Threshold Crossing Models and Bounds on Treatment Effects: A Nonparametric Analysis *

Azeem M. Shaikh
Department of Economics
Stanford University

Edward Vytlacil
Department of Economics
Stanford University

February 21, 2005

Abstract

This paper considers the evaluation of the average treatment effect of a binary endogenous regressor on a binary outcome when one imposes a threshold crossing model on both the endogenous regressor and the outcome variable but without imposing parametric functional form or distributional assumptions. Without parametric restrictions, the average effect of the binary endogenous variable is not generally point identified. This paper constructs sharp bounds on the average effect of the endogenous variable that exploit the structure of the threshold crossing models and any exclusion restrictions. We also develop methods for inference on the resulting bounds.

JEL Numbers: C14, C25, C35.
KEYWORDS: Binary Response, Probit, Endogeneity, instrumental variables, sample selection models, social program evaluation

*Stanford University, Department of Economics. We would like to thank Lars Hansen, Peter Hansen, Aprajit Mahajan, Chuck Manski, John Pepper, Elie Tamer and Jim Powell for helpful comments. We would also like to thank seminar participants at the University of California at Berkeley, University of Chicago, Columbia, Harvard/MIT, Michigan Ann-Arbor, Northwestern, Stanford Statistics Department, and at the ZEW 2nd Conference on Evaluation Research in Mannheim. Edward Vytlacil would especially like to acknowledge his gratitude to James Heckman for his continued support. This research was conducted in part while Edward Vytlacil was a W. Glenn Campbell and Rita Ricardo-Campbell Hoover National Fellow. Correspondence: Edward Vytlacil, Landau Economics Building, 579 Serra Mall, Stanford CA 94305; Email: vytlacil@stanford.edu; Phone: 650-725-7836; Fax: 650-725-5702.

1 Introduction

This paper considers the evaluation of the average treatment effect of a binary endogenous regressor on a binary outcome when one imposes threshold crossing models on both the endogenous variable and the outcome variable but without imposing parametric functional form or distributional assumptions. Without parametric restrictions, the average effect of the binary endogenous variable is not generally point identified even in the presence of exclusion restrictions. This paper constructs sharp bounds on the average effect of the endogenous variable that exploit the structure of the threshold crossing models and any exclusion restrictions.

As an example, suppose the researcher wishes to evaluate the average effect of job training on later employment. The researcher will often impose a threshold crossing model for the employment outcome and include a dummy variable for receipt of training as a regressor in the model. The researcher may believe that individuals self-select into job training in such a way that job training is endogenous within the outcome equation: It might be the case, for example, that those individuals with the worst job prospects are the ones who self-select into training. The researcher might model job training as also being determined by a threshold crossing model, resulting in a triangular system of equations for the joint determination of job training and later employment, as in Heckman (1978).

If the researcher is willing to impose parametric assumptions, then the researcher can estimate the model described above by maximum likelihood. In the classic case of Heckman (1978), linear index and joint normality assumptions are imposed so the resulting model is in the form of a bivariate probit.[1] However, suppose that the researcher does not wish to impose such strong parametric functional form or distributional assumptions. What options are available under these circumstances? If the researcher has access to an instrument, a standard approach to estimate the effect of an endogenous regressor is to use a two-stage least squares (TSLS) estimator.[2,3] But in the example described above, a classic TSLS approach is invalid due to the fact that the error term is not additively separable from the regressors in the outcome equation. Likewise, semiparametric "control function" approaches, such as that developed by Blundell and Powell (2004), are relevant if the endogenous variable is continuous but do not extend to the current context of a binary endogenous regressor. The recent analysis of Altonji and Matzkin (2005) is not applicable unless one believes that the error term is independent of the regressors conditional on some external variables. If the researcher has access to an instrument with "large support", then the researcher can follow Heckman (1990) in using identification-at-infinity arguments. The support condition required for this approach to work, however, is very strong, and the researchers might not have

[1] A number of other estimators are also available if the endogenous variable is continuous instead of being discrete, see Amemiya (1978), Lee (1981), Newey (1986), and Rivers and Vuong (1988). See also Blundell and Smith (1986, 1989) for closely related analysis when the outcome equation is given by a tobit model.

[2] See Amemiya (1974) for a classic treatment of endogenous variables in regression equations that are non-linear in both variables and parameters. See Blundell and Powell (2003) for a recent survey of instrumental variable techniques in semiparametric regression models. See Angrist (1991) and Bhattacharya, McCaffrey, and Goldman (2005) for related monte carlo evidence on the properties of TSLS in this context.

[3] As argued by Angrist (2001), standard linear TSLS will still identify the "local average treatment effect" (LATE) parameter of Imbens and Angrist (1994) if there are no other covariates in the regression, if the other covariates are all discrete and the model is fully saturated, or if the LATE Instrumental Variable assumptions hold even without conditioning on covariates. See Heckman and Vytlacil (2001) for the relationship between the LATE parameter and other mean treatment parameters including the average treatment effect. Also, see Vytlacil (2002) for the equivalence between the assumptions imposed in LATE analysis and imposing a threshold crossing model on the endogenous variable.

access to an instrument with large support.[4,5] Another option is to follow Vytlacil and Yildiz (2004), but their approach requires a continuous co-variate that enters the outcome equation but which does not enter the model for the endogenous variable.

The researcher can also construct bounds on the average effect of a binary endogenous variable using, e.g., Manski (1990, 1994), Balke and Pearl (1997), Manski and Pepper (2000), and Heckman and Vytlacil (2001).[6] Yet, these methods may only exploit a subset of the assumptions the researcher is willing to make, and, as a result, the bounds may not be as informative as possible. In particular, these methods would not allow the researcher to exploit the fact that the outcome variable might be determined by a threshold crossing model, as in the scenario described above.

This paper constructs sharp bounds on the average effect of the binary endogenous variable inside of a threshold crossing model under the assumptions that the binary endogenous variable itself is given by a threshold crossing model. We assume that there exists at least one variable that directly enters the first stage equation determining the endogenous treatment variable but does not directly enter the second stage outcome equation. The analysis will also exploit any variables that enter the outcome equation but not the treatment equation if such variables are present, but the bounds will hold even in the absence of such variables. The analysis will exploit the assumption of a threshold crossing model on the outcome equation and a threshold crossing model on the treatment equation, but does not impose any parametric functional form or distributional assumptions such as linear indices or normality assumptions.

This paper proceeds as follows. In Section 2, we introduce our model and assumptions, and define some notation that will be used in the later sections. We define and discuss our average treatment parameters of interest in Section 3. We then proceed with our bounding analysis in Sections 4 and 5, first considering the bounds without covariates in the outcome equation and then showing how covariates in the outcome equation can be exploited to narrow the width of the bounds. In Section 6, we compare and contrast our bounds with the bounds of Manski (1990), Manski and Pepper (2000), and Heckman and Vytlacil (2001). We develop inference for the bounds in Section 7. Section 8 concludes.

2 Model, Assumptions, and Notation

Assume that for each individual there are two potential outcomes, (Y_0, Y_1), corresponding respectively to the potential outcomes in the untreated and treated states. Let $D = 1$ denote the receipt

[4]Heckman (1990) assumed that the outcome equation is additively separable in the regressors and the error term, but his analysis extends immediately to the case without additive separability. See also Cameron and Heckman (1998), Aakvik, Heckman, and Vytlacil (1998), and Chen, Heckman, and Vytlacil (1999) for identification-at-infinity arguments in the context of a system of discrete choice equations. See also Lewbel (2005) for identification and estimation using large support assumptions on a "special regressor."

[5]While this paper examines the average effect of a dummy endogenous variable on the outcome of interest without imposing any linear latent index structure on the model, a separate problem arises in the literature which imposes a linear latent index for the binary choice model and then seeks to identify and estimate the slope coefficients on the linear index. Identification of the slope coefficients of the linear index does not imply identification of the average effect of covariates on the outcome of interest. Recent contributions to that literature with endogenous regressors include Hong and Tamer (2003), Lewbel (2000), and Magnac and Maurin (2005).

[6]Chesher (2003) provides a related bounding analysis, but his bounds are not applicable to the current model since his results do not extend to the case in which the endogenous regressor takes only two values.

of treatment and $D = 0$ denote nonreceipt. Let Y be the measured outcome variable so that

$$Y = DY_1 + (1 - D)Y_0.$$

For example, in labor economics D might be an indicator variable for receipt of job training, Y_1 an indicator variable for whether the individual would have been employed had she received training, Y_0 an indicator variable for whether the individual would have been employed had she not received training, and Y an indicator variable for observed employment status. In health economics, on the other hand, D might be an indicator variable for receiving a particular medical intervention, Y_1 an indicator variable for survival given the medical intervention, Y_0 an indicator variable for survival without the medical intervention, and Y an indicator variable for survival. We impose the following latent index model on Y_1, Y_0 and D:

$$\begin{array}{rcl} Y_1 & = & \mathbf{1}[Y_1^* \geq 0] \\ Y_0 & = & \mathbf{1}[Y_0^* \geq 0] \\ D & = & \mathbf{1}[D^* \geq 0], \end{array} \tag{1}$$

with

$$\begin{array}{rcl} Y_1^* & = & \nu_1(X) - \epsilon_1 \\ Y_0^* & = & \nu_0(X) - \epsilon_0 \\ D^* & = & \vartheta(Z) - U, \end{array} \tag{2}$$

where $(X, Z) \in \Re^{K_X} \times \Re^{K_Z}$ is a random vector of observed covariates, $\epsilon_1, \epsilon_0, U$ are unobserved random variables, and $\mathbf{1}[\cdot]$ is the logical indicator function taking the value 1 if its argument is true and the value 0 otherwise. The model for Y can be rewritten as

$$\begin{array}{rcl} Y & = & \mathbf{1}[Y^* \geq 0] \\ Y^* & = & \nu_0(X) + D(\nu_1(X) - \nu_0(X) - \epsilon_1 + \epsilon_0) - \epsilon_0 \end{array}$$

D^*, Y_1^*, and Y_0^* are latent indices. The model for Y and D are threshold-crossing models. Here, $\vartheta(Z) + U$ is interpreted as net utility to the agent from choosing $D = 1$. In the labor supply example, Y_1^* and Y_0^* might be offered wage minus reservation wage with and without job training, respectively. In the health example, Y_1^* and Y_0^* might be latent measures of health with and without the medical intervention, respectively. We are considering threshold crossing models with additive separability in the latent index between observables and unobservables. These models are more general than they may at first appear: It is shown in Vytlacil (2004) that a wide class of threshold crossing models without the additive structure on the latent index will have a representation with the additive structure on the latent index.[7]

We will maintain the following assumptions:

(A-1) The distribution of U is absolutely continuous with respect to Lebesgue measure;

(A-2) $(U, \epsilon_1, \epsilon_0)$ is independent of (Z, X);

(A-3) $\epsilon_j \mid U \sim \epsilon \mid U$, for $j = 0, 1$;

(A-4) $\vartheta(Z)$ is nondegenerate conditional on X;

[7]See also Vytlacil (2002) for an equivalence result between the threshold-crossing model on D with independence between Z and $(U, \epsilon_1, \epsilon_0)$ and the independence and monotonicity assumptions of Imbens and Angrist (1994).

(A-5) The distribution of ϵ conditional on U has a strictly positive density with respect to Lebesgue measure on $\Re$; and

(A-6) The support of the distribution of (X, Z) is compact, and $\vartheta(\cdot)$, $\nu_1(\cdot)$, and $\nu_0(\cdot)$ are continuous.

Assumption (A-1) is a regularity condition imposed to guarantee smoothness of the relevant conditional expectation functions. Assumption (A-2) is a critical independence condition, that the observed covariates (with the exception of the binary endogenous variable of interest) are independent of the unobserved covariates. Assumption (A-3) is the assumption that ϵ_1 and ϵ_0 have the same distribution conditional on U. This assumption that ϵ_1 and ϵ_0 have the same distribution conditional on U will be critical to the following analysis. This assumption makes the analysis more restrictive than the Roy-model/switching regression framework considered in Heckman (1990). The assumption would be implied by a model where $\epsilon_1 = \epsilon_0$ in which case the effect of D on the latent index for Y^* is the same for all individuals with given X covariates. The assumption will also be satisfied if $\epsilon_1 \neq \epsilon_0$ with restrictions on what information is available to the agent when deciding whether to receive treatment. Assumption (A-4) requires an exclusion restriction – there is at least one variable in Z that is not a component of X. Assumption (A-5) is a standard regularity condition that aids in the exposition of the analysis. It is implied by most standard parametric assumptions on (ϵ, U), for example, by $(\epsilon, U) \sim BVN$ as long as $\text{Corr}(\epsilon, U) \neq 1$. The assumption can be removed at the cost of somewhat weaker results.[8]. Assumption (A-6) also eases the exposition by ensuring that certain supremums and infimums are obtained. This assumption can be easily relaxed for the identification analysis.

As a normalization, we will set $U \sim \text{Unif}[0, 1]$ and $\vartheta(Z) = P(Z)$, where $P(Z) = \Pr(D = 1|Z)$. This normalization is innocuous given assumptions (A-1) and (A-2). Given the model of equations (1)-(2) and assumptions (A-2) and (A-3) we also have the following index sufficiency restriction:

$$\begin{array}{rcl} E(DY|X, Z) & = & E(DY|X, P(Z)), \\ E((1-D)Y|X, Z) & = & E((1-D)Y|X, P(Z)). \end{array} \tag{3}$$

It will often be more convenient as a result to condition on $P(Z)$ instead of conditioning on Z directly. We will sometimes suppress the Z argument and write P as a shorthand for the variable $P(Z)$.

We do not impose any parametric structure on $\nu_1(\cdot)$, $\nu_0(\cdot)$, or $\vartheta(\cdot)$, and we do not impose a parametric distribution on ϵ_1, ϵ_0 or U. Many classic latent index models that impose specific parametric distributional and functional form assumptions are nested within the assumptions considered here, even though we do not impose any such parametric structure. For example, the classical bivariate probit with structural shift described in Heckman (1978) can be written in the form (1) and (2) by taking

$$D^* = Z\gamma - U \geq 0,$$
$$Y_1^* = X\beta - \epsilon,$$
$$Y_0^* = X\beta + \alpha - \epsilon,$$

so that

$$D = \mathbf{1}[Z\gamma - U \geq 0]$$

[8]See the discussion in footnotes 16, 18, 20 and 24.

$$Y = 1[X\beta + \alpha D - \epsilon \geq 0],$$

and (ϵ, U) to be distributed bivariate normal. Notice that the classical bivariate probit model has much more structure than is imposed in this paper, including: the linear structure on $\vartheta(\cdot)$, $\nu_1(\cdot)$, and $\nu_0(\cdot)$; $\epsilon_1 = \epsilon_0 = \epsilon$ and the parametric distributional assumption on (ϵ, U); $\nu_1(\cdot) = \nu_0(\cdot) + \alpha$ so that the effect of treatment on the latent index does not depend on X. In comparison, our analysis does not impose any parametric functional form assumption on $\vartheta(\cdot)$, $\nu_1(\cdot)$, $\nu_0(\cdot)$, allows $\epsilon_1 \neq \epsilon_0$ as long as $\epsilon_1|U \sim \epsilon_0|U$, and allows the effect of D on the latent index to depend on X.

We conclude this section by defining some additional notation. For any random variables A and B, let $F_A(\cdot)$ denote the cumulative distribution function of A and let $F_{A|B}(\cdot|b)$ denote the cumulative distribution function of A conditional on $B \leq b$. For any random vector A, let Ω_A denote the support of the distribution of A and let Ω_A^j denote the support of the distribution of A conditional of $D = j$. Thus, using this notation, we have that $\Omega_{X,P}$ denotes the support of the distribution of (X, P) and $\Omega_{X,P}^j$ denote the support of the distribution of (X, P) conditional on $D = j$.

Let $\text{sgn}[t]$ denote the sign function, defined as follows:

$$\text{sgn}[t] = \begin{cases} 1 & \text{if } t > 0 \\ 0 & \text{if } t = 0 \\ -1 & \text{if } t < 0. \end{cases}$$

Define[9]

$$\begin{array}{rcl} m_0(x, p, \tilde{p}) & = & p^{-1}\big(\Pr[D = 0, Y = 1|X = x, P = \tilde{p}] - \Pr[D = 0, Y = 1|X = x, P = p]\big), \\ m_1(x, p, \tilde{p}) & = & (1-p)^{-1}\big(\Pr[D = 1, Y = 1|X = x, P = \tilde{p}] - \Pr[D = 1, Y = 1|X = x, P = p]\big). \end{array} \tag{4}$$

For $j = 0, 1$, define

$$q_j(p, \tilde{p}) = \left[\frac{\tilde{p}}{p}\right]^{1-j} \left[\frac{1-\tilde{p}}{1-p}\right]^{j}. \tag{5}$$

For scalar evaluation points p_0, p_1 with $p_0 > p_1$, define

$$\begin{array}{rcl} h_0(p_0, p_1, x) & = & p m_0(x, p_0, p_1) \\ & = & \Pr\big[D = 0, Y = 1\big|X = x, P = p_1\big] - \Pr\big[D = 0, Y = 1\big|X = x, P = p_0\big] \\ h_1(p_0, p_1, x) & = & (1-p) m_1(x, p_1, p_0) \\ & = & \Pr\big[D = 1, Y = 1\big|X = x, P = p_0\big] - \Pr\big[D = 1, Y = 1\big|X = x, P = p_1\big], \end{array} \tag{6}$$

$$\begin{aligned} h(p_0, p_1, x) &\equiv h_1(p_0, p_1, x) - h_0(p_0, p_1, x) \\ &= \Pr[Y = 1 \mid X = x, P = p_0] - \Pr[Y = 1 \mid X = x, P = p_1], \end{aligned} \tag{7}$$

and

$$\begin{aligned} &H(x_0, x_1) \\ &\quad = \int_0^1 \int_0^{p_0} [h_1(p_0, p_1, x_1) - h_0(p_0, p_1, x_0)]\, \mathbf{1}[(x_i, p_j) \in \Omega_{X,P},\ i, j = 0, 1] dF_P(p_1) dF_P(p_0). \end{aligned} \tag{8}$$

[9]For ease of exposition, we will often leave implicit that m_0 and m_1 are only well defined for appropriate values of $(x, p, \tilde{p})$, i.e., for $(x, p, \tilde{p})$ such that $(x, p), (x, \tilde{p}) \in \Omega_{X,P}$.

Define

$$\begin{array}{rcl} \mathcal{X}_0^U(x_0) & = & \{x_1 : \ H(x_0, x_1) \geq 0\} \\ \mathcal{X}_0^L(x_0) & = & \{x_1 : \ H(x_0, x_1) \leq 0\} \\ \mathcal{X}_1^L(x_1) & = & \{x_0 : \ H(x_0, x_1) \geq 0\} \\ \mathcal{X}_1^U(x_1) & = & \{x_0 : \ H(x_0, x_1) \leq 0\}. \end{array} \tag{9}$$

Define

$$\begin{array}{rcl} p^l & = & \inf\{p \in \Omega_P\} \\ p^u & = & \sup\{p \in \Omega_P\} \\ p^l(x) & = & \inf\{p : (x, p) \in \Omega_{X,P}\} \\ p^u(x) & = & \sup\{p : (x, p) \in \Omega_{X,P}\}. \end{array} \tag{10}$$

3 Parameters of Interest and the Identification Problem

Let Δ denote the treatment effect on the given individual:

$$\Delta = Y_1 - Y_0 = \mathbf{1}[\epsilon_1 \leq \nu_1(X)] - \mathbf{1}[\epsilon_0 \leq \nu_0(X)].$$

For example, suppose that Y is mortality and D is a medical intervention. In this case, $\Delta = 1$ if the individual would have died without the medical intervention but lives with the intervention (the intervention saves the individual's life); $\Delta = 0$ if the individual would die with or without the intervention, or would survive with or without the intervention (the intervention has no effect); and $\Delta = -1$ if the individual would have lived without the intervention but not with the intervention (the intervention kills the individual). In general, Δ will vary even among individuals with the same observed characteristics. For example, the model allows the possibility that the intervention saves the lives of some individuals while costing the lives of other individuals with the same observed characteristics.

Y_1 is only observed for individuals who received the treatment, and Y_0 is only observed for individuals who did not receive the treatment. Thus $\Delta = Y_1 - Y_0$ is never observed for any individual. We do not attempt to recover Δ for each individual but rather consider two averaged versions of Δ.[10] The first parameter that we consider is the average effect of treatment on person with given observable characteristics. This parameter is known as the average treatment effect (ATE) and is given by

$$\Delta^{ATE}(x) \equiv E(Y_1 - Y_0 | X = x) = \Pr[Y_1 = 1 | X = x] - \Pr[Y_0 = 1 | X = x] = F_\epsilon(\nu_1(x)) - F_\epsilon(\nu_0(x)),$$

where the final equality is exploiting our independence assumption (A-2) and that $\epsilon_1, \epsilon_0 \sim \epsilon$ from assumption (A-3). For example, $\Delta^{ATE}(x)$ might represent the change in probability of survival resulting from the medical intervention among those individuals with specified X characteristics.

The second parameter that we consider is the average effect of treatment on individuals who selected into treatment and have given observable characteristics. This parameter is known as

[10] See Heckman and Vytlacil (2001) for a discussion of treatment parameters and the connections among them within selection models.

treatment on the treated (TT) and is given by:[11]

$$\begin{aligned}\Delta^{TT}(x,p) &\equiv E(\Delta|X=x,P=p,D=1)\\ &= \Pr[Y_1=1|X=x,P=p,D=1]-\Pr[Y_0=1|X=x,P=p,D=1]\\ &= F_{\epsilon|U}(\nu_1(x)\mid p)-F_{\epsilon|U}(\nu_0(x)|p), \quad (11)\end{aligned}$$

where the final equality is exploiting our independence assumption (A-2) and that $\epsilon_j|U \sim \epsilon|U$ for $j=0,1$ from assumption (A-3). For example, $\Delta^{TT}(x,p)$ might represent the change in probability of survival resulting from the medical intervention among those individuals who did receive the medical intervention and have the specified covariates.

Neither ATE nor TT are immediately identified from the population distribution of (Y,D,X,Z). Knowledge of the distribution of (Y,D,X,Z) implies identification of $P(z) \equiv \Pr[D=1 \mid Z=z]$ for z in the support of the distribution of Z, and of the following conditional expectations,

$$\begin{array}{rcccl} E(Y|D=1,X=x,P=p) & = & \Pr[Y_1=1|D=1,X=x,P=p] & = & F_{\epsilon|U}(\nu_1(x)|p) \\ E(Y|D=0,X=x,P=p) & = & \Pr[Y_0=1|D=0,X=x,P=p] & = & F_{\epsilon|-U}(\nu_0(x)|p), \end{array} \quad (12)$$

where the first equation is identified for $(x,p) \in \Omega^1_{X,P}$ and the second equation is identified for $(x,p) \in \Omega^0_{X,P}$.[12]

While knowledge of the population distribution of (Y,D,X,Z) immediately implies identification of equations (12), it does not immediately imply identification of the treatment parameters. First consider the treatment on the treated parameter. Recall from equation (11) that TT is comprised of the sum of two terms. The first term, $\Pr[Y_1=1 \mid D=1,X=x,P=p]$, is the average with treatment outcome among those who did receive the treatment and is immediately identified from the data using equation (12). However, the second term, $\Pr[Y_0=1 \mid D=1,X=x,P=p]$, is the average without treatment outcome among those who did receive the treatment. This term corresponds to a counterfactual value: What would have happened to treated individuals if they had, counter to fact, not received the treatment? This term is not directly identified from the data. The analysis will proceed by bounding this term, which in turn will imply bounds on the treatment on the treated parameter.

[11] Note that we define the average treatment effect conditional on X while we define treatment on the treated conditional on (X,P). From our model and independence assumptions, we have that the treatment effect is mean independent of P conditional on X so that $E(\Delta \mid X,P) = E(\Delta \mid X)$. In contrast, in general the treatment effect is not independent of P conditional on $(X,D=1)$ so that in general $E(\Delta \mid X,P,D=1) \neq E(\Delta \mid X,D=1)$. Also note that while we define the treatment on the treated parameter conditional on P instead of conditional on Z, we have that $E(\Delta|X,P(Z),D=1) = E(\Delta \mid X,Z,D=1)$.

[12] For ease of exposition, we will often leave implicit that we are only identifying and evaluating these conditional expectations over the appropriate support. We will explicitly state the support condition when failure to do so might reasonably lead to confusion or ambiguity.

Likewise, consider the average treatment effect. For any p, we have that[13]

$$\begin{aligned}
\Delta^{ATE}(x) &= E(Y_1 - Y_0 \mid X = x) \\
&= E(Y_1 - Y_0 \mid X = x, P = p) \\
&= \Pr[Y_1 = 1 \mid X = x, P = p] - \Pr[Y_0 = 1 \mid X = x, P = p] \\
&= \left[p \Pr[Y_1 = 1 \mid D = 1, X = x, P = p] + (1-p) \Pr[Y_1 = 1 \mid D = 0, X = x, P = p]\right] \\
&\quad - \left[p \Pr[Y_0 = 1 \mid D = 1, X = x, P = p] + (1-p) \Pr[Y_0 = 1 \mid D = 0, X = x, P = p]\right],
\end{aligned}$$

where the second equality follows from our independence assumption (A-2). Thus, for the average treatment effect, we will again consider bounds on $\Pr[Y_0 = 1 \mid D = 1, X = x, P = p]$ and will also need to bound $\Pr[Y_1 = 1 \mid D = 0, X = x, P = p]$. Bounds on $\Pr[Y_0 = 1 \mid D = 1, X = x, P = p]$ and $\Pr[Y_1 = 1 \mid D = 0, X = x, P = p]$ will imply bounds on the average treatment effect.

We now turn to our bounding analysis. For the bounding analysis we assume that the population distribution of (Y, D, X, Z) is known and consider bounds on the average treatment effect and treatment on the treated. We first consider the analysis with no X covariates in Section 4 and then proceed to consider how X covariates can allow one to shrink the bounds in Section 5.

4 Analysis With No X Covariates

Consider the model with no X covariates. In this case the model has a very simple structure:

$$\begin{aligned}
Y_1 &= \mathbf{1}[\nu_1 - \epsilon_1 \geq 0] \\
Y_0 &= \mathbf{1}[\nu_0 - \epsilon_0 \geq 0] \\
D &= \mathbf{1}[\vartheta(Z) - U \geq 0],
\end{aligned} \tag{13}$$

As discussed in the previous section, our goal is to bound $\Pr[Y_0 = 1|D = 1, P = p] = F_{\epsilon|U}(\nu_1|p)$ and $\Pr[Y_1 = 1|D = 0, P = p] = F_{\epsilon|-U}(\nu_0|p)$ which in turn allow us to bound $\Delta^{ATE} = E(Y_1 - Y_0) = F_\epsilon(\nu_1) - F_\epsilon(\nu_0)$ and $\Delta^{TT}(p) = E(Y_1 - Y_0|D = 1, P = p) = F_{\epsilon|U}(\nu_1|p) - F_{\epsilon|U}(\nu_0|p)$.

Our analysis exploits two central ideas. First, we use a strategy similar to Heckman and Vytlacil (2001), to express $\Pr[Y_0 = 1|D = 1, P = p]$ and $\Pr[Y_1 = 1|D = 0, P = p]$ as a sum of an identified term and an unidentified term. The result underlying this part of the analysis is formally stated in Lemma 4.1. Second, we use an instrumental variables type of expression to identify the sign of $\nu_1 - \nu_0$, which then provides bounds on unidentified terms. The central result needed for this part of the analysis is formally stated in Lemma 4.2.

We now state the first lemma, using the notation m_j and q_j, $j = 0, 1$, introduced above in equations (4) and (5).

Lemma 4.1. *Assume that (D, Y_0, Y_1) are generated according to equation (13). Assume conditions (A-1), (A-2) and (A-4). Then, for $j = 0, 1$ and for any $p, \tilde{p}$ evaluation points,*[14]

$$\Pr[Y_j = 1 \mid D = 1 - j, P = p] = m_j(p, \tilde{p}) + q_j(p, \tilde{p}) \Pr[Y_j = 1 \mid D = 1 - j, P = \tilde{p}].$$

[13] Recall that we are leaving implicit that we are only evaluating the conditional expectations where the conditional expectations are well defined. Thus, the following equalities hold for any p such that $(x, p) \in \Omega^1_{X,P} \bigcap \Omega^0_{X,P}$.

[14] Recall that we are leaving implicit the standard support condition. Thus, this lemma holds for $p, \tilde{p} \in \Omega_P$.

Proof. Consider the case where $\tilde{p} > p$ (the case where $\tilde{p} < p$ is symmetric, and the case $p = \tilde{p}$ is immediate). Consider $\Pr[Y_1 = 1 \mid D = 0, P = p]$ (the analysis for $\Pr[Y_0 = 1 \mid D = 1, P = p]$ is symmetric). We have

$$\begin{aligned}
\Pr[Y_1 = 1|D = 0, P = p] &= \Pr[\epsilon \leq \nu_1 | U > p] \\
&= \tfrac{1}{1-p} \Pr[U > p, \epsilon \leq \nu_1] \\
&= \tfrac{1}{1-p} \big\{ \Pr[p < U \leq \tilde{p}, \epsilon \leq \nu_1] + \Pr[U > \tilde{p}, \epsilon \leq \nu_1] \big\} \\
&= \tfrac{1}{1-p} \big\{ \Pr[U \leq \tilde{p}, \epsilon \leq \nu_1] - \Pr[U \leq p, \epsilon \leq \nu_1] + \Pr[U > \tilde{p}, \epsilon \leq \nu_1] \big\} \\
&= \tfrac{\Pr[D=1,Y=1|P=\tilde{p}] - \Pr[D=1,Y=1|P=p]}{1-p} + \tfrac{1-\tilde{p}}{1-p} \Pr[Y_1 = 1|D = 0, P = \tilde{p}]
\end{aligned}$$

where the first equality is using our model of equation (13) and our independence assumption (A-2); the second equality is using our normalization that $U \sim$Unif$[0, 1]$; and the final equality is again using our model of equation (13), our independence assumption (A-2), and the equivalence of the events $(D = 1, Y_1 = 1)$ and $(D = 1, Y = 1)$. □

Since m_j and q_j, $j = 0, 1$, are identified from the population distribution of (Y, D, Z), and since any probability is bounded by zero and one, we can now use Lemma 4.1 to bound $\Pr[Y_0 = 1 \mid D = 1, P = p]$ and $\Pr[Y_1 = 1 \mid D = 0, P = p]$ by

$$\Pr[Y_0 = 1 \mid D = 1, P = p] \in [m_0(p, \tilde{p}), m_0(p, \tilde{p}) + q_0(p, \tilde{p})]$$

$$\Pr[Y_1 = 1 \mid D = 0, P = p] \in [m_1(p, \tilde{p}), m_1(p, \tilde{p}) + q_1(p, \tilde{p})].$$

Since these bounds hold for any $\tilde{p}$ evaluation point, we have[15]

$$\begin{aligned}
\Pr[Y_0 = 1 \mid D = 1, P = p] &\in \bigcap_{\tilde{p}} [m_0(p, \tilde{p}), m_0(p, \tilde{p}) + q_0(p, \tilde{p})] \\
\Pr[Y_1 = 1 \mid D = 0, P = p] &\in \bigcap_{\tilde{p}} [m_1(p, \tilde{p}), m_1(p, \tilde{p}) + q_1(p, \tilde{p})].
\end{aligned} \tag{14}$$

It is possible to further improve upon these bounds using an IV-like strategy to identify the sign of $\nu_1 - \nu_0$. For any p_0, p_1 with $p_0 > p_1$, consider

$$h(p_0, p_1) = \Pr[Y = 1|P = p_0] - \Pr[Y = 1|P = p_1].$$

$h(p_0, p_1)$ is the numerator of the population analog of the instrumental variables estimator. Let H denote an integrated version of $h(p_0, p_1)$,

$$H = \int_0^1 \int_0^{p_0} h(p_0, p_1) dF_P(p_1) dF_P(p_0).$$

Using this notation, we have the following lemma:

Lemma 4.2. *Assume that (D, Y_0, Y_1) are generated according to equation (13). Assume conditions (A-1)-(A-5).*[16] *Then, for any p_0, p_1, with $p_0 > p_1$,*[17]

$$sgn\big[H\big] = sgn\big[h(p_0, p_1)\big] = sgn\big[\nu_1 - \nu_0\big].$$

[15] The following intersections are implicitly taken over $\tilde{p} \in \Omega_P$.

[16] A weaker version of the lemma holds without assumption (A-5). Without assuming (A-5), we still have that $h(p_0, p_1) > 0 \Rightarrow \nu_1 > \nu_0$ and $h(p_0, p_1) < 0 \Rightarrow \nu_1 < \nu_0$, but are no longer able to infer the sign of $\nu_1 - \nu_0$ if $h(p_0, p_1) = 0$.

[17] Recall that we are leaving implicit that we are only evaluating expressions where they are well defined. Thus, the following assertion holds for $p_0, p_1 \in \Omega_P$ with $p_0 > p_1$.

Proof. We have

$$\begin{aligned}\Pr[Y=1|P=p] &= \Pr[D=1,Y=1|P=p]+\Pr[D=0,Y=1|P=p]\\ &= \Pr[D=1,Y_1=1|P=p]+\Pr[D=0,Y_0=1|P=p]\\ &= \Pr[U\leq p,\epsilon\leq\nu_1]+\Pr[U>p,\epsilon\leq\nu_0]\end{aligned}$$

where the last equality is using our independence assumption and that $\epsilon_j|U \sim \epsilon|U$ for $j=0,1$. Thus, for any p_0, p_1 with $p_0 > p_1$,

$$h(p_0,p_1)=\begin{cases}\Pr[p_1<U\leq p_0,\nu_0<\epsilon\leq\nu_1] & \text{if } \nu_1>\nu_0\\ 0 & \text{if } \nu_1=\nu_0\\ -\Pr[p_1<U\leq p_0,\nu_1<\epsilon\leq\nu_0] & \text{if } \nu_1<\nu_0.\end{cases}$$

Using (A-5), we thus have that $h(p_0,p_1)$ will be strictly positive if $\nu_1-\nu_0>0$, $h(p_0,p_1)$ will equal zero if $\nu_1-\nu_0=0$, and $h(p_0,p_1)$ will be strictly negative if $\nu_1-\nu_0<0$. Thus $\text{sgn}[h(p_0,p_1)]=\text{sgn}[\nu_1-\nu_0]$. Since the sign of $h(p_0,p_1)$ does not depend on the p_0,p_1 evaluation points provided that $p_0>p_1$, we have $\text{sgn}[H]=\text{sgn}[h(p_0,p_1)]$. □

Figure 4, below, provides a graphical illustration of the proof of Lemma 4.2 for an example with $\nu_1>\nu_0$. $\Pr[Y=1|P=p_0]$ corresponds to the probability that (U,ϵ) lies in one of two rectangles: (1) the set of all (U,ϵ) values lying southwest of (p_0,ν_1), which is the set of (U,ϵ) values resulting in $(D=1,Y=1)$, and (2) the set of all (U,ϵ) values lying southeast of (p_0,ν_0), which is the set of all (U,ϵ) values resulting in $(D=0,Y=1)$. Likewise, $\Pr[Y=1|P=p_1]$ corresponds to the probability that (U,ϵ) lies in one of two rectangles, the set of all (U,ϵ) values lying southwest of (p_1,ν_1) and the set of all (U,ϵ) values lying southeast of (p_1,ν_0). Thus, if $\nu_1>\nu_0$ (as in the figure), then $h(p_0,p_1)=\Pr[Y=1|P=p_0]-\Pr[Y=1|P=p_1]$ corresponds to the probability that (ϵ,U) lies in a particular rectangle with positive Lebesgue measure on $\Re^2$ and thus the probability that (ϵ,U) lies in this rectangle is strictly positive by our Assumption (A-5). The figure is done for an example with $\nu_1>\nu_0$. In contrast, if $\nu_1=\nu_0$, then $h(p_0,p_1)=\Pr[Y=1|P=p_0]-\Pr[Y=1|P=p_1]$ corresponds to the probability that (ϵ,U) lies along a line which has zero Lebesgue measure on $\Re^2$ and thus the probability that (ϵ,U) lies along the line is zero by Assumption (A-5). Finally, if $\nu_1<\nu_0$ then $h(p_0,p_1)$ corresponds to the negative of the probability of (U,ϵ) lying in a particular rectangle.

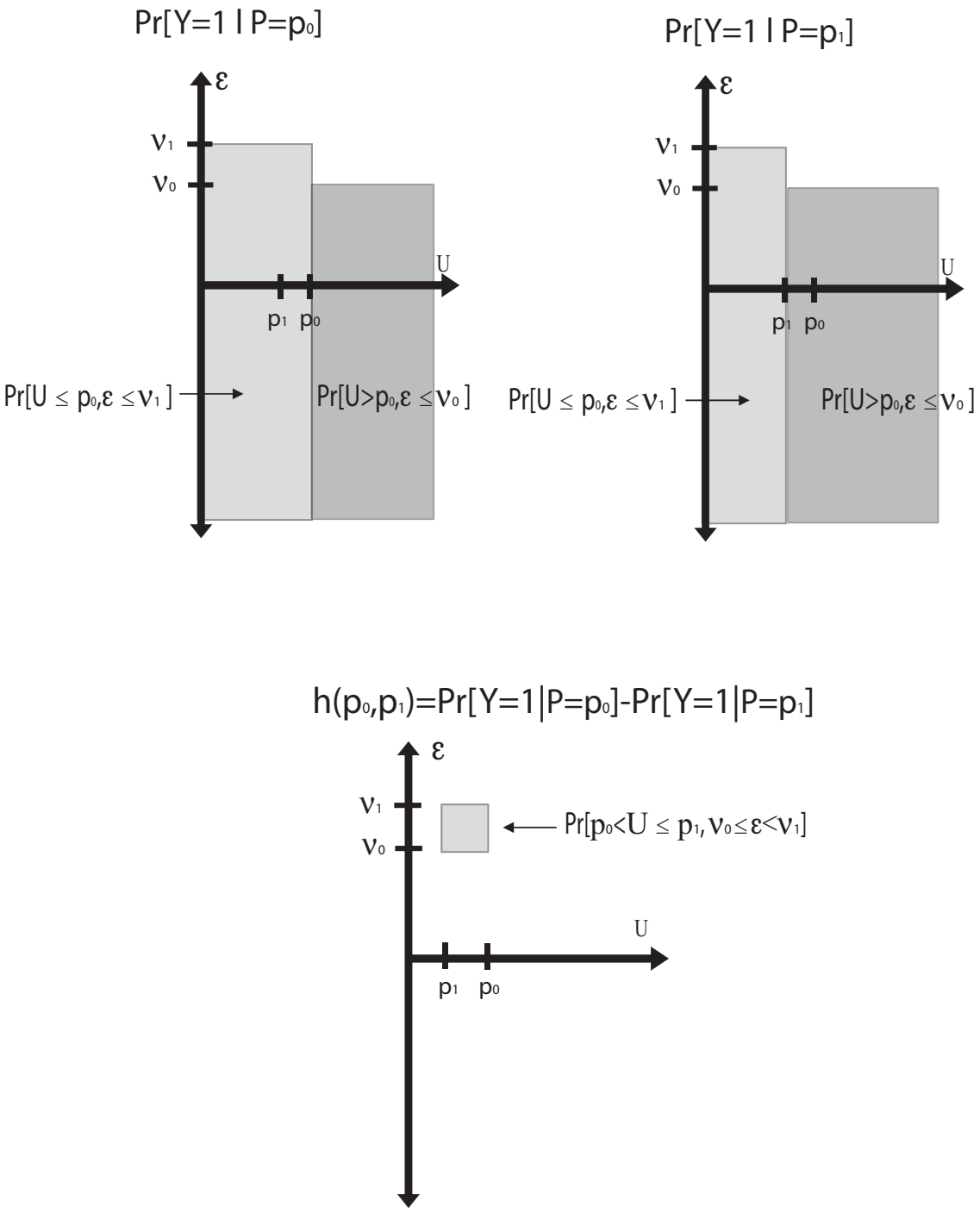
We can use Lemma 4.2 to bound $\Pr[Y_0=1|D=1,P=p]=F_{\epsilon|U}(\nu_1|p)$ and $\Pr[Y_1=1|D=0,P=p]=F_{\epsilon|-U}(\nu_0|p)$. For example, suppose that $H>0$. Then we know that $\nu_1>\nu_0$, and thus

$$\begin{array}{rcccccl}\Pr[Y_1=1|D=0,P=p] &=& F_{\epsilon|-U}(\nu_1|p) &>& F_{\epsilon|-U}(\nu_0|p) &=& \Pr[Y=1|D=0,P=p]\\ \Pr[Y_0=1|D=1,P=p] &=& F_{\epsilon|U}(\nu_0|p) &<& F_{\epsilon|U}(\nu_1|p) &=& \Pr[Y=1|D=1,P=p],\end{array}$$

where the strict inequalities follow from assumption (A-5) implying that both conditional cumulative distribution functions are strictly increasing. Thus, if $H>0$, we can bound the unidentified term $\Pr[Y_1=1|D=0,P=p]$ from below by the identified term $\Pr[Y=1|D=0,P=p]$, and we can bound the unidentified term $\Pr[Y_0=1|D=1,P=p]$ from above by the identified term $\Pr[Y=1|D=1,P=p]$. Since any probability must lie in the unit interval, we then have

$$\begin{array}{rcccccl}\Pr[Y=1|D=0,P=p] &<& \Pr[Y_1=1|D=0,P=p] &\leq& 1 & & \\ 0 &\leq& \Pr[Y_0=1|D=1,P=p] &<& \Pr[Y=1|D=1,P=p]. & & \end{array} \tag{15}$$

Figure 1: Graphical Illustration of Lemma 4.2, Example with $\nu_1 > \nu_0$



Parallel bounds hold if $H < 0$. If $H = 0$, then $\Pr[Y_1 = 1|D = 0, P = p] = \Pr[Y = 1|D = 0, P = p] =$ and $\Pr[Y_0 = 1|D = 1, P = p] = \Pr[Y = 1|D = 1, P = p]$.

Combining the results of Lemma 4.1 and Lemma 4.2, we immediately have:

$$\Pr[Y_0 = 1 \mid D = 1, P = p] \in \begin{cases} \bigcap_{\tilde{p}} \left[m_0(p,\tilde{p}), \quad m_0(p,\tilde{p}) + q_0(p,\tilde{p}) \Pr[Y = 1|D = 1, P = \tilde{p}] \right] & \text{if } H > 0 \\ \{\Pr[Y = 1 \mid D = 1, P = p]\} & \text{if } H = 0 \\ \bigcap_{\tilde{p}} \left[m_0(p,\tilde{p}) + q_0(p,\tilde{p}) \Pr[Y = 1|D = 1, P = \tilde{p}], \quad m_0(p,\tilde{p}) + q_0(p,\tilde{p}) \right] & \text{if } H < 0 \end{cases} \tag{16}$$

$$\Pr[Y_1 = 1 \mid D = 0, P = p] \in \begin{cases} \bigcap_{\tilde{p}} \left[m_1(p,\tilde{p}) + q_1(p,\tilde{p}) \Pr[Y = 1|D = 0, P = \tilde{p}], \quad m_1(p,\tilde{p}) + q_1(p,\tilde{p}) \right] & \text{if } H > 0 \\ \{\Pr[Y = 1|D = 0, P = p]\} & \text{if } H = 0 \\ \bigcap_{\tilde{p}} \left[m_1(p,\tilde{p}), \quad m_1(p,\tilde{p}) + q_1(p,\tilde{p}) \Pr[Y = 1|D = 0, P = \tilde{p}] \right] & \text{if } H < 0. \end{cases} \tag{17}$$

We can simplify this expression somewhat. Consider the case where $H > 0$ and consider the bounds on $\Pr[Y_1 = 1 \mid D = 0, P = p]$. Given our model of equation (13) and independence assumption (A-2), we have

$$m_1(p,\tilde{p}) + q_1(p,\tilde{p}) = \begin{cases} \frac{1}{1-p} \left\{ \Pr[p < U \leq \tilde{p}, \epsilon \leq \nu_1] + \Pr[U > \tilde{p}] \right\} & \text{if } \tilde{p} > p \\ \frac{1}{1-p} \Pr[U > \tilde{p}] & \text{if } \tilde{p} = p \\ -\frac{1}{1-p} \left\{ \Pr[\tilde{p} < U \leq p, \epsilon \leq \nu_1] - \Pr[U > \tilde{p}] \right\} & \text{if } \tilde{p} < p, \end{cases}$$

and

$$\begin{aligned} &m_1(p,\tilde{p}) + q_1(p,\tilde{p}) \Pr[Y = 1|D = 0, P = \tilde{p}] \\ &\qquad = \begin{cases} \frac{1}{1-p} \left\{ \Pr[p < U \leq \tilde{p}, \epsilon \leq \nu_1] + \Pr[U > \tilde{p}, \epsilon \leq \nu_0] \right\} & \text{if } \tilde{p} > p \\ \frac{1}{1-p} \Pr[U > \tilde{p}, \epsilon \leq \nu_0] & \text{if } \tilde{p} = p \\ -\frac{1}{1-p} \left\{ \Pr[\tilde{p} < U \leq p, \epsilon \leq \nu_1] - \Pr[U > \tilde{p}, \epsilon \leq \nu_0] \right\} & \text{if } \tilde{p} < p. \end{cases} \end{aligned}$$

For any $c > 0$

$$\begin{aligned} &(m_1(p,\tilde{p}+c) + q_1(p,\tilde{p}+c)) - (m_1(p,\tilde{p}) + q_1(p,\tilde{p})) \\ &\qquad = \frac{1}{1-p} \left\{ \Pr[\tilde{p} < U \leq \tilde{p} + c, \epsilon \leq \nu_1] - \Pr[\tilde{p} < U \leq \tilde{p} + c] \right\} \\ &\qquad = -\frac{1}{1-p} \Pr[\tilde{p} < U \leq \tilde{p} + c, \epsilon > \nu_1] \\ &\qquad < 0, \end{aligned}$$

so that $m_1(p,\tilde{p}) + q_1(p,\tilde{p})$ is decreasing in $\tilde{p}$. From $H > 0$, we have $\nu_0 < \nu_1$, and thus for any $c > 0$,

$$\begin{aligned} &(m_1(p,\tilde{p}+c) + q_1(p,\tilde{p}+c) \Pr[Y = 1|D = 0, P = \tilde{p} + c]) \\ &\qquad - (m_1(p,\tilde{p}) + q_1(p,\tilde{p}) \Pr[Y = 1|D = 0, P = \tilde{p}]) \\ &\qquad = \frac{1}{1-p} \left\{ \Pr[\tilde{p} < U \leq \tilde{p} + c, \epsilon \leq \nu_1] - \Pr[\tilde{p} < U \leq \tilde{p} + c, \epsilon \leq \nu_0] \right\} \\ &\qquad = \frac{1}{1-p} \Pr[\tilde{p} < U \leq \tilde{p} + c, \nu_0 < \epsilon \leq \nu_1] \\ &\qquad > 0, \end{aligned}$$

so that $m_1(p,\tilde{p}) + q_1(p,\tilde{p}) \Pr[Y=1|D=0,P=\tilde{p}]$ is increasing in $\tilde{p}$. Thus, if $H > 0$,

$$\begin{aligned}
&\Pr[Y_1 = 1 \mid D=0, P=p] \\
&\quad \in \bigcap_{\tilde{p}} \left[m_1(p,\tilde{p}) + q_1(p,\tilde{p}) \Pr[Y=1|D=0,P=\tilde{p}], \quad m_1(p,\tilde{p}) + q_1(p,\tilde{p}) \right] \\
&\quad = \left[m_1(p,p^u) + q_1(p,p^u) \Pr[Y=1|D=0,P=p^u], \quad m_1(p,p^u) + q_1(p,p^u) \right],
\end{aligned}$$

where p^u was defined by equation (10) as the supremum of the support of the distribution of P. Furthermore, note that

$$\begin{aligned}
q_1(p,p^u) \Pr[Y=1|D=0,P=p^u] &= \frac{1-p^u}{1-p} \Pr[Y=1|D=0,P=p^u] \\
&= (1-p)^{-1} \Pr[D=0,Y=1|P=p^u],
\end{aligned}$$

so that

$$\begin{aligned}
&\left[m_1(p,p^u) + q_1(p,p^u) \Pr[Y=1|D=0,P=p^u], \quad m_1(p,p^u) + q_1(p,p^u) \right] \\
&\qquad = \left[m_1(p,p^u) + (1-p)^{-1} \Pr[D=0,Y=1|P=p^u], \quad m_1(p,p^u) + q_1(p,p^u) \right].
\end{aligned}$$

Following the analogous arguments for $\Pr[Y_0 = 1 \mid D=1, P=p]$ and following the analogous argument for both $\Pr[Y_1 = 1 \mid D=0, P=p]$ and $\Pr[Y_0 = 1 \mid D=1, P=p]$ for the case of $H < 0$, we see that the bounds of equations (16) and (17) simplify to

$$\Pr[Y_0 = 1 \mid D=1, P=p] \in \mathcal{B}_0(p)$$

$$\Pr[Y_1 = 1 \mid D=0, P=p] \in \mathcal{B}_1(p)$$

with

$$\mathcal{B}_0(p) = \begin{cases} \left[m_0(p,p^l), \quad m_0(p,p^l) + p^{-1} \Pr[D=1,Y=1|P=p^l] \right] & \text{if } H > 0 \\ \{ \Pr[Y=1 \mid D=1, P=p] \} & \text{if } H = 0 \\ \left[m_0(p,p^l) + p^{-1} \Pr[D=1,Y=1|P=p^l], \quad m_0(p,p^l) + q_0(p,p^l) \right] & \text{if } H < 0, \end{cases} \tag{18}$$

$$\mathcal{B}_1(p) = \begin{cases} \left[m_1(p,p^u) + (1-p)^{-1} \Pr[D=0,Y=1|P=p^u], \quad m_1(p,p^u) + q_1(p,p^u) \right] & \text{if } H > 0 \\ \{ \Pr[Y=1|D=0,P=p] \} & \text{if } H = 0 \\ \left[m_1(p,p^u), \quad m_1(p,p^u) + (1-p)^{-1} \Pr[D=0,Y=1|P=p^u] \right] & \text{if } H < 0, \end{cases} \tag{19}$$

where p^l was defined by equation (10) as infimum of the support of the distribution of P.

The following theorem uses these bounds on $\Pr[Y_0 = 1 \mid D=1, P=p]$ and $\Pr[Y_1 = 1 \mid D=0, P=p]$ to bound the effect of treatment on the treated and the average treatment effect.

Theorem 4.1. *Assume that (D, Y_0, Y_1) are generated according to equation (13). Assume conditions (A-1)-(A-6).*[18] *Then,*

$$\begin{aligned}
\Delta^{TT}(p) &\in \mathcal{B}^{TT}(p) \\
\Delta^{ATE} &\in \mathcal{B}^{ATE},
\end{aligned}$$

[18] A weaker version of the theorem holds without assumption (A-5). Without assuming (A-5), we still have that the stated bounds hold when $H \neq 0$, though the stated bounds no longer hold when $H = 0$.

where

$$\mathcal{B}^{TT}(p) = \begin{cases} \left[p^{-1}h(p,p^l), \quad p^{-1}\left(h(p,p^l) + \Pr[D=1,Y=1 \mid P=p^l]\right)\right] & \text{if } H>0 \\ \{0\} & \text{if } H=0 \\ \left[p^{-1}\left(h(p,p^l) - \Pr[D=1,Y=0 \mid P=p^l]\right), \quad p^{-1}h(p,p^l)\right] & \text{if } H<0, \end{cases}$$

$$\mathcal{B}^{ATE} = \begin{cases} \left[h(p^u,p^l), \quad h(p^u,p^l) + \Pr[D=1,Y=1|P=p^l] + \Pr[D=0,Y=0|P=p^u]\right] & \text{if } H>0 \\ \{0\} & \text{if } H=0 \\ \left[h(p^u,p^l) - \Pr[D=1,Y=0|P=p^l] - \Pr[D=0,Y=1|P=p^u], \quad h(p^u,p^l)\right] & \text{if } H<0. \end{cases}$$

The bounds are sharp, they cannot be improved without additional restrictions.

Proof. First consider the bounds for TT. From $\Pr[Y_0=1|D=1,P=p] \in \mathcal{B}_0(p)$, we immediately have

$$\Delta^{TT}(p) \in \{\Pr[Y=1 \mid D=1,P=p] - s : s \in \mathcal{B}_0(p)\}.$$

By plugging in the definitions of m_0 and q_0 and rearranging terms, one can easily show that

$$\Pr[Y=1 \mid D=1,P=p] - m_0(p,p^l) - p^{-1}\Pr[D=1,Y=1|P=p^l] = p^{-1}h(p,p^l),$$

and

$$\Pr[Y=1 \mid D=1,P=p] - m_0(p,p^l) - q_0(p,p^l) = p^{-1}\left(h(p,p^l) - \Pr[D=1,Y=0|P=p^l]\right).$$

The stated bounds on TT now immediately follow.

Now consider the bounds for ATE. From $\Pr[Y_0=1|D=1,P=p] \in \mathcal{B}_0(p)$ and $\Pr[Y_1=1|D=0,P=p] \in \mathcal{B}_1(p)$, we have

$$\Delta^{ATE} \in \bigcap_{p,\tilde{p}} \{\Pr[D=1,Y=1 \mid P=p] + (1-p)t \\ - \Pr[D=0,Y=1 \mid P=\tilde{p}] - \tilde{p}s : s \in \mathcal{B}_0(\tilde{p}), t \in \mathcal{B}_1(p)\}.$$

Following reasoning analogous to how we simplified from equations (16) and (17) to equations (18) and (19), one can show that

$$\bigcap_{p,\tilde{p}} \{\Pr[D=1,Y=1 \mid P=p] + (1-p)t \\ - \Pr[D=0,Y=1 \mid P=\tilde{p}] - \tilde{p}s : s \in \mathcal{B}_0(\tilde{p}), t \in \mathcal{B}_1(p)\} \\ = \{\Pr[D=1,Y=1 \mid P=p^u] + (1-p^u)t \\ - \Pr[D=0,Y=1 \mid P=p^l] - p^l s : s \in \mathcal{B}_0(p^l), t \in \mathcal{B}_1(p^u)\}. \quad (20)$$

Using the definitions of m_1, m_0, q_1, and q_0, we have $m_1(p^u,p^u) = m_0(p^l,p^l) = 0$ and $q_1(p^u,p^u) = q_0(p^l,p^l) = 1$. By adding and subtracting terms, one can easily show that

$$\Pr[D=1,Y=1|P=p^u] + (1-p^u) - \Pr[D=0,Y=1|P=p^l] \\ = h(p^u,p^l) + \Pr[D=1,Y=1|P=p^l] + \Pr[D=0,Y=0|P=p^u]$$

and

$$\begin{aligned}\Pr[D=1,Y=1|P=p^u]-\Pr[D=0,Y=1|P=p^l]-p^l \\ = h(p^u,p^l)-\Pr[D=1,Y=0|P=p^l]-\Pr[D=0,Y=1|P=p^u].\end{aligned}$$

The stated bounds on ATE now immediately follow.

We now show that the constructed bounds are sharp. First consider the bounds on TT. Let (ϵ^*,U^*) denote a random vector with $(\epsilon^*,U^*) \perp\!\!\!\perp Z$ and with (ϵ^*,U^*) having density $f^*_{\epsilon,U}$ with respect to Lebesgue measure on $\Re^2$. Let f^*_U denote the corresponding marginal density of U^* and let $f^*_{\epsilon|U}$ denote the corresponding density of ϵ^* conditional on U^*. We show that for any fixed $\tilde{p} \in \Omega_p$, and s in the interior of $\mathcal{B}^{TT}(\tilde{p})$, there exists a density function $f^*_{\epsilon,U}$ such that: (1) $f^*_{\epsilon|U}$ is strictly positive on $\Re$; (2) $\Pr[D=1 \mid P=p]=\Pr[U^* \leq p]$, $\Pr[Y=1|D=1,P=p]=\Pr[\epsilon^* \leq \nu_1 \mid U^* \leq p]$, and $\Pr[Y=1|D=0,P=p]=\Pr[\epsilon^* \leq \nu_0 \mid U^* > p]$ for all $p \in \Omega_P$ (i.e., the, proposed model is consistent with the observed data); (3) $\Pr[\epsilon^* \leq \nu_1 \mid U^* \leq \tilde{p}]-\Pr[\epsilon^* \leq \nu_0 \mid U^* \leq \tilde{p}]=s$ (i.e., the proposed model is consistent with the specified value of TT). If we can construct a density $f^*_{\epsilon,U}$ satisfying conditions (1)-(3) for any s in the interior of $\mathcal{B}^{TT}(\tilde{p})$, we can conclude that any value in $\mathcal{B}^{TT}(\tilde{p})$ can be rationalized by a model consistent both with the observed data and our assumptions, and thus $\mathcal{B}^{TT}(\tilde{p})$ are sharp bounds.

Take the case where $H>0$. The case with $H<0$ is symmetric, and the case with $H=0$ is immediate. Fix some $\tilde{p} \in \Omega_p$ and some s in the interior of $\mathcal{B}^{TT}(\tilde{p})$. Let $s^*=\tilde{p}[\Pr[Y=1|D=1,P=\tilde{p}]-m_0(\tilde{p},p^l)]-s]$, and note that s being in the interior of $\mathcal{B}^{TT}(\tilde{p})$ implies $s^* \in (0,F_{\epsilon,U}(\nu_1,p^l))$. Note that $\nu_1>\nu_0$ since $H>0$ by assumption. Construct the proposed $f^*_{\epsilon,U}$ as follows. Let $f^*_{\epsilon,U}(t_1,t_2)=f^*_{\epsilon|U}(t_1|t_2)f^*_U(t_2)$, where $f^*_U(t_2)=f_U(t_2)=\mathbf{1}[0 \leq t_2 \leq 1]$ and

$$f^*_{\epsilon|U}(t_1|t_2)=\begin{cases} f_{\epsilon|U}(t_1|t_2) & \text{if } t_1 \geq \nu_1 \text{ or } t_2 \geq p^l \\ b(t_2)f_{\epsilon|U}(t_1|t_2) & \text{if } \nu_0 < t_1 < \nu_1 \text{ and } t_2 < p^l \\ a(t_2)f_{\epsilon|U}(t_1|t_2) & \text{if } t_1 < \nu_0 \text{ and } t_2 < p^l \end{cases} \tag{21}$$

with

$$\begin{aligned} a(t_2) &= \frac{\Pr[\epsilon \leq \nu_1|U=t_2]}{\Pr[\epsilon \leq \nu_0|U=t_2]}\frac{s^*}{F_{\epsilon,U}(\nu_1,p^l)} \\ b(t_2) &= \frac{\Pr[\epsilon \leq \nu_1|U=t_2]-a(t_2)\Pr[\epsilon \leq \nu_0|U=t_2]}{\Pr[\nu_0 \leq \epsilon < \nu_1|U=t_2]}. \end{aligned} \tag{22}$$

First consider whether $f^*_{\epsilon|U}$ integrates to one and is strictly positive on $\Re$. For $t_2 \geq p_l$, $f^*_{\epsilon|U}(\cdot|t_2)=f_{\epsilon|U}(\cdot|t_2)$ and thus trivially $\int f^*_{\epsilon|U}(t_1|t_2)dt_1=\int f_{\epsilon|U}(t_1|t_2)dt_1=1$. For $t_2<p_l$,

$$\begin{aligned}\int_{-\infty}^{\infty} f^*_{\epsilon|U}(t_1|t_2)dt_1 &= a(t_2)\int_{-\infty}^{\nu_0} f_{\epsilon|U}(t_1|t_2)dt_1 + b(t_2)\int_{\nu_0}^{\nu_1} f_{\epsilon|U}(t_1|t_2)dt_1 + \int_{\nu_1}^{\infty} f_{\epsilon|U}(t_1|t_2)dt_1 \\ &= \Pr[\epsilon \leq \nu_1|U=t_2]+\Pr[\epsilon > \nu_1|U=t_2]=1.\end{aligned}$$

Since $f_{\epsilon|U}$ is strictly positive on $\Re$, we have that $f^*_{\epsilon|U}$ is strictly positive on $\Re$ if $a(t_2)>0$ and $b(t_2)>0$. Recall that $s^* \in (0,F_{\epsilon,U}(\nu_1,p^l))$. $s^*>0$ implies $a(t_2)>0$. Using that $s^* \in (0,F_{\epsilon,U}(\nu_1,p^l))$, we have $\Pr[\epsilon \leq \nu_1|U=t_2]-a(t_2)\Pr[\epsilon \leq \nu_0|U=t_2]=\Pr[\epsilon \leq \nu_1|U=t_2](1-s^*/F_{\epsilon,U}(\nu_1,p^l))>0$ and thus that $b(t_2)>0$. We have thus shown that $f^*_{\epsilon|U}$ is a proper density satisfying part (1) of the assertion.

Now consider part (2) of the assertion. $f^*_U = f_U$ implies that

$$\Pr[U^* \leq p] = \int_0^p f^*_U(t)dt = \int_0^p f_U(t)dt = \Pr[U \leq p] = \Pr[D = 1|P = p] \quad \forall p \in \Omega_P.$$

$f^*_U = f_U$ and $f^*_{\epsilon|U}(t_1|t_2) = f_{\epsilon|U}(t_1|t_2)$ for $t_2 \geq p^l$ imply that $f^*_{\epsilon,U}(t_1, t_2) = f_{\epsilon,U}(t_1, t_2)$ for all $t_2 \geq p^l$, and thus

$$\begin{aligned}
\Pr[\epsilon^* \leq \nu_0|U^* > p] &= \frac{1}{1-p}\int_p^1 \int_{-\infty}^{\nu_0} f^*_{\epsilon,U}(t_1, t_2)dt_1dt_2 \\
&= \frac{1}{1-p}\int_p^1 \int_{-\infty}^{\nu_0} f_{\epsilon,U}(t_1, t_2)dt_1dt_2 \\
&= \Pr[\epsilon \leq \nu_0|U > p] = \Pr[Y = 1|D = 0, P = p]
\end{aligned}$$

for all $p \in \Omega_P$. Likewise,

$$\begin{aligned}
&\Pr[\epsilon^* \leq \nu_1|U^* \leq p] \\
&= \tfrac{1}{p}\textstyle\int_0^p \int_{-\infty}^{\nu_1} f^*_{\epsilon,U}(t_1, t_2)dt_1dt_2 \\
&= \tfrac{1}{p}\left\{\textstyle\int_{p^l}^p \int_{-\infty}^{\nu_1} f_{\epsilon,U}(t_1, t_2)dt_1dt_2 + \int_0^{p^l}\left[b(t_2)\int_{\nu_0}^{\nu_1} f_{\epsilon,U}(t_1, t_2)dt_1 + a(t_2)\int_{-\infty}^{\nu_0} f_{\epsilon,U}(t_1, t_2)dt_1\right]dt_2\right\} \\
&= \tfrac{1}{p}\left\{\Pr[\epsilon \leq \nu_1, p^l < U \leq p] + \Pr[\epsilon \leq \nu_1, U \leq p^l]\right\} \\
&= \Pr[\epsilon \leq \nu_1|U \leq p] = \Pr[Y = 1 \mid D = 1, P = p]
\end{aligned}$$

for all $p \in \Omega_P$. We have thus established part (2) of the assertion. Consider part (3) of the assertion. We have already shown $\Pr[\epsilon^* \leq \nu_1|U^* \leq \tilde{p}] = \Pr[\epsilon \leq \nu_1|U \leq \tilde{p}]$ since $\tilde{p} \in \Omega_P$. Consider $\Pr[\epsilon^* \leq \nu_0|U^* \leq \tilde{p}]$,

$$\begin{aligned}
\Pr[\epsilon^* \leq \nu_0|U^* \leq \tilde{p}] &= \frac{1}{\tilde{p}}\int_0^{\tilde{p}} \int_{-\infty}^{\nu_0} f^*_{\epsilon,U}(t_1, t_2)dt_1dt_2 \\
&= \frac{1}{\tilde{p}}\left\{\int_0^{p^l} \int_{-\infty}^{\nu_0} f^*_{\epsilon,U}(t_1, t_2)dt_1dt_2 + \int_{p^l}^{\tilde{p}} \int_{-\infty}^{\nu_0} f^*_{\epsilon,U}(t_1, t_2)dt_1dt_2\right\} \\
&= \frac{1}{\tilde{p}}\left\{\int_0^{p^l} a(t_2)\int_{-\infty}^{\nu_0} f_{\epsilon,U}(t_1, t_2)dt_1dt_2 + \int_{p^l}^{\tilde{p}} \int_{-\infty}^{\nu_0} f_{\epsilon,U}(t_1, t_2)dt_1dt_2\right\} \\
&= \frac{1}{\tilde{p}}\left\{s^* + \tilde{p}m_0(\tilde{p}, p^l)\right\} = \Pr[\epsilon \leq \nu_1|U^* \leq \tilde{p}] - s
\end{aligned}$$

and thus $\Pr[\epsilon^* \leq \nu_1|U^* \leq \tilde{p}] - \Pr[\epsilon^* \leq \nu_0|U^* \leq \tilde{p}] = s$.

Now consider the bounds on ATE. Take the case where $H > 0$. The case with $H < 0$ is symmetric, and the case with $H = 0$ is immediate. Fix some $b \in \mathcal{B}^{ATE}$. Fix some s, t pair, s in the interior of $\mathcal{B}_0(p^l)$ and t in the interior of $\mathcal{B}_1(p^u)$, such that

$$b = \Pr[D = 1, Y = 1 \mid P = p^u] + (1 - p^u)t - \Pr[D = 0, Y = 1 \mid P = p^l] - p^l s.$$

(The existence of such an s, t pair follows from equation (20)). Let $s^* = p^l s$, $t^* = (1-p^u)t$. Construct

$f^*_{\epsilon,U}$ as follows. Let $f^*_{\epsilon,U}(t_1,t_2) = f^*_{\epsilon|U}(t_1|t_2)f^*_U(t_2)$, where $f^*_U(t_2) = f_U(t_2) = \mathbf{1}[0 \leq t_2 \leq 1]$ and

$$f^*_{\epsilon|U}(t_1|t_2) = \begin{cases} f_{\epsilon|U}(t_1|t_2) & \text{if } (p^l < t_2 \leq p^u) \text{ or } (t_1 > \nu_1, t_2 < p^l) \text{ or } (t_1 \leq \nu_0, t_2 > p^u) \\ b(t_2)f_{\epsilon|U}(t_1|t_2) & \text{if } \nu_0 < t_1 < \nu_1 \text{ and } t_2 < p^l \\ a(t_2)f_{\epsilon|U}(t_1|t_2) & \text{if } t_1 < \nu_0 \text{ and } t_2 < p^l \\ d(t_2)f_{\epsilon|U}(t_1|t_2) & \text{if } t_1 > \nu_1 \text{ and } t_2 > p^u \\ c(t_2)f_{\epsilon|U}(t_1|t_2) & \text{if } \nu_0 < t_1 < \nu_1, \text{ and } t_2 > p^u \end{cases}$$

with $a(t_2)$ and $b(t_2)$ defined by equation (27) and $c(t_2)$ and $d(t_2)$ defined by:

$$\begin{aligned} c(t_2) &= \frac{t^* - \Pr[\epsilon \leq \nu_0, U > p^u]}{\Pr[\epsilon > \nu_0, U > p^u]} \frac{\Pr[\epsilon > \nu_0 | U = t_2]}{\Pr[\nu_0 < \epsilon \leq \nu_1, U > p^u | U = t_2]} \\ d(t_2) &= \frac{\Pr[\epsilon > \nu_0 | U = t_2] - c(t_2)\Pr[\nu_0 < \epsilon < \nu_1 | U = t_2]}{\Pr[\epsilon > \nu_1 | U = t_2]}. \end{aligned}$$

Following arguments closely analogous to those given above for the TT bounds, one can now proceed to show that proposed density has the desired properties and that the bounds on ATE are sharp. □

By Lemma 4.1, the sign of H equals the sign of $h(p_0, p_1)$ for all p_0, p_1 with $p_0 > p_1$. Thus, the constructed bounds on Δ^{ATE} and $\Delta^{TT}(p)$ always identify whether these parameters are positive, zero, or negative. For example, consider ATE. If $H > 0$, then ATE is identified to be positive and the width of the bounds on ATE is $\Pr[D = 1, Y = 1 | P = p^l] + \Pr[D = 0, Y = 0 | P = p^u]$; if $H = 0$, then ATE is point identified to be zero; if $H < 0$ then ATE is identified to be negative and the width of the bounds on ATE is $\Pr[D = 1, Y = 0 | P = p^l] + \Pr[D = 0, Y = 1 | P = p^u]$. Sufficient conditions for the bounds on ATE to collapse to point identification are either $H = 0$ or $p^u = 1$ and $p^l = 0$. The width of the bounds on ATE if $H > 0$ are shrinking in $(1 - p^u)$, p^l, $\Pr[Y = 1 | D = 1, P = p^l]$ and $\Pr[Y = 0 | D = 0, P = p^u]$. The width of the bounds on ATE if $H < 0$ are shrinking in $(1 - p^u)$, p^l, $\Pr[Y = 0 | D = 1, P = p^l]$ and $\Pr[Y = 1 | D = 0, P = p^u]$. We will show in the next section that any covariates that directly affect Y can also be exploited to further narrow the bounds on ATE and TT. We will show further in Section 6 that these bounds are expected to be substantially narrower than alternative bounds that exploit an instrument but do not impose or exploit the threshold crossing structure on D and Y.

5 Analysis With X Covariates

We now consider analysis with X covariates. If there is variation in X conditional on $P(Z)$, then the X covariates can be used to substantially decrease the width of the bounds compared to the case with no X covariates. A sufficient condition for X to vary conditional on $P(Z)$ is that there exists an element of X that is not contained in Z. However, this condition is not required: even if all elements of X are contained in Z, then it is still possible to have X nondegenerate conditional on $P(Z)$. We first generalize Lemmas 4.1 and 4.2 to allow for X regressors.

Lemma 5.1. *Assume that (D, Y_0, Y_1) are generated according to equations (1)-(2). Assume conditions (A-1)-(A-4). Then, for $j = 0, 1$ and for any $(x, p), (x, \tilde{p})$,*[19]

$$\Pr[Y_j = 1 \mid D = 1 - j, X = x, P = p] = m_j(x, p, \tilde{p}) + q_j(p, \tilde{p}) \Pr[Y_j = 1 \mid D = 1 - j, X = x, P = \tilde{p}].$$

Proof. Follows from a trivial modification to the proof of Lemma 4.1. □

Lemma 5.2. *Assume that (D, Y_0, Y_1) are generated according to equations (1)-(2). Assume conditions (A-1)-(A-5).*[20] *Then, for any x_0, x_1, p_0, p_1 evaluation points with $p_0 > p_1$,*[21]

$$sgn\big[H(x_0, x_1)\big] = sgn\big[h_1(p_0, p_1, x_1) - h_0(p_0, p_1, x_0)\big] = sgn\big[\nu_1(x_1) - \nu_0(x_0)\big].$$

Proof. We have

$$\begin{aligned} \Pr[D = 1, Y = 1 | X = x, P = p] &= \Pr[D = 1, Y_1 = 1 | X = x, P = p] \\ &= \Pr[U \leq p, \epsilon \leq \nu_1(x)] \end{aligned}$$

and

$$\begin{aligned} \Pr[D = 0, Y = 1 | X = x, P = p] &= \Pr[D = 0, Y_0 = 1 | X = x, P = p] \\ &= \Pr[U > p, \epsilon \leq \nu_0(x)], \end{aligned}$$

where we are using our independence assumption and that $\epsilon_j | U \sim \epsilon | U$ for $j = 0, 1$. Thus

$$h_1(p_0, p_1, x_1) = \Pr[p_1 < U \leq p_0, \epsilon \leq \nu_1(x_1)]$$

$$h_0(p_0, p_1, x_0) = \Pr[p_1 < U \leq p_0, \epsilon \leq \nu_0(x_0)]$$

and thus

$$h_1(p_0, p_1, x_1) - h_0(p_0, p_1, x_0) = \begin{cases} \Pr[p_1 < U \leq p_0, \nu_0(x_0) < \epsilon \leq \nu_1(x_1)] & \text{if } \nu_1(x_1) > \nu_0(x_0) \\ 0 & \text{if } \nu_1(x_1) = \nu_0(x_0) \\ -\Pr[p_1 < U \leq p_0, \nu_1(x_1) < \epsilon \leq \nu_0(x_0)] & \text{if } \nu_1(x_1) < \nu_0(x_0). \end{cases}$$

Using (A-5), we thus have that $h_1(p_0, p_1, x_1) - h_0(p_0, p_1, x_0)$ will be strictly positive if $\nu_1(x_1) - \nu_0(x_0) > 0$, $h_1(p_0, p_1, x_1) - h_0(p_0, p_1, x_0)$ will be strictly negative if $\nu_1(x_1) - \nu_0(x_0) < 0$, and thus $\text{sgn}[h_1(p_0, p_1, x_1) - h_0(p_0, p_1, x_0)] = \text{sgn}[\nu_1(x_1) - \nu_0(x_0)]$. Since the sign of $h_1(p_0, p_1, x_1) - h_0(p_0, p_1, x_0)$ does not depend on the p_0, p_1 evaluation points provided that $p_0 > p_1$, we have $\text{sgn}[H(x_0, x_1)] = \text{sgn}[h_1(p_0, p_1, x_1) - h_0(p_0, p_1, x_0)]$. □

[19] Recall that we are leaving implicit the standard support condition. Thus, this lemma holds for $(x, p), (x, \tilde{p}) \in \Omega_{X,P}$.

[20] A weaker version of the lemma holds without assumption (A-5). Without assuming (A-5), we still have that $h_1(p_0, p_1, x_1) - h_0(p_0, p_1, x_0) > 0 \Rightarrow \nu_1(x_1) > \nu_0(x_0)$ and $h_1(p_0, p_1, x_1) - h_0(p_0, p_1, x_0) < 0 \Rightarrow \nu_1(x_1) < \nu_0(x_0)$, but we are no longer able to infer the sign of $\nu_1(x_1) - \nu_0(x_0)$ if $h_1(p_0, p_1, x_1) - h_0(p_0, p_1, x_0) = 0$.

[21] Recall that we are leaving implicit that we are only evaluating expressions where they are well defined. Thus, this lemma holds for $(x_i, p_j) \in \Omega_{X,P}$ for $i = 0, 1$, $j = 0, 1$, with $p_0 > p_1$.

Figures 2 and 3, below, provide a graphical illustration of the proof of Lemma 5.2 for an example with $\nu_1(0) > \nu_0(1)$. $\Pr[D = 1, Y = 1|X = 0, P = p_0]$ and $\Pr[D = 1, Y = 1|X = 0, P = p_1]$ correspond to the probability that (U, ϵ) lies southwest of $(p_0, \nu_1(0))$ and $(p_1, \nu_1(0))$, respectively, so that $h_1(p_0, p_1, 0) = \Pr[D = 1, Y = 1|X = 0, P = p_0] - \Pr[D = 1, Y = 1|X = 0, P = p_1]$ corresponds to the probability that (U, ϵ) lies in a particular rectangle. In comparison, $\Pr[D = 0, Y = 1|X = 1, P = p_0]$ and $\Pr[D = 0, Y = 1|X = 1, P = p_1]$ correspond to the probability that (U, ϵ) lies southeast of $(p_0, \nu_0(1))$ and $(p_1, \nu_0(1))$, respectively, so that $h_0(p_0, p_1, 1) = \Pr[D = 0, Y = 1|X = 1, P = p_1] - \Pr[D = 0, Y = 1|X = 1, P = p_0]$ corresponds to the probability that (U, ϵ) lies in a particular rectangle. Thus, in this example with $\nu_1(0) > \nu_0(1)$, we have that $h_1(p_0, p_1, 0) - h_0(p_0, p_1, 1)$ corresponds to the probability that (ϵ, U) lies in a particular rectangle (with positive Lebesgue measure on $\Re^2$), and thus the probability that (ϵ, U) lies this rectangle is positive by Assumption (A-5).

Now consider bounds on ATE and TT. As discussed in Section 3, our goal is to bound $E(Y_0|D = 1, X = x, P = p)$ and $E(Y_1|D = 0, X = x, P = p)$ which in turn will allow us to bound $E(Y_1 - Y_0|X = x)$ and $E(Y_1 - Y_0|D = 1, X = x, P = p)$. First consider $\Pr[Y_0 = 1 \mid D = 1, X = x, P = p]$. From Lemma 5.1 we can decompose $\Pr[Y_0 = 1 \mid D = 1, X = x, P = p]$ as

$$\Pr[Y_0 = 1 \mid D = 1, X = x, P = p] = m_0(x, p, \tilde{p}) + p^{-1}\tilde{p}\Pr[Y_0 = 1 \mid D = 1, X = x, P = \tilde{p}].$$

While the first term, $m_0(x, p, \tilde{p})$, is identified, the second term, $\Pr[Y_0 = 1 \mid D = 1, X = x, P = \tilde{p}] = F_{\epsilon|U}(\nu_0(x)|\tilde{p})$, is not immediately identified. However, we can use Lemma 5.2 to bound this term. For example, suppose that $H(x, \tilde{x}) \geq 0$, i.e., $\tilde{x} \in X_0^U(x)$ where $\mathcal{X}_0^U$ was defined by equation (9). Then we know that $\nu_1(\tilde{x}) > \nu_0(x)$, and thus

$$\begin{aligned}\Pr[Y_0 = 1|D = 1, X = x, P = p] &= F_{\epsilon|U}(\nu_0(x)|p) \\ &< F_{\epsilon|U}(\nu_1(\tilde{x})|p) = \Pr[Y = 1|D = 1, X = \tilde{x}, P = p].\end{aligned}$$

Thus, we can bound the unidentified term $\Pr[Y_0 = 1|D = 1, X = x, P = p]$ from above by the identified term $\Pr[Y = 1|D = 1, X = \tilde{x}, P = p]$ for any $\tilde{x}$ such that $H(x, \tilde{x}) \geq 0$, i.e., for any $\tilde{x} \in \mathcal{X}_0^U(x)$. Symmetrically, we can bound $\Pr[Y_0 = 1|D = 1, X = x, P = p]$ from below by the identified term $\Pr[Y = 1|D = 1, X = \tilde{x}, P = p]$ for any $\tilde{x}$ such that $H(x, \tilde{x}) \leq 0$, i.e., $\tilde{x} \in \mathcal{X}_0^L(x)$. We thus have[22]

$$\begin{aligned}&\sup_{\tilde{x} \in \mathcal{X}_0^L} \{\Pr[Y = 1 \mid D = 1, X = \tilde{x}, P = p]\} \\ &\qquad \leq \Pr[Y_0 = 1 \mid D = 1, X = x, P = p] \\ &\qquad \leq \inf_{\tilde{x} \in \mathcal{X}_0^U} \{\Pr[Y = 1 \mid D = 1, X = \tilde{x}, P = p]\}, \quad (23)\end{aligned}$$

assuming that $\mathcal{X}_0^L$ and $\mathcal{X}_0^U$ are nonempty. If either $\mathcal{X}_0^L$ or $\mathcal{X}_0^U$ is empty, we can replace the upper and lower bounds by zero or one, respectively. For ease of exposition, we will write $\sup_{\tilde{x} \in \mathcal{X}_0^L} \{\Pr[Y = 1 \mid D = 1, X = \tilde{x}, P = p]\}$ with the implicit understanding that this term is 0 when $\mathcal{X}_0^L$ is empty, and $\inf_{\tilde{x} \in \mathcal{X}_0^U} \{\Pr[Y = 1 \mid D = 1, X = \tilde{x}, P = p]\}$ with the implicit understanding that this term is 1 when $\mathcal{X}_0^U$ is empty. This notation corresponds to the adopting the convention that the supremum over the empty set is zero and the infimum over the empty set is one.

[22]Recall that we are implicitly only evaluating terms where they are well defined. Thus, the following surpremum and infimum are over $\tilde{x}$ such that $(\tilde{x}, p) \in \Omega_{X,P}$.

Figure 2: Graphical Illustration of Lemma 5.2, Example with $\nu_1(0) > \nu_0(1)$

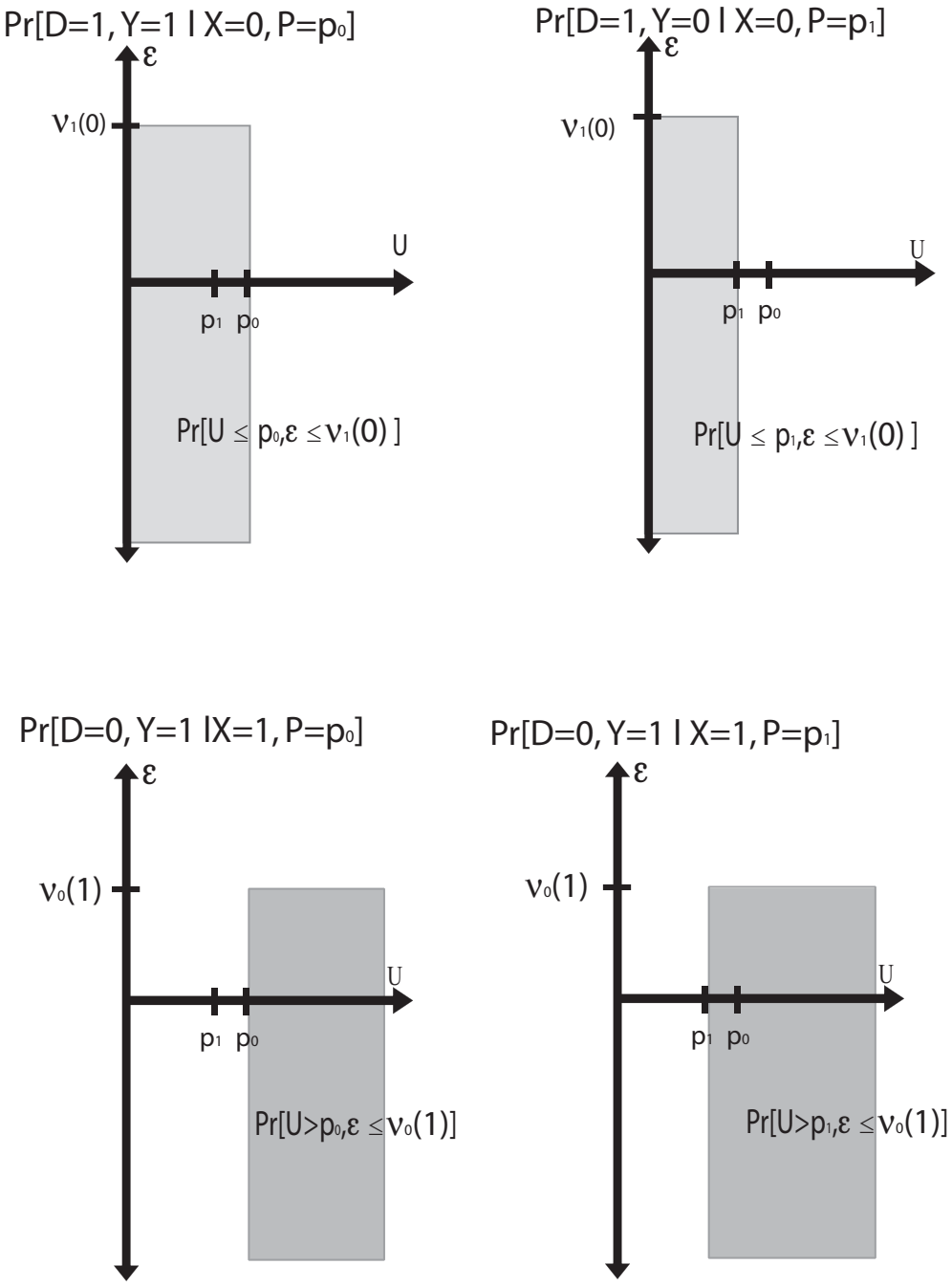
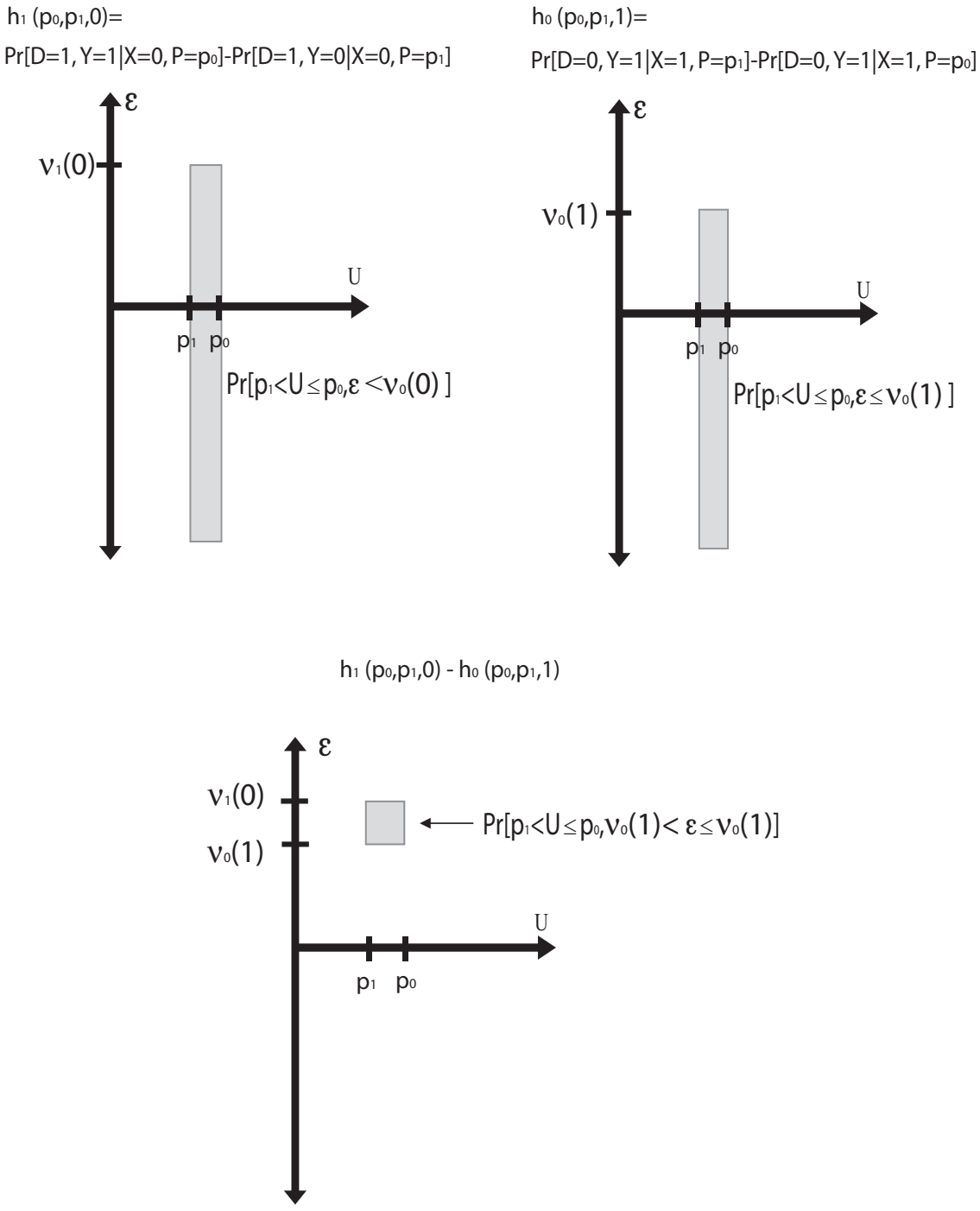


Figure 3: Graphical Illustration of Lemma 5.2, Example with $\nu_1(0) > \nu_0(1)$



The parallel argument can be made to construct bounds on $\Pr[Y_1 = 1 \mid D = 0, X = x, P = p]$. Thus, combining the results of Lemmas 5.1 and 5.2, we have, for $j = 0, 1$,

$$\Pr[Y_j = 1 \mid D = 1 - j, X = x, P = p] \in \mathcal{B}_j(x, p),$$

where[23]

$$\mathcal{B}_j(x,p) = \left[\sup_{\tilde{p}} \left\{ m_j(x,p,\tilde{p}) + q_j(p,\tilde{p}) \sup_{\tilde{x} \in \mathcal{X}_j^L(x)} \{\Pr[Y = 1 \mid D = 1 - j, X = \tilde{x}, P = \tilde{p})]\} \right\}, \right.$$
$$\left. \inf_{\tilde{p}} \left\{ m_j(x,p,\tilde{p}) + q_j(p,\tilde{p}) \inf_{\tilde{x} \in \mathcal{X}_j^U(x)} \{\Pr[Y = 1 \mid D = 1 - j, X = \tilde{x}, P = \tilde{p}]\} \right\} \right]. \quad (24)$$

We now use these bounds on $\Pr[Y_0 = 1 \mid D = 1, X = x, P = p]$ and $\Pr[Y_1 = 1 \mid D = 0, X = x, P = p]$ to form bounds on our objects of interest, the treatment on the treated and average treatment effect parameters.

Theorem 5.1. *Assume that (D, Y_0, Y_1) are generated according to equations (1)-(2). Assume conditions (A-1)-(A-6).*[24] *Then,*

$$\Delta^{TT}(x,p) \in [L^{TT}(x,p), U^{TT}(x,p)]$$
$$\Delta^{ATE}(x) \in [L^{ATE}(x), U^{ATE}(x)],$$

where[25]

$$L^{TT}(x,p) = \frac{1}{p} \sup_{\tilde{p}} \left\{ h(p,\tilde{p},x) + \Pr[D = 1, Y = 1 | X = x, P = \tilde{p}] \right.$$
$$\left. - \tilde{p} \inf_{\tilde{x} \in \mathcal{X}_0^U(x)} \{\Pr[Y = 1 \mid D = 1, X = \tilde{x}, P = \tilde{p}]\} \right\}$$

$$U^{TT}(x,p) = \frac{1}{p} \inf_{\tilde{p}} \left\{ h(p,\tilde{p},x) + \Pr[D = 1, Y = 1 | X = x, P = \tilde{p}] \right.$$
$$\left. - \tilde{p} \sup_{\tilde{x} \in \mathcal{X}_0^L(x)} \{\Pr[Y = 1 \mid D = 1, X = \tilde{x}, P = \tilde{p}]\} \right\}$$

[23] Recall that we are implicitly only evaluating terms where they are well defined. Thus, for example, the first surpremum is over $\tilde{p}$ such that $(x, \tilde{p}) \in \Omega_{X,P}$, and the second supremum is over $\tilde{x} \in \mathcal{X}_j^L(x)$ such that $(\tilde{x}, \tilde{p}) \in \Omega_{X,P}$. Further recall that we are implicitly adopting the convention that the supremum over the empty set is zero and the infimum over the empty set is one.

[24] A weaker version of the theorem holds without assumption (A-5). Without assuming (A-5), the stated bounds still hold but we must redefine the sets $\mathcal{X}_j^k(x)$ for $k \in \{U, L\}$, $j = 0, 1$, to be defined in terms of strict inequalities instead of weak inequalities.

[25] Recall our notational convention that the supremum over the empty set is zero and the infimum over the empty set is one.

$$L^{ATE}(x) =$$
$$\sup_p \left\{ \Pr[D=1,Y=1|X=x,P=p] + (1-p) \sup_{\tilde{x}\in\mathcal{X}_1^L(x)} \{\Pr[Y=1 \mid D=0, X=\tilde{x}, P=p]\} \right\}$$
$$- \inf_{\tilde{p}} \left\{ \Pr[D=0,Y=1|X=x,P=\tilde{p}] + \tilde{p} \inf_{\tilde{x}\in\mathcal{X}_0^U(x)} \{\Pr[Y=1|D=1,X=\tilde{x},P=\tilde{p}]\} \right\}$$

$$U^{ATE}(x) =$$
$$\inf_p \left\{ \Pr[D=1,Y=1|X=x,P=p] + (1-p) \inf_{\tilde{x}\in\mathcal{X}_1^U(x)} \{\Pr[Y=1 \mid D=0, X=\tilde{x}, P=p]\} \right\}$$
$$- \sup_{\tilde{p}} \left\{ \Pr[D=0,Y=1|X=x,P=\tilde{p}] + \tilde{p} \sup_{\tilde{x}\in\mathcal{X}_0^L(x)} \{\Pr[Y=1|D=1,X=\tilde{x},P=\tilde{p}]\} \right\}.$$

If $\Omega_{X,P} = \Omega_X \times \Omega_P$, then the bounds are sharp, they cannot be improved without additional restrictions.

Proof. First consider TT. From our previous analysis, we have

$$\Delta^{TT}(x,p) \in \{\Pr[Y=1 \mid D=1, X=x, P=p] - s : s \in \mathcal{B}_0(x,p)\}.$$

Rearranging terms, one can easily show that

$$\Pr[Y=1 \mid D=1, X=x, P=p] - m_0(x,p,\tilde{p})$$
$$= \frac{1}{p} h(p,\tilde{p},x) + \Pr[D=1,Y=1|X=x,P=\tilde{p}]$$

where $h(p,\tilde{p},x)$ was defined by equation (7). The stated bounds on TT now immediately follow.

Now consider ATE. From our previous analysis, we have:[26]

$$\Delta^{ATE}(x) \in \bigcap_{p,p^*} \{\Pr[D=1,Y=1 \mid X=x, P=p] + (1-p)s$$
$$- \Pr[D=0,Y=1 \mid X=x, P=p^*] - p^* t : s \in \mathcal{B}_1(x,p), t \in \mathcal{B}_0(x,p^*)\}.$$

The stated result now follows by rearranging terms.

We now show that the stated bounds are sharp if $\Omega_{X,P} = \Omega_X \times \Omega_P$. Impose $\Omega_{X,P} = \Omega_X \times \Omega_P$. Consider the bounds on TT (the proof that the bounds on ATE are sharp follows from an analogous argument). Let (ϵ^*, U^*) denote a random vector with $(\epsilon^*, U^*) \perp\!\!\!\perp (X, Z)$ and with (ϵ^*, U^*) having density $f^*_{\epsilon,U}$ with respect to Lebesgue measure on $\Re^2$. Let f^*_U denote the corresponding marginal density of U^* and let $f^*_{\epsilon|U}$ denote the corresponding density of ϵ^* conditional on U^*. We show that for any fixed $(\tilde{x},\tilde{p}) \in \Omega_{X,P}$, and $s \in (L^{TT}(\tilde{x},\tilde{p}), U^{TT}(\tilde{x},\tilde{p}))$, there exists a density function

[26] Recall that we are leaving implicit that we are only evaluating the conditional expectations where the conditional expectations are well defined. Thus, e.g., the following intersection is over all p, p^* such that $(x,p),(x,p^*) \in \Omega^1_{X,P} \bigcap \Omega^0_{X,P}$.

$f^*_{\epsilon,U}$ such that: (1) $f^*_{\epsilon|U}$ is strictly positive on $\Re$; (2) $\Pr[D=1 \mid X=x, P=p] = \Pr[U^* \leq p]$, $\Pr[Y=1|D=1, X=x, P=p] = \Pr[\epsilon^* \leq \nu_1(x) \mid U^* \leq p]$, and $\Pr[Y=1|D=0, X=x, P=p] = \Pr[\epsilon^* \leq \nu_0(x) \mid U^* > p]$ for all $(x,p) \in \Omega_{X,P}$ (i.e., the, proposed model is consistent with the observed data); (3) $\Pr[\epsilon^* \leq \nu_1(\tilde{x}) \mid U^* \leq \tilde{p}] - \Pr[\epsilon^* \leq \nu_0(\tilde{x}) \mid U^* \leq \tilde{p}] = s$ (i.e., the proposed model is consistent with the specified value of TT). If we can construct a density $f^*_{\epsilon,U}$ satisfying conditions (1)-(3) for any $s \in (L^{TT}(\tilde{x},\tilde{p}), U^{TT}(\tilde{x},\tilde{p}))$, we can conclude that any value in $(L^{TT}(\tilde{x},\tilde{p}), U^{TT}(\tilde{x},\tilde{p}))$ can be rationalized by a model consistent both with the observed data and our assumptions, and thus $(L^{TT}(\tilde{x},\tilde{p}), U^{TT}(\tilde{x},\tilde{p}))$ are sharp bounds.

Fix some $(\tilde{x},\tilde{p}) \in \Omega_{X,P}$ and some $s \in (L^{TT}(\tilde{x},\tilde{p}), U^{TT}(\tilde{x},\tilde{p}))$. Take the case where $\mathcal{X}_0^L(\tilde{x}), \mathcal{X}_0^U(\tilde{x})$ are both nonempty and are disjoint. The proof for the case where $\mathcal{X}_0^L(\tilde{x})$ or $\mathcal{X}_0^U(\tilde{x})$ are empty follows from an analogous argument, and the case where $\mathcal{X}_0^L(\tilde{x}) \cap \mathcal{X}_0^U(\tilde{x}) \neq \emptyset$ is immediate. Using that $\mathcal{X}_0^L(\tilde{x}), \mathcal{X}_0^U(\tilde{x})$ are both nonempty and are disjoint, we have that

$$\mathcal{B}_0(\tilde{x},\tilde{p}) = \left[m_0(\tilde{x},\tilde{p},p^l) + \frac{1}{\tilde{p}} F_{\epsilon,U}(\nu_1(x_0^l(\tilde{x})), p^l), \quad m_0(\tilde{x},\tilde{p},p^l) + \frac{1}{\tilde{p}} F_{\epsilon,U}(\nu_1(x_0^u(\tilde{x})), p^l) \right] \tag{25}$$

where $x_0^l(\tilde{x})$, $x_0^u(\tilde{x})$ denote evaluation points such that[27]

$$\begin{aligned}
\Pr[D=1, Y=1|X=x_0^l(\tilde{x}), P=p^l] &= \sup_{x^* \in \mathcal{X}_0^L(\tilde{x})} \left\{ \Pr[D=1, Y=1|X=x^*, P=p^l] \right\} \\
\Pr[D=1, Y=1|X=x_0^u(\tilde{x}), P=p^l] &= \inf_{x^* \in \mathcal{X}_0^U(\tilde{x})} \left\{ \Pr[D=1, Y=1|X=x^*, P=p^l] \right\}.
\end{aligned}$$

Let $s^* = \tilde{p}[\Pr[Y=1|D=1, P=\tilde{p}] - m_0(\tilde{x},\tilde{p},p^l)] - s]$. Using equation (25), we have that $s \in (L^{TT}(\tilde{x},\tilde{p}), U^{TT}(\tilde{x},\tilde{p}))$ implies $s^* \in (F_{\epsilon,U}(\nu_1(x_0^l(\tilde{x})), p^l), F_{\epsilon,U}(\nu_1(x_0^u(\tilde{x})), p^l))$. Note that $\nu_1(x_0^u(\tilde{x})) > \nu_0(\tilde{x}) > \nu_1(x_0^l(\tilde{x}))$ with the strict inequalities following from our assumption that $\mathcal{X}_0^L(\tilde{x}), \mathcal{X}_0^U(\tilde{x})$ are disjoint. Further notice that given the definitions of $x_0^u(\tilde{x}), x_0^l(\tilde{x})$, we have $\nu_1(x) \notin (\nu_1(x_0^l(\tilde{x})), \nu_1(x_0^u(\tilde{x})))$ for any $x \in \Omega_X$. Construct the proposed $f^*_{\epsilon,U}$ as follows. Let $f^*_{\epsilon,U}(t_1,t_2) = f^*_{\epsilon|U}(t_1|t_2) f^*_U(t_2)$, where $f^*_U(t_2) = f_U(t_2) = \mathbf{1}[0 \leq t_2 \leq 1]$ and

$$f^*_{\epsilon|U}(t_1|t_2) = \begin{cases} f_{\epsilon|U}(t_1|t_2) & \text{if } t_1 \geq \nu_1(x_0^u(\tilde{x})) \text{ or } t_2 \geq p^l \text{ or } t_1 \leq \nu_1(x_0^l(\tilde{x})) \\ b(t_2) f_{\epsilon|U}(t_1|t_2) & \text{if } \nu_0(\tilde{x}) < t_1 < \nu_1(x_0^u(\tilde{x})) \text{ and } t_2 < p^l \\ a(t_2) f_{\epsilon|U}(t_1|t_2) & \text{if } \nu_1(x_0^l(\tilde{x})) < t_1 < \nu_0(\tilde{x}) \text{ and } t_2 < p^l \end{cases} \tag{26}$$

with

$$\begin{aligned}
a(t_2) &= \frac{\Pr[\nu_1(x_0^l(\tilde{x})) < \epsilon < \nu_1(x_0^u(\tilde{x})) | U=t_2]}{\Pr[\nu_1(x_0^l(\tilde{x})) < \epsilon < \nu_0(\tilde{x}) | U=t_2]} \frac{s^* - F_{\epsilon,U}(\nu_1(x_0^l(\tilde{x})), p^l)}{F_{\epsilon,U}(\nu_1(x_0^u(\tilde{x})), p^l) - F_{\epsilon,U}(\nu_1(x_0^l(\tilde{x})), p^l)} \\
b(t_2) &= \frac{\Pr[\nu_1(x_0^l(\tilde{x})) < \epsilon < \nu_1(x_0^u(\tilde{x})) | U=t_2] - a(t_2) \Pr[\nu_1(x_0^l(\tilde{x})) < \epsilon < \nu_0(\tilde{x}) | U=t_2]}{\Pr[\nu_0(\tilde{x}) < \epsilon < \nu_1(x_0^u(\tilde{x})) | U=t_2]}.
\end{aligned} \tag{27}$$

First consider whether $f^*_{\epsilon|U}$ integrates to one and is strictly positive on $\Re$. For $t_2 \geq p_l$, $f^*_{\epsilon|U}(\cdot|t_2) =$

[27] The existence of such evaluation points follows from our assumption (A-6).

$f_{\epsilon|U}(\cdot|t_2)$ and thus trivially $\int f^*_{\epsilon|U}(t_1|t_2)dt_1 = \int f_{\epsilon|U}(t_1|t_2)dt_1 = 1$. For $t_2 < p_l$,

$$\begin{aligned}
&\int_{-\infty}^{\infty} f^*_{\epsilon|U}(t_1|t_2)dt_1 \\
&= \int_{-\infty}^{\nu_1(x_0^l(\tilde{x}))} f_{\epsilon|U}(t_1|t_2)dt_1 + a(t_2)\int_{\nu_1(x_0^l(\tilde{x}))}^{\nu_0(\tilde{x})} f_{\epsilon|U}(t_1|t_2)dt_1 + b(t_2)\int_{\nu_0(\tilde{x})}^{\nu_1(x_0^u(\tilde{x}))} f_{\epsilon|U}(t_1|t_2)dt_1 \\
&\quad + \int_{\nu_1(x_0^u(\tilde{x}))}^{\infty} f_{\epsilon|U}(t_1|t_2)dt_1 \\
&= \Pr[\epsilon \le \nu_1(x_0^l(\tilde{x}))|U = t_2] + \Pr[\nu_1(x_0^l(\tilde{x})) < \epsilon \le \nu_1(x_0^u(\tilde{x}))|U = t_2] + \Pr[\epsilon > \nu_1(x_0^u(\tilde{x}))|U = t_2] \\
&= 1.
\end{aligned}$$

Since $f_{\epsilon|U}$ is strictly positive on $\Re$, we have that $f^*_{\epsilon|U}$ is strictly positive on $\Re$ if $a(t_2) > 0$ and $b(t_2) > 0$. Recall that $s^* \in (F_{\epsilon,U}(\nu_1(x_0^l(\tilde{x})), p^l), F_{\epsilon,U}(\nu_1(x_0^u(\tilde{x})), p^l))$. $s^* > F_{\epsilon,U}(\nu_1(x_0^l(\tilde{x})), p^l)$ implies $a(t_2) > 0$. $s^* < F_{\epsilon,U}(\nu_1(x_0^u(\tilde{x})), p^l)$ implies that

$$\frac{s^* - F_{\epsilon,U}(\nu_1(x_0^l(\tilde{x})), p^l)}{F_{\epsilon,U}(\nu_1(x_0^u(\tilde{x})), p^l) - F_{\epsilon,U}(\nu_1(x_0^l(\tilde{x})), p^l)} < 1$$

and thus

$$\begin{aligned}
&\Pr[\nu_1(x_0^l(\tilde{x})) < \epsilon < \nu_1(x_0^u(\tilde{x}))|U = t_2] - a(t_2)\Pr[\nu_1(x_0^l(\tilde{x})) < \epsilon < \nu_0(\tilde{x})|U = t_2] \\
&\quad = \Pr[\nu_1(x_0^l(\tilde{x})) < \epsilon < \nu_1(x_0^u(\tilde{x}))|U = t_2]\left(1 - \frac{s^* - F_{\epsilon,U}(\nu_1(x_0^l(\tilde{x})), p^l)}{F_{\epsilon,U}(\nu_1(x_0^u(\tilde{x})), p^l) - F_{\epsilon,U}(\nu_1(x_0^l(\tilde{x})), p^l)}\right) > 0
\end{aligned}$$

so that $b(t_2) > 0$. We have thus shown that $f^*_{\epsilon|U}$ is a proper density satisfying part (1) of the assertion.

Now consider part (2) of the assertion. $f^*_U = f_U$ implies that

$$\Pr[U^* \le p] = \int_0^p f^*_U(t)dt = \int_0^p f_U(t)dt = \Pr[U \le p] = \Pr[D = 1|P = p] \quad \forall p \in \Omega_P.$$

$f^*_U = f_U$ and $f^*_{\epsilon|U}(t_1|t_2) = f_{\epsilon|U}(t_1|t_2)$ for $t_2 \ge p^l$ imply that $f^*_{\epsilon,U}(t_1, t_2) = f_{\epsilon,U}(t_1, t_2)$ for all $t_2 \ge p^l$, and thus

$$\begin{aligned}
\Pr[\epsilon^* \le \nu_0(x)|X = x, U^* > p] &= \frac{1}{1-p}\int_p^1 \int_{-\infty}^{\nu_0(x)} f^*_{\epsilon,U}(t_1, t_2)dt_1dt_2 \\
&= \frac{1}{1-p}\int_p^1 \int_{-\infty}^{\nu_0(x)} f_{\epsilon,U}(t_1, t_2)dt_1dt_2 \\
&= \Pr[\epsilon \le \nu_0(x)|U > p] = \Pr[Y = 1|D = 0, X = x, P = p]
\end{aligned}$$

for all $(x, p) \in \Omega_{X,P}$.

Consider $\Pr[\epsilon^* \le \nu_1(x)|U^* \le p]$. By the definition of $x_0^l(\tilde{x})$ and $x_0^u(\tilde{x})$, we have that $\nu_1(x) \le \nu_1(x_0^l(\tilde{x}))$ or $\nu_1(x) \ge \nu_1(x_0^u(\tilde{x}))$ for any $x \in \Omega_X$. For x such that $\nu_1(x) \le \nu_1(x_0^l(\tilde{x}))$, and for any $p \in \Omega_P$,

$$\begin{aligned}
\Pr[\epsilon^* \le \nu_1(x)|U^* \le p] &= \frac{1}{p}\int_0^p \int_{-\infty}^{\nu_1(x)} f^*_{\epsilon,U}(t_1, t_2)dt_1dt_2 \\
&= \frac{1}{p}\int_0^p \int_{-\infty}^{\nu_1(x)} f_{\epsilon,U}(t_1, t_2)dt_1dt_2 \\
&= \Pr[\epsilon \le \nu_1(x)|U \le p] = \Pr[Y = 1 \mid D = 1, X = x, P = p].
\end{aligned}$$

For x such that $\nu_1(x) \geq \nu_1(x_0^u(\tilde{x}))$, and for any $p \in \Omega_P$,

$$
\begin{aligned}
&\Pr[\epsilon^* \leq \nu_1(x) | U^* \leq p] \\
&= \tfrac{1}{p} \int_0^p \int_{-\infty}^{\nu_1(x)} f^*_{\epsilon,U}(t_1,t_2) dt_1 dt_2 \\
&= \tfrac{1}{p} \Bigg\{ \int_{p^l}^p \int_{-\infty}^{\nu_1(x)} f_{\epsilon,U}(t_1,t_2) dt_1 dt_2 + \int_0^{p^l} \Bigg[\int_{-\infty}^{\nu_1(x_0^l(\tilde{x}))} f_{\epsilon|U}(t_1|t_2) dt_1 + a(t_2) \int_{\nu_1(x_0^l(\tilde{x}))}^{\nu_0(\tilde{x})} f_{\epsilon|U}(t_1|t_2) dt_1 \\
&\qquad + b(t_2) \int_{\nu_0(\tilde{x})}^{\nu_1(x_0^u(\tilde{x}))} f_{\epsilon|U}(t_1|t_2) dt_1 + \int_{\nu_1(x_0^u(\tilde{x}))}^{\nu_1(x)} f_{\epsilon|U}(t_1|t_2) dt_1 \Bigg] dt_2 \Bigg\} \\
&= \tfrac{1}{p} \Big\{ \Pr[\epsilon \leq \nu_1(x), p^l < U \leq p] + \Pr[\epsilon \leq \nu_1(x), U \leq p^l] \Big\} \\
&= \Pr[\epsilon \leq \nu_1(x) | U \leq p] = \Pr[Y = 1 \mid D = 1, X = x, P = p].
\end{aligned}
$$

We thus have that $\Pr[\epsilon^* \leq \nu_1(x) | U^* \leq p] = \Pr[Y = 1 \mid D = 1, X = x, P = p]$ for all $(x,p) \in \Omega_{X,P}$. We have thus established part (2) of the assertion. Consider part (3) of the assertion. We have already shown $\Pr[\epsilon^* \leq \nu_1(\tilde{x}) | U^* \leq \tilde{p}] = \Pr[\epsilon \leq \nu_1(\tilde{x}) | U \leq \tilde{p}]$ since $(\tilde{x}, \tilde{p}) \in \Omega_{X,P}$. Consider $\Pr[\epsilon^* \leq \nu_0(\tilde{x}) | U^* \leq \tilde{p}]$,

$$
\begin{aligned}
&\Pr[\epsilon^* \leq \nu_0(\tilde{x}) | U^* \leq \tilde{p}] \\
&= \tfrac{1}{\tilde{p}} \int_0^{\tilde{p}} \int_{-\infty}^{\nu_0(\tilde{x})} f^*_{\epsilon,U}(t_1,t_2) dt_1 dt_2 \\
&= \tfrac{1}{\tilde{p}} \Bigg\{ \int_0^{p^l} \Big(\int_{-\infty}^{\nu_1(x_0^l(\tilde{x}))} f^*_{\epsilon,U}(t_1,t_2) dt_1 + \int_{\nu_1(x_0^l(\tilde{x}))}^{\nu_0(\tilde{x})} f^*_{\epsilon,U}(t_1,t_2) dt_1 \Big) dt_2 + \int_{p^l}^{\tilde{p}} \int_{-\infty}^{\nu_0(\tilde{x})} f^*_{\epsilon,U}(t_1,t_2) dt_1 dt_2 \Bigg\} \\
&= \tfrac{1}{\tilde{p}} \Bigg\{ \int_0^{p^l} \Big(\int_{-\infty}^{\nu_1(x_0^l(\tilde{x}))} f_{\epsilon,U}(t_1,t_2) dt_1 + a(t_2) \int_{\nu_1(x_0^l(\tilde{x}))}^{\nu_0(\tilde{x})} f_{\epsilon,U}(t_1,t_2) dt_1 \Big) dt_2 + \int_{p^l}^{\tilde{p}} \int_{-\infty}^{\nu_0(\tilde{x})} f_{\epsilon,U}(t_1,t_2) dt_1 dt_2 \Bigg\} \\
&= \tfrac{1}{\tilde{p}} \Big\{ s^* + \tilde{p} m_0(\tilde{x}, \tilde{p}, p^l) \Big\} = \Pr[\epsilon \leq \nu_1(\tilde{x}) | U^* \leq \tilde{p}] - s
\end{aligned}
$$

and thus $\Pr[\epsilon^* \leq \nu_1(\tilde{x}) | U^* \leq \tilde{p}] - \Pr[\epsilon^* \leq \nu_0(\tilde{x}) | U^* \leq \tilde{p}] = s$.

□

Corollary 5.1. *The bounds on ATE and TT defined in Theorem 5.1 always identify whether these parameters are positive, zero, or negative.*

Proof. Consider the assertion for TT. An analogous argument proves the assertion for ATE. Suppose $E(Y_1 - Y_0 | D = 1, X = x, P = p) > 0$ so that $\nu_1(x) > \nu_0(x)$. Then, by Lemma 5.2, $H(x,x) > 0$ and thus $x \in \mathcal{X}_0^U(x)$ and $h(p, \tilde{p}, x) > 0$ for any $\tilde{p} < p$. Thus, fixing any arbitary $\tilde{p} < p$,

$$
\begin{aligned}
L^{TT}(x,p) > &\\
&\frac{1}{p} \left\{ h(p, \tilde{p}, x) + \Pr[D = 1, Y = 1 | X = x, P = \tilde{p}] - \Pr[D = 1, Y = 1 | X = x, P = \tilde{p}] \right\} \\
&= \frac{1}{p} \left\{ h(p, \tilde{p}, x) \right\} > 0.
\end{aligned}
$$

The symmetric argument shows that $E(Y_1 - Y_0 | D = 1, X = x, P = p) < 0$ implies $U^{TT}(x,p) < 0$, and $E(Y_1 - Y_0 | D = 1, X = x, P = p) = 0$ implies $L^{TT}(x,p) = U^{TT}(x,p) = 0$. □

Clearly, having X covariates allows us to narrow the bounds compared to the case considered in Section 4 without X covariates. The extent to which the X covariates are able to narrow the bounds depends on the extent to which X varies conditional on P. For example, suppose that X is degenerate conditional on P. Then one can easily show that the bounds of Theorem 5.1 collapse down to the same form as the bounds of Theorem 4.1. In contrast, consider the case where $\Omega_{X,P} = \Omega_X \times \Omega_P$, i.e., when the support of the distribution of (X, P) equals the products of the support of the distributions of X and P. In addition, suppose that $\mathcal{X}_j^L, \mathcal{X}_j^U$ are nonempty for $j = 0, 1$. Then, following the same type of argument used to simplify from equations (16) and (17) to (18) and (19), one can show

$$U^{TT}(x,p) - L^{TT}(x,p) = p^{-1}\Bigg(\inf_{\tilde{x}\in\mathcal{X}_0^U(x)}\Big\{\Pr[D=1,Y=1|X=\tilde{x},P=p^l]\Big\} - \sup_{\tilde{x}\in\mathcal{X}_0^L(x)}\Big\{\Pr[D=1,Y=1|X=\tilde{x},P=p^l]\Big\}\Bigg).$$

Notice that the width of the bounds collapse to zero if there exists any $\tilde{x}$ such that $H(\tilde{x}, x) = 0$, in which case $\tilde{x} \in \mathcal{X}_0^U(x) \bigcap \mathcal{X}_0^L(x)$. In other words, if there exists any $\tilde{x}$ such that $\nu_0(\tilde{x}) = \nu_1(x)$ (i.e., not receiving the treatment and having $X = \tilde{x}$ leads to the same value of the latent index as receiving treatment but having $X = x$), then the bounds provide point identification. This point identification result is a special case of our bounding analysis, and is essentially the same as the central identification result from Vytlacil and Yildiz (2004).

To further analyze the bounds on TT, let $x_0^l(x)$, $x_0^u(x)$ denote evaluation points such that $\Pr[D = 1, Y = 1|X = x_0^u(x), P = p^l] = \inf_{\tilde{x}\in\mathcal{X}_0^U(x)}\{\Pr[D = 1, Y = 1|X = \tilde{x}, P = p^l]\}$, and $\Pr[D = 1, Y = 1|X = x_0^l(x), P = p^l] = \sup_{\tilde{x}\in\mathcal{X}_0^L(x)}\{\Pr[D = 1, Y = 1|X = \tilde{x}, P = p^l]\}$. Then the width of the bounds can be rewritten as

$$U^{TT}(x,p) - L^{TT}(x,p) = p^{-1}\Pr[U \le p^l, \nu_1(x_0^l(x)) < \epsilon \le \nu_1(x_0^u(x))]$$

if $\nu_1(x_0^l(x))$ is strictly smaller than $\nu_1(x_0^u(x))$, and the width of the bounds equals zero if $\nu_1(x_0^l(x)) = \nu_1(x_0^u(x))$. Clearly, the closer $\nu_1(x_0^l(x))$ is to $\nu_1(x_0^u(x))$, the narrower the resulting bounds on TT. The more variation there is in X the smaller we expect the difference to be between $\nu_1(x_0^l(x))$ and $\nu_1(x_0^u(x))$.

Now consider ATE. Again following the same type of argument used to simplify from equations (16) and (17) to (18) and (19), one can show

$$\begin{aligned} U^{ATE}(x) - L^{ATE}(x) = {} & \inf_{\tilde{x}\in\mathcal{X}_1^U(x)}\{\Pr[D=0,Y=1|X=\tilde{x},P=p^u]\} - \sup_{\tilde{x}\in\mathcal{X}_1^L(x)}\{\Pr[D=0,Y=1|X=\tilde{x},P=p^u]\} \\ & + \inf_{\tilde{x}\in\mathcal{X}_0^U}\Big\{\Pr[D=1,Y=1|X=\tilde{x},P=p^l]\Big\} - \sup_{\tilde{x}\in\mathcal{X}_0^L}\Big\{\Pr[D=1,Y=1|X=\tilde{x},P=p^l]\Big\}. \end{aligned}$$

Notice that the width of the bounds collapse to zero if there exists a $\tilde{x}$ such that $H(\tilde{x}, x) = 0$ and a x^* such that $H(x, x^*) = 0$, in which case $\tilde{x} \in \mathcal{X}_0^U(x) \bigcap \mathcal{X}_0^L(x)$ and $x^* \in \mathcal{X}_1^U(x) \bigcap \mathcal{X}_1^L(x)$. In other words, if there exists any $\tilde{x}$, x^* such that $\nu_0(\tilde{x}) = \nu_1(x)$, $\nu_0(x) = \nu_1(x^*)$ (i.e., not receiving the treatment and having $X = \tilde{x}$ leads to the same value of the latent index as receiving treatment but having $X = x$, and not receiving the treatment and having $X = x$ leads to the same value of the latent index as receiving treatment but having $X = x^*$), then the bounds provide point identification on ATE. Again, following the same analysis as for TT, we expect the bounds on ATE to be narrower the greater the variation in X.

6 Comparison to Other Bounds

This paper is related to a large literature on the use of instrumental variables to bound treatment effects. Particularly relevant are the IV bounds of Manski (1990), Heckman and Vytlacil (2001), and Manski and Pepper (2000).[28] We now compare our assumptions and resulting bounds. First consider the Manski (1990) mean-IV bounds. Manski (1990) imposes a mean-independence assumption: $E(Y_1|X,Z) = E(Y_1 \mid X)$, and $E(Y_0|X,Z) = E(Y_0 \mid X)$. This assumption is strictly weaker than the assumptions imposed in this paper.[29] The mean independence assumption and the assumption that the outcomes are bounded imply that

$$B_M^L(x) \leq E(Y_1 - Y_0 \mid X = x) \leq B_M^U(x),$$

with[30]

$$B_M^L(x) = \sup_z \{\Pr[D=1, Y=1|X=x, Z=z]\} - \inf_z \{\Pr[D=0, Y=1|X=x, Z=z] + P(z)\},$$

$$B_M^U(x) = \inf_z \{\Pr[D=1, Y=1|X=x, Z=z] + (1 - P(z))\} - \sup_z \{\Pr[D=0, Y=1|X=x, Z=z]\}.$$

As discussed by Manski (1994), these bounds are sharp under the mean-independence condition. Note that these bounds neither impose nor exploit the full statistical independence assumptions considered in this paper, the structure of the threshold crossing model on the outcome equation, or the structure of the threshold crossing model on the treatment selection equations.

Now consider the analysis of Heckman and Vytlacil (2001). They strengthen the assumptions imposed by Manski (1990) by imposing statistical independence instead of mean independence, and imposing a threshold crossing model on the treatment equation. In particular, they assume that $D = 1[\vartheta(Z) - U \geq 0]$ and that Z is statistically independent of (Y_1, Y_0, U) conditional on X. Given these assumptions, they derive the following bounds on the average treatment effect:

$$B_{HV}^L(x) \leq E(Y_1 - Y_0 \mid X = x) \leq B_{HV}^U(x),$$

with

$$B_{HV}^U(x) = \Pr[D=1, Y=1|X=x, P=p^u(x)] + (1 - p^u(x)) - \Pr[D=0, Y=1|X=x, P=p^l(x)]$$

$$B_{HV}^L(x) = \Pr[D=1, Y=1|X=x, P=p^u(x)] - \Pr[D=0, Y=1|X=x, P=p^l(x)] - p^l(x).$$

[28] The bounds of Chesher (2003) do not apply to the problem of this paper with a binary endogenous regressor since his bounds are only relevant when the endogenous regressor takes at least three values. Other IV bounds not considered in this paper include the contaminated IV bounds of Hotz, Mullins, and Sanders (1997), the IV bounds of Balke and Pearl (1997) and Blundell, Gosling, Ichimura, and Meghir (2004), and the bounds on policy effects of Ichimura and Taber (1999). We do not attempt a review or survey of the entire bounding literature or even of the entire literature on bounds that exploits exclusion restrictions. Surveys of the bounding approach include Manski (1995, 2003). Heckman, LaLonde, and Smith (1999) includes an alternative survey of the bounding approach.

[29] In particular, note that our model and assumption (A-2) immediately implies Manski's mean independence assumption.

[30] Recall that we are leaving implicit that we are only evaluating the conditional expectations where the conditional expectations are well defined. Thus, e.g., the supremum and infimum in the following expressions are over z in the support of the distribution of Z conditional on $X = x$.

The width of the bounds is

$$B_S^U(x) - B_S^L(x) = ((1 - p^u(x)) + p^l(x)).$$

Trivially, $p^u(x) = 1$ and $p^l(x) = 0$ is necessary and sufficient for the bounds to collapse to point identification.

Heckman and Vytlacil (2001) analyze how these bounds compare to the Manski (1990) mean independence bounds, and analyze whether these bounds are sharp. They show that the selection model imposes restrictions on the observed data such that the Manski (1990) mean independence bounds collapse to the simpler Heckman and Vytlacil (2001) bounds. Furthermore, Heckman and Vytlacil (2001) establish that their bounds are sharp given their assumptions. Thus, somewhat surprisingly, imposing a threshold crossing model on the treatment equation does not narrow the bounds when compared to the case of imposing only the weaker assumption of mean independence, but does impose structure on the observed data such that the mean-independence bounds simplify substantially. By the Vytlacil (2002) equivalence result, the same conclusion holds for the Local Average Treatment Effect (LATE) assumptions of Imbens and Angrist (1994) – imposing the LATE assumptions does not narrow the bounds compared to only imposing the weaker assumption of mean independence, but does impose restrictions on the observed data that substantially simplifies the form of the bounds.[31] Note that the Heckman and Vytlacil (2001) bounds do not exploit a threshold crossing structure on the outcome equation.

In comparison, the analysis of this paper imposes and exploits more structure than either the Manski (1990) or Heckman and Vytlacil (2001) bounds. In return for this additional structure we obtain substantially narrower bounds. First, consider the case of no X covariates and the resulting comparison of their bounds with the bounds of Theorem 4.1. Imposing our assumptions, including that the first stage model for D is given by a threshold crossing model, we have that the Manski (1990) and Heckman and Vytlacil (2001) bounds coincide. By adding and subtracting terms, we can rewrite B_{HV}^L and B_{HV}^U as

$$B_{HV}^L = h(p^u, p^l) - \Pr[D = 0, Y = 1 | P = p^u] - \Pr[D = 1, Y = 0 | P = p^l]$$

$$B_{HV}^U = h(p^u, p^l) + \Pr[D = 1, Y = 1 | P = p^l] + \Pr[D = 0, Y = 0 | P = p^u],$$

where $h(p^u, p^l)$ was defined as $\Pr[Y = 1|P = p^u] - \Pr[Y = 1|P = p^l]$. First consider the case when $h(p^u, p^l) > 0$. Then the upper bound on ATE of Theorem 4.1 coincides with the Manski/Heckman-Vytlacil upper bound, while the lower bound of Theorem 4.1 is $h(p^u, p^l)$. Thus, if $h(p^u, p^l) > 0$, then imposing the threshold crossing structure on the outcome equation does not improve the upper bound but does increase the lower bound by the quantity $\Pr[D = 0, Y = 1|P = p^u] + \Pr[D = 1, Y = 0|P = p^l]$. The improvement in the lower bound will be a strict improvement except in the special case of point identification for Manski/Heckman-Vytlacil when $p^u = 1$ and $p^l = 0$, and in general can be expected to be a considerable improvement. Symmetrically, if $h(p^u, p^l) < 0$, then the lower bound on ATE of Theorem 4.1 coincides with the Manski/Heckman-Vytlacil upper bound, while the upper bound of Theorem 4.1 is $h(p^u, p^l)$. Thus, if $h(p^u, p^l) < 0$, then imposing the threshold crossing structure on the outcome equation does not improve the lower bound but does improve the upper

[31] This same essential result was shown previously by Balke and Pearl (1997) for the special case of a binary outcome variable and binary instrument. They show that imposing statistical independence generally results in more informative and more complex bounds than the Manski (1990) mean independence bounds. However, they show that under the Imbens and Angrist (1994) assumptions, the full independence bounds simplify to the Manski mean-independence bounds.

bound. Finally, if $h(p^u, p^l) = 0$, then the bounds of Theorem 4.1 collapse to point identification at zero while the Manski/Heckman-Vytlacil bounds will still have width $(1 - p^u) + p^l$. Notice only in the case where $p^u = 1$, $p^l = 0$ will the Manski/Heckman-Vytlacil bounds and the bounds of Theorem 4.1 coincide, and otherwise the bounds of Theorem 4.1 will offer a strict improvement over the Manski/Heckman-Vytlacil bounds. The cost of this improvement are the added assumptions required for this analysis, in particular imposing the threshold crossing structure on both D and on Y.

To illustrate the differences in the bounds, consider the following special case of our model:

$$Y = \mathbf{1}[\alpha D - \epsilon \geq 0]$$

$$D = \mathbf{1}[\delta Z - U \geq 0],$$

with $(\epsilon, U) \sim N(0, I)$, Z taking values in $\{-1, 1\}$, and $\delta > 0$. Figures 4 sets $\alpha = 1/4$ and plots ATE, the Manski/Heckman-Vytlacil bounds, and the bounds of Theorem 4.1 for δ in $(0, 2)$. The upper bounds from Theorem 4.1 coincide with the Manski/Heckman-Vytlacil upper bounds in this example (since $\alpha > 0$), while the lower bounds from Theorem 4.1 are substantially higher than the lower bounds from Manski/Heckman-Vytlacil. The width of the Manski/Heckman-Vytlacil bounds and the width of the bounds of Theorem 4.1 are both decreasing in δ, and both widths asymptote to zero as δ goes to infinity. The bounds of Theorem 4.1 provide an improvement over the bounds of Manski/Heckman-Vytlacil for any value of δ, decreasing the width of the bounds by almost half and providing the most substantial improvement for low values of δ.

Figure 5 sets $\delta = 1/4$ and plots ATE, the Manski/Heckman-Vytlacil bounds, and the bounds of Theorem 4.1 for $\alpha \in (-2, 2)$. The lower bounds of Theorem 4.1 coincide with the Manski/Heckman-Vytlacil lower bounds when $\alpha < 0$, while the upper bounds coincide when $\alpha > 0$. The width of the Manski/Heckman-Vytlacil bounds do not depend on α while the bounds of Theorem 4.1 are decreasing as α approaches zero with a discontinuity at the point $\alpha = 0$ (they provide point identification at $\alpha = 0$). The bounds of Theorem 4.1 cut the width of the Manski/Heckman-Vytlacil bounds by approximately half for any value of α (as long $\alpha \neq 0$), with the improvement most substantial for α close to zero. Notice that in this example, the bounds of Theorem 4.1 are always narrower than the bounds of Manski/Heckman-Vytlacil, with the reduction being most substantial when both α and δ are close to zero, i.e., when the treatment has only a small effect on the outcome variable and when the instrument has only a small effect on selection into the treatment.

Now consider the case with X covariates and the resulting comparison of the Manski and Heckman-Vytlacil bounds with the bounds of Theorem 5.1. Imposing our assumptions, including that the first stage model for D is given by a threshold crossing model, we again have that the Manski (1990) and Heckman and Vytlacil (2001) bounds coincide. To simplify the comparison, suppose that $\Omega_{X,P} = \Omega_X \times \Omega_P$, i.e., that the support of the distribution of (X, P) equals the product of the supports of the marginal distributions of X and P. In addition, suppose that $\mathcal{X}_j^L, \mathcal{X}_j^U$ are nonempty for $j = 0, 1$. Exploiting $\Omega_{X,P} = \Omega_X \times \Omega_P$ and that $\mathcal{X}_j^L, \mathcal{X}_j^U$ are assumed to be nonempty for $j = 0, 1$, the bounds on ATE of Theorem 5.1 become

$$\begin{aligned} L^{ATE}(x) = \Pr[D = 1, Y = 1 | X = x, P = p^u] + \sup_{\tilde{x} \in \mathcal{X}_1^L(x)} \{\Pr[D = 0, Y = 1 \mid X = \tilde{x}, P = p^u]\} \\ - \Pr[D = 0, Y = 1 | X = x, P = p^l] - \inf_{\tilde{x} \in \mathcal{X}_0^U} \left\{\Pr[D = 1, Y = 1 | X = \tilde{x}, P = p^l]\right\} \end{aligned}$$

Figure 4: Bounds, For Model with No X Covariates, $\alpha = 1/4$

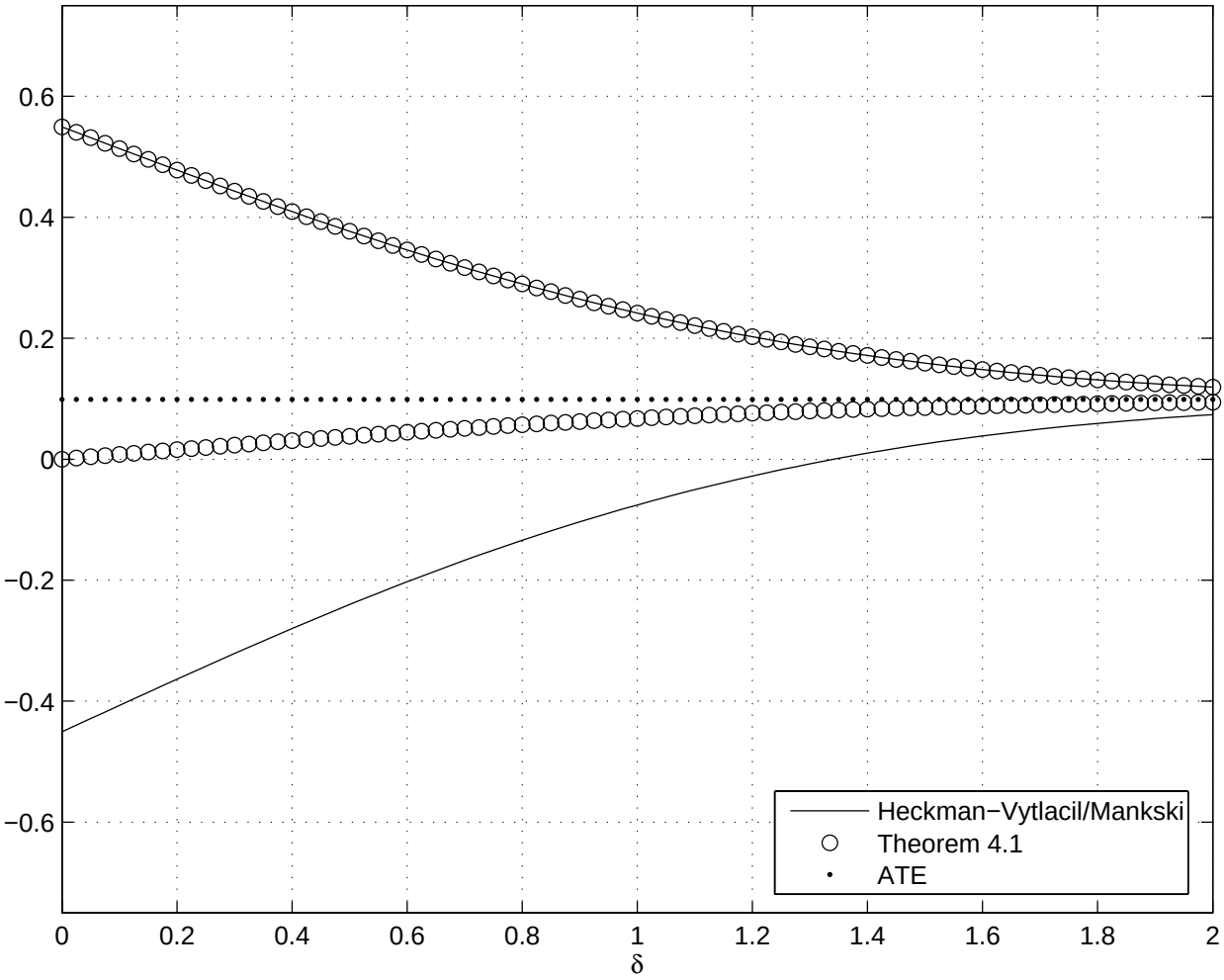
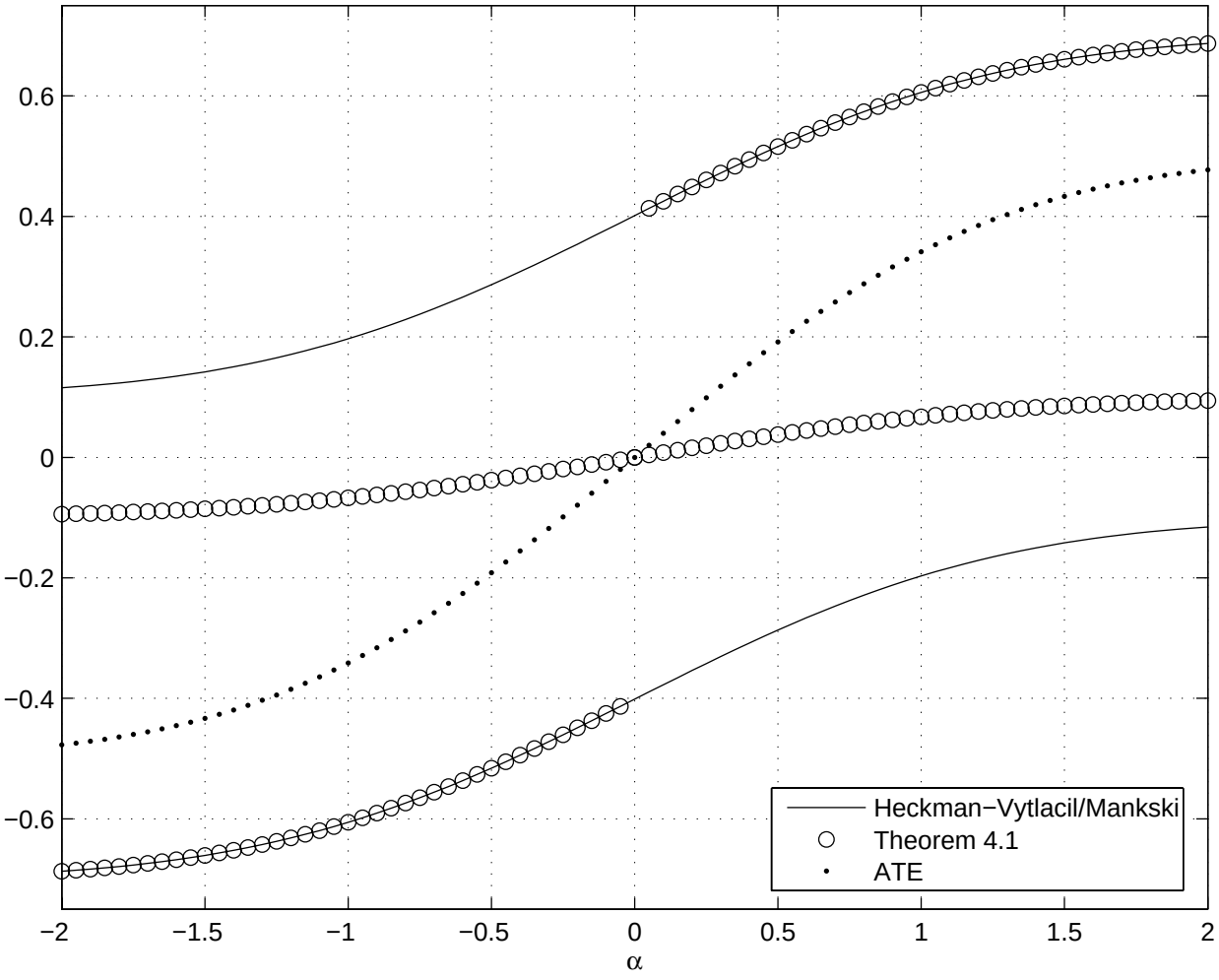


Figure 5: Bounds, For Model with No X Covariates, $\delta = 1/4$



$$U^{ATE}(x) = \Pr[D=1, Y=1|X=x, P=p^u] + \inf_{\tilde{x}\in\mathcal{X}_1^U(x)} \{\Pr[D=0, Y=1 \mid X=\tilde{x}, P=p^u]\}$$
$$- \Pr[D=0, Y=1|X=x, P=p^l] - \sup_{\tilde{x}\in\mathcal{X}_0^L} \left\{\Pr[D=1, Y=1|X=\tilde{x}, P=p^l]\right\}.$$

By adding and subtracting terms, one can show

$$L^{ATE}(x) = B_{HV}^L(x) + p^l$$
$$+ \sup_{\tilde{x}\in\mathcal{X}_1^L(x)} \{\Pr[D=0, Y=1 \mid X=\tilde{x}, P=p^u]\} - \inf_{\tilde{x}\in\mathcal{X}_0^U} \left\{\Pr[D=1, Y=1|X=\tilde{x}, P=p^l]\right\}$$

$$U^{ATE}(x) = B_{HV}^U(x) + (1-p^u)$$
$$+ \sup_{\tilde{x}\in\mathcal{X}_1^U(x)} \{\Pr[D=0, Y=1 \mid X=\tilde{x}, P=p^u]\} - \inf_{\tilde{x}\in\mathcal{X}_0^L} \left\{\Pr[D=1, Y=1|X=\tilde{x}, P=p^l]\right\},$$

and by further rearranging terms that

$$L^{ATE}(x) = B_{HV}^L(x)$$
$$+ \sup_{\tilde{x}\in\mathcal{X}_1^L(x)} \{\Pr[D=0, Y=1 \mid X=\tilde{x}, P=p^u]\} + \sup_{\tilde{x}\in\mathcal{X}_0^U} \left\{\Pr[D=1, Y=0|X=\tilde{x}, P=p^l]\right\}$$

$$U^{ATE}(x) = B_{HV}^U(x)$$
$$- \inf_{\tilde{x}\in\mathcal{X}_1^U(x)} \{\Pr[D=0, Y=0 \mid X=\tilde{x}, P=p^u]\} - \inf_{\tilde{x}\in\mathcal{X}_0^L} \left\{\Pr[D=1, Y=1|X=\tilde{x}, P=p^l]\right\}.$$

We thus see that in this case the bounds of Theorem 5.1 provide a strict improvement in both the lower and upper bounds on ATE compared to the bounds on Manski/Heckman-Vytlacil unless $p^u = 1$ and $p^l = 0$. The improvement in the bounds is expected to be substantial.

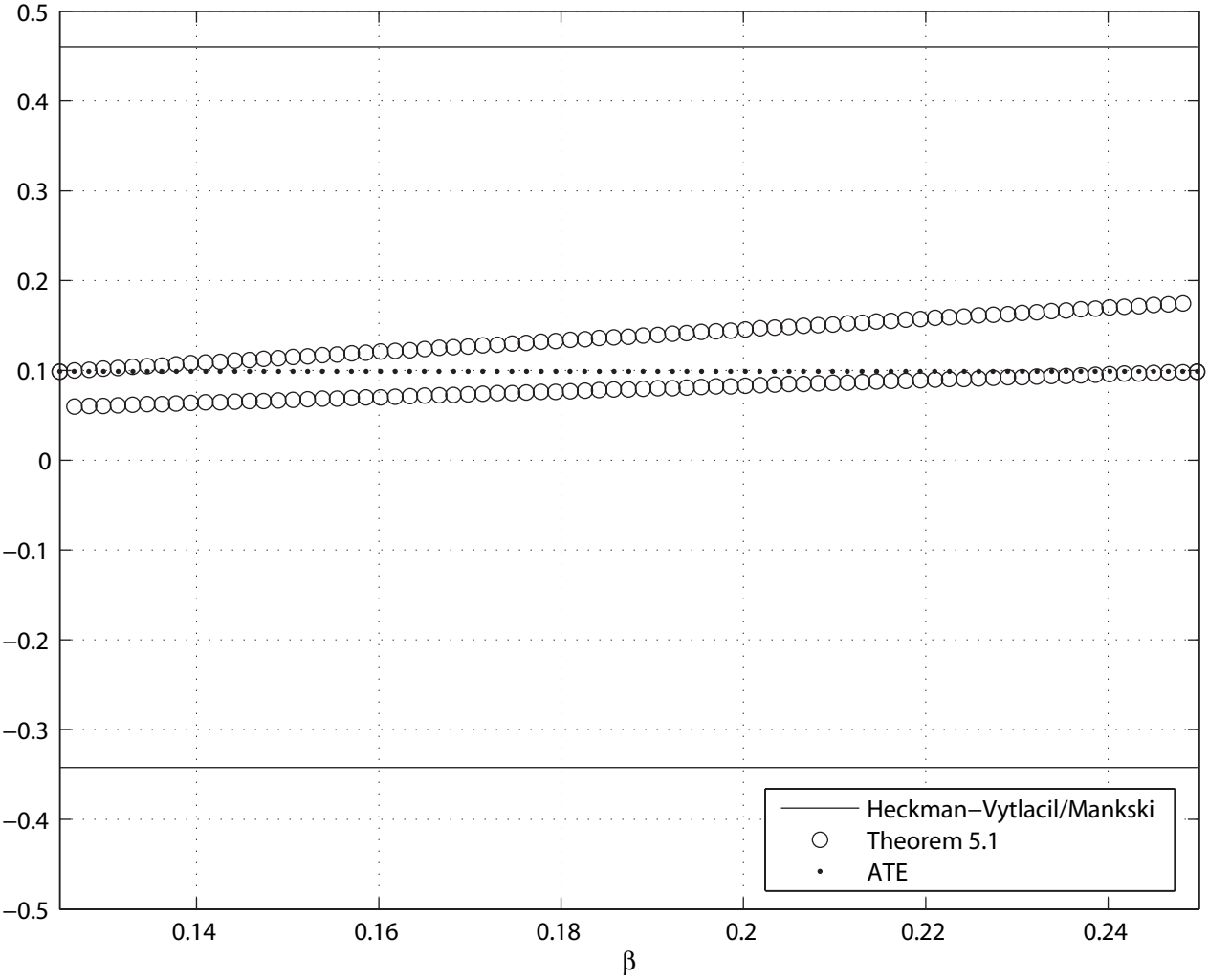
To illustrate the differences in the bounds with X covariates, consider the following special case of our model:

$$Y = \mathbf{1}[\beta X + \alpha D - \epsilon \geq 0]$$
$$D = \mathbf{1}[\delta Z - U \geq 0],$$

with $(\epsilon, U) \sim N(0, I)$, Z taking values in $\{-1, 1\}$, and X takes the values $-2, -1, 0, 1, 2$. Figures 6 plots ATE, the Manski/Heckman and Vytlacil bounds, and the bounds of Theorem 4.1. In this figure, we set $\delta = 1/4$, $\alpha = 1/4$, and plot over $\beta \in [1/8, 1/4]$. The width of the Manski/Heckman-Vytlacil bounds do not depend on β while the bounds of Theorem 4.1 do depend on β. In this example, the bounds of Theorem 4.1 provide a dramatic improvement over the bounds of Manski/Heckman-Vytlacil, and provide point identification when $\beta = 1/8$ or $1/4$. The tradeoff for this improvement in the bounds is the need to impose more structure, in particular, imposing the threshold crossing model on both D and Y.

Finally, consider the relationship of the bounding analysis of this paper with the bounding analysis of Manski and Pepper (2000). Manski and Pepper consider combining a weakened instrumental variable assumption ("monotone instrumental variables", MIV) with a "monotone treatment response" (MTR) assumption. The MTR assumption is that one knows a priori that $Y_1 \geq Y_0$ for

Figure 6: Bounds, For Model with X Covariates



all individuals or one knows a priori that $Y_0 \geq Y_1$ for all individuals. In comparison, our analysis identifies the sign of the average treatment effect from the data and does not impose it a priori. However, our analysis imposes the threshold crossing model on D and Y while no such assumption is imposed by Manski and Pepper. Consider the case of no X regressors and the Manski and Pepper bounds that would result from imposing $Y_1 \geq Y_0$ (MTR) and the Manski IV assumption that $\Pr[Y_1 = 1|Z] = \Pr[Y_1 = 1]$, $\Pr[Y_0 = 1|Z] = \Pr[Y_0 = 1]$. Modifying Proposition 2 of Manski and Pepper, the MTR assumption and Manski-IV assumption jointly imply

$$\sup_z\{\Pr[Y = 1|Z = z]\} \leq E(Y_1) \leq \inf_z\{\Pr[D = 1, Y = 1|Z = z] + \Pr[D = 0|Z = z]\}$$

$$\sup_z\{\Pr[D = 0, Y = 1|Z = z]\} \leq E(Y_0) \leq \inf_z\{\Pr[Y = 1|Z = z]\}.$$

Then following the same type of argument used to simplify from equations (16) and (17) to (18) and (19), given our assumptions including the threshold crossing structure on D, these bounds simplify to

$$\Pr[Y = 1|P = p^u]\} \leq E(Y_1) \leq \Pr[D = 1, Y = 1|P = p^u] + (1 - p^u)$$

$$\Pr[D = 0, Y = 1|P = p^l] \leq E(Y_0) \leq \Pr[Y = 1|P = p^l]\}.$$

Combining the bounds on $E(Y_1)$ and $E(Y_0)$ to obtain bounds on $E(Y_1 - Y_0)$, and rearranging terms, results in the same bounds as in Theorem 4.1 for the case of no X regressors and $H > 0$ (i.e., for the case when $Y_1 \geq Y_0$). Thus, if there are no X regressors and our assumptions hold, and the treatment effect is positive, then our bounds coincide with the Manski and Pepper bounds that result from imposing a priori a positive effect and the Manski IV assumption. Likewise, one can show that if there are no X regressors, the treatment effect is negative, and our assumptions hold, then our bounds of Theorem 4.1 coincide with the Manski and Pepper bounds that impose a priori a negative effect and impose the Manski IV assumption. Thus, in the case of no X covariates, there is a tight link between the Manski and Pepper bounds and the bounds of this paper, with the tradeoff that the Manski and Pepper bounds require that one knows a priori the sign of the treatment effect but does not impose the threshold crossing structure imposed in this paper. This discussion, however, has taken the case of no X regressors. With X regressors, the width of our bounds can shrink substantially while the Manski and Pepper bounds are unaffected by the presence of X regressors. Thus, the link of our bounds with the bounds of Manski and Pepper breaks down in the presence of X regressors.

7 Confidence Sets

We now turn to the construction of confidence sets for the bounds on $\Delta^{TT}(x, p)$ and $\Delta^{ATE}(x)$ described in Sections 4 and 5. We focus on the construction of confidence sets instead of consistent estimation of the bounds because of difficulties caused by the discontinuity in the form of the bounds at $H = 0$. It is possible to estimate consistently the bounds by equating all values of $H_n \in (-\epsilon_n, \epsilon_n)$ with zero for $\epsilon_n \searrow 0$ at a rate slower than $1/\sqrt{n}$. However, consistency places no restrictions on the level of ϵ_n and so estimation of the bounds becomes quite arbitrary. The confidence set approach we describe below circumvents this difficulty altogether.

Specifically, in this section we will describe a construction of random sets $\mathcal{C}_n^{TT}$ and $\mathcal{C}_n^{ATE}$ that will asymptotically contain each point in the sets described in Theorems 4.1 and 5.1, respectively, with probability at least $1 - \alpha$ for a researcher-specified $\alpha \in (0, 1)$. We will do this in the special

case in which both X and Z are discrete random variables. Concretely, our analysis will assume the following structure on the data generation process:

(B-1) The observed data $\{Y_i, D_i, Z_i, X_i\}_{1 \leq i \leq n}$ are i.i.d;

(B-2) $\Omega_{X,Z} = \{(x_1, z_1), \ldots, (x_L, z_L)\}$; and

(B-3) $z \neq z' \Rightarrow \Pr[D = 1|Z = z] \neq \Pr[D = 1|Z = z']$.

Assumtion (B-2) is not essential, but makes the analysis considerably simpler by avoiding the need to resort to more sophisticated smoothing-based estimators of certain objects. Given Assumption (B-2), Assumption (B-3) is not especially restrictive, but makes the exposition of our results much easier. It can be relaxed at the expense of somewhat more notationally involved arguments below.

As before, we will first analyze the situation in which there are no X covariates so as not to obscure the main ideas behind our construction. We will then generalize our results to allow for X covariates.

7.1 Analysis With No X Covariates

In order to make the notation less cumbersome, we will use the following shorthand for some of the terms that appear in the statement of Theorem 4.1:

$$\begin{aligned}
A &= \frac{1}{p} h(p, p^l) \\
B^+ &= \frac{1}{p}(h(p, p^l) + \Pr[D = 1, Y = 1|P = p^l]) \\
B^- &= \frac{1}{p}(h(p, p^l) - \Pr[D = 1, Y = 0|P = p^l]) \\
C &= h(p^u, p^l) \\
D^+ &= h(p^u, p^l) + \Pr[D = 1, Y = 1|P = p^l] + \Pr[D = 0, Y = 0|P = p^u] \\
D^- &= h(p^u, p^l) - \Pr[D = 1, Y = 0|P = p^l] - \Pr[D = 0, Y = 1|P = p^u] \ ,
\end{aligned}$$

where we have suppressed the dependence of both A, B^+, and B^- on p. Define the function

$$P(z) = \Pr[D = 1|Z = z] \tag{28}$$

and let z^l satisfy $P(z^l) = p^l$, z^u satisfy $P(z^u) = p^u$, and z satisfy $P(z) = p$. Thus, with z^l, z^u, and z so defined, we have, as a result of index sufficiency, that

$$\begin{aligned}
A &= \frac{1}{P(z)} h^*(z, z^l) \\
B^+ &= \frac{1}{P(z)}(h^*(z, z^l) + \Pr[D = 1, Y = 1|Z = z^l]) \\
B^- &= \frac{1}{P(z)}(h^*(z, z^l) - \Pr[D = 1, Y = 0|Z = z^l]) \\
C &= h^*(z^u, z^l) \\
D^+ &= h^*(z^u, z^l) + \Pr[D = 1, Y = 1|Z = z^l] + \Pr[D = 0, Y = 0|Z = z^u] \\
D^- &= h^*(z^u, z^l) - \Pr[D = 1, Y = 0|Z = z^l] - \Pr[D = 0, Y = 1|P = z^u] \ ,
\end{aligned}$$

where we have defined

$$h^*(z_0, z_1) = \Pr[Y = 1|Z = z_0] - \Pr[Y = 1|Z = z_1] \ .$$

For the purposes of constructing confidence sets, we will therefore think of the parameter Δ^{TT} and the bounds $\mathcal{B}^{TT}$ for it as functions of z rather than p.

It is natural to define estimators A_n, B_n^+, B_n^-, C_n, D_n^+, and D_n^- of their population counterparts by simply replacing conditional population means with conditional sample means. Consistency of these estimators follows from assumption (B-1) using conventional arguments, which we omit here. As an example, we have that

$$A_n = \frac{1}{\hat{P}(z)} \hat{h}^*(z, \hat{z}^l) \ ,$$

where

$$\begin{aligned} \hat{P}(z) &= \frac{1}{|\{i : Z_i = z\}|} \sum_{i:Z_i=z} D_i \\ \hat{h}^*(z, z^l) &= \frac{1}{|\{i : Z_i = z\}|} \sum_{i:Z_i=z} Y_i - \frac{1}{|\{i : Z_i = \hat{z}^l\}|} \sum_{i:Z_i=\hat{z}^l} Y_i \end{aligned}$$

and $\hat{z}^l$ solves $\min_z \hat{P}(z)$. Note that as a result of assumptions (B-2) and (B-3), we have that $\hat{z}^l = z^l$ with arbitrarily high probability for all sufficiently large n. Thus, asymptotically we need not worry about the estimation of z^l. A similar remark holds for z^u.

It follows from assumption (B-3) that the mapping $P(z) = z$ is invertible. Therefore, using index sufficiency again, we also have that a consistent estimator of

$$H = \int_0^1 \int_0^{p_0} h(p_0, p_1) dF_P(p_1) dF_P(p_0)$$

is given by

$$H_n = \frac{1}{|\Omega_Z|^2} \sum_{(z_0,z_1):\hat{P}(z_0)<\hat{P}(z_1)} h^*(z_0, z_1) \ .$$

Our construction of confidence intervals will rely on the fact that for each of the terms described above, it is possible to construct asymptotically valid confidence regions. In this case, since each of the estimators has the form of a sum of a fixed number of sample means, it is easy to show, for example, that

$$\sqrt{n}(A_n - A) \xrightarrow{d} N(0, \sigma_A^2) \ . \tag{29}$$

Analogous statements hold for B_n^+, B_n^-, C_n, D_n^+, D_n^-, and H_n. Thus, it is possible to construct asymptotically valid confidence regions in a number of different ways. For example, if one denotes by $\hat{\sigma}_A^2$ a consistent estimate of the asymptotic variance of (29) and by $q_{1-\alpha}$ the $1-\alpha$ quantile of a standard normal distribution, it follows that

$$\Pr[A > A_n - \frac{\hat{\sigma}_A q_{1-\alpha}}{\sqrt{n}}] \to 1 - \alpha \ .$$

We are now prepared to describe our construction of the confidence sets $\mathcal{C}_n^{TT}$ and $\mathcal{C}_n^{ATE}$ in the case in which there are no X covariates.

Algorithm 7.1

1. Construct a two-sided $1-\alpha$ confidence interval for H as

$$I_n = [H_n - \frac{\hat{\sigma}_H q_{1-\alpha/2}}{\sqrt{n}}, H_n + \frac{\hat{\sigma}_H q_{1-\alpha/2}}{\sqrt{n}}] \ .$$

2. If

$$H_n - \frac{\hat{\sigma}_H q_{1-\alpha/2}}{\sqrt{n}} > 0 \ ,$$

then let

$$\begin{aligned} \mathcal{C}_n^{TT} &= [A_n - \frac{\hat{\sigma}_A q_{1-\alpha}}{\sqrt{n}}, B_n^+ + \frac{\hat{\sigma}_{B^+} q_{1-\alpha}}{\sqrt{n}}] \\ \mathcal{C}_n^{ATE} &= [C_n - \frac{\hat{\sigma}_C q_{1-\alpha}}{\sqrt{n}}, D_n^+ + \frac{\hat{\sigma}_{D^+} q_{1-\alpha}}{\sqrt{n}}] \ . \end{aligned}$$

3. If

$$H_n + \frac{\hat{\sigma}_H q_{1-\alpha/2}}{\sqrt{n}} < 0 \ ,$$

then let

$$\begin{aligned} \mathcal{C}_n^{TT} &= [B_n^- - \frac{\hat{\sigma}_{B^-} q_{1-\alpha}}{\sqrt{n}}, A_n + \frac{\hat{\sigma}_A q_{1-\alpha}}{\sqrt{n}}] \\ \mathcal{C}_n^{ATE} &= [D_n^- - \frac{\hat{\sigma}_{D^-} q_{1-\alpha}}{\sqrt{n}}, C_n + \frac{\hat{\sigma}_C q_{1-\alpha}}{\sqrt{n}}] \ . \end{aligned}$$

4. If $0 \in I_n$, then let

$$\begin{aligned} \mathcal{C}_n^{TT} &= [B_n^- - \frac{\hat{\sigma}_{B^-} q_{1-\alpha}}{\sqrt{n}}, B_n^+ + \frac{\hat{\sigma}_{B^+} q_{1-\alpha}}{\sqrt{n}}] \\ \mathcal{C}_n^{ATE} &= [D_n^- - \frac{\hat{\sigma}_{D^-} q_{1-\alpha}}{\sqrt{n}}, D_n^+ + \frac{\hat{\sigma}_{D^+} q_{1-\alpha}}{\sqrt{n}}] \ . \end{aligned}$$

We now show that the confidence sets constructed in this way satisfy the desired coverage property.

Theorem 7.1. *The sets $\mathcal{C}_n^{TT}$ and $\mathcal{C}_n^{ATE}$ constructed according to Algorithm 7.1 satisfy for each $\theta^{TT} \in \mathcal{B}^{TT}(z)$ and each $\theta^{ATE} \in \mathcal{B}^{ATE}$*

$$\begin{aligned} \liminf \Pr[\theta^{TT} \in \mathcal{C}_n^{TT}] &\geq 1-\alpha \\ \liminf \Pr[\theta^{ATE} \in \mathcal{C}_n^{ATE}] &\geq 1-\alpha \ . \end{aligned}$$

Proof. We describe the proof for the bounds on the TT parameter in detail. The argument for the ATE parameter is entirely analogous.

First consider the case in which $H > 0$, so $\Delta^{TT}(p) = [A, B^+]$. Then with probability approaching 1, we have that

$$H_n - \frac{\hat{\sigma}_H q_{1-\alpha/2}}{\sqrt{n}} > 0 \ .$$

Thus, with probability approaching 1,

$$\mathcal{C}_n^{TT} = [A_n - \frac{\hat{\sigma}_A q_{1-\alpha}}{\sqrt{n}}, B_n^+ + \frac{\hat{\sigma}_{B^+} q_{1-\alpha}}{\sqrt{n}}] \ .$$

Hence, for each $\theta^{TT} \in [A, B^+]$, we have that

$$\liminf \Pr[\theta^{TT} \in \mathcal{C}_n^{TT}] \geq 1 - \alpha \ ,$$

as desired. The proof for the case in which $H < 0$ is symmetric. Now consider the case in which $H = 0$. Then, with probability at least $1 - \alpha$ asymptotically, $0 \in I_n$. Therefore, with probability at least $1 - \alpha$ asymptotically, we have that

$$\mathcal{C}_n^{TT} = [B_n^- - \frac{\hat{\sigma}_{B^-} q_{1-\alpha}}{\sqrt{n}}, B_n^+ + \frac{\hat{\sigma}_{B^+} q_{1-\alpha}}{\sqrt{n}}] \ .$$

Since $B^- < 0 < B^+$, we have that $B_n^- < 0 < B_n^+$ with probability approaching 1. It follows that when $H = 0$, we have that $0 \in \mathcal{C}_n^{TT}$ with probability at least $1 - \alpha$ asymptotically, as desired. □

7.2 Analysis With X Covariates

We begin as before by noting that as a result of index sufficiency we have that $L^{TT}(x, p) = L^{TT*}(x, z)$ and $U^{TT}(x, p) = U^{TT*}(x, z)$ where

$$L^{TT*}(x,z) = \frac{1}{P(z)} \sup_{\tilde{z}} \Big\{ h(z, \tilde{z}, x) + \Pr[D = 1, Y = 1 | X = x, Z = \tilde{z}] \\ - P(\tilde{z}) \inf_{\tilde{x} \in \mathcal{X}_0^U(x)} \{\Pr[Y = 1 \mid D = 1, X = \tilde{x}, Z = \tilde{z}]\} \Big\}$$

$$U^{TT*}(x,z) = \frac{1}{P(z)} \inf_{\tilde{z}} \Big\{ h(z, \tilde{z}, x) + \Pr[D = 1, Y = 1 | X = x, Z = \tilde{z}] \\ - P(\tilde{z}) \sup_{\tilde{x} \in \mathcal{X}_0^L(x)} \{\Pr[Y = 1 \mid D = 1, X = \tilde{x}, Z = \tilde{z}]\} \Big\}$$

where $h^*(z_0, z_1, x) = \Pr[Y = 1 \mid X = x, Z = z_0] - \Pr[Y = 1 \mid X = x, Z = z_1]$, for any z such that $P(z) = p$. By analogy with the case in which there were no X covariates, for the purposes of constructing confidence sets we will think of Δ^{TT} as a function of x and z rather than x and p and thus define the bounds for it in terms of L^{TT*} and U^{TT*} rather than L^{TT} and U^{TT}.

Similarly, we have for any z such that $P(z) = p$ that $L^{ATE}(x) = L^{ATE*}(x)$ and $U^{ATE}(x) = U^{ATE*}(x)$ where

$$L^{ATE*}(x) = \\ \sup_z \Big\{ \Pr[D = 1, Y = 1 | X = x, Z = z] + (1 - P(z)) \sup_{\tilde{x} \in \mathcal{X}_1^L(x)} \{\Pr[Y = 1 \mid D = 0, X = \tilde{x}, Z = z]\} \Big\} \\ - \inf_{\tilde{z}} \Big\{ \Pr[D = 0, Y = 1 | X = x, Z = \tilde{z}] + P(\tilde{z}) \inf_{\tilde{x} \in \mathcal{X}_0^U(x)} \{\Pr[Y = 1 | D = 1, X = \tilde{x}, Z = \tilde{z}]\} \Big\}$$

$$U^{ATE*}(x) =$$
$$\inf_z \left\{ \Pr[D=1,Y=1|X=x,Z=z] + (1-P(z)) \inf_{\tilde{x}\in\mathcal{X}_1^U(x)} \{\Pr[Y=1 \mid D=0, X=\tilde{x}, Z=z]\} \right\}$$
$$- \sup_{\tilde{z}} \left\{ \Pr[D=0,Y=1|X=x,Z=\tilde{z}] + P(\tilde{z}) \sup_{\tilde{x}\in\mathcal{X}_0^L(x)} \{\Pr[Y=1|D=1,X=\tilde{x},Z=\tilde{z}]\} \right\}.$$

Our analysis for the parameter Δ^{ATE} will hereafter be based on L^{ATE*} and U^{ATE*} rather than L^{ATE} and U^{ATE}.

Note that the four quantities L^{TT*}, U^{TT*}, L^{ATE*}, and U^{ATE*} depend on the sets $\mathcal{X}_0^L(x)$, $\mathcal{X}_0^U(x)$, $\mathcal{X}_1^L(x)$, and $\mathcal{X}_1^U(x)$. It will be useful for us to make this dependence explicit by writing $L^{TT*}(x,z,\mathcal{X}_0^L(x))$, $U^{TT*}(x,z,\mathcal{X}_0^U(x))$, $L^{ATE*}(x,z,\mathcal{X}_1^L(x),\mathcal{X}_0^U(x))$, and $U^{ATE*}(x,z,\mathcal{X}_0^L(x),\mathcal{X}_1^U(x))$ and thereby think of these quantities as functions not only of x and z, but also of the underlying sets $\mathcal{X}_0^L(x)$, $\mathcal{X}_0^U(x)$, $\mathcal{X}_1^L(x)$, and $\mathcal{X}_1^U(x)$. In order to avoid later confusion, let us write $\mathcal{A}_0^L(x)$, $\mathcal{A}_0^U(x)$, $\mathcal{A}_1^L(x)$ and $\mathcal{A}_1^U(x)$ for arbitrary such sets.

Because of our assumption (B-2), for fixed values of x, z, $\mathcal{A}_0^L(x)$, $\mathcal{A}_0^U(x)$, $\mathcal{A}_1^L(x)$ and $\mathcal{A}_1^U(x)$, it is straightforward to construct consistent estimates of the quantities L^{TT*}, U^{TT*}, L^{ATE*}, and U^{ATE*} by simply replacing the conditional population means in the above expressions with their sample counterparts. Let us denote the estimators obtained in this way as L_n^{TT*}, U_n^{TT*}, L_n^{ATE*}, and U_n^{ATE*}.

As before, assumption (B-3) and index sufficiency together enable us to construct a consistent estimator of the quantity

$$H(x_0,x_1) = \int_0^1 \int_0^{p_0} [h_1(p_0,p_1,x_1) - h_0(p_0,p_1,x_0)]\, \mathbf{1}[(x_i,p_j)\in\Omega_{X,P},\ i,j=0,1] dF_P(p_1) dF_P(p_0)$$

as

$$H_n(x_0,x_1) =$$
$$\frac{1}{|\Omega_Z|^2} \sum_{(z_0,z_1)\ :\ \hat{P}(z_1)<\hat{P}(z_0)} [\hat{h}_1^*(z_0,z_1,x_1) - \hat{h}_0^*(z_0,z_1,x_0)] \mathbf{1}[(x_i,z_j)\in\Omega_{X,Z}, i,j\in\{0,1\}] \ ,$$

where $\hat{h}_0^*(z_0,z_1,x_0)$ and $\hat{h}_1^*(z_0,z_1,x_1)$ are, respectively, the consistent estimates of the quantities

$$\begin{aligned} h_0^*(z_0,z_1,x_0) &= \Pr[D=0,Y=1|X=x_0,Z=z_1] - \Pr[D=0,Y=1|X=x_0,Z=z_0] \\ h_1^*(z_0,z_1,x_1) &= \Pr[D=1,Y=1|X=x_1,Z=z_0] - \Pr[D=1,Y=1|X=x_1,Z=z_1] \end{aligned}$$

formed by replacing conditional population means with their sample counterparts.

We will need in our construction of confidence sets in the with X case, for fixed values of x, z, $\mathcal{A}_0^L(x)$, $\mathcal{A}_0^U(x)$, $\mathcal{A}_1^L(x)$ and $\mathcal{A}_1^U(x)$, asymptotically valid confidence intervals for each of the quantities L^{TT*}, U^{TT*}, L^{ATE*}, and U^{ATE*}. Note that the estimator L_n^{TT*} of L^{TT*} described above can easily be shown to satisfy under our assumptions

$$\sqrt{n}(L_n^{TT*} - L^{TT*}) \xrightarrow{d} \mathcal{N} \ ,$$

where $\mathcal{N}$ is a continuous transformation of a multivariate normal random variable. We omit the details, which are completely straightforward, here. Analogous statements hold for the estimators

U_n^{TT*}, L_n^{ATE*}, and U_n^{ATE*}. Thus, using subsampling, for example, it is possible to construct asymptotically valid confidence regions for each of the quantities L^{TT*}, U^{TT*}, L^{ATE*}, and U^{ATE*}.[32] Denote by $\mathcal{L}_n^{TT}(1-\alpha)$ the lower bound of a one-sided $1-\alpha$ confidence interval for L^{TT*} and by $\mathcal{U}_n^{TT}(1-\alpha)$ the upper bound of a one-sided $1-\alpha$ confidence interval for U^{TT*}. Define $\mathcal{L}_n^{ATE}(1-\alpha)$ and $\mathcal{U}_n^{ATE}(1-\alpha)$ analogously. Note that we have suppressed the dependence of these quantities on x, z, $\mathcal{A}_0^L(x)$, $\mathcal{A}_0^U(x)$, $\mathcal{A}_1^L(x)$ and $\mathcal{A}_1^U(x)$.

We will also require an asymptotic confidence band for $H(x_0, x_1)$, both when viewed as a function of x_0 for fixed x_1 and when viewed as a function of x_1 for fixed x_0. To see how this might be achieved, first fix x_0 and consider the random variable given by

$$\sqrt{n}\sup_{x_1}(H_n(x_0,x_1) - H(x_0,x_1)) \xrightarrow{d} \mathcal{N} \ ,$$

where $\mathcal{N}$ is a continuous transformation of a multivariate normal random variable (distinct from the one used in the preceding paragraph). Again, using subsampling it is therefore possible to estimate the $1-\alpha$ quantile of this limiting distribution. Denote this estimate by $\epsilon_{0n}(1-\alpha)$. Symmetrically, for fixed x_1, we will define $\epsilon_{1n}(1-\alpha)$ to be the subsampling estimate of the $1-\alpha$ quantile of the limiting distribution of $\sqrt{n}\sup_{x_0}(H_n(x_0,x_1) - H(x_0,x_1))$. Note that $\epsilon_{0n}(1-\alpha)$ depends on the x_0 evaluation point and $\epsilon_{0n}(1-\alpha)$ depends on the x_1 evaluation point, but we have suppressed this dependence.

Using this notation, we may now describe our construction of confidence sets for the case in which there are X covariates.

Algorithm 7.2

1. Construct

$$\begin{aligned}
\mathcal{A}_0^L(x) &= \{x' | \hat{H}(x,x') < -\frac{\epsilon_{0n}(1-\alpha)}{\sqrt{n}}\} \\
\mathcal{A}_0^U(x) &= \{x' | \hat{H}(x,x') > \frac{\epsilon_{0n}(1-\alpha)}{\sqrt{n}}\} \\
\mathcal{A}_1^L(x) &= \{x' | \hat{H}(x',x) > \frac{\epsilon_{1n}(1-\alpha)}{\sqrt{n}}\} \\
\mathcal{A}_1^U(x) &= \{x' | \hat{H}(x',x) < -\frac{\epsilon_{1n}(1-\alpha)}{\sqrt{n}}\} \ .
\end{aligned}$$

2. Set

$$\begin{aligned}
\mathcal{C}_n^{TT} &= [\mathcal{L}_n^{TT}(1-\alpha), \mathcal{U}_n^{TT}(1-\alpha)] \\
\mathcal{C}_n^{ATE} &= [\mathcal{L}_n^{ATE}(1-\alpha), \mathcal{U}_n^{ATE}(1-\alpha)] \ .
\end{aligned}$$

We now show that the confidence sets constructed in this way have the desired coverage property.

Theorem 7.2. *The confidence sets $\mathcal{C}_n^{TT}$ and $\mathcal{C}_n^{ATE}$ defined in Algorithm 7.2 satisfy for all $\theta^{TT} \in [L^{TT*}(x,z,\mathcal{X}_0^L(x)), U^{TT*}(x,z,\mathcal{X}_0^U(x))]$ and $\theta^{ATE} \in [L^{ATE*}(x,z,\mathcal{X}_1^L,\mathcal{X}_0^U), U^{ATE*}(x,z,\mathcal{X}_0^L,\mathcal{X}_1^U)]$, we have that*

$$\begin{aligned}
\liminf \Pr[\theta^{TT} \in \mathcal{C}^{TT}] &\geq 1-\alpha \\
\liminf \Pr[\theta^{ATE} \in \mathcal{C}^{ATE}] &\geq 1-\alpha \ .
\end{aligned}$$

[32] See Politis, Romano, and Wolf (1999), Theorem 2.2.1.

Proof. We describe in detail the proof for the bounds on the TT parameter. The argument for ATE is entirely analogous.

First consider the case in which $\mathcal{X}_0^L(x) \cap \mathcal{X}_0^U(x) = \emptyset$. Then, for all x', either $H(x,x') > 0$ or $H(x,x') < 0$. Therefore, with arbitrarily high probability for large enough n, we have that $\mathcal{A}_0^L(x) = \mathcal{X}_0^L(x)$ and $\mathcal{A}_0^U(x) = \mathcal{X}_0^U(x)$. Thus, for any $\theta \in [L^{TT*}(x,z,\mathcal{X}_0^L(x)), U^{TT*}(x,z,\mathcal{X}_0^U(x))]$, we have that

$$\liminf \Pr[\theta \in [\mathcal{L}_n^{TT}(1-\alpha), \mathcal{U}_n^{TT}(1-\alpha)]] \geq 1-\alpha \ .$$

Now consider the case in which $\mathcal{X}_0^L(x) \cap \mathcal{X}_0^U(x) \neq \emptyset$. Let θ denote the common value $L^{TT*}(x,z,\mathcal{X}_0^L(x)) = U^{TT*}(x,z,\mathcal{X}_0^U(x))$. Note that $H(x,x') = 0$ for all $x' \in \mathcal{X}_0^L(x) \cap \mathcal{X}_0^U(x)$, and thus

$$\liminf \Pr[\hat{H}(x,x') - \frac{\epsilon_{0n}(1-\alpha)}{\sqrt{n}} \leq 0 \leq \hat{H}(x,x') + \frac{\epsilon_{0n}(1-\alpha)}{\sqrt{n}} \ \ \forall x' \in \mathcal{X}_0^L(x) \cap \mathcal{X}_0^U(x)] \geq 1-\alpha \ .$$

Thus, with probability at least $1-\alpha$ asymptotically $\mathcal{A}_0^L(x)$ and $\mathcal{A}_0^U(x)$ both exclude all values of x' in $\mathcal{X}_0^L(x) \cap \mathcal{X}_0^U(x)$. Note that for such $\mathcal{A}_0^L(x)$ and $\mathcal{A}_0^U(x)$ we have that

$$L^{TT*}(x,z,\mathcal{A}_0^L(x)) \leq \theta \leq U^{TT*}(x,z,\mathcal{A}_0^U(x)) \ .$$

As a result, we have that for such $\mathcal{A}_0^L(x)$ and $\mathcal{A}_0^U(x)$ with probability approaching 1, $\theta \in [\mathcal{L}_n^{TT}(1-\alpha), \mathcal{U}_n^{TT}(1-\alpha)]$. Thus, the desired coverage property holds in this case as well. □

8 Conclusion

This paper has constructed sharp bounds for the effect of a binary endogenous variable on a binary outcome variable under the assumption that the endogenous variable and outcome variable are jointly determined by triangular system of threshold-crossing models. We have also provided methods for inference for the resulting bounds. The assumptions considered in this paper are substantially weaker than those underlying, for example, the traditional bivariate probit model, since no parametric distributional assumptions are imposed. On the other hand, we impose more structure relative to the earlier analyses of Manski (1990) or Heckman and Vytlacil (2001).

Relaxing the parametric assumptions of a bivariate probit model comes at a cost: While the average effect of the binary endogenous variable is point-identified under the parametric distributional and functional form assumptions of the traditional bivariate probit model, the average effect is in general only set-identified without such assumptions. There is no loss of identifying power from removing these assumptions if the average effect of the treatment is zero or if there is variation in other regressors that directly compensates for variation in the endogenous variable. In these instances, even without the parametric assumptions, our analysis also point-identifies the parameter of interest. Moreover, even when the average treatment effect is not point-identified, we are still able to identify the sign of the average effect.

Strengthening the assumptions of Manski (1990) and Heckman and Vytlacil (2001) has a benefit: The width of the bounds are narrower if one imposes the threshold crossing structure, and, as noted above, always identify the sign of the average treatment effect. The narrowing of the bounds relative to Manski (1990) and Heckman and Vytlacil (2001) is particularly dramatic if there are regressors that enter the outcome equation that do not enter the selection equation for the endogenous variable.

This is of practical significance because the Manski (1990) and Heckman and Vytlacil (2001) bounds are sometimes too wide to allow applied researchers to make meaningful inferences in the context of their application (see, e.g., our empirical example of Bhattacharya, Shaikh, and Vytlacil (2005)). In exchange for imposing the threshold crossing model (but without having to impose any parametric assumptions), the techniques developed in this paper circumvent this difficulty, especially in cases where the sign of the average effect is of primary interest.

References

Aakvik, A., J. Heckman, and E. Vytlacil (1998): "Semiparametric Program Evaluation Lessons from an Evaluation of a Norwegian Training Program," mimeo, University of Chicago.

Altonji, J., and R. L. Matzkin (2005): "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *Econometrica*, forthcoming.

Amemiya, T. (1974): "The Nonlinear Two-Stage Least-Squares Estimator," *Journal of Econometrics*, 2, 105–110.

——— (1978): "The Estimation of a Simultaneous Equation Generalized Probit Model," *Econometrica*, 46, 1193–1205.

Angrist, J. (1991): "Instrumental Variables Estimation of Average Treatment Effects in Econometrics and Epidemiology," NBER Technical Working Paper No. 115.

——— (2001): "Estimation of Limited-Dependent Variable Models with Binary Endogenous Regressors: Simple Strategies for Empirical Practice," *Journal of Business and Economic Statistics*, 19, 2–16.

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

Bhattacharya, J., D. McCaffrey, and D. Goldman (2005): "Estimating Probit Models with Endogenous Covariates," *Statistics in Medicine*, forthcoming.

Bhattacharya, J., A. Shaikh, and E. Vytlacil (2005): "Treatment Effect Bounds: An Application to Swan-Ganz Catherization," mimeo, Stanford University.

Blundell, R., A. Gosling, H. Ichimura, and C. Meghir (2004): "Changes in the Distribution of Male and Female Wage Accounting for Employment Composition Using Bounds," mimeo, Institute for Fiscal Studies.

Blundell, R., and J. Powell (2003): "Endogeneity in Nonparametric and Semiparametric Regression Models," in *Advances in Economics and Econometrics: Theory and Applications, Eigth World Congress*, ed. by M. Dewatripont, L. Hansen, and S. Turnovsky. Cambridge University Press, Cambridge.

——— (2004): "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies*, 71, 655–79.

Blundell, R., and R. Smith (1986): "An Exogeneity Test for a Simultaneous Tobit Model," *Econometrica*, 54, 679–685.

——— (1989): "Estimation in a Class of Simultaneous Equation Limited Dependent Variable Models," *Review of Economic Studies*, 56, 37–58.

Cameron, S., and J. Heckman (1998): "Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males," *Journal of Political Economy*, 106, 262–333.

Chen, X., J. Heckman, and E. Vytlacil (1999): "Identification and Root-N Estimability of Semiparametric Panel Data Models with Binary Dependent Variables and a Latent Factor," mimeo, University of Chicago.

Chesher, A. (2003): "Nonparametric Identification under Discrete Variation," mimeo, University College London.

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

Heckman, J. (1990): "Varieties of Selection Bias," *American Economic Review*, 80, 313–318.

Heckman, J., R. LaLonde, and J. Smith (1999): "The Economics and Ecconometrics of Active Labor Market Programs," in *Handbook of Labor Economics*, ed. by A. Orley, and D. Card. Elsevier Science, North Holland, Amsterdam, New York and Oxford.

Heckman, J., and E. Vytlacil (2001): "Local Instrumental Variables," in *Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya*, ed. by C. Hsiao, K. Morimune, and J. Powell. Cambridge University Press, Cambridge.

Hong, H., and E. Tamer (2003): "Endogenous Binary Choice Model with Median Restrictions," *Economics Letters*, 80, 219–225.

Hotz, J., C. Mullins, and S. Sanders (1997): "Bounding Causal Effects Using Data from a Contaminated Natural Experiment: Analyzing the Effects of Teenage Childbearing," *Review of Economic Studies*, 64, 575–603.

Ichimura, H., and C. Taber (1999): "Estimation of Policy Effects Under Limited Support Conditions," mimeo, Northwestern University and University College London.

Imbens, G., and J. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(467-476).

Lee, L. F. (1981): "Simultaneous Equation Models with Discrete and Censored Dependent Variables," in *Structural Analysis of Discrete Data with Economic Applications*, ed. by C. Manski, and D. McFadden. MIT Press, Cambridge, MA.

Lewbel, A. (2000): "Semiparametric Qualitative Response Model Estimation with Unknown Heteroskedasticity or Instrumental Variables," *Journal of Econometrics*, 97, 145–77.

——— (2005): "Simple Estimators for Endogenous Selection Models," mimeo, Boston College.

Magnac, T., and E. Maurin (2005): "Identification and Information in Monotone Binary Models," *Journal of Econometrics*, Forthcoming.

Manski, C. (1990): "Nonparametric Bounds on Treatment Effects," *American Economic Review, Papers and Proceedings*, 80, 319–323.

——— (1994): "The Selection Problem," in *Advances in Econometrics: Sixth World Congress*, ed. by C. Manski, and D. McFadden. Cambridge University Press, Cambridge.

——— (1995): *Identification Problems in the Social Sciences*. Harvard University Press, Cambridge and London.

——— (2003): *Partial Identification of Probabilities Distributions*. Springer, New York and Heidelberg.

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

Newey, W. (1986): "Linear Instrumental Variable Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables," *Journal of Econometrics*, 32, 127–141.

Politis, D. N., J. P. Romano, and M. Wolf (1999): *Subsampling*. Springer, New York.

Rivers, D., and Q. Vuong (1988): "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models," *Journal of Econometrics*, 39, 347–366.

Vytlacil, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70(1), 331–341.

——— (2004): "A Note on Additive Separability and Latent Index Models of Binary Choice: Representation Results," mimeo, Stanford University.

Vytlacil, E., and N. Yildiz (2004): "Dummy Endogenous Variables in Weakly Separable Models," mimeo, Stanford University.