}(p),\Phi^{-1}(\alpha);\rho\big]\Big\},\\
&\equiv (1-\alpha)^{-1}\big[p-C(p,\alpha;\rho)\big],
\end{aligned}$$

and note that these two conditional distributions satisfy (2.2).

Thus, using (3.4), we have

$$\begin{aligned}\mathbb{E}[D|Z=z] &= (1-\alpha)F_{V|\varepsilon=0}\big(P(z)\big)+\alpha[1-F_{V|\varepsilon=1}\big(P(z)\big)],\\ &= P(z)+\alpha-2C(P(z),\alpha;\rho),\end{aligned}$$

and

$$\begin{aligned}\Delta_{YZ}(z',z) &= \mathbb{E}[Y|Z=z]-\mathbb{E}[Y|Z=z'],\\ &= \int_{P(z')}^{P(z)}\mathbb{E}[\beta|V^*=\Phi^{-1}(v)]dv,\\ \Delta_{DZ}(z',z) &= \mathbb{E}[D|Z=z]-\mathbb{E}[D|Z=z'],\\ &= P(z)-P(z')-2\Big\{C\big(P(z),\alpha;\rho\big)-C\big(P(z'),\alpha;\rho\big)\Big\},\end{aligned}$$

Hence, we have the upper bound of MTE[12] as follows:[13]

$$\lim_{z'\uparrow z}\frac{\Delta_{YZ}(z',z)}{|\Delta_{DZ}(z',z)|}=\lim_{p'\uparrow p}\frac{\int_{p'}^{p}\mathbb{E}[\beta|V^*=\Phi^{-1}(v)]dv}{|p-p'-2[C(p,\alpha;\rho)-C(p',\alpha;\rho)]|},$$

or using L'Hôpital rule and properties of the copula C (Meyer, 2013),

$$\lim_{z'\uparrow z}\frac{\Delta_{YZ}(z',z)}{|\Delta_{DZ}(z',z)|}=\lim_{p'\uparrow p}\frac{\Delta_{DZ}(z',z)}{|\Delta_{DZ}(z',z)|}\cdot\frac{-\frac{1}{2}\Phi^{-1}(p')-2}{\Big|-1+2\Phi\big[\frac{\Phi^{-1}(\alpha)-\rho\Phi^{-1}(p')}{\sqrt{1-\rho^2}}\big]\Big|},$$

whenever the limit exists. Note that the limit does not exist if $p=\alpha=\frac{1}{2}$ or if $\Phi^{-1}(\alpha)=\rho\Phi^{-1}(p)$.

G.2. **Details on the numerical illustration of the special case.** We assume in this illustration that the researcher knows that the rate of misclassification α is less than $1/2$. Consider the same example from the previous section (3.14) where $\rho=0$ (i.e., ε is independent of V). Note that

$$\begin{aligned}\mathbb{P}(D=1|Z=z) &= P(z)+\alpha-2P(z)\alpha\\ &= P(z)(1-2\alpha)+\alpha\end{aligned}$$

[12]Note that this upper bound does not exploit the information from $TV_{(Y,D=d)}(z',z)$.

[13]This implicitly assumes that $P(z)$ is known to be monotone because it states that as z' increases p' also increases (same direction).

because we have $\lim_{\rho\to 0^+} C(P(z),\alpha;\rho) = P(z)\alpha$ (Meyer, 2013). Hence, the following can be verified for the identification region for α:

$$\begin{aligned}\inf_z \mathbb{P}(D=1|Z=z) = \inf_z \mathbb{P}(D=0|Z=z) = \alpha,\\ \sup_z \mathbb{P}(D=1|Z=z) = \sup_z \mathbb{P}(D=0|Z=z) = 1-\alpha.\end{aligned}$$

Moreover, because we have

$$\begin{aligned}\mathbb{E}[Y|\mathbb{P}(D=1|Z)=p] &= \mathbb{E}[\beta D^* + U | P(Z)(1-2\alpha)+\alpha = p]\\ &= \mathbb{E}\left[\beta D^* + U \Big| Z = \frac{1}{2}\Phi^{-1}\Big(\frac{1}{1-2\alpha}(p-\alpha)\Big)\right]\\ &= \int_0^{P(\frac{1}{2}\Phi^{-1}(\frac{1}{1-2\alpha}(p-\alpha)))} \mathbb{E}[\beta|V^* = \Phi^{-1}(v)]dv + \mathbb{E}[U]\end{aligned}$$

and

$$\begin{aligned}LIV(p) &= \frac{\partial \mathbb{E}[Y|\mathbb{P}(D=1|Z)=p]}{\partial p}\\ &= \frac{1}{1-2\alpha}\left(\frac{1}{2}\Phi^{-1}\Big(\frac{1}{1-2\alpha}(p-\alpha)\Big)+2\right)\end{aligned}$$

for $\alpha \neq \frac{1}{2}$, we verify the true MTE lies within the identification region as follows:

$$\begin{aligned}MTE(p;\tilde{\alpha}) &= (1-2\tilde{\alpha})LIV\big((1-2\tilde{\alpha})p+\tilde{\alpha}\big)\\ &= \frac{1-2\tilde{\alpha}}{1-2\alpha}\left(\frac{1}{2}\Phi^{-1}\Big(\frac{1}{1-2\alpha}\big((1-2\tilde{\alpha})p+\tilde{\alpha}-\alpha\big)\Big)+2\right)\\ &= \frac{1}{2}\Phi^{-1}(p)+2 \quad \text{if} \quad \tilde{\alpha}=\alpha\end{aligned}$$

References

Acerenza, S. 2021. "Partial Identification of Marginal Treatment Effects with Discrete Instruments and Misreported Treatment." *Working Paper* .

Aigner, D. J. 1973. "Regression with a Binary Independent Variable Subject to Errors of Observation." *Journal of Econometrics* 1:49–60.

Battistin, E., M. De Nadai, and B. Sianesi. 2014. "Misreported schooling, multiple measures and returns to educational qualifications." *Journal of Econometrics* 181:136–150.

Battistin, E. and B. Sianesi. 2011. "Misclassified Treatment. Status and Treatment Effects: An application to Returns to Education in the United Kingdom." *The Review of Economics and Statistics* 93 (2):495–509.

Black, D., S. Sanders, and L. Taylor. 2003. "Measurement of Higher Education in the Census and CPS." *Journal of the American Statistical Association* 98:463:545–554.

Bollinger, C. R. 1996. "Bounding mean regressions when a binary regressor is mismeasured." *Journal of Econometrics* 73 (2):387–399.

Calonico, Sebastian, Matias D. Cattaneo, and Max H. Farrell. 2019. "nprobust: Nonparametric Kernel-Based Estimation and Robust Bias-Corrected Inference." *arXiv e-prints* :arXiv:1906.00198.

Calvi, C., A. Lewbel, and D. Tommasi. 2018. "Women's Empowerment and Family Health: Estimating LATE with Mismeasured Treatment." *Working Paper* .

Carneiro, P., M. Lokshin, and N. Umapathi. 2017. "Average and Marginal Returns to Upper Secondary Schooling in Indonesia." *Journal of Applied Econometrics* 32:16–36.

Carneiro, Pedro, James J Heckman, and Edward Vytlacil. 2010. "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin." *Econometrica* 78 (1):377–394.

———. 2011. "Estimating Marginal Returns to Education." *American Economic Review* 101 (6):2754–2781.

Carneiro, Pedro and Sokbae Lee. 2009. "Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality." *Journal of Econometrics* 149:191–208.

Chalak, K. 2017. "Instrumental Variables Methods with Heterogeneity and Mismeasured Instruments." *Econometric Theory* 33:69—-104.

DiTraglia, Francis J. and Camilo Garcia-Jimeno. 2019. "Identifying the effect of a misclassified, binary, endogenous regressor." *Journal of Econometrics* 209:376–390.

Fan, Jianqing and Irene Gijbels. 1996. *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, vol. 66. CRC Press.

Haider, S. J. and M. Stephens Jr. 2020. "Correcting for Misclassified Binary Regressors Using Instrumental Variables." *NBER Working Paper No. 27797* .

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

Heckman, James J and Edward Vytlacil. 1999. "Local Instrumental Variable and Latent Variable Models for Identifying and Bounding Treatment Effects." *Proceedings of the National Academy of Sciences* 96:4730–4734.

———. 2001. "Local Instrumental Variables." *in Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya* :1–46.

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

Horowitz, Joel L. and Charles F. Manski. 1995. "Identification and Robustness with Contaminated and Corrupted Data." *Econometrica* 63 (2):pp. 281–302.

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

Hu, Y. and Arthur Lewbel. 2012. "Returns to lying? identifying the effects of misreporting when the truth is unobserved." *Front. Econ. China* 7 (2):163–192.

Hu, Y. and S. M. Schennach. 2008. "Instrumental variable treatment of nonclassical measurement error models." *Econometrica* 76 (1):195–216.

Jiang, Z. and P. Ding. 2020. "Measurement errors in the binary instrumental variable model." *Biometrika* 107 (1):238–245.

Kasahara, H. and K. Shimotsu. 2021. "Identification of Regression Models with a Misclassified and Endogenous Binary Regressor." *Econometric Theory (forthcoming)* .

Kreider, B. 2010. "Regression coefficient identification decay in the presence of infrequent classification errors." *The Review of Economics and Statistics* 92 (4):1017–1023.

Kreider, B. and J. V. Pepper. 2007. "Disability and Employment: Reevaluating the Evidence in Light of Reporting Errors." *Journal of American Statistical Association* 102 (478):432–441.

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 American Statistical Association* 107:958–975.

Lewbel, Arthur. 2007. "Estimation of Average Treatment Effects with Misclassification." *Econometrica* 75 (2):537–551.

Mahajan, Aprajit. 2006. "Identification and Estimation of Regression Models with Misclassification." *Econometrica* 74 (3):631–665.

Meyer, Christian. 2013. "The bivariate normal copula." *Communications in Statistics-Theory and Methods* 42 (13):2402–2422.

Millimet, D. 2011. "The elephant in the corner: a cautionary tale about measurement error in treatment effects models." *In Missing Data Methods: Cross-Sectional Methods and Applications. In: Advances in Econometrics Advances in Econometrics, Emerald Group Publishing Limited* 27:1–39.

Mogstad, Magne, Andres Santos, and Alexander Torgovitsky. 2018. "Using Instrumental Variables for Inference about Policy Relevant Treatment Effects." *Econometrica* 86 (5):pp. 1589–1619.

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

Mourifié, I., M. Henry, and R. Méango. 2020. "Sharp Bounds and Testability of a Roy Model of STEM Major Choices." *Journal of Political Economy* 8 (128):3220–3283.

Nguimkeu, P., A. Denteh, and R. Tchernis. 2019. "On the estimation of treatment effects with endogenous misreporting." *Journal of Econometrics* 208:487–506.

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." *Discussion Working Paper series, IZA DP No. 13430* .

Ura, Takuya. 2018. "Heterogeneous Treatment Effects with Mismeasured Endogenous Treatment." *Quantitative Economics* 9 (3):1335–1370.

———. 2020. "Instrumental variable quantile regression with misclassification." *Econometric Theory (forthcoming)* .

Yanagi, T. 2019. "Inference on local average treatment effects for misclassified treatment." *Econometric Reviews* 38:938–960.