Bounds for Average Treatment Effect: A Comparison of Non-parametric and Quasi Maximum Likelihood Estimators

Chuhui Li, D.S. Poskitt and Xueyan Zhao

December 2018

Working Paper 26/18
Bounds for Average Treatment Effect: A Comparison of Non-parametric and Quasi Maximum Likelihood Estimators

Chuhui Li, D. S. Poskitt, and Xueyan Zhao

Department of Econometrics and Business Statistics, Monash University

April 17, 2018

Abstract

This paper examines issues arising from the application of partial identification for models involving binary endogenous treatment and binary response variables. Monte Carlo experimentation constructed using Australian population data is employed to investigate the performance of parametric and nonparametric estimates of average treatment effect (ATE) bounds. Quasi maximum likelihood estimation (QMLE) and local QMLE based on the bivariate probit model is shown to outperform nonparametric estimation and other QMLE derived from two equation triangular systems driven by alternative copulae. The effects of instrument strength on the size and estimation of identified sets is also studied. The relationship between alternative ATE identified sets that have appeared in the literature is examined, suggesting the general applicability of the findings.

Keywords: partial identification, binary models, treatment effect, bounds.

1 Introduction

Following the potential outcome framework in Neyman (1923), Rubin (1974, 2006) and Heckman (1990), let $Y_i$ represent the outcome observed for individual $i$, and let $Y_{1i}$ denote the response when individual $i$ is treated ($D_i = 1$) and $Y_{0i}$ the response when untreated ($D_i = 0$), where the dummy variable $D_i \in \{0, 1\}$ indicates the individuals treatment status. A basic quantity of interest in many policy analysis and program evaluation problems is the average treatment effect ($ATE$): $ATE_i = E[Y_{1i} - Y_{0i}]$. We wish to estimate the $ATE$, but with observational data of the type typically used in socio-economic empirical studies, we usually never observe $Y_{0i}$ and $Y_{1i}$ for the same individual. How then can the $ATE$ be identified, if at all?

An early contribution to the problem of identifying and estimating treatment effects was that of Heckman (1978), who developed a class of econometric models for simultaneous equation systems with dummy endogenous variables, and a typical approach to estimating the effect of a binary treatment variable on a binary response, where both may be driven by common observable and unobservable factors, is to assume a parametric model, such as a bivariate probit, together with the use of instrumental variables to force point identification (see Angrist and Pischke (2008) for example, and Jones (2007) for a discussion in the context of health economics). However, the underlying assumptions, including a parametric distributional form, a threshold crossing rule for the binary variables, and a separable error structure, are unlikely to be true of real data. Moreover, the identification relies heavily upon the parametric and distributional assumptions, and following the work of Manski (1988) researchers became aware that the assumptions underlying the simultaneous equations model need not point identify parametric functions of interest such as the $ATE$. Nevertheless, meaningful restrictions on the values that these functions may take can still be ascertained, i.e. they can be partially identified.

To be specific, following Chesher (2005, 2010), consider a model $M$ for a scalar binary response variable $Y$ of the form $Y = h(D, X, U)$ (a structural equation) where the structure function $h$ is weakly monotonic in $U$ and is normalized so that the marginal distribution of $U \in (0, 1)$ is uniform, $X$ denotes a vector of exogenous explanatory variables, and $D$ is a binary endogenous regressor. Suppose also that there exists a vector of instruments $Z$ such that $P[U \leq \tau|Z = z] = \tau$ for all $\tau \in (0, 1)$ and all $z \in \Omega_Z$, the support of $Z$. 

1
Since $Y$ is binary $h(d, x, u)$ must be a non-decreasing step function, and recognizing that $Y$ is a Bernoulli random variable, it follows that $h(d, x, u)$ can be characterized by a probability threshold function $p(d, x)$ such that

$$Y = h(D, X, U) = \begin{cases} 0, & 0 < U \leq p(D, X) \\ 1, & p(D, X) < U \leq 1 \end{cases}$$

(1)

where, without loss of generality, the support of $Y$ and the dummy treatment variable $D$ is taken as $\{0, 1\}$. An explicit example is where $Y$ indicates if an individual has visited a dentist in the last year, $X$ measures various socioeconomic and demographic characteristics of the individual, and $D$ denotes whether the individual possesses private health insurance. From equation (1) we have, trivially, that the distribution of the potential outcomes $Y_1$ and $Y_0$ is given by

$$Y_1 = h(1, X, U) = \begin{cases} 0, & 0 < U \leq p(1, X) \\ 1, & p(1, X) < U \leq 1 \end{cases}$$

(2)

and

$$Y_0 = h(0, X, U) = \begin{cases} 0, & 0 < U \leq p(0, X) \\ 1, & p(0, X) < U \leq 1 \end{cases}$$

(3)

It follows that the $ATE$ for an individual with characteristics $x$ is

$$ATE(x) = \mathbb{E}[Y_1 - Y_0 | X = x]$$

$$= P(\{Y = 1\}|D = 1, x) - P(\{Y = 1\}|D = 0, x)$$

$$= \mathbb{E}[Y_1 | X = x] - \mathbb{E}[Y_0 | X = x]$$

$$= 1 - p(1, x) - (1 - p(0, x)) = p(0, x) - p(1, x),$$

(4)

and it is apparent that the $ATE$ can be partially identified (bounded) by constructing bounds on the probability threshold function.

In a seminal contribution to this field Manski (1990) derived nonparametric bounds under very general assumptions, but unfortunately the resulting bounds proved to be uninformative for most practical purposes since they can be very wide, although with additional assumptions such as monotonicity of treatment response or monotonicity of treatment selection the bounds can be improved; see Manski (1997) and Manski and Pepper (2000). Shaikh and Vytlacil (2005) studied the partial identification and nonparametric estima-
ton of the relationship between a binary response variable and a binary treatment when both the response and the treatment are determined by threshold crossing rules. Chiburis (2010) extended the work of Shaikh and Vytlacil (2005) and considered the effects of imposing further constraints, such as linear latent index restrictions and parametric distributional assumptions. Kitagawa (2009) partially identified the ATE using exclusion restrictions and treatment monotonicity, and Hahn (2010) compared the bounds of Kitagawa with those of Chesher. Bhattacharya et al. (2012) exploited a further nonparametric, structural assumption (positive quadrant dependent error terms) to narrow the bounds of Manski (1990) and Shaikh and Vytlacil (2005). Huber et al. (2017) used monotonicity and stochastic dominance assumptions to derive bounds on the ATEs for different groups (treated population, entire population, the compliers, the always takers and the never takers). Jun et al. (2011) studied a nonparametric triangular system with potentially discrete endogenous regressors and nonseparable errors, and imposed a global exclusion and exogeneity condition to achieve a tighter identified interval than Chesher (2005). All of these different bounds rely on the imposition of additional constraints, however, and bearing in mind the adage in Manski (2003) that the credibility of inference decreases with the strength of the assumptions maintained (the law of decreasing credibility), we wish to examine the partial identification of the ATE under as weak a set of assumptions as possible. Our intention therefore is to investigate the practical importance of the results on the partial identification of ATEs by building upon the analysis of Chesher (2005, 2010).

Estimating the ATE identified set involves the estimation of various conditional probabilities. These can be estimated using different parametric and nonparametric estimation methods, and in this paper our interest will focus on how such estimators perform under different scenarios. Whilst parametric methods often impose restrictive assumptions, they can be less prohibitive than nonparametric methods in the sense of demanding less computational effort and not requiring large data sets. On the other hand, whereas nonparametric estimators generally impose fewer assumptions upon the DGP, numerical constraints and paucity of data can limit their application. In this paper, we employ simulation (Monte Carlo) techniques to investigate the performance of six alternative methods for estimating the identified set of the ATE; raw nonparametric methods, nonparametric methods based on kernel smoothing techniques, and different quasi maximum likelihood estimation (QMLE) and local-QMLE. We will examine the effects that changes in the probabilistic structure of the DGP, such as the level of endogeneity and variations in sample
size, have upon these different techniques and thereby evaluate their relative advantages and disadvantages.

In our investigations we will also examine the effects of changes in the strength of the instruments $z \in \Omega_Z$. Research on weak instruments has to date confined its attention, by and large, to standard simultaneous equations models (see Poskitt and Skeels, 2012). Here we generalise the associated ideas to models involving dummy endogenous treatment variables and binary responses, and construct a measure of instrument strength that is relevant to such models. This we do by employing an adaptation of the pseudo-$R^2$ statistic proposed by Veall and Zimmermann (1992), designated $R^2_{VZ}$ in what follows. For a discussion of various measures of fit suggested for qualitative response models see Windmeijer (1995). The measure $R^2_{VZ}$ allows us to control the weakness of the instrumental variables in our experimental design, hence enabling us to examine the impact of alterations in instrument strength on the size of the $ATE$ identified set, and the ability of the different estimation techniques to identify the set in practice.

In order that our experimental DGP should reflect the type of data sets likely to be encountered in empirical applications, the Monte-Carlo observations are generated from a generic experimental DGP that imposes no explicit restrictions either on the underlying relationships or the probability distributions. This is achieved by basing the experimental DGP upon the 2004/5 Australian National Health Survey (ANHS) data and using observed regressors $X$ and instruments $Z$ that determine a hypothetical $ATE$ of private health insurance status on the probability of visiting a dentist. Thus the experimental data is generated in such a way that it mimics an actual Australian population distribution profile, rather than being based on a simple and stylized parametric specification of the type commonly used in the existing literature. This allows us to impose as little structure as possible and hopefully implies that our conclusions will be more meaningful to practitioners. Furthermore, using a DGP derived from a real world data set involving a number of regressors representing individual socioeconomic and demographic characteristics, our paper also serves as an illustration of different ways $ATE$ bounds might be estimated in empirical applications.

Finally, although our examination and comparison of the performance of the different estimators is couched in terms of the estimation of the $ATE$ identified set due to Chesher (2005, 2010), our results and conclusions can be generalised to other definitions of the $ATE$ identified set that incorporate alternative, more restrictive assumptions. We demon-
strate this by showing that various conditional probabilities underlying the ATE identified sets of Shaikh and Vytlacil (2005), for example, are equivalent to those that underly the Chesher ATE identified sets that are the focus of this paper, and that the identified sets of Bhattacharya et al. (2012) are also equivalent to those of Chesher when the latter are subject to the imposition of additional monotone constraints.

The rest of this paper is arranged as follows. Section 2 reviews the derivation of the ATE partially identified set of Chesher (2005, 2010). Section 3 presents the different estimators that are the subject of our investigation. Section 4 introduces the details of our Monte Carlo design based on an Australian population survey. The experimental results are presented and the performance of the different estimation methods is examined in Section 5. In Section 6 it is shown that the conceptual framework used throughout the paper is sufficiently general to encompass a broad range of models and ATE bounds considered in the literature. Section 7 provides a summary of the paper. Proofs of results presented in the paper are assembled in an appendix.

2 The ATE Identified Set

It is clear from equation (4) that the values obtained by \( p(0, X) \) and \( p(1, X) \) determine the sign and magnitude of the ATE. In order to construct bounds on \( p(0, X) \) and \( p(1, X) \), and hence ATE, consider first the case where \( p(0, X) < p(1, X) \). The event \( \{Y < h(D, X, U)\} \) occurs if and only if \( h(D, X, U) = 1 \) and \( Y = 0 \), and since \( h(D, X, U) = 1 \iff p(D, X) < U \) we have, using an obvious notation,

\[
P\left(\{Y < h(D, X, U)\}\mid Z\right) = P\left(\{Y = 0\} \cap \{p(D, X) < U\}\right) \mid Z.
\] (5)

Calculating the probability \( P\left(\{Y < h(D, X, U)\}\mid Z\right) \) as given in (5), recognising that there are three possibilities for \( U \): \( 0 \leq U \leq p(0, X) \), \( p(0, X) < U \leq p(1, X) \), and \( p(1, X) < U \leq 1 \), yields for each case:

- \[
P\left(\{Y = 0, D = 0\}\mid X, Z\right) \quad \text{when } 0 \leq U \leq p(0, X),
\]
- \[
P\left(\{Y = 0, D = 0\}\mid X, Z\right) + P\left(\{Y = 0, D = 1\}\mid X, Z\right) \quad \text{when } p(0, X) < U \leq p(1, X) \text{ and } p(1, X) < U \leq 1,
\]
respectively. From the inequality $P\left(\{Y < h(D, X, U)\} | Z\right) < U$ we can now deduce that

\[
p(0, X) \geq P(\{Y = 0, D = 0\} | X, Z) \tag{6}
\]
\[
p(1, X) \geq P(\{Y = 0, D = 0\} | X, Z) + P(\{Y = 0, D = 1\} | X, Z)
\]

Next consider the event $\{Y \leq h(D, X, U)\}$. This event occurs if and only if $h(D, X, U) = 1$, when any value of $Y \in \{0, 1\}$ is admissible, or $h(D, X, U) = 0$ and $Y = 0$, and we have

\[
P\left(\{Y \leq h(D, X, U)\} | Z\right) = P\left(\{Y = 0\} \cap \{U \leq p(D, X)\} | Z\right) + P\left(\{p(D, X) < U\} | Z\right).
\]

Calculating $P\left(\{Y \leq h(D, X, U)\} | Z\right)$ for each of $0 \leq U \leq p(0, X)$, $p(0, X) < U \leq p(1, X)$, and $p(1, X) < U \leq 1$, we have

\[
\begin{cases}
P(\{Y = 0, D = 0\} | X, Z) + P(\{Y = 0, D = 1\} | X, Z) & \text{when } 0 \leq U \leq p(0, X), \\
P(\{D = 0\} | X, Z) + P(\{Y = 0, D = 1\} | X, Z) & \text{when } p(0, X) < U \leq p(1, X) \text{ and} \\
1 & \text{when } p(1, X) < U \leq 1,
\end{cases}
\]

and from the inequality $P\left(\{Y \leq h(D, X, U)\} | Z\right) \geq U$ we can conclude that

\[
p(0, X) \leq P(\{Y = 0, D = 0\} | X, Z) + P(\{Y = 0, D = 1\} | X, Z) \tag{7}
\]
\[
p(1, X) \leq P(\{D = 0\} | X, Z) + P(\{Y = 0, D = 1\} | X, Z).
\]

Thus $ATE(x)$ can be bounded by bounding $p(0, x)$ and $p(1, x)$ to give the identified set of the $ATE$ as the interval

\[
\left[\sup_{z \in \Omega_Z} P(\{Y = 0, D = 0\} | x, z) - \inf_{z \in \Omega_Z} \{P(\{D = 0\} | x, z) + P(\{Y = 0, D = 1\} | x, z)\}, \right.
\]
\[
\left.\inf_{z \in \Omega_Z} \{P(\{Y = 0\} | x, z)\} - \sup_{z \in \Omega_Z} \{P(\{Y = 0\} | x, z)\}\right]\]

\[
(8)
\]

in the case where $p(0, x) < p(1, x)$ and the $ATE$ is negative.
Similarly, \( ATE(x) \) can be bounded by the interval

\[
\left[ \sup_{z \in \Omega_Z} \{ P(\{ Y = 0 \} | x, z) \} - \inf_{z \in \Omega_Z} \{ P(\{ Y = 0 \} | x, z) \} , \right. \\
\left. \inf_{z \in \Omega_Z} \{ P(\{ D = 1 \} | x, z) + P(\{ Y = 0, D = 0 \} | x, z) \} - \sup_{z \in \Omega_Z} P(\{ Y = 0, D = 1 \} | x, z) \right]
\]

(9)
in the case where \( p(0, x) > p(1, x) \) and the \( ATE \) is positive.

From the expressions in (8) and (9) it is apparent that the \( ATE \) can be only be bounded or partially identified even if the various conditional probability functions are known. Estimation of the \( ATE \) identified set obviously turns on the estimation of these conditional probabilities.

3 Estimation Methods

3.1 The raw nonparametric estimator

The simplest and most straightforward estimator to contemplate is the naïve estimator obtained by replacing the unknown conditional probabilities in the bound formulas (8) and (9) by their empirical counterparts — the observed relative frequencies. This method is straightforward to implement, but it requires that all the variables be measured on a discrete scale (as is often the case for age, gender, marital status, for example) for calculation of conditional sample relative frequencies. In addition, in empirical applications the calculation of sample analogues of the conditional probabilities can become questionable, and even unfeasible, because in finite samples the data may not be sufficiently rich and there may not be ‘enough’ observations for each possible value of \( X \) and \( Z \) for the sample relative frequencies to be either reliable or calculable.

3.2 The smoothed nonparametric estimator

Rather than using the raw frequencies, we can use a nonparametric smoothing method to estimate the conditional probabilities in the bound formulas. This method is based on the fact that the conditional probability of a dichotomous variable coincides with its
conditional mean. The Nadaraya-Watson kernel regression estimator (Nadaraya 1964; Watson 1964) can therefore be employed to estimate the required probabilities expressed as conditional mean functions, using adaptations of kernel smoothing methods for mixed data types, as studied in Li and Racine (2003, 2007).

In the simulation experiments reported below we used the kernel introduced in Aitchison and Aitken (1976) for the discrete regressors:

$$K_A(x_i, x, h) = \begin{cases} 
1 - h, & \text{if } x_i = x, \\
h/(c - 1), & \text{if } x_i \neq x,
\end{cases}$$

where $c$ is the number of discrete outcomes assumed by the regressor $X$ and the bandwidth $h$ must lie between 0 and $(c - 1)/c$. For the continuous regressors we used the Gaussian kernel $K_G(z) = \exp(-z^2/2)/\sqrt{2\pi}$ with $z = (x_i - x)/h$, $h > 0$. The bandwidths were computed for the mixed continuous and discrete data types using the least-squares cross validation method introduced in Li and Racine (2004) and Racine and Li (2004).\(^1\)

### 3.3 The QMLE estimator

The QMLE estimates of $ATE$ bounds studied in this paper are obtained by inserting maximum likelihood estimates of the required conditional probabilities into (8) and (9), where the maximum likelihood estimates are calculated by fitting a parametric bivariate binary response model to the data. It should perhaps be emphasised that the model underlying the likelihood calculations is only being used as a tool to construct a QMLE estimator; the DGP is not assumed to be governed by the probability structure of the fitted model but is viewed as an unknown member of the broad class of processes that satisfy the assumptions of the structural equation specification.

Consider, for example, the recursive bivariate probit (RBVP) model. This is a linear index, threshold crossing model with two equations, with both the discrete response variable and the endogenous treatment dummy variable determined by a threshold crossing rule and a

\(^1\)See the np package from R\(^\copyright\) developed by Jeffrey S. Racine and Tristen Hayfield. Details are available at https://github.com/JeffreyRacine/R-Package-np/.
linear index equation with a separable error:

\[
\begin{aligned}
D^* &= X\alpha + Z'\delta + E, \quad D = 1 \text{ if } D^* > 0 \text{ and } 0 \text{ otherwise}; \\
Y^* &= X\beta + D\gamma + V, \quad Y = 1 \text{ if } Y^* > 0 \text{ and } 0 \text{ otherwise},
\end{aligned}
\]

(10)

where \(Y, D, X\) and \(Z\) are as previously defined, and \(E\) and \(V\) are correlated standard normal random variables.

That the RBVP model falls within the ambit of the structural equation model follows by letting the random variable \(U = \Phi(V)\) where \(\Phi(\cdot)\) is the standard normal distribution function. Now set

\[
h(D, X, U) = I\{U > \Phi(-X\beta - D\gamma)\}
\]

where \(I\{A\}\) denotes the indicator function for the event \(A\). Then \(h(D, X, U)\) defines a structural function that is weakly monotonic in \(U\) where \(P(U \leq u|Z = z) = u\) for all \(u \in (0, 1)\) and all \(z \in \Omega_Z\) since by assumption \(U\) is independent of \(Z\). Thus the assumptions of the structural equation model are satisfied and \(Y\) can be expressed as in (1) with a threshold function given by \(p(D, X) = 1 - \Phi(X\beta + D\gamma)\). Thus the RBVP model corresponds to a structural equation specification augmented with a reduced form for the dummy treatment variable and additional parametric assumptions.

The RBVP model belongs to the general class of models considered in Han and Vytlacil (2017). Han and Vytlacil examine a two equation triangular system for binary endogenous variables where the bivariate normality assumption on the latent error terms of a RBVP model is generalised through the use of monotone regression dependent parametric copulae. The RBVP model corresponds to the use of probit marginal distributions in conjunction with the Gaussian copula (see Han and Vytlacil, 2017, Appendix 3.3). Henceforth we will therefore label the QMLE based upon the RBVP model QMLE \((G^P)\).

It is obvious that different members of the class of models considered in Han and Vytlacil (2017) can also be used to construct QMLE estimates, and in this paper we also study the performance of other copula-based bivariate binary response model QMLE estimators obtained using different copulae and marginals. In particular, we will consider the two equation triangular systems for binary endogenous variables where the latent error terms are characterised by the Frank copula with logit marginals, and the Joe copula with probit marginals, which we will denote as the QMLE \((F^L)\) and QMLE \((J^P)\) estimators respectively.
The local-QMLE estimator

The local-QMLE is a semi-parametric counterpart of the QMLE. This estimator is obtained by using an adaptation of the local likelihood technique due to Tibshirani and Hastie (1987), Fan et al. (1998) and Kauermann and Opsomer (2003) for fitting generalized linear models and regression equations.

For a given model with parameter $\theta$, let $L(\theta : y_i, d_i, x_i, z_i)$ denote the log-likelihood value obtained at each data point $(y_i, d_i, x_i, z_i), i = 1, \ldots, n$. Then for data in a neighborhood of the point $(x_0, z_0)$ each log-likelihood value is weighted by a kernel function value $K_H(H^{-1}(X_i - x_0), H^{-1}(Z_i - z_0))$ where $H_X$ and $H_Z$ are diagonal matrices of bandwidth parameters for the exogenous regressors $X$ and the instruments $Z$ (see Frölich, 2006). The locally weighted log-likelihood function for a response $Y$ with an endogenous treatment $D$ is thus given by

$$L(\theta ; H, (x_0, z_0)) = \sum_{i=1}^{n} L(\theta : y_i, d_i, x_i, z_i)K_H(H^{-1}(X_i - x_0), H^{-1}(Z_i - z_0)) \quad (11)$$

The local-QMLE estimator is then defined as the maximiser of (11). This in turn yields a (conditional) local-QMLE estimator of $\theta$ at the location $(x_0, z_0)$, given by

$$\hat{\theta}_{(x_0, z_0)} = \arg \max_{\theta} \{L(\theta ; H, (x_0, z_0))\} \quad (12)$$

In the simulation experiments that follow, the kernel function in (12) is taken as being multiplicative, that is to say, $K_H(H^{-1}(X_i - x_0), H^{-1}(Z_i - z_0))$ is set equal to the product of

$$\prod_{q=1}^{k_1} K_A\left(\frac{x_{qi} - x_{q0}}{h_q}\right) \prod_{r=1}^{k_2} K_G\left(\frac{x_{(k_1+r)i} - x_{(k_1+r)0}}{h_{(k_1+r)}}\right)$$

and

$$\prod_{m=1}^{l_1} K_A\left(\frac{z_{mi} - z_{m0}}{h_m}\right) \prod_{s=1}^{l_2} K_G\left(\frac{z_{(l_1+s)i} - z_{(l_1+s)0}}{h_{(l_1+s)}}\right)$$

where $x = (x_1, \ldots, x_{k_1}, x_{k_1+1}, \ldots, x_{k_1+k_2})$ and $z = (z_1, \ldots, z_{l_1}, z_{l_1+1}, \ldots, z_{l_1+l_2})$ denote divisions of the exogenous regressors and the instruments into discrete and continuous variables, and $K_A(\cdot)$ and $K_G(\cdot)$ are the discrete and continuous regressor kernels used for the smoothed nonparametric estimator outlined in Section 3.2. It is not unreasonable to suppose, therefore, that bandwidth choices made for the local-QMLE would-and-should be similar to those used for the smooth nonparametric estimator, so bandwidths derived
for the smoothed nonparametric estimator were used for the local-QMLE. This latter allo-
cation also seems reasonable since in the simulation study we are interested in comparing
the properties of the different estimators when applied to the same data sets and the use of
common bandwidths avoids fluctuations in performance that might arise due to variation
in user assigned tuning parameters.

4 Experimental Design

In order to investigate the finite sample properties of the different estimation methods
we generate data from a generic DGP in which commonly employed assumptions such as
known functional form, additive errors, and so on are not imposed. To give some guidance
on the estimators’ likely empirical performance, and ensure that the simulated DGP has
some practical relevance, the ANHS data will be used as the basis for mimicking a real
world problem, in which the endogenous treatment variable is an individual’s private
health insurance status and the response of interest is his/her dental service utilisation.

For our explanatory variables $X$ we include age (divided into 11 age groups), gender,
marital status, and education. Altogether there are 88 different possible combinations of
values of $X$, referred to subsequently as cells. All these variables are typical of those
used in health and labour economics applications. To construct instruments $Z$ similar to
t hose that might be used in applied research, whilst not overcomplicating the simulation
design, we generate two dichotomous instrumental variables $Z_1$ and $Z_2$. The variable
$Z_1$ is designed as an indicator of the insurance premium paid to acquire private health
insurance, with $Z_1 = 1$ indicating that the individual has paid an above average premium
to purchase health insurance. The variable $Z_2$ is supposed to be a variable measuring a
person’s attitude towards risk, with $Z_2 = 1$ indicating that the individual is risk averse.
This variable was constructed using the ANHS data on a persons life style variables such
as drinking, smoking and exercising.\(^2\)

The probabilistic structure of the experimental DGP is based upon the sample relative
frequencies in ANHS, with values of the regressors $X$ and instruments $Z$ in the simula-
tions, $x_i$ and $z_i$, $i = 1, \ldots, n$, being drawn from the empirical distribution of the variables
in the data. To begin, we generate a set of $n$ values of $X$ and $Z$, and we then generate

\(^2\)See Li (2015) for a more detailed description of the variable selection and construction used here.
realizations \( d_i \) and \( y_i \), \( i = 1, \ldots, n \), of \( D \) and \( Y \) from their joint (conditional) distribution \( P(Y, D|X, Z) \). The latter distribution is determined as follows. Since \( Y, D, Z_1, \) and \( Z_2 \) are binary, their joint probability mass functions can be displayed as in Table 1, wherein for convenience we introduce a simplified notation that omits reference to the exogenously determined \( X \). The correlation coefficient between \( D \) and \( Z_j \), \( j = 1, 2 \), is

\[
\rho(D, Z_j) = \frac{\text{cov}(D, Z_j)}{\sqrt{\text{var}(D) \text{var}(Z_j)}} = \frac{p(D, Z_j) - p(D) p(Z_j)}{\sqrt{p(D) p(Z_j) p(D) p(Z_j)}}.
\] (13)

which can be rearranged so as to express the joint probability \( P(D = 1, Z_j = 1) \) as

\[
p(1, Z_j = 1) = \rho(D, Z_j) \sqrt{p(D) p(Z_j) p(D) p(Z_j)} + p(D) p(Z_j).
\] (14)

Set \( \rho_1 = \rho(D, Z_1) \) and \( \rho_2 = \rho(D, Z_2) \). The correlation coefficients \( \rho_1 \) and \( \rho_2 \) determine the strength of the two instrumental variables, and we shall control the values of \( \rho_1 \) and \( \rho_2 \) as part of our experimental design. It is apparent that given the joint probability distribution of \( Z_j \), \( j = 1, 2 \), and the probability distribution of \( D \), we can manipulate Equation (14) so as to generate joint and conditional probability mass functions for the pairs \((D, Z_1)\) and \((D, Z_2)\) that are consistent with any chosen values of \( \rho_1 \) and \( \rho_2 \).

In order to obtain the joint probability function of the quadruple \((Y, D, Z_1, Z_2)\) we will suppose that \( Z_1 \) and \( Z_2 \) are independent conditional on \( D \) \( (Z_1 \perp \perp Z_2|D) \) so that we can derive the joint probability function of the triple \((D, Z_1, Z_2)\) as

\[
P(D = d, Z_1 = z_1, Z_2 = z_2) = P(Z_1 = z_1, Z_2 = z_2|D = d) P(D = d) \]
\[
= P(Z_1 = z_1|D = d) P(Z_2 = z_2|D = d) P(D = d).
\] (15)

This then allows us to obtain the joint probability function of \((Y, D, Z_1, Z_2)\) as

\[
P(Y = y, D = d, Z_1 = z_1, Z_2 = z_2) = P(Y = y|D = d, Z_1 = z_1, Z_2 = z_2) P(D = d, Z_1 = z_1, Z_2 = z_2)
\]
\[
= P(Y = y|D = d) P(D = d, Z_1 = z_1, Z_2 = z_2). \quad (16)
\]

Equations (15) and (16) allow us to use the relative frequencies observed in the ANHS data to assign probabilities to the joint and conditional probability distributions of \( Y, D, Z_1 \) and \( Z_2 \). Thus, the Monte-Carlo replications of the experimental DGP are generated
Table 1. Joint distributions of $Y, D, Z_1, Z_2$

(a) $P(D, Z_j), j = 1, 2$

<table>
<thead>
<tr>
<th>$Z_j$</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$p^D_{11}$</td>
<td>$p^D_{10}$</td>
</tr>
<tr>
<td>0</td>
<td>$p^D_{01}$</td>
<td>$p^D_{00}$</td>
</tr>
</tbody>
</table>

(b) $P(D, Z_1, Z_2)$

<table>
<thead>
<tr>
<th>$Z_1Z_2$</th>
<th>11</th>
<th>10</th>
<th>01</th>
<th>00</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$p_D^{11}$</td>
<td>$p_D^{10}$</td>
<td>$p_D^{01}$</td>
<td>$p_D^{00}$</td>
</tr>
<tr>
<td>0</td>
<td>$p_D^{01}$</td>
<td>$p_D^{00}$</td>
<td>$p_D^{10}$</td>
<td>$p_D^{11}$</td>
</tr>
</tbody>
</table>

(c) $P(Y, D, Z_1, Z_2)$

<table>
<thead>
<tr>
<th>$YD$</th>
<th>11</th>
<th>10</th>
<th>01</th>
<th>00</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>$p_Y^{1111}$</td>
<td>$p_Y^{1101}$</td>
<td>$p_Y^{1011}$</td>
<td>$p_Y^{1001}$</td>
</tr>
<tr>
<td>10</td>
<td>$p_Y^{1110}$</td>
<td>$p_Y^{1100}$</td>
<td>$p_Y^{1010}$</td>
<td>$p_Y^{1000}$</td>
</tr>
<tr>
<td>01</td>
<td>$p_Y^{0111}$</td>
<td>$p_Y^{0101}$</td>
<td>$p_Y^{0011}$</td>
<td>$p_Y^{0001}$</td>
</tr>
<tr>
<td>00</td>
<td>$p_Y^{0100}$</td>
<td>$p_Y^{0010}$</td>
<td>$p_Y^{0000}$</td>
<td>$p_Y^{0000}$</td>
</tr>
</tbody>
</table>

by first evaluating the required probabilities using the probability structure of the experimental DGP as shown schematically in Table 1, and then simulating random draws of $D$ and $Y$ as follows:

(1) Set $P(Z_1 = 1) = p^Z_1$ and $P(Z_1 = 0) = p^Z_0 = 1 - p^Z_1$, and $P(Z_2 = 1) = p^Z_2$ and $P(Z_2 = 0) = p^Z_0 = 1 - p^Z_2$, equal to the relative frequencies of the instrumental variable values observed in the real data. Similarly, set the marginal and joint distributions of $Y$ and $D$ equal to those given by the observed relative frequencies in the real data.

(2) Assign values to $\rho_1$ and $\rho_2$ to build Table 1(a) by using Equation (14).

(3) Assume that $Z_1 \perp Z_2 | D$ and calculate the probabilities $P(D, Z_1, Z_2)$ in Table 1(b)
from Equation (15).

(4) From Equation (16) determine $P(Y,D,Z_1,Z_2)$ in Table 1(c), and hence evaluate the conditional probability distribution $P(Y,D|Z_1,Z_2)$.

Before continuing, recall that all the probabilities evaluated in the previous steps that relate to events involving $Y$ and $D$ are conditional on $X$. Reintroducing $X$ into the notation, the final step of the experimental DGP is to generate realisations of $D$ and $Y$ in the following manner.

(5) Set

$$q_{00}(x,z) = P\left(\{Y = 0, D = 0\}|x,z\right),$$
$$q_{01}(x,z) = P\left(\{Y = 0, D = 0\}|x,z\right) + P\left(\{Y = 0, D = 1\}|x,z\right) \text{ and}$$
$$q_{11}(x,z) = P\left(\{Y = 0\}|x,z\right) + P\left(\{Y = 1, D = 0\}|x,z\right).$$

Now simulate a value $u$ of a random variable $U$ uniformly distributed on the interval $(0,1)$ and allocate the value $(y,d)$ to the pair $(Y,D)$ where:

$$(y,d) = \begin{cases} (0,0), & \text{if } u \leq q_{00}(x,z), \\ (0,1), & \text{if } q_{00}(x,z) < u \leq q_{01}(x,z), \\ (1,0), & \text{if } q_{01}(x,z) < u \leq q_{11}(x,z), \\ (1,1), & \text{if } q_{11}(x,z) < u \leq 1. \end{cases}$$

As indicated above, this mechanism is implemented by first generating simple random samples of $n$ values of the regressors $X$ and instruments $Z$ using the relative frequency distributions calculated from the ANHS data. The corresponding values of $Y$ and $D$ are then generated as described in steps 1 through 5 once values have been assigned to $(\rho_1, \rho_2)$, with the values of $X$ and $Z$ held fixed throughout those replications associated with any particular combination of $n$ and $(\rho_1, \rho_2)$.

As an indication of the overall instrument strength associated with a given choice of the correlation coefficients $\rho_1$ and $\rho_2$, we will calculate a theoretical pseudo-$R^2$ value. Veall and Zimmermann (1992) construct a pseudo-$R^2$ statistic for qualitative response models using the maximised log-likelihood values of the model with and without regressors, and our pseudo-$R^2$ value is obtained by employing an adaptation of the rationale used by
Veall and Zimmermann. The log-likelihood function of a simple random sample \( d_i, i = 1, \ldots, n \), of the treatment variable \( D \) is

\[
\ln L_D = \ln \left( \prod_{i=1}^{n} P(d_i = 1)^{d_i} P(d_i = 0)^{1-d_i} \right) = \sum_{i=1}^{n} d_i \ln P(d_i = 1) + (1 - d_i) \ln P(d_i = 0).
\]

By the strong law of large numbers \( \ln L_D/n \) will converge almost surely to \( \mathcal{L}_{XZ} \) where

\[
\mathcal{L}_{XZ} = \mathbb{E} \left[ P(D = 1|X, Z) \ln P(D = 1|X, Z) + P(D = 0|X, Z) \ln P(D = 0|X, Z) \right],
\]

and \( \lim_{n \to \infty} | \ln L_D/n - \mathcal{L}_X | = 0 \) where

\[
\mathcal{L}_X = \mathbb{E} \left[ P(D = 1|X) \ln P(D = 1|X) + P(D = 0|X) \ln P(D = 0|X) \right] \quad (17)
\]

under the assumption that \( D \) is independent of \( Z \). The various probabilities required to evaluate (17) and (18) are those derived from the probability structure of the DGP used to generate the experimental data. Our theoretical, or “population”, pseudo-\( R^2 \) value is now defined as

\[
\mathcal{R}^2_{XZ} = \frac{\delta - 1}{\delta - \mathcal{L}_{XZ}} \mathcal{L}_{XZ}, \quad \delta = \frac{1}{2 \mathcal{L}_X}, \quad (19)
\]

where

\[
\mathcal{L}_{XZ} = 1 - \frac{\mathcal{L}_{XZ}}{\mathcal{L}_X} \quad (20)
\]

is a theoretical DGP analog of the likelihood ratio index of McFadden (1974).

5 Simulation Results

In this section we present results on the performances of six alternative estimators for ATE bounds due to Chesher (2010). We have conducted a range of experiments for different combinations of \( n \) and \((\rho_1, \rho_2)\), but due to space considerations we only present specific results for \( n = 1000 \) and \( n = 10000 \) in conjunction with a representative range of values for pseudo-\( R^2 \) representing instrument strength.\(^3\) The experimental results reported here are based on replicating each experimental DGP \( R = 1000 \) times.

\(^3\)The ANHS data set used in the construction of the experimental DGP contains 17,187 observations.
5.1 Estimation of ATE Bounds

Figure 1 shows the true ATE bounds together with the estimates produced by six estimators when \( n = 1000 \) and \( (\rho_1, \rho_2) = (-0.5, 0.5) \). The value \( (\rho_1, \rho_2) = (-0.5, 0.5) \) was chosen to mimic instruments having weak to moderate strength, and gives a value of \( R^2_{VZ} = 0.208 \). In each sub-figure, the x-axis denotes the X cells numbered from 1 to 88, and the y-axis the ATE values. The red and blue diamond markers are the true ATE upper and lower bounds respectively for each X cell. The green and pink square markers denote respectively the upper and lower bound estimates averaged across the \( R = 1000 \) experimental replications. All 1000 estimates of the upper and lower bounds are also plotted as green and pink dots in each sub-figure.

Perhaps the first notable feature of Figure 1 is the sparsity of green and pink dots, and the sparsity of the associated green and pink square markers in sub-figure Figure 1(d) for raw nonparametric estimator. Across all \( R = 1000 \) replications a total of 71 of the 88 X cells appear void, indicating that for many replications and cells the raw nonparametric method broke down as the required conditional probabilities could not be calculated due to a lack of observations with the required instrument values. For other cells low numbers of observations resulted in the raw nonparametric estimator producing an upper bound estimate near 1 and a lower bound estimate close to 0, and although their entries are not empty the pink and green dots for the upper and lower bounds significantly overlap. Simple visual inspection of Figure 1 also suggests that whereas the QMLE \((G^P)\) and local-QMLE \((G^P)\) estimators have reasonably small variances, as does the QMLE \((J^P)\) estimator, with all the lower and upper bound estimates well separated and contained within the interval \((0, 1)\), both nonparametric estimates and the QMLE \((F^L)\) estimates are far more widely dispersed. Overall, QMLE \((G^P)\), local-QMEL \((G^P)\) and QMLE \((J^P)\) estimators have the best performances without exhibiting any systematic bias, whilst this is far from the case for raw nonparametric and QMLE \((F^L)\) estimators. QMLE \((F^L)\) seems to overestimate the lower bounds for most cells.

The features seen in Figure 1 are repeated in Figure 2, when sample size is increased to \( n = 10000 \). This ten-fold increase in sample size has clearly reduced the variances of all the estimators, and appears to have gone some way in addressing the paucity issue previously observed with the raw nonparametric estimator; cf. Figure 1(d) and Figure 2(d). Nevertheless, the overall performance of the QMLE \((G^P)\) and local-QMLE \((G^P)\) esti-
Figure 1. Six estimated ATE (Chesher) bounds against the true bounds: scatter plots of estimated upper and lower ATE bounds from the 1,000 replications for different estimation methods for sample size 1,000. $\rho_1 = -0.5$, $\rho_2 = 0.5$, $V_Z = 0.208$. 

- (a) QMLE ($G^p$)
- (b) Local-QMLE ($G^p$)
- (c) Smoothed nonparametric estimator
- (d) Raw nonparametric estimator
- (e) QMLE ($F^L$)
- (f) QMLE ($J^P$)
Figure 2. Six estimated ATE (Chesher) bounds against the true bounds: scatter plots of estimated upper and lower ATE bounds from the 1000 replications for different estimation methods for sample size 10000. $\rho_1 = -0.5$, $\rho_2 = 0.5$ ($R_{VZ}^2 = 0.208$).
Table 2. Bias, variance, and root mean squared error of the upper and lower ATE bound estimates ($n = 10000$)

<table>
<thead>
<tr>
<th>ATE Bound</th>
<th>Upper Bound</th>
<th>Lower Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BIAS</strong></td>
<td><strong>SE</strong></td>
<td><strong>RMSE</strong></td>
</tr>
<tr>
<td>True</td>
<td>[0.1047, 0.4367]</td>
<td>-</td>
</tr>
<tr>
<td>QMLE ($G^P$)</td>
<td>[0.0966, 0.4591]</td>
<td>0.0238</td>
</tr>
<tr>
<td>Local-QMLE ($G^P$)</td>
<td>[0.0942, 0.4605]</td>
<td>-0.0379</td>
</tr>
<tr>
<td>Smoothed Nonparametric</td>
<td>[0.1119, 0.3988]</td>
<td>-0.1590</td>
</tr>
<tr>
<td>Raw Nonparametric</td>
<td>[0.2110, 0.2777]</td>
<td>0.0355</td>
</tr>
<tr>
<td>QMLE ($F^L$)</td>
<td>[0.1709, 0.4722]</td>
<td>-0.0177</td>
</tr>
</tbody>
</table>

The latter observation is confirmed in Table 2, which presents the bias, standard error, and root mean squared error for each of the ATE upper and lower bound estimates shown in Figure 2 for $n = 10000$ over $R = 1000$ replications. The superiority of the QMLE ($G^P$) and local-QMLE ($G^P$) estimators is apparent from this summary table. The magnitudes of the biases and standard errors are such that the root mean squared errors (RMSEs) for the nonparametric estimator are four to ten times of those of the QMLE ($G^P$) and local-QMLE ($G^P$) estimators. The smoothed nonparametric, QMLE ($J^P$) and QMLE ($F^L$) estimators also all have more than twice the RMSEs than those for the QMLE ($G^P$) and local-QMLE ($G^P$) estimators.

In studies of partial identification it is not uncommon to examine the properties and performance of estimators by evaluating the Hausdorff distance between the true identified set and the estimated identified set (Manski and Tamer, 2002; Chernozhukov et al., 2007; Beresteanu and Molinari, 2008; Menzel, 2014). Figure 3 plots the Hausdorff distance between the true ATE identified set and the estimated ATE identified set for each cell, with the cells arranged by frequency of occurrence for $X$, for the six estimators examined here, averaged over the $R = 1000$ replications. Given that the QMLE ($G^P$) and local-QMLE ($G^P$) estimators exhibited significantly better performance on the basis of the traditional measures presented in Table 2, it is perhaps not surprising to find that overall they also perform better than the other estimators on the basis of Hausdorff distance. Interestingly enough, whereas the performance of the QMLE ($J^P$) estimator parallels that of the QMLE ($G^P$) estimator reasonably closely, the QMLE ($F^L$) estimator performs badly and
Figure 3. Hausdorff distances between the the estimated ATE identified set and the true ATE identified set for different estimation methods when $\rho_1 = -0.5$ and $\rho_2 = 0.5$ ($R_{VZ}^2 = 0.208$). Cells ordered by relative frequency of occurrence, most frequent to least frequent from left to right.

exhibits a performance between that of the raw nonparametric estimator and the smooth nonparametric estimator. This latter feature demonstrates the broad range in performance of the QMLE estimator based upon different copulae compared to the much more stable superior performance of the QMLE ($GP$) estimator.

REMARK: Our additional experiments show that overall, out of all copula based QMLE estimators from the class of models considered in Han and Vytlacil (2017), the QMLE ($JP$) estimator was the second best after the best performing QMLE ($GP$) estimator, and the QMLE ($FL$) estimator was the worst performing. This ranking was not changed by the local-QMLE estimator, whose performance was generally very similar to that produced by the original model on which it was based, as seen with the local-QMLE ($GP$) and QMLE ($GP$) estimators.
Among the estimators examined here, the raw nonparametric estimator clearly exhibited the worst performance and our experiments demonstrate the practical issues associated with using the raw nonparametric method in finite sample empirical applications. The extent to which this may be a function of the sparsity issue associated with the raw nonparametric method is indicated in Table 3, which shows the percentage of times out of the 88000 cell-replications that the raw nonparametric method broke down due to data paucity. From this table we can see that when \( n = 1000 \), in excess of 96% of the replications and cells resulted in data for which the raw nonparametric estimator could not be used to construct the ATE bounds, and when \( n = 10000 \) this figure still exceeded 50%. Such numbers are obviously related to the absence of green pentagrams (stars) seen in Figure 3(a), and the increase in Hausdorff distance of the raw nonparametric estimator as the cell frequency decreases as seen in Figure 3(b). Somewhat counter intuitively, the number of times the raw nonparametric method breaks down appears to increase slightly as the instrument strength increases, irrespective of sample size, an issue to which we will now turn.

### 5.2 The Impact of Instrument Strength

From the expressions in (8) and (9) it is clear that the ATE bounds are the results of intersecting sets over all possible instrument values for \( Z \), and that variations in the value of the instruments will impact upon the ATE identified set via their influence on the underlying probabilities. The ATE bounds are thus related to the support and strength of the instruments, and our aim here is to investigate the impact of these on the size and estimation of the ATE identified set.

In Figure 4 we plot the true ATE bounds and the average values of the estimates over...
Figure 4. True and estimated ATE upper and lower bounds, arranged in descending order from left to right according to the width of the true ATE identified set, sample size \( n = 10000 \). For each subplot, the \( R^2 \) values are 0.1, 0.4, and 0.7 from left to right.

1000 replications obtained from the QMLE \( (G^P) \) and local-QMLE \( (G^P) \), and the raw nonparametric and smoothed nonparametric, estimators. The 88 X cells are sorted in decreasing order according to the width of the true ATE identified set, so that the left most cell has the widest theoretical interval and the right most cell the narrowest. The true ATE bounds are marked with navy diamonds and the estimates are marked with cyan squares. The actual estimates of the bounds from all 1000 replications are depicted in green and pink dots for the upper and lower bound estimates respectively. For each of the four estimators, the three subplots, from left to right, correspond to \( R^2 \) values of
0.1, 0.4, and 0.7, representing increasing IV strength.

The first obvious observation to make of Figure 4 is that as expected the widths of the true bounds and their estimates decrease as the IV strength increases for all $X$ cells. Secondly, the average upper and lower ATE bound estimates are in close accord with the true values across all $X$ cells for the QMLE ($G_p$) and local-QMLE ($G_p$) estimators without obvious biases, but there appear to be systematic biases for the two nonparametric estimators, especially when the instruments are not strong ($R_{VZ}^2 = 0.1, 0.4$). From a comparison of Figure 4(a) and Figure 4(b), with Figure 4(c) and Figure 4(d), it is also apparent that the spread of the ATE upper and lower bound estimates over the 1000 replications (the green and pink dots) is much larger for the nonparametric estimation methods than that for the QMLE ($G_p$) and local-QMLE ($G_p$) estimators, uniformly across all $X$ cells and at each level of instrument strength. The impact of data paucity is also seen in Figure 4(d), with the raw nonparametric estimator having fewer available estimates as the instrument strength increases.

![Figure 5. Hausdorff distance between different estimators and the true bound of ATE for sample size $N = 10000$](image)

In order to better understand the later paradoxical result, and noting from Figure 4 that the behaviour of all the estimators is roughly invariant with respect to the width of the ATE identified set, in Figure 5 we graph the average value of the Hausdorff distance between the true and the estimated ATE identified set when averaged over all observations and replications, plotted as a function of $R_{VZ}^2$. As can be seen from the figure, the distance...
between the QMLE \((G^P)\) and local-QMLE \((G^P)\), as well as the QMLE \((J^P)\), estimates and the true ATE identified set is always significantly smaller than the Hausdorff distance for the smoothed nonparametric estimator and QMLE \((F^L)\), uniformly in \(\mathcal{R}^2_{VZ}\). The raw nonparametric estimator is clearly dominated by all the other estimators, and whereas the Hausdorff distance decreases slightly as the instrument strength increases for the other estimators, as \(\mathcal{R}^2_{VZ}\) increases beyond 0.6 the distance between the raw nonparametric ATE identified set estimate and the true set increases rapidly.

A detailed examination of the simulation results reveals that the relative frequencies observed in individual \(X\) cells exhibit large fluctuations. This is prevalent irrespective of the sample size or the true ATE identified set width. Even with sample sizes as big as \(n = 10000\) there can be as few as 3 observations in some cells, and the probabilities required to estimate the ATE bounds could not be calculated for all \(X\) cells on at least 7% of occasions. When \(\mathcal{R}^2_{VZ} = 0.9\) almost all lower bound estimates were at or near 0. The reason for this appears to be that in order to obtain a large value of \(\mathcal{R}^2_{VZ}\) the correlation coefficients \(\rho_1\) and \(\rho_2\) must be large. Consequently, the conditional probabilities \(P(D|X, Z)\), \(P(Y|X, Z)\) and \(P(Y \cap D|X, Z)\) varied significantly for different combinations of \(X\) and \(Z\), with the joint probability often forced thereby to be small. As a result, the frequency of occurrence of some combinations can be very small, resulting in a situation not dissimilar from unbalanced grouped data. Very few observations fell into the cell corresponding to young females with low educational status (women aged between 19 and 24 who did not complete year twelve education), for example, and when \(N = 1000\) there were replications where such cells contained no observations. The sample conditional probability estimates for such cells do not reflect their true values and the precision of the raw nonparametric estimator is correspondingly decreased, resulting in large fluctuations in the estimates of the ATE bounds, and hence the increase in the metric. This highlights the unstable nature of the raw nonparametric estimator relative to the other methods.

Localised paucity of observations can be anticipated to impact upon the precision of the nonparametric estimators, and the local-QMLE estimators, more dramatically than it will the QMLE estimators. Figure 5 suggests, however, that although the smooth nonparametric estimator is not effected in the same manner as the raw non-parametric estimator, it is not capable of matching the the performance of the local-QMLE estimator when applied using common bandwidths. Indeed, the performance of the local-QMLE estimator is very similar to that of the QMLE itself. Our additional experiments show that, when
the sample size was increased beyond 10000, the precision of the QMLE and local-QMLE estimators improved dramatically relative to the improvement seen with the nonparametric estimators. And at all sample sizes considered the ranking of the estimators did not change.\(^4\)

6 Alternative Treatment Effect Bounds

Heretofore we have used the structural equation model of Chesher (2005, 2010) as the vehicle with which to present our results, both as a framework for our theoretical development, and as the basis for our numerical algorithms and evaluations. It seems natural therefore at this point to pose the question of whether our results and conclusions can be applied to other models and definitions of the \(ATE\) identified set in the literature that are derived under different assumptions. Bearing in mind the law of decreasing credibility, the extent to which our findings are applicable depends on the extent to which the Chesher structural equation model encompasses other specifications that incorporate alternative or more restrictive assumptions underlying the \(ATE\) bounds.

Consider, for example, the model discussed in Shaikh and Vytlacil (2005). Both the outcome variable \(Y\) and the treatment variable \(D\) are determined by threshold crossing rules with nonlinear index equations with a separable error:

\[
\begin{align*}
Y^* &= r(X, D) - V, & Y &= \mathbb{I}\{Y^* \geq 0\}; \\
D^* &= s(X, Z) - E, & D &= \mathbb{I}\{D^* \geq 0\}.
\end{align*}
\]

(21)

Both \(X\) and \(Z\) are assumed to have compact support, and to be exogenous in the sense that \((X, Z) \perp (V, E)\). The distribution of \((V, E)\) is assumed to have a strictly positive density with respect to Lebesgue measure on \(R^2\). The functions \(r(\cdot, \cdot)\) and \(s(\cdot, \cdot)\) are assumed to be continuous, and \(s(X, \cdot)\) is non-degenerate given \(X\). From this model Shaikh and

\(^4\)When the instruments were strong local data scarcity was still a problem for the raw nonparametric estimator despite substantial increases in sample size. Even for a sample size as large as \(n = 50000\) it was still possible for there not to be enough observations in all of the \(X\) cells to reliably estimate the required probabilities.

25
Vytlacil (2005) derive the ATE identified set as

\[
\max_{z \in \Omega_Z} P(\{Y = 1\} | x, z) - \min_{z \in \Omega_Z} P(\{Y = 1\} | x, z),
\]

\[
\min_{z \in \Omega_Z} \left\{ P(\{Y = 1, D = 1\} | x, z) + P(\{D = 0\} | x, z) \right\} - \max_{z \in \Omega_Z} P(\{Y = 1, D = 0\} | x, z) \]

(22) for a positive ATE and

\[
\max_{z \in \Omega_Z} P(\{Y = 1, D = 1\} | x, z) - \min_{z \in \Omega_Z} \left\{ P(\{Y = 1, D = 0\} | x, z) + P(\{D = 1\} | x, z) \right\},
\]

\[
\min_{z \in \Omega_Z} P(\{Y = 1\} | x, z) - \max_{z \in \Omega_Z} P(\{Y = 1\} | x, z) \]

(23) for a negative ATE.

It is obvious that the RBVP model in (10) represents a special case of the Shaikh-Vytlacil model in (21). Given that the RBVP model corresponds to a structural equation model that has been augmented with a reduced form equation for the treatment variable, we can anticipate that the formulation and assumptions of the Shaikh-Vytlacil model will be either contained within or implied by the Chesher structural model as outlined in Section 1 and Section 2. That the former is the case follows on noting that the Shaikh-Vytlacil threshold crossing specification for the outcome variable, namely, \(Y^* = r(X, D) - V\), \(Y = I\{Y^* \geq 0\}\), is equivalent to

\[
Y = I(r(X, D) \geq V)
\]

\[= I(h(X, D) \geq U) \quad (24)\]

where \(h(X, D) = F_V(r(X, D))\) and \(F_V(\cdot)\) denotes the unspecified marginal probability distribution of \(V\), and \(P(U \leq u | Z = z) = u\) for all \(u \in (0, 1)\) and all \(z \in \Omega_Z\).

Thus it is perhaps not too surprising to find that there is a direct correspondence between the ATE bounds derived by Shaikh and Vytlacil (2005) and those of Chesher (2005, 2010). Take the case of a positive ATE for example. The lower limit of the Chesher identified set
is
\[
\sup_{z \in \Omega_{y}} \{ P(Y = 0|X, z) \} - \inf_{z \in \Omega_{y}} \{ P(Y = 0|X, z) \} \\
= \sup_{z \in \Omega_{y}} \{ 1 - P(Y = 1|X, z) \} - \inf_{z \in \Omega_{y}} \{ 1 - P(Y = 1|X, z) \} \\
= 1 - \inf_{z \in \Omega_{y}} \{ P(Y = 1|X, z) \} - \left\{ 1 - \sup_{z \in \Omega_{y}} \{ P(Y = 1|X, z) \} \right\} \\
= \sup_{z \in \Omega_{y}} \{ P(Y = 1|X, z) \} - \inf_{z \in \Omega_{y}} \{ P(Y = 1|X, z) \} 
\]

The upper limit of the Chesher identified set is
\[
\inf_{z \in \Omega_{y}} \{ P(D = 1|X, z) + P(Y = 0, D = 0|X, z) \} - \sup_{z \in \Omega_{y}} \{ P(Y = 0, D = 1|X, z) \} \\
= \inf_{z \in \Omega_{y}} \{ 1 - P(Y = 1, D = 0|X, z) \} - \sup_{z \in \Omega_{y}} \{ 1 - P(Y = 1, D = 1|X, z) - P(D = 0|X, z) \} \\
= 1 - \sup_{z \in \Omega_{y}} \{ P(Y = 1, D = 0|X, z) \} - \left\{ 1 - \inf_{z \in \Omega_{y}} \{ P(Y = 1, D = 1|X, z) + P(D = 0|X, z) \} \right\} \\
= \inf_{z \in \Omega_{y}} \{ P(Y = 1, D = 1|X, z) + P(D = 0|X, z) \} - \sup_{z \in \Omega_{y}} \{ P(Y = 1, D = 0|X, z) \} 
\]

The reformulations in Equation (25) and Equation (26) clearly reproduce the Shaikh-Vytlacil identified set in Equation (22). Similarly, in the case of a negative ATE the upper limit of the Chesher identified set is
\[
\inf_{z \in \Omega_{y}} \{ P(Y = 0|X, z) \} - \sup_{z \in \Omega_{y}} \{ P(Y = 0|X, z) \} \\
= \inf_{z \in \Omega_{y}} \{ 1 - P(Y = 1|X, z) \} - \sup_{z \in \Omega_{y}} \{ 1 - P(Y = 1|X, z) \} \\
= 1 - \sup_{z \in \Omega_{y}} \{ P(Y = 1|X, z) \} - \left\{ 1 - \inf_{z \in \Omega_{y}} \{ P(Y = 1|X, z) \} \right\} \\
= \inf_{z \in \Omega_{y}} \{ P(Y = 1|X, z) \} - \sup_{z \in \Omega_{y}} \{ P(Y = 1|X, z) \} 
\]

and the lower limit of the Chesher identified set for a negative ATE is
\[
\sup_{z \in \Omega_{y}} \{ P(Y = 0, D = 0|X, z) \} - \inf_{z \in \Omega_{y}} \{ P(D = 0|X, z) + P(Y = 0, D = 1|X, z) \} \\
= \sup_{z \in \Omega_{y}} \{ 1 - P(Y = 1, D = 0|X, z) - P(D = 1|X, z) \} - \inf_{z \in \Omega_{y}} \{ 1 - P(Y = 1, D = 1|X, z) \} \\
= 1 - \inf_{z \in \Omega_{y}} \{ P(Y = 1, D = 0|X, z) + P(D = 1|X, z) \} - \left\{ 1 - \sup_{z \in \Omega_{y}} \{ P(Y = 1, D = 1|X, z) \} \right\} \\
= \sup_{z \in \Omega_{y}} \{ P(Y = 1, D = 1|X, z) \} - \inf_{z \in \Omega_{y}} \{ P(Y = 1, D = 0|X, z) + P(D = 1|X, z) \} 
\]

(28)
Equations (27) and (28) yield the Shaikh-Vytlacil identified set in Equation (23). Thus we have established that the Shaikh-Vytlacil (2005??) identified sets and bounds for the ATE are equivalent to the identified sets and bounds of Chesher (2005, 2010), and we have thereby completed what amounts to a proof of the following proposition.

**Proposition 1** The Shaikh-Vytlacil threshold crossing model (Shaikh and Vytlacil, 2005) is encompassed by the structural equation model of Chesher (Chesher, 2005, 2010), and the bounds are identical.

Bhattacharya et al. (2012) show that the bounds of Shaikh and Vytlacil (2005) (and those of Manski (1990)) can be narrowed by imposing a positive quadrant dependence assumption on the joint distribution of $E$ and $V$. Under their positive quadrant condition Bhattacharya et al. (2012) established that the ATE bounds of Shaikh and Vytlacil (2005) can be improved (i.e. the ATE identified set narrowed) to give a lower bound of

$$\max_{z \in \Omega_Z} P(\{Y = 1, D = 1\} | x, z) - \min_{z \in \Omega_Z} \{P(\{Y = 1, D = 0\} | x, z) + P(\{D = 1\} | x, z)\}$$

(29)

and an upper bound of

$$\min_{z \in \Omega_Z} \{P(\{Y = 1, D = 1\} | x, z) + P(\{D = 0\} | x, z) \cdot p_{\min}(1, x, z)\} - \max_{z \in \Omega_Z} \{P(\{Y = 1, D = 0\} | x, z) + P(\{D = 1\} | x, z) \cdot p_{\max}(1, x, z)\},$$

(30)

where

$$p_{\min}(y, x, z) = \min\{P(\{Y = y\} | D = 1, x, z), P(\{Y = y\} | D = 0, x, z)\} \quad \text{and}$$

$$p_{\max}(y, x, z) = \max\{P(\{Y = y\} | D = 1, x, z), P(\{Y = y\} | D = 0, x, z)\},$$

for a negative ATE, and a lower bound of

$$\max_{z \in \Omega_Z} \{P(Y = 1 | x, z)\} - \min_{z \in \Omega_Z} \{P(Y = 1 | x, z)\}$$

(31)

and an upper bound of

$$\min_{z \in \Omega_Z} \{P(Y = 1 | D = 1, x, z)\} - \max_{z \in \Omega_Z} \{P(Y = 1 | D = 0, x, z)\}$$

(32)

for a positive ATE.
Manski and Pepper (2000) showed that the bounds of Manski (1990) can also be improved using a monotone treatment selection assumption. Monotone treatment selection states that 
\[ \mathbb{E}[Y_0 | D = d, X = x] \text{ and } \mathbb{E}[Y_1 | D = d, X = x] \] are weakly increasing in the realised treatment \( d \), implying that individuals who are treated have a mean response function not less than that of individuals who are not treated. In Appendix A we establish that if the monotone treatment selection assumption is imposed upon the structural equation model of Chesher (2005, 2010) then the bounds for a negative ATE can be modified to give lower and upper bounds of

\[
\sup_{z \in \Omega} P(\{Y = 0, D = 0\} | x, z) - \inf_{z \in \Omega} \{P(\{Y = 0, D = 1\} | x, z) + P(\{D = 0\} | x, z)\}
\]

and

\[
\inf_{z \in \Omega} \{P(\{Y = 0, D = 0\} | x, z) + P(\{D = 1\} | x, z) \cdot p_{\min}(0, x, z)\} - \sup_{z \in \Omega} \{P(\{Y = 0, D = 1\} | x, z) + P(\{D = 0\} | x, z) \cdot p_{\max}(0, x, z)\}
\]

respectively, and lower and upper bounds of

\[
\sup_{z \in \Omega} \{P(\{Y = 0\} | x, z)\} - \inf_{z \in \Omega} \{P(\{Y = 0\} | x, z)\}
\]

and

\[
\inf_{z \in \Omega} P(\{Y = 0\} | D = 0, x, z) - \sup_{z \in \Omega} P(\{Y = 0\} | D = 1, x, z)
\]

respectively for a positive ATE.

**Proposition 2** The ATE bounds generated by the Bhattacharya-Shaikh-Vytlacil positive quadrant dependent threshold crossing model (Bhattacharya et al., 2012) are identical to the ATE bounds obtained when the structural equation model of Chesher (Chesher, 2005, 2010) is augmented with a monotone treatment selection assumption.

That the bounds presented in equations (29) - (32) are equivalent to those presented in equations (33) - (36) is proved in Appendix A.

Additional simulation results not reported in detail here illustrate the numerical equivalence of the different ATE bounds as indicated in Propositions 1 and 2. We can therefore infer that the previous conclusions concerning the relative merits of the different estimators that were obtained using the structural equation framework of Chesher will hold good
for these and other models considered in the literature that are equivalent. Additionally, as other ATE bounds for the binary outcome and binary treatment models are also expressed in terms of similar conditional probabilities, it is reasonable to expect that similar results will hold for other bounds. In particular, it seems reasonable to conjecture that the performance of the QMLE \((G^P)\) and local-QMLE \((G^P)\) estimators will prove to be significantly superior to that of non-parametric estimators in a broad range of circumstances.

7 Summary

Recent developments on the analysis of treatment effects indicate that ATE partially identified sets can be characterised by constructing bounds on different probability threshold functions. Such bounds are determined by various conditional probabilities and it is natural to consider estimating these probabilities using nonparametric methods. Whilst nonparametric methods generally impose few assumptions upon the DGP, numerical constraints and paucity of data can limit their application. Whereas parametric methods often impose specific assumptions, they can be less prohibitive in the sense of demanding less computational effort and not requiring large data sets. In this paper we have used Monte Carlo experimentation to compare the finite sample performance of raw nonparametric and smoothed nonparametric estimates of ATE bounds with those obtained from QMLE and local-QMLE estimators based upon the RBVP model and other members of the family of triangular two equation systems for binary endogenous variables where the latent error terms of the RBVP model are generalised through the use of monotone regression dependent parametric copulae.

Our simulation results indicate that on the basis of bias, variance and mean squared error, the performance of the RBVP QMLE and local-QMLE estimators was significantly better than that of the nonparametric methods for sample sizes \(n < 10000\). The values of these conventional performance measures for the RBVP QMLE and local-QMLE estimates of the ATE bounds were significantly less than those observed for the nonparametric methods. For example, the RMSEs for the raw nonparametric method for the ATE bounds were four to ten times larger than the RMSEs from the QMLE and local-QMLE estimates based on RBVP. Our results also show that the QMLE estimators based on other members of the family of triangular two equation systems can perform either very poorly or reasonably well depending on the specific choice of the copula, but the RBVP or QMLE
(G^P) based estimator consistently out performed all other QMLE candidates.

Instrument strength plays an important role in determining the size of the ATE identified set, and the Hausdorff distance between the RBVP based QMLE and local-QMLE estimates and the true ATE identified set was relatively small uniformly in R^2_{VZ} (our measure of instrument strength), and they are around one half of the Hausdorff distance of the raw nonparametric estimator. The raw nonparametric estimator was clearly dominated by the other estimators, and the distance between the raw nonparametric estimate of the ATE identified set and the true set increased as R^2_{VZ} increased beyond 0.6, reflecting that the raw nonparametric estimator can be rendered infeasible due to data paucity when the instruments are strong and the sample size is not overly large. The differences between the nonparametric methods and the QMLE and local-QMLE estimators declined slowly as sample size n was increased, but in the simulations examined here the nonparametric methods were never able to outperform either the RBVP parametric or the RBVP semiparametric estimator even for sample sizes n as large as 50000.

Unlike traditional simulation exercises our results are based on generating data from a generic probabilistic structure derived from population representative ANHS data. Consequently, our experimental design imposes no parametric or structural restrictions on the experimental DGP and the generated data. Given that the QMLE and local-QMLE estimators are based on a two equation triangular system for binary endogenous variables where the probability structure of the latent error terms is characterised as a member of the monotone regression dependent parametric copulae family considered in Han and Vytlacil (2017), and the underlying assumptions of such a model – including a parametric distributional form, a threshold crossing rule for the binary variables, and a separable error structure – are unlikely to be true of real data, the superior performance of the RBVP based QMLE and local-QMLE estimators as compared to the raw nonparametric estimator observed here might be viewed as surprising. Such results are in accord with the findings of Li et al. (2017) however.

In an analysis of the performance of the QMLE derived from such models Li et al. (2017) showed that ATE estimates based upon the miss-specified RBVP model offer the practitioner a sensible choice as they generally out perform other QMLE estimators. Li et al. (2017) also demonstrated the existence of compensating effects between parameter estimates and estimates of parametric functions such as the predicted probabilities and the ATE. In particular, they showed that the QMLE (G^P) estimator generates pseudo-true
parameter values that yield estimates of the $ATE$ that lie alongside the value determined by the true DGP with 95% confidence intervals that fall within the true $ATE$ bounds. Their results suggest that when estimating treatment effects the concepts and ideas of partial identification can be critical to the analysis of such measures, and that practitioners should take steps to estimate the $ATE$ bounds.

Finally, we have shown that although our results concerning the relative merits of the QMLE and local-QMLE estimators and non-parametric methods were obtained using the structural equation framework of Chesher (2005, 2010), they will hold good for other specifications considered in the literature, and it seems reasonable to suppose that the RBVP based QMLE and local-QMLE estimators will exhibit strong performance for a broad range of partially identified models and data sets. These findings may offer some comfort to applied workers. The RBVP model is commonly employed in situations where the response variable of interest is a dichotomous indicator and the determinants of the probable outcome includes qualitative information in the form of an endogenous dummy or treatment variable. It is frequently used by empirical researchers in policy evaluations because it allows for the straightforward estimation of the $ATE$. Our results suggest that use of the RBVP model, i.e. the QMLE ($G^P$) estimator, represents a conservative but rational choice for the practitioner in the estimation of treatment effects and $ATE$ bounds.

References


A Proof of Proposition 2

A.1 Chesher Bounds with Monotone Treatment Selection

To begin, substitute
\[ P(\{ Y = 0, D = d \}|X, Z) = P(\{ Y = 0 \}|D = d, X, Z)P(\{ D = d \}|X, Z) \]
for \( d = 0, 1 \), into (6) and (7) and rewrite the bounds on the probability threshold functions as
\[ P(\{ Y = 0, D = 0 \}|X, Z) \leq p(0, X) \leq P(\{ Y = 0 \}|D = 1, X, Z)P(\{ D = 1 \}|X, Z) + P(\{ Y = 0, D = 0 \}|X, Z) \]
(A.1)

and
\[ P(\{ Y = 0, D = 1 \}|X, Z) + P(\{ Y = 0 \}|D = 0, X, Z)P(\{ D = 0 \}|X, Z) \leq p(1, X) \leq P(\{ D = 0 \}|X, Z) + P(\{ Y = 0, D = 1 \}|X, Z). \]
(A.2)

Monotone treatment selection implies that
\[ P(\{ Y = 0 \}|D = 0, X, Z) \geq P(\{ Y = 0 \}|D = 1, X, Z), \]
(A.3)

and a comparison of the terms in expressions (A.1) and (A.2) with (A.3) leads to the conclusion that under monotone treatment selection the probability threshold function bounds can be reformulated as
\[ P(\{ Y = 0, D = 0 \}|X, Z) \leq p(0, X) \leq p_{\min}(0, X, Z) \cdot P(\{ D = 1 \}|X, Z) + P(\{ Y = 0, D = 0 \}|X, Z) \]
(A.4)

where \( p_{\min}(0, X, Z) = \min\{ P(\{ Y = 0 \}|D = 0, X, Z), P(\{ Y = 0 \}|D = 1, X, Z)\} \), and
\[ P(\{ Y = 0, D = 1 \}|X, Z) + p_{\max}(0, X, Z) \cdot P(\{ D = 0 \}|X, Z) \leq p(1, X) \leq P(\{ D = 0 \}|X, Z) + P(\{ Y = 0, D = 1 \}|X, Z). \]
(A.5)

where \( p_{\max}(0, X, Z) = \max\{ P(\{ Y = 0 \}|D = 0, X, Z), P(\{ Y = 0 \}|D = 1, X, Z)\} \).

Appropriate application of lower and upper bounds with respect to \( z \in \Omega_Z \) on the left
and right hand sides of (A.4) and (A.5) now yields the Chesher monotone treatment selection (Chesher-MTS) bounds as in equations (33) and (34), or equations (35) and (36), as required.

A.2 Chesher-MTS and BSV Bound Equivalence

To establish the equivalence of the Chesher-MTS bounds to the Bhattacharya-Shaikh-Vytalacil (BSV) bounds, first consider the case of a positive ATE. The Chesher-MTS lower bound in (35) is

\[
\sup_{z \in \Omega} \{ P(\{Y = 0\}|x, z) \} - \inf_{z \in \Omega} \{ P(\{Y = 0\}|x, z) \} \\
= \sup_{z \in \Omega} \{ 1 - P(\{Y = 1\}|x, z) \} - \inf_{z \in \Omega} \{ 1 - P(\{Y = 1\}|x, z) \} \\
= 1 - \inf_{z \in \Omega} \{ P(\{Y = 1\}|x, z) \} - \left\{ 1 - \sup_{z \in \Omega} \{ P(\{Y = 1\}|x, z) \} \right\} \\
= \sup_{z \in \Omega} \{ P(\{Y = 1\}|x, z) \} - \inf_{z \in \Omega} \{ P(\{Y = 1\}|x, z) \},
\]

and the Chesher-MTS upper bound in (36) is

\[
\inf_{z \in \Omega} P(\{Y = 0\}|D = 0, x, z) - \sup_{z \in \Omega} P(\{Y = 0\}|D = 1, x, z) \\
= \inf_{z \in \Omega} \{ 1 - P(\{Y = 1\}|D = 0, x, z) \} - \sup_{z \in \Omega} \{ 1 - P(\{Y = 1\}|D = 1, x, z) \} \\
= \{ 1 - \sup_{z \in \Omega} P(\{Y = 1\}|D = 0, x, z) \} - \left\{ 1 - \inf_{z \in \Omega} \{ P(\{Y = 1\}|D = 1, x, z) \} \right\} \\
= \inf_{z \in \Omega} P(\{Y = 1\}|D = 1, x, z) - \sup_{z \in \Omega} P(\{Y = 1\}|D = 0, x, z).
\]

The Chesher-MTS lower and upper bounds as reexpressed in (A.6) and (A.7) clearly reproduce the BSV bounds in (31) and (32).

Now consider the case of a negative ATE. If \(p_{\text{max}}(0, x, z) = P(\{Y = 0\}|D = 0, x, z)\) then \(p_{\text{min}}(0, x, z) = P(\{Y = 0\}|D = 1, x, z)\), obviously, and the Chesher-MTS bounds in (33) and (34) are exactly the same as those in the Chesher unconstrained ATE identified.
therefore follows directly from Proposition 1. The equivalence between the BSV and the Chesher-MTS bounds in (29) and (30) are the same as those in the Shaikh-Vytlacil unconstrained ATE set in (8). From the equality

\[
\begin{align*}
\min & \{P(\{Y = 1 - y\}|D = 1, x, z), P(\{Y = 1 - y\}|D = 0, x, z)\} \\
& = \min\{1 - P(\{Y = y\}|D = 1, x, z), 1 - P(\{Y = y\}|D = 0, x, z)\} \\
& = 1 - \max\{P(\{Y = y\}|D = 1, x, z), P(\{Y = y\}|D = 0, x, z)\} \\
& = 1 - p_{\text{Max}}(y, x, z).
\end{align*}
\]

Equation (A.8)

it also follows that \(p_{\text{Max}}(1, x, z) = P(\{Y = 1\}|D = 1, x, z)\) and \(p_{\text{Min}}(1, x, z) = P(\{Y = 1\}|D = 0, x, z)\) when \(p_{\text{Max}}(0, x, z) = P(\{Y = 0\}|D = 0, x, z)\). But when \(p_{\text{Max}}(1, x, z) = P(\{Y = 1\}|D = 1, x, z)\) and \(p_{\text{Min}}(1, x, z) = P(\{Y = 1\}|D = 0, x, z)\), Equation (30) reduces to \(\min_{z \in \Omega_Z} P(\{Y = 1\}|x, z) - \max_{z \in \Omega_Z} p(\{Y = 1\}|x, z)\) and the BSV bounds in (29) and (30) are the same as those in the Shaikh-Vytlacil unconstrained ATE identified set in (23). The equivalence between the BSV and the Chesher-MTS bounds therefore follows directly from Proposition 1.

If the ATE is negative and \(p_{\text{Min}}(0, x, z) = P(\{Y = 0\}|D = 1, x, z)\) – which means that \(p_{\text{Min}}(0, x, z) = P(\{Y = 0\}|D = 0, x, z)\) and via Equation (A.8) that \(p_{\text{Max}}(1, x, z) = P(\{Y = 1\}|D = 1, x, z)\), and \(p_{\text{Min}}(1, x, z) = P(\{Y = 1\}|D = 0, x, z)\) – then for the Chesher-MTS lower bound in (33) we have

\[
\begin{align*}
\sup_{z \in \Omega_Z} P(\{Y = 0, D = 0\}|x, z) - \inf_{z \in \Omega_Z} \{P(\{Y = 0, D = 1\}|x, z) + P(\{D = 0\}|x, z)\} \\
& = \sup_{z \in \Omega_Z} \left\{1 - P(\{Y = 1, D = 0\}|x, z) - P(\{D = 1\}|x, z)\right\} - \inf_{z \in \Omega_Z} \left\{1 - P(\{Y = 1, D = 1\}|x, z)\right\} \\
& = 1 - \inf_{z \in \Omega_Z} \left\{P(\{Y = 1, D = 0\}|x, z) + P(\{D = 1\}|x, z)\right\} - \left\{1 - \sup_{z \in \Omega_Z} P(\{Y = 1, D = 1\}|x, z)\right\} \\
& = \sup_{z \in \Omega_Z} P(\{Y = 1, D = 1\}|x, z) - \inf_{z \in \Omega_Z} \left\{P(\{Y = 1, D = 0\}|x, z) + P(\{D = 1\}|x, z)\right\},
\end{align*}
\]

(A.9)
and the Chesher-MTS upper bound in (34) can be rewritten as

\[
\inf_{z \in \Omega_Z} \{ P(\{Y = 1, D = 1\}|x, z) + P(\{D = 0\}|x, z) \cdot P(\{Y = 1\}|D = 1, x, z) \} \\
- \sup_{z \in \Omega_Z} \{ P(\{Y = 1, D = 0\}|x, z) + P(\{D = 1\}|x, z) \cdot P(\{Y = 1\}|D = 0, x, z) \}
\]

\[
= \inf_{z \in \Omega_Z} \{ P(\{Y = 0\}|D = 0, x, z) - \sup_{z \in \Omega_Z} P(\{Y = 0\}|D = 1, x, z) \}
\]

\[
= \inf_{z \in \Omega_Z} P(\{Y = 1\}|D = 1, x, z) - \sup_{z \in \Omega_Z} P(\{Y = 1\}|D = 0, x, z).
\]

(A.10)

Now, the Chesher-MTS lower bound as expressed in (A.9) equals the BSV lower bound in (29), and the BSV upper bound in (30) can be reformulated as

\[
\min_{z \in \Omega_Z} \{ P(\{Y = 1, D = 1\}|x, z) + P(\{D = 0\}|x, z) \cdot P(\{Y = 1\}|D = 1, x, z) \}
\]

\[
- \max_{z \in \Omega_Z} \{ P(\{Y = 1, D = 0\}|x, z) + P(\{D = 1\}|x, z) \cdot P(\{Y = 1\}|D = 0, x, z) \}
\]

\[
= \min_{z \in \Omega_Z} \{ P(\{Y = 1\}|x, z) + P(\{D = 0\}|x, z) \cdot [P(\{Y = 1\}|D = 1, x, z) - P(\{Y = 1\}|D = 0, x, z)] \}
\]

\[
- \max_{z \in \Omega_Z} \{ P(\{Y = 1\}|x, z) + P(\{D = 1\}|x, z) \cdot [P(\{Y = 1\}|D = 0, x, z) - P(\{Y = 1\}|D = 1, x, z)] \}
\]

\[
= \min_{z \in \Omega_Z} P(\{Y = 1\}|D = 1, x, z) - \max_{z \in \Omega_Z} P(\{Y = 1\}|D = 0, x, z),
\]

(A.11)

which is the same as the Chesher-MTS lower bound as expressed in (A.10).

Equality between the BSV bounds and the Chesher-MTS bounds in all possible cases has thus been established, and the proof of the equivalence as stated in Proposition 2 is thereby completed.