THE CENTRE FOR MARKET AND PUBLIC ORGANISATION

Instrumental Variable Estimators for Binary Outcomes*

Paul Clarke and Frank Windmeijer

June 2010

Working Paper No. 10/239 *(update of 09/209)

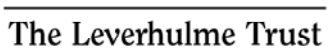
Centre for Market and Public Organisation
Bristol Institute of Public Affairs
University of Bristol
2 Priory Road
Bristol BS8 1TX
http://www.bristol.ac.uk/cmpo/

Tel: (0117) 33 10799
Fax: (0117) 33 10705
E-mail: cmpo-office@bristol.ac.uk

The Centre for Market and Public Organisation (CMPO) is a leading research centre, combining expertise in economics, geography and law. Our objective is to study the intersection between the public and private sectors of the economy, and in particular to understand the right way to organise and deliver public services. The Centre aims to develop research, contribute to the public debate and inform policy-making.

CMPO, now an ESRC Research Centre was established in 1998 with two large grants from The Leverhulme Trust. In 2004 we were awarded ESRC Research Centre status, and CMPO now combines core funding from both the ESRC and the Trust.

ISSN 1473-625X



CMPO Working Paper Series No. 10/239

Instrumental Variable Estimators for Binary Outcomes*

Paul Clarke[1]
and
Frank Windmeijer[2]

[1]*CMPO, University of Bristol, UK*
[2]*Department of Economics and CMPO, University of Bristol, UK*

June 2010

***Please Note: This is a substantially revised version of the original January 2009 paper (09/209)**

Abstract
Instrumental variables (IVs) can be used to construct estimators of exposure effects on the outcomes of studies affected by non-ignorable selection of the exposure. Estimators which fail to adjust for the effects of non-ignorable selection will be biased and inconsistent. Such situations commonly arise in observational studies, but even randomised controlled trials can be affected by non-ignorable participant non-compliance. In this paper, we review IV estimators for studies in which the outcome is binary. Recent work on identification is interpreted using an integrated structural modelling and potential outcomes framework, within which we consider the links between different approaches developed in statistics and econometrics. The implicit assumptions required for bounding causal effects and point-identification by each estimator are highlighted and compared within our framework. Finally, the implications for practice are discussed.

Keywords: bounds, causal inference, generalized method of moments, local average treatment effects, marginal structural models, non-compliance, parameter identification, potential outcomes, structural mean models, structural models.

JEL Classification: C13, C14

Electronic version: www.bristol.ac.uk/cmpo/publications/papers/2010/wp239.pdf

Acknowledgements
This work was funded by UK Economic & Social Research Council grant RES-060-23-0011 and UK Medical Research Council grant G0601625. The authors thank Vanessa Didelez, Roger Harbord, Koen Jochmans, Tom Palmer and Nuala Sheehan for their helpful comments on earlier drafts.

Address for correspondence
CMPO, Bristol Institute of Public Affairs
University of Bristol
2 Priory Road
Bristol BS8 1TX
paul.clarke@bristol.ac.uk, f.windmeijer@bristol.ac.uk
www.bristol.ac.uk/cmpo/

1 Introduction

The estimation of causal exposure effects on study outcomes is almost always complicated by non-random selection of exposure. The problem is well known to affect observational studies, but it also affects randomised controlled trials, which are rarely perfectly conducted and usually affected by issues like participant non-compliance. If the selection mechanism is non-random then inferences based on estimators that fail to adjust for its effects will be misleading. For example, in epidemiology, the impact of non-random selection is termed 'confounding' bias, which arises if confounding variables C associated with outcome Y and exposure X are omitted from the analysis. Exposure selection is ignorable if all the confounding variables C are observed and conditioned on appropriately in the analysis, but selection is non-ignorable if there are unobserved confounding variables (e.g., bias due to 'residual confounding'). In economics, the problem is commonly framed in terms of a regression model from which important regressor variables have been omitted and so become part of the model's error term. In this context, the exposure is termed 'exogenous' if it is not associated with the error, and 'endogenous' if it is, even after conditioning on C.

Instrumental variables are widely used in economics to solve the problems posed by endogenous X, and more generally, those problems arising from non-ignorable selection. An instrumental variable (IV) Z is associated with X but associated with Y only indirectly through its association with X. IVs are also used in disciplines other than economics. For example, there has recently been great interest in the use of IVs based on genetic information to exploit the 'Mendelian randomisation' hypothesis (e.g., Lawlor et al., 2008); and in the analysis of randomised experiments with non-compliance, the IV is the randomisation indicator of the experimental group to which each experimental unit is randomised (e.g., Angrist et al., 1996; Greenland, 2000).

In this paper, we review the problems associated with 'binary IV' estimators, that is,

estimators for the causal effects of exposures on binary outcomes which are based on IVs. It has already been recognised that binary IV estimators cannot identify causal effects without additional assumptions concerning the nature of the data generating process (Chesher, 2010). In two recent papers, extensive simulation studies are used to compare the performance of different binary IV estimators under specific data generating processes (Didelez et al., 2010; Vansteelandt et al., 2010). However, our focus is somewhat different: we use a general causal modelling framework to compare the different IV estimators proposed in the literature, make clear the underlying identification assumptions of each, and explore the links between these estimators. Our survey includes estimators of 'local', or 'complier-specific' causal effects (Imbens and Angrist, 1994), and thus links in with the literature on 'principal stratification' (Fragakis and Rubin, 2002), and we also aim to emphasise the implications of our findings for practitioners looking to apply these methods. While our focus here is on non-ignorable selection, we note that the important problem of measurement error can also be addressed using binary IV estimators (e.g., Carroll et al., 2006; Vansteelandt et al., 2008).

The paper is organised as follows. We start by setting out in Section 2 the framework within which the different estimators are to be assessed. To simplify the presentation and to emphasise concepts, we focus on setting out this framework for the simplest possible scenario with X and Z both binary and no covariates. In Section 3, we explicate the assumptions required to identify causal effects and bounds for these effects. The various estimators are considered in Sections 4-7 where we again consider only the simplest possible scenarios to facilitate a comparison between the identifying assumptions made by each. Finally, in Section 8 we make concluding remarks about recent developments in this area, and make recommendations for practice.

2 Causal Framework

2.1 Study design

It is first helpful to clarify the nature of the studies for which the estimators we consider are appropriate. If we set causal inference as the analytical goal, then we take the ideal study to be a randomised experiment in which randomisation determines the exposure level received by each study unit. However, randomised experiments are not always perfectly conducted or even feasible, and it is for studies falling below the ideal standard that IVs can be used to obtain causal inference.

The first class of studies we consider are called 'encouragement designs'. These are experiments which involve an initial selection stage wherein exposure is randomly assigned to the study units, followed by a second stage in which the study units select whether or not to comply with this assignment; the outcome is measured at some point following selection. More generally, the first stage involves a selection mechanism that is known to be ignorable given pre-study covariates C (Rosenbaum and Rubin, 1983). Special cases of encouragement design impose constraints on stage-two selection. For example, in randomised placebo-controlled trials, those assigned to the control group who non-comply cannot take the active treatment, only a placebo (e.g., Greenland, 2000; Nagelkerke et al., 2000); a more extreme example of this type directly 'forces' compliance among those assigned to the control group by denying access to any treatment, be it the active treatment or a placebo (e.g., Somer and Zeger, 1991).

The second class of studies we refer to simply as 'observational studies'. These can be of cross-sectional or longitudinal surveys of a population or cohort in which X and Y are measured. We follow Rubin (2008) and argue that a prerequisite for causal inference from observational studies is that X can plausibly be conceived as the result of some selection mechanism driven by factors causally antecedent to X. If these factors are known and part of C then causal inference is possible. However, the scenario of interest here is one

where important factors are omitted from C and so causal inference requires the use of IVs.

Choosing IVs for observational studies is far from simple because, as we discuss below, an IV must be associated with exposure *and* independent of the unobserved factors driving selection. Potentially successful strategies involve exploiting 'natural experiments', such as administrative differences between two otherwise homogeneous areas, and the Mendelian randomisation hypothesis that X is a phenotype for a randomly determined genotype Z (e.g., Didelez and Sheehan, 2007). Other study designs that work on similar principles, like regression discontinuity designs (e.g., Imbens and Lemieux, 2008), will not be considered here.

2.2 Structural models

We begin by considering the classical application of IVs from econometrics, namely, estimation of the linear model for the regression of outcome Y on exposure X when the exposure and the model's residual error term are correlated. For illustration, we allow Y to have any measurement scale provided that the linear model

$$Y = \beta_0 + X\beta_1 + U, \tag{1}$$

holds, where U represents the combined contribution of the omitted variables such that $E(U) = 0$ and $\mathrm{Var}(U) = \sigma^2$. Non-ignorable selection in this case results in endogenous X where $\mathrm{Cov}(X, U) \neq 0$. To make causal inferences, we interpret model (1) as structural in the sense that the target parameter β_1 is the *ceterus paribus* effect of X, that is, the effect on Y of a unit change in X if U is held fixed (Goldberger, 1972).

The structural modelling approach involves finding suitable estimators for the model parameters. In this example, the ordinary least squares (OLS) estimator of β_1 is always consistent for $\mathrm{Cov}(Y, X)/\mathrm{Var}(X)$, but is consistent for β_1 only if $\mathrm{Cov}(X, U) = 0$. However, by expanding our data set to include a suitable IV Z, the classical IV estimator

$\widehat{\beta}_1^{IV} = \text{Cov}(Y,Z)/\text{Cov}(X,Z)$ is consistent for β_1 even if $\text{Cov}(X,U) \neq 0$. For binary Z, the classical IV estimator is

$$\widehat{\beta}_1^{IV} = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(X|Z=1) - E(X|Z=0)}. \tag{2}$$

To be a suitable IV, Z must be chosen so that the directed acyclic graph (DAG) for the true joint distribution of Z, U, X and Y satisfies the following constraints:

C1. Independence between the IV and the omitted variables: $Z \perp U$.

C2. Conditional independence of the outcome and IV: $Y \perp Z | X, U$.

Figure 1 displays the DAG corresponding to these constraints. Note that $\perp$ indicates independence between random variables, but for semi-parametric estimators the stochastic independence assumption can somtimes be relaxed to conditional mean independence. To ensure that the denominator of the IV estimator is non-zero, a further requirement is needed:

C3. Causal effect of Z on X: $E(X|Z=1) - E(X|Z=0) \neq 0$ for binary Z.

Didelez and Sheehan (2007) call these three requirements the IV 'core conditions'. Robins (2006) discuss an exception to C3 in which Z is a 'surrogate' IV and there is no arrow between Z and X in Figure 1, but we will assume throughout that C3 holds.

Two-stage least squares (2SLS) is the most widely used IV estimator for linear models (e.g., Wooldridge, 2002, ch. 5). It is a generalisation of (2) to multiple regression models and multiple IVs. In this simple set-up, the two stages of the 2SLS estimator are defined as follows: first, fit the 'reduced-form' linear regression model $E(X|Z) = \alpha_0 + Z\alpha_1$ using OLS to obtain $\widehat{X} = \widehat{\alpha}_0 + Z\widehat{\alpha}_1$; and second, fit $E(Y|\widehat{X}) = \beta_0 + \widehat{X}\beta_1$ using OLS. The 2SLS estimator is consistent (but not unbiased) provided that (1) holds and the IV satisfies core conditions C1-C3. Moreover, there is a certain degree of robustness because the estimator is consistent even if the reduced-form model is mis-specified (e.g., if X is binary).

2.3 Potential outcomes

The IV conditions can also be stated in terms of potential outcomes (e.g., Angrist et al., 1996; Robins and Rotnitzky, 2004). For each individual, define the potential exposures $X(z)$ and potential outcomes $Y(z, x)$ for, respectively, the exposures and outcomes which *would* have been obtained if the exposure had been set to x and the IV to z by external intervention rather than by the true data generating process.

The consistency assumption linking observed and potential outcomes is $Y = Y(Z, X)$ and $X = X(Z)$, which is trivially taken to hold; all the other potential outcomes are unobservable and thus counterfactual. More importantly, the IV must satisfy three conditions:

P1. Independence of the potential outcomes and IV: $X(z), Y(z, x) \perp Z$.

P2. Exclusion restriction: $Y(z, x) = Y(x)$.

P3. Causal effect of IV on exposure: $E\{X(1) - X(0)\} \neq 0$ (for binary Z).

Condition P1 corresponds to independence between the IV and all the potential outcomes and exposures; the exclusion restriction P2 ensures that Z has no direct effect on the potential outcome. Together with condition P1, these conditions ensure that the IV affects the outcome only indirectly through its effect on the exposure.

The linear structural model (1) can be written in terms of potential outcomes as

$$Y(x) = \beta_0 + x\beta_1 + U, \tag{3}$$

where exposure x is set by external intervention and so irrespectively of U, which means that β_1 has the same interpretation as in structural model (1). Clearly, the right-hand side of model (3) satisfies the exclusion restriction, and the conditional independence constraints on U and Z ensure that condition P1 holds. In other words, the original core conditions C1-C3 can be viewed as a special case of conditions P1-P3 as specified within the structural framework.

Model (3) is restrictive because it constrains the exposure effect for each individual to be constant, that is, $Y(1)-Y(0)=\beta_1$. An attraction of the potential outcomes approach is that explicit modelling assumptions like these are not necessary. Instead, inferences are made directly about meaningful expectations of potential outcomes. For binary exposure, an important causal parameter is the average causal effect $\text{ACE} = E\{Y(1)\} - E\{Y(0)\}$, which is sometimes known as the average treatment effect (ATE). Other population causal parameters are the causal risk ratio $\text{CRR} = E\{Y(1)\}/E\{Y(0)\}$ and the causal odds ratio

$$\text{COR} = \frac{E\{Y(1)\}/E\{1-Y(1)\}}{E\{Y(0)\}/E\{1-Y(0)\}};$$

also of interest are causal parameters among the exposed group like the average causal effect among the exposed, $E\{Y(1)-Y(0)|X=1\}$, and covariate-conditional effects like $E\{Y(1)-Y(0)|C\}$. Under model (3) it follows that $\beta_1 = \text{ACE}$, but structural model parameters do not always correspond to causal effects.

A consequence of conditions P1-P3 is the 'randomisation assumption'

$$E\{Y(x)|Z\} = E\{Y(x)\}, \tag{4}$$

which is also known as 'conditional mean independence' (CMI). CMI plays an important role in the identification of causal effects using IV estimators (see Section 3).

2.4 Models for binary outcomes

To link the structural model and potential outcomes approaches, we assume that all the potential outcomes and exposures are the result of an underlying data generating process. In an abstract but intuitive fashion, we can represent this process by the 'generating model'

$$X(z) = f_X(z,V), Y(x) = f_Y(x,U), \tag{5}$$

where U and V are latent random variables (or vectors) representing omitted variables (Clarke and Windmeijer, 2010). Hernán and Robins (2006) refer to (5) as a *non-*

parametric structural equation model, in the sense that no constraints are placed on its unknown form.

In the structural framework, $f_Y(x, U)$ is referred to as the structural model, so to avoid confusion with other frameworks we refer to the $f_X(z, V)$ component as the 'selection' model and the $f_Y(x, U)$ component as the 'scientific' model (Rubin, 2008). The functions f_X and f_Y index the effects of *ceterus paribus* variation in z and x, respectively, with V and U indexing variation between individuals. Throughout we informally assume that every combination of generating model and IV $\{Z, f_X, f_Y, U, V\}$ corresponds to a well-defined distribution for $X(z), Y(x)$ given Z.

Using the generating model notation, we can interchange between the usual structural model and potential outcomes representations. The usual structural model representation follows from the consistency assumption, namely, $X = f_X(Z, V)$ and $Y = f_Y(X, U)$. Clearly, if follows that $E(Y|X, U) = E\{Y(X)|X, U\} = Y(X) = Y$, which emphasises that X and U alone determine each individual's outcome. For binary outcomes, we consider scientific models of the form

$$Y(x) = f_Y(x, U) = I\{f_Y^*(x, U) > 0\}, \tag{6}$$

where I is the indicator function, and f_Y^* is defined on the latent scale. A simple model is

$$Y(x) = I(\beta_0 + x\beta_1 + U > 0), \tag{7}$$

constraining the exposure effect on the latent scale to be constant as in (1).

We can now introduce three important examples of scientific model. First, if U is specified to be a scalar random variable following the standard logistic distribution, then integrating U out of (7) leads to the logistic model

$$E\{Y(x)\} = \text{expit}(\beta_0 + x\beta_1), \tag{8}$$

where $\text{expit}(z) = \exp(z)/\{1 + \exp(z)\}$ is the cumulative distribution function (cdf) of

the standard logistic distribution; this model has the convenient property that $\exp(\beta_1) =$ COR. Second, if U is assumed to follow the standard normal distribution then integrating it out of (7) leads to the probit model

$$E\{Y(x)\} = \Phi(\beta_0 + x\beta_1), \tag{9}$$

where Φ is the cdf of the standard normal distribution. Neither β_0 nor β_1 have obvious interpretations as causal parameters, but the causal parameters can be obtained by construction (e.g., ACE $= \Phi(\beta_0 + \beta_1) - \Phi(\beta_0)$ and CRR $= \Phi(\beta_0 + \beta_1)/\Phi(\beta_0)$).

The last model we consider is based on the latent random vector $U = (U_1, U_2)'$ such that

$$Y(x) = I(\beta_0 + x\beta_1 + U_1 + U_2 > 0), \tag{10}$$

where U_2 represents the effect of those omitted variables which are independent of X, and U_1 represents those variables which are associated with X. If U_2 follows the standard logistic distribution then it can be integrated out to give the 'mixed effects' logistic model

$$E\{Y(x)|U_1\} = \text{expit}(\beta_0 + x\beta_1 + U_1). \tag{11}$$

A mixed effects probit model is similarly obtained. Commonly in mixed effects modelling, U_1 is assumed to be normally distributed, but no such parametric assumption will be made here unless it is explicitly stated.

An important feature of the class of scientific models we consider here is that causally implausible models like

$$Y(x) = I(\beta_0 + x\beta_1 + U_1 > 0) + U_2,$$

are excluded from consideration. The implausibility of this model stems from the support of U_2 needing to depend on x to ensure that $Y(x) \in \{0, 1\}$, but from Figure 1 it is clear that U_2 is causally antecedent to the exposure. Furthermore, the specifc examples (7-11) are all 'symmetric' in that U and X both act on the outcome through the latent scale

in the same way, which is a desirable property if we view U as representing the effects of omitted variables.

The second component of a generating model is the selection model. A simple linear selection model is

$$X(z) = \alpha_0 + z\alpha_1 + V, \tag{12}$$

which is analogous to the reduced-form model for 2SLS. This model has the additional property of being 'monotonic', which is discussed further in Section 7. If X is binary then a more appropriate selection model is

$$X(z) = I(\alpha_0 + z\alpha_1 + V > 0), \tag{13}$$

which is also monotonic.

Finally, to complete specification of the generating model, denote the cdf of (U, V) by F_{uv}. Selection is ignorable only if U and V are independent. To index the dependence of the latent variables, let 'correlation' parameter ρ be a notational device indexing non-zero moments of the joint distribution which involve cross-products of U^k and V^k $(k = 1, 2, \ldots)$. Hence, $\rho = 0$ corresponds to ignorable selection and $\rho \neq 0$ corresponds to non-ignorable selection.

3 Identification of Population Causal Effects

3.1 Bounds

In this section, we review how IVs are used to identify bounds for causal effects for the entire population based only on assumptions P1-P3. More generally, sets containing the causal effect can be identified: hence the term 'set-identification'.

Manski (1990) (see also Robins (1989)) propose bounds for the ACE using the following argument. The conditional expectation of the exposure-free potential outcome given

the IV can be written as

$$E\{Y(0)|Z\} = \Pr(X=0|Z)q^0_{0Z} + \Pr(X=1|Z)q^0_{1Z},$$

where $q^0_{xz} = E\{Y(0)|X=x, Z=z\}$ is the expected potential outcome. The counterfactual component of the right-hand side is q^0_{1Z}, and the observed component is $q^0_{0Z} \equiv q_{0Z} = E(Y|X=0, Z)$. In the absence of prior information, we can say only that q^0_{1Z} lies in the closed interval $(0,1)$, in which case $E\{Y(0)|Z\} > \Pr(X=0|Z)q_{0Z}$ and $E\{Y(0)|Z\} < \Pr(X=0|Z)q_{0Z} + \Pr(X=1|Z)$, or

$$E\{Y(0)|Z\} \in (p_{10.Z}, 1-p_{00.Z}),$$

where $p_{yx.z} = \Pr(Y=y, X=x|Z=z)$; similarly, it follows that $E\{Y(1)|Z\} \in (p_{11.Z}, 1-p_{01.Z})$. CMI (4) constrains $E\{Y(x)|Z=1\} = E\{Y(x)|Z=0)$, and so bounds for ACE, CRR and COR can be constructed. However, these bounds are not 'sharp' because not all information about the observed distribution is used.

Balke and Pearl (1997) construct sharp bounds by using linear programming techniques to find all

$$\text{ACE} = E_U\{\Pr(Y=1|X=1,U) - \Pr(Y=1|X=0,U)\}$$

satisfying the constraints $E_U\{\Pr(Y=y, X=x|Z=z, U)\} = p_{yx.z}$; Dawid (2003) gives an equivalent geometrical interpretation of the same problem. The sharp bounds are

$$E\{Y(0)\} \in \left(\max \begin{pmatrix} p_{11.0} \\ p_{11.1} \\ p_{00.1} + p_{11.1} - p_{00.0} - p_{01.0} \\ p_{10.1} + p_{11.1} - p_{01.0} - p_{10.0} \end{pmatrix}, \min \begin{pmatrix} 1 - p_{01.1} \\ 1 - p_{01.0} \\ p_{00.0} + p_{11.0} + p_{10.1} + p_{11.1} \\ p_{10.0} + p_{11.0} + p_{00.1} + p_{11.1} \end{pmatrix} \right),$$

$$E\{Y(1)\} \in \left(\max \begin{pmatrix} p_{10.1} \\ p_{10.0} \\ p_{10.0} + p_{11.0} - p_{00.1} - p_{11.1} \\ p_{01.0} + p_{10.0} - p_{00.1} - p_{01.1} \end{pmatrix}, \min \begin{pmatrix} 1 - p_{00.1} \\ 1 - p_{00.0} \\ p_{01.0} + p_{10.0} + p_{10.1} + p_{11.1} \\ p_{10.0} + p_{11.0} + p_{01.1} + p_{10.1} \end{pmatrix} \right),$$

from which bounds for ACE, CRR and COR can be established. Balke and Pearl (1997) given an expression for bounds on ACE, where the width of these bounds is itself bounded by the probability of non-compliance, and generally includes zero. As would be expected, the bounds are very wide if the causal effect of Z on X is weak, and so cannot identify the direction of the causal effect.

Chesher (2010) considers an alternative approach based on a wide class of non-linear scientific models for outcomes with discrete support. The class of binary outcome scientific models he considers can be written as

$$Y(x) = I(\widetilde{U} \geq c_x), \tag{14}$$

where c_x is the cut-off point determining the value of the potential outcome, and $\widetilde{U}$ is (marginally) a uniformly distributed scalar random variable on the $(0, 1)$ interval; the other requirement is that the cdf of $\widetilde{U}$ given Z must not depend on Z (c.f., condition C1). Latent $\widetilde{U}$ is a 'normalisation' of random vector U, and is generally associated with X so that the conditional distribution of $\widetilde{U}$ given X is not uniform. This corresponds to a scientific model (6) in which $f_Y^*(x, U)$ must be a separable function where, for example, an additively separable function satisfies $f_Y^*(x, U) = f_1^*(x) + f_2^*(U)$. The marginal distribution of U determines the functional form of c_x; it also determines, together with the selection model, the conditional distribution of $\widetilde{U}$ given X and Z.

The binary scientific models introduced in Section 2.4 are all in this class: the cut-off for the simple logistic model (8) is $c_x = \text{expit}(-\beta_0 - x\beta_1)$, and for the simple probit model (9) it is $c_x = \Phi(-\beta_0 - x\beta_1)$. The focus is on the cut-offs as $E\{Y(x)\} = 1 - c_x$, but an identification problem arises because

$$\Pr(Y = y|X = x, Z = z) = F_{\widetilde{U}|XZ}(c_x|x, z),$$

where $F_{\widetilde{U}|XZ}(c_x|x, z)$ is the cdf of the conditional distribution of $\widetilde{U}$ given X and Z. The left-hand side is observed but the function determining the right-hand side is unobservable. Equality clearly holds for the true value of c_x and the correct function $F_{\widetilde{U}|XZ}$, but

it also holds for $c_x^* \neq c_x$ because $F^*_{\widetilde{U}|XZ} \neq F_{\widetilde{U}|XZ}$ can be found explicitly to satisfy the equality. Hence, the data cannot distinguish between distinct but *observationally equivalent* scientific models, and so the causal effect is non-identified. However, the range of c_x^* is constrained by the requirement that $F^*_{\widetilde{U}|XZ}$ must satisfy the IV core conditions, which enables the true causal effect to be set-identified, or bounded, in a non-trivial sense.

Chesher (2010) shows that all observationally equivalent models for which the IV core conditions hold must satisfy the sharp inequalities $\Pr\{Y < h(X,\tau)|Z\} < \tau, \Pr\{Y \leq h(X,\tau)|Z\} \geq \tau$, for all $\tau \in (0,1)$ and all Z. In the case where X and Z are both binary, these inequalities yield the following bounds for c_0 and c_1:

$$\begin{aligned} p_{01.z} &\leq c_0 < p_{00.z} + p_{01.z} \leq c_1 < 1 - p_{11.z}, \\ p_{00.z} &\leq c_1 < p_{00.z} + p_{01.z} \leq c_0 < 1 - p_{10.z}, \end{aligned}$$

where $p_{yx.z}$ is defined above. As the inequalities must be satisfied for all $Z \in \{0,1\}$, the resulting set can be written

$$\begin{aligned} E\{Y(0)\} &\in B_0 \bigcup A_0 \equiv \left\{ \bigcap_{z=0,1} (p_{10.z}, p_{10.z} + p_{11.z}) \right\} \bigcup \left\{ \bigcap_{z=0,1} (p_{11.z} + p_{10.z}, 1 - p_{00.z}) \right\}, \\ E\{Y(1)\} &\in B_1 \bigcup A_1 \equiv \left\{ \bigcap_{z=0,1} (p_{11.z} + p_{10.z}, 1 - p_{01.z}) \right\} \bigcup \left\{ \bigcap_{z=0,1} (p_{11.z}, p_{11.z} + p_{10.z}) \right\}. \end{aligned}$$

Each set comprises the union of two regions corresponding to above and below the $E\{Y(0)\} = E\{Y(1)\}$ (ACE = 0; CRR = COR = 1) line: the sets B_0 and B_1 together define the region of $[E\{Y(0)\}, E\{Y(1)\}]$ pairs lying below the ACE = 0 line; and A_0 and A_1 define the equivalent region above the ACE = 0 line.

Chesher (2010, sec. 3) illustrates the geometry of these sets using a numerical example. An IV strongly associated with exposure will eventually have one or both of $B_0, B_1 = \oslash$ or one or both of $A_0, A_1 = \oslash$, thus identifying the sign of the effect (e.g., if $B_1 = \varnothing$ then the positive causal effect region contributes nothing to the set and the causal effect is identified as negative). In its limit, if $\Pr(X = x|Z = z) = 1$ for some pair

$x, z \in \{0, 1\}$ then c_x is point-identified. We discuss this situation again in the context of encouragement designs with 'no-contamination' restrictions in Section 8.

Bounds for ACE, CRR and COR can be calculated straightforwardly based on the sets above. Both 'Chesher bounds' and those of Balke and Pearl (1997) are sharp, but we expect the former to be narrower because the structure of (14) excludes structurally implausible scientific models from consideration. Results from a limited simulation study show Chesher bounds to be marginally narrower (details available from the authors), but a more formal comparison is on-going. However, neither evaluating bounds for more complex scenarios nor interval estimation is straightforward.

3.2 Identification

In this section, we consider the identification (or more precisely, point-identification) of population causal effects. To begin, we return to the classical result introduced in Section 2.2, namely, if the true scientific model is linear with constant exposure effects, then the IV core conditions C1-C3 identify ACE. Despite the convention within statistics of modelling binary outcomes using non-linear models, linear models can be used if the outcome probabilities are bounded away from 0 and 1. In applications with no, or coarsely defined, covariates, this assumption can be verified for the observed outcomes (although the constant exposure effects assumption cannot), in which case the 2SLS estimator may only have small bias in large samples. Arguments supporting the linear probability model are not unknown (e.g., Angrist, 2001), but generally either the bounded probability assumption demonstrably fails or cannot be verified (e.g., Imbens, 2001).

We now move on to consider identification in situations where the assumptions behind linear IV estimators are implausible. Using the framework developed by Chesher (2010), suppose that the conditional distribution of $\widetilde{U}$ given X has known cdf $G(\tau, x) = \Pr(\widetilde{U} \leq \tau | X = x)$. For the simple example with a binary exposure and a binary IV, the cut-offs

are identified if the inverse of $G(c_x|x) = \Pr(Y = 0|X = x)$ exists because

$$c_x = G^{-1}\{\Pr(Z = 0|X = x)(1 - q_{x0}) + \Pr(Z = 1|X = x)(1 - q_{x1}), x\},$$

where $G^{-1}(p, x)$ is the inverse cdf and $q_{xz} = E(Y|X = x, Z = z)$. Such approaches are identified by the functional form of G, which follows from fully parametric assumptions about the generating model, including those about the joint distribution of U and V. Maximum likelihood estimators explicitly incorporate such assumptions to obtain identification (see Sections 6 and 7).

Another set of identifying assumptions, which does not rely on functional form, concern the expected potential outcomes $q_{xz}^{x^*} = E\{Y(x^*)|X = x, Z = z\}$, where clearly $q_{xz}^{x} \equiv q_{xz}$ is identified. Any assumption that identifies the counterfactual expectation $q_{xz}^{(1-x)}$ $(x = 0, 1)$ also identifies the cut-offs because

$$c_x = \Pr(X = 0|Z = z)(1 - q_{0z}^{x}) + \Pr(X = 1|Z = z)(1 - q_{1z}^{x}),$$

follows under CMI (4).

Up until this point, the focus has been on identifying population causal effects, and so it has been necessary to make unverifiable assumptions about both $Y(0)$ and $Y(1)$ via CMI. However, identification of causal effects among the exposed group (e.g., the average causal effect for the exposed $E\{Y(1) - Y(0)|X = 1\}$) requires only unverifiable assumptions about $Y(0)$ through $E\{Y(0)|Z\} = E\{Y(0)\}$ because $Y(1)$ is observed among those exposed and its distribution identified. The parameters of structural mean models are identified in this way (see Section 5).

4 The Generalized Method of Moments

The 2SLS estimator is based on the moment conditions

$$E(R) = E(RZ) = 0, \tag{15}$$

where $R = Y - \beta_0 - X\beta_1$ is the residual of the linear scientific model (1). Under this model it follows that $R = U$, and so (15) trivially holds under the core conditions and β_1 is identified.

Estimators for non-linear models can be obtained using the generalized method of moments (GMM). GMM estimators solve the same basic moment condition (15), but where R is a generalised residual function that satisfies $E(R|Z) = 0$. Johnston et al. (2008) give a concise overview of GMM estimators, while Wooldridge (2002, ch. 14) gives a more complete account. As the name suggests, GMM is a generalisation of the method of moments to allow for more than one endogenous covariate and multiple IVs for each. Only situations involving one endogenous exposure and one IV are considered here, but the points we make also apply to the general case.

To construct a GMM estimator that exploits the IV core conditions it must be possible to separate U from the parameters of the underlying scientific model. However, scientific models like (6) are not mean separable because of the indicator function. For example, the additive residual $R = Y - E(Y|X) \neq U$, which means that $E(R|Z) \neq 0$ and any GMM estimator based on this residual cannot be consistent.

We now review two GMM estimators based on the assumption that the scientific model is a logistic mixed model (11), which can be written as

$$E(Y|X, U_1) = \text{expit}(\beta_0 + X\beta_1 + U_1),$$

recalling that U_1 represents the effect of the omitted variables that are associated with X. The error structure is slightly more complex than for standard logistic and probit models, and changes the interpretation of β_1: it is now the conditional log-odds ratio given U_1 and does not correspond to COR because of non-collapsibility (e.g., Greenland et al., 1999). Neither estimator is consistent but both estimators are approximations based on different assumptions; the key issue for practice is how good each approximation is.

4.1 Additive residual approximation

Johnston et al. (2008) consider the scientific model

$$E(Y|X,\widetilde{U}) = \text{expit}(\beta_0 + X\beta_1) + \widetilde{U}, \tag{16}$$

which is asymmetric and causally implausible in the sense we discussed in Section 2.4. They argue that it is a first-order approximation of (11), that is, $\text{expit}(\beta_0 + x\beta_1 + u_1) \simeq \mu(x) + u_1\mu(x)\{1-\mu(x)\}$, which in turn is approximately $\mu(x) + u_1\mu(\overline{x})\{1-\mu(\overline{x})\}$, where $\overline{x} = E(X)$. However, we expect the first-step approximation alone to be poor because

$$\begin{aligned} E\{Y - \mu(X)|Z\} &= E\{\text{expit}(\beta_0 + X\beta_1 + U_1) - \mu(X)|Z\} \\ &\simeq E\left[U_1\mu(X)\{1-\mu(X)\} + \frac{1}{2}U_1^2\mu(X)\{1-\mu(X)\}\{1-2\mu(X)\}|Z\right], \end{aligned}$$

which equals zero only trivially if X and U_1 are independent; moreover, the U_1^2 term indicates that the approximation will be good only if the variance of U_1 is small.

Ten Have et al. (2003) propose the closely related 'marginal' estimator, based on a marginal structural model (MSM) specification for the scientific model (e.g., Robins et al., 2000; Hogan and Lancaster, 2004). The logistic MSM is

$$E\{Y(x)\} = \text{expit}(\beta_0 + x\beta_1), \tag{17}$$

which follows from integrating U out of (6) with respect to its unspecified marginal distribution. An advantage of this specification over (11) is that, in the absence of covariates, $\beta_1 = \text{COR}$; covariates can be added through extending the linear predictor to include C, with the proviso that the effect of X is now conditional on C and so β_1 will not equal COR, again due to non-collapsibility. No explicit assumption of constant exposure effects has been made.

The marginal estimator comes from two moment conditions, one of which is

$$E\left[\{Z - E(Z)\}R\right] = 0,$$

where $R = Y - \text{expit}(\beta_0 + X\beta_1)$ is the additive residual based on (17). If $E(R|Z) = 0$ then it follows that these moment conditions are equivalent to $E(R) = E(RZ) = 0$, and so the marginal estimator is equivalent to that proposed by Johnston et al. (2008).

Ten Have et al. (2003) argue that the moment condition holds if the scientific model follows a mixed model (10) that satisfies (17). Clarke and Windmeijer (2009, app. 2) show that their justification relies implicitly on R having the same properties as U_1, but that $E(R|Z) \neq E(U_1|Z) = 0$ unless exposure selection is ignorable. Ten Have et al. (2003) present simulation results which show their estimator's bias depends both on the association between X and U_1 and between U_1 and Y, but these findings cannot be interpreted merely as finite sample bias because the estimator is inconsistent.

4.2 Multiplicative residual approximation

If Y (or $1 - Y$) is a rare outcome then the logistic mixed model can be approximated by the exponential mean model

$$E(Y|X, U_1) \simeq \exp(\beta_0 + X\beta_1 + U_1). \tag{18}$$

In practice, the exponential mean model is mainly used for the estimation of risk ratios from count data when X is endogenous (e.g., Mullahy, 1997). A GMM estimator for these models is based on the multiplicative residual

$$R = \frac{Y}{\exp(\alpha + X\beta_1)} - 1 \tag{19}$$

from which the multiplicative moment condition $E(RZ) = 0$ identifies $e^{\alpha} = e^{\beta_0}E(e^{U_1})$ and β_1. Hence, the multiplicative GMM estimator is consistent for ACE $= (e^{\beta_1} - 1)e^{\alpha}$ and CRR $= e^{\beta_1} \simeq$ COR provided that model (18) holds.

For applications in which the outcome is rare, the GMM estimator based on (19) is a sensible way to proceed. However, if we assume that $\exp(\beta_0 + x\beta_1 + u_1) \in (0, \delta)$ for small $\delta > 0$ and for all (x, u_1), then it can be shown that $E(R|Z) = O(\delta)$, where expectation

is taken with respect to the logistic mixed model (11) and $a = O(\delta)$ implies that $|a/\delta|$ is bounded above. This indicates that the moment condition error is of the same order as the event probability itself. If X is exogenous (or equivalently selection is ignorable) then the model can also be fitted using a GMM estimator based on the same model but using $R = Y - \exp(\beta_0 + X\beta_1)$ and $E(R|X) = 0$, i.e., the additive (or Poisson first-order) moment condition. The additive moment condition satisfies $E(R|X) = O(\delta^2)$ so the error is an order of magnitude smaller than the event probability itself. It follows from this that the bias of the multiplicative GMM increases more quickly than the additive estimator as the event becomes less rare (Clarke and Windmeijer, 2009, app. 1). However, this estimator is useful as a first-order aproximation to CRR (or COR) for Mendelian randomisation studies.

5 Structural Mean Model Estimators

5.1 Structural mean models

Robins (1989, 1994) introduced the class of semi-parametric structural mean models (SMMs) and 'G-estimation' for causal effects of treatment regimes on outcomes from randomised controlled trials affected by non-compliance. The parameters of SMMs correspond to meaningful functions of expected potential outcomes for the population of participants exposed to the treatment. For example, additive SMMs are specified in terms of average treatment (or causal) effects, and multiplicative SMMs in terms of causal risk ratios (Hernán and Robins, 2006). Vansteelandt and Goetghebeur (2003) developed the generalised SMM from which we consider two important special cases: the logistic and probit SMMs (see also Goetghebeur and Vansteelandt (2005)).

We again consider SMMs in the simplest possible set-up, namely, an encouragement design for a randomised controlled trial where IV Z is the randomisation assignment indicator of a binary treatment/exposure, X is the corresponding indicator for the actual

treatment chosen by the patient, and $X \neq Z$ is possible due to non-compliance. The generalised SMM for this design is written

$$b\{E\left(Y|X,Z\right)\} - b[E\left\{Y\left(0\right)|X,Z\right\}] = \left(\psi_0 + \psi_1 Z\right) X,$$

where $Y\left(0\right)$ is the exposure-free potential outcome and b is a suitable link function. This model is saturated, or non-parametric, but more generally the right hand side can be a parametric function incorporating the effect of C and/or variable exposure dose, provided that $X = 0$ is equivalent to the exposure for those who comply in the control group $Z = 0$. For instance, the link function for the logistic SMM is $b = \text{logit}$, where $\text{logit}(a) = \log\left\{a/\left(1-a\right)\right\}$ is the inverse cdf of the standard logistic distribution; the parameters of the logistic SMM are thus the causal odds ratios among those who are assigned to Z in the exposed group:

$$\exp(\psi_0 + \psi_1 Z) = \frac{E\left\{Y(1)|X=1,Z\right\}/E\left\{1-Y(1)|X=1,Z\right\}}{E\left\{Y(0)|X=1,Z\right\}/E\left\{1-Y(0)|X=1,Z\right\}}.$$

The link functions for the additive, multiplicative and probit SMMs are, respectively, the identity function, the natural logarithm, and the inverse cdf of the standard normal distribution.

It is important to recognise that this specification assumes nothing explicitly about the underlying generating model, and so in principle all four SMMs can be applied to binary outcomes. Recalling Section 3.2, it is clear that the SMM can be viewed as an identifying assumption: it explicitly links the expected counterfactual $q^0_{1z} = E\{Y(0)|X=1,Z=z\}$ to the observed expectation $q_{1z} = E(Y|X=1,Z=z)$ via a semi-parametric model. It remains now to establish the conditions under which the parameters of the SMM are identified.

5.2 SMM estimation

Estimators for the additive and multiplicative SMMs are based on the moment condition

$$E\left\{Y\left(0\right)|Z=1\right\} = E\left\{Y\left(0\right)|Z=0\right\},$$

which follows under CMI (4). For example, under the multiplicative SMM ($b = \log$) the moment condition is

$$E\left[Y \exp\left\{-\left(\psi_0 + \psi_1\right) X\right\} | Z = 1\right] = E\left\{Y \exp\left(-\psi_0 X\right) | Z = 0\right\}. \tag{20}$$

It is clear that the SMM parameters are not directly identified by this moment condition because it constitutes a system with two unknowns and one equation. Therefore, further assumptions are required.

Hernán and Robins (2006) highlight the importance of the 'no effect modification by Z' (NEM) assumption that $\psi_1 = 0$. Under NEM, the target parameter for the additive SMM is $\psi_0 = E\left\{Y(1) - Y(0) | X = 1\right\}$, the average causal effect among those exposed; and for the multiplicative SMM it is $\exp(\psi_0) = E\left\{Y(1) | X = 1\right\} / E\left\{Y(0) | X = 1\right\}$, the causal risk ratio among those exposed. The estimator $\widehat{\psi}_0$ under the additive SMM is easily shown to equal the classical IV estimator (2), whereas for the multiplicative SMM it is

$$\widehat{\exp(\psi_0)} = 1 - \frac{E(Y|Z=1) - E(Y|Z=0)}{E\left\{(1-X)Y|Z=1\right\} - E\left\{(1-X)Y|Z=0\right\}} \tag{21}$$

(Hernán and Robins, 2006).

Robins (1994) developed G-estimation for non-saturated semi-parametric additive and multiplicative SMMs. G-estimators are asymptotically normal, semi-parametrically efficient and obtain uniform convergence. Generally, the variance of a G-estimator's asymptotically normal distribution is difficult to evaluate, but this approximation has good finite sample properties (see the arguments given by Robins and Ritov (1997)). For non-saturated models, these SMMs are not ideal because the model does not constrain probabilities to lie in $(0, 1)$, but van der Laan et al. (2007) develop two alternative strategies for multiplicative SMMs which overcome this problem.

The logistic and probit SMMs are considered separately because no G-estimator can be found for ψ_0 for either model (e.g., Robins and Rotnitzky, 2004). The double-logistic

estimator was developed by exploiting the result that ψ_0 can be identified if the researcher additionally specifies a logistic 'association model' $E(Y|X,Z) = \text{expit}(\eta_0 + X\eta_1 + Z\eta_2 + ZX\eta_3)$ (Vansteelandt and Goetghebeur, 2003). In some circumstances, the double specification of the SMM and association model as logistic can be uncongenial, but this does not affect the saturated SMMs considered here (Robins and Rotnitzky, 2004; Vansteelandt et al., 2010).

Identification of ψ_0 under the double-logistic SMM is based on the moment condition

$$E\left[\text{expit}\left\{\eta_0 + \eta_2 + (\eta_1 + \eta_3 - \psi_0)X\right\}|Z=1\right] = E\left[\text{expit}\left\{\eta_0 + (\eta_1 - \psi_0)X\right\}|Z=0\right], \tag{22}$$

where an estimate of $(\eta_0, \eta_1, \eta_2, \eta_3)$ is obtained at the first stage by fitting the logistic association model. Similarly, the double-probit estimator ($b = \Phi^{-1}$) is based on association model $E(Y|X,Z) = \Phi(\eta_0 + X\eta_1 + Z\eta_2 + ZX\eta_3)$, and is similarly calculated. Both 'double' estimators have similar asymptotic properties to G-estimators when the association model is correctly specified, with an additional 'local robustness' property if it is mis-specified, namely, it is always consistent under the null hypothesis $\psi_0 = 0$ (Vansteelandt and Goetghebeur, 2003). However, as with G-estimators, the expression for the asymptotic variance is generally difficult to evaluate, so for simple saturated models the non-parametric bootstrap is recommended instead (Didelez et al., 2010). We also note that an approximate version of the double-logistic model has been developed (Vansteelandt et al., 2010).

An alternative assumption to NEM can be used for encouragement designs with constraints on participant selection following assignment. For instance, in a randomised placebo-controlled trial, patients in the control group cannot receive the treatment because non-compliers ($Z = 0, X = 1$) receive only the placebo and so $\Pr(X = 0|Z = 0) = 1$. Cuzick et al. (2007) refer to designs with this property as having 'no contamination' restrictions; Robins and Rotnitzky (2004) discuss the role played by such restrictions in

identifying the parameters of logistic and probit SMMs.

5.3 The NEM assumption

Clarke and Windmeijer (2010) investigate the validity of NEM using the generating model framework introduced in Section 2.4. The SMM parameters are simple functions of $E\{Y(x)|X=1,Z\}$, which can be written in terms of the generating model as

$$E\left\{Y(x)|X=1,Z=z\right\}=\Pr\left\{f_Y^*\left(x,U\right)>0|f_X^*\left(z,V\right)>0\right\}.$$

All members of this class automatically satisfy the CMI assumption and so we can focus on NEM for each SMM in turn. For a specific example, suppose that the generating model follows the 'bivariate probit' model

$$Y(x)=I\left(\alpha+\beta x-U>0\right),\ X(z)=I\left(\gamma+\delta z-V>0\right), \tag{23}$$

where $F_{uv}=\Phi_\rho$ is the cdf of the standard bivariate normal distribution and ρ is its correlation parameter; the bivariate probit generating model is considered again in Section 6. In this case,

$$E\left\{Y\left(x\right)|X=1,Z=1\right\}=\Phi_\rho\left(\alpha+x\beta,\gamma+Z\delta\right)/\Phi\left(\gamma+Z\delta\right),$$

where $\Phi\left(v\right)$ is the cdf of V. Clearly, if non-compliance is ignorable then $\rho=0$ and NEM automatically holds for every SMM. However, if $\rho\neq0$ then NEM does not necessarily hold. For example, NEM fails for the logistic SMM because

$$\begin{aligned}\exp\left(\psi_0+\psi_1\right) &= \frac{\Phi_\rho\left(\alpha+\beta,\gamma+\delta\right)/\left\{\Phi\left(\gamma+\delta\right)-\Phi_\rho(\alpha+\beta,\gamma+\delta)\right\}}{\Phi_\rho\left(\alpha,\gamma+\delta\right)/\left\{\Phi\left(\gamma+\delta\right)-\Phi_\rho(\alpha,\gamma+\delta)\right\}}\\ &\neq \frac{\Phi_\rho\left(\alpha+\beta,\gamma\right)/\left\{\Phi\left(\gamma\right)-\Phi_\rho\left(\alpha+\beta,\gamma\right)\right\}}{\Phi_\rho\left(\alpha,\gamma\right)/\left\{\Phi\left(\gamma\right)-\Phi_\rho\left(\alpha,\gamma\right)\right\}}=\exp\left(\psi_0\right)\end{aligned}$$

almost everywhere. Perhaps this is not surprising, but NEM also fails for the probit SMM because

$$\begin{aligned}\Phi^{-1}[E\{Y(1)|X &= 1,Z=1\}]-\Phi^{-1}[E\{Y(0)|X=1,Z=1\}]\\ &\neq \Phi^{-1}[E\{Y(1)|X=1,Z=0\}]-\Phi^{-1}[E\{Y(0)|X=1,Z=0\}]\end{aligned}$$

almost everywhere; the corresponding NEM assumption for the additive and multiplicative SMMs also fail under the bivariate probit model.

Of course, this does not mean that generating models under which NEM holds for a particular SMM cannot be found, but it is almost impossible to write-down an explicit generating model satisfying these requirements, and NEM clearly places considerable constraints on the family of models that satisfy it. It is possible to generate data from a logistic or probit SMM indirectly using the association model parameterisation $E(Y|X,Z) = b^{-1}(\eta_0 + X\eta_1 + Z\eta_2 + ZX\eta_3)$, because the SMM and NEM together imply that $E\{Y(0)|X,Z\} = b^{-1}\{\eta_0 + X(\eta_1 - \psi_0) + Z\eta_2 + ZX\eta_3\}$ (Robins and Rotnitzky, 2004; Vansteelandt et al., 2010). However, this is not a true scientific model and $(\eta_0, \eta_1, \eta_2, \eta_3)$ must be constrained explicitly to satisfy CMI.

To finish this discussion, we note that parametric assumptions can be used to relax NEM and identify interactions between Z and X for generalised SMMs where Z is not binary and there are covariates (Vansteelandt and Goetghebeur, 2005).

5.4 Links with other estimators

There is a clear link between the additive and multiplicative SMM estimators and the GMM estimators considered in the previous section. To show this link, let us assume that the true scientific model is the same as that used to construct the multiplicative GMM estimator in Section 4.2, that is, $E\{Y(x)|U_1\} = \exp(\beta_0 + x\beta_1 + U_1)$. It follows that $E\{Y(0)|Z\} = \exp(\alpha) = E\{Y(0)\}$, where α is defined in equation (19) and so CMI holds. Furthermore, the model satisfies NEM for the multiplicative SMM and has the desirable property $\psi_0 = \text{CRR}$ because the causal risk ratios among the exposed and control groups

are equal. The moment conditions for the multiplicative SMM can be thus written

$$\begin{aligned} E\left[E\left\{Y(0)|X,Z\right\}|Z\right] &= \exp(\alpha), \\ E\left\{\exp(-\alpha - X\psi_0)E(Y|X,Z) - 1|Z\right\} &= 0, \\ E\left\{\exp(-\alpha - X\psi_0)Y - 1|Z\right\} &= 0, \end{aligned}$$

which are equivalent to the moment conditions for the multiplicative GMM estimator. In other words, the multiplicative GMM estimator is a special case of multiplicative SMM G-estimator that exploits the additional assumptions embodied by the scientific model.

For the logistic and probit SMMs, the SMM defines $E\{Y(0)|X,Z\}$ to be a residual on the scale of the link function: $b\{E(Y(0)|X,Z)\} = b\{E(Y|X,Z)\} - \psi_0 X$. This residual can be interpreted as that of a non-orthogonal projection of $b\{E(Y|X,Z)\}$ onto X, where ψ_0 is chosen to satisfy zero expectation with respect to the conditional distribution of X given Z, and identification requires NEM and semi-parametric assumptions about the association model. As we have already discussed, it is very difficult to explicitly specify a generating model for binary data that satisfies NEM for either the logistic or probit SMMs, and hence to derive an expression for $E\{Y(0)\}$. Hence, there is no equivalent relationship between GMM and the double-SMM estimators.

If the design has a no-contamination restriction or the selection model is monotonic then there is also a correspondence between SMM estimators and local effect estimators: this connection is discussed in Section 7.

6 Parametric Likelihood Estimators

6.1 Probit models

In Section 3, we saw that causal parameters can be identified under parametric assumptions about the generating model. Specifically, these assumptions relate to the conditional distribution of U given X and Z and augment the IV core conditions. Likelihood theory

offers a natural way to incorporate such assumptions and develop consistent estimators. The cost is that U is unobserved and so any assumptions about its association with X are unverifiable from the observed data.

Heckman (1979) and Lee (1981) proposed ML estimators for generating models in which the latent variables are normally distributed. Suppose that the scientific model is $Y(x) = \beta_0 + x\beta_1 + U$ and that there are two basic types of selection models: if X is binary then $X(z) = I(\alpha_0 + x\alpha_1 + V > 0)$; and if X and Z are linearly related then $X(z) = \alpha_0 + x\alpha_1 + V$. Furthermore, the latent variables are assumed to follow the bivariate normal distribution

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma_v \\ \rho\sigma_v & \sigma_v^2 \end{pmatrix} \right\},$$

where ρ is the correlation coefficient; the usual identification constraint $\sigma_v^2 = 1$ is used if X is binary. Then, for example, when X is binary the ML estimator is

$$\begin{aligned} \widehat{\theta} &= \arg\max \prod_i \Phi_\rho(-\alpha_0 - \alpha_1 Z_i, -\beta_0)^{I(X_i=Y_i=0)} \Phi_\rho(\alpha_0 + \alpha_1 Z_i, -\beta_0 - \beta_1)^{I(X_i=1,Y_i=0)} \\ &\quad \times \Phi_\rho(-\alpha_0 - \alpha_1 Z_i, \beta_0)^{I(X_i=0,Y_i=1)} \Phi_\rho(\alpha_0 + \alpha_1 Z_i, \beta_0 + \beta_1)^{I(X_i=Y_i=1)}, \end{aligned}$$

where $\Phi_\rho(u, v)$ is the cdf of (U, V), and $\theta = (\alpha_0, \alpha_1, \beta_0, \beta_1, \rho)'$. ML estimates of the important causal effects can be estimated by construction.

It is important to distinguish the role played by the selection model here from the role played by the reduced-form model for 2SLS: the reduced-form model yields a consistent 2SLS estimator if the scientific model is linear *even if* the true selection model is non-linear, whereas the selection model here implicitly encodes parametric identification assumptions for the generating model parameters. More precisely, these assumptions determine the crucial U given X distribution:

$$\begin{aligned} F_{u|x}(u|0) &= (1-\mu_x)\frac{\Phi_\rho(u, -\alpha_0)}{\Phi(-\alpha_0)} + \mu_x \frac{\Phi_\rho(u, -\alpha_0 - \alpha)}{\Phi(-\alpha_0 - \alpha_1)}, \\ F_{u|x}(u|1) &= (1-\mu_x)\frac{\Phi(u) - \Phi_\rho(u, -\alpha_0)}{\Phi(\alpha_0)} + \mu_x \frac{\Phi(u) - \Phi_\rho(u, -\alpha_0 - \alpha_1)}{\Phi(\alpha_0 + \alpha_1)}, \end{aligned}$$

where $F_{u|x}(u|x) = \Pr(U \leq u|X = x)$ is identified because Φ_ρ and its marginal Φ are both known, and $\mu_x = E(Z|X = x)$, α_0 and α_1 are all identified by the IV core conditions. The ML estimator based on the model above is clearly sensitive to functional form Φ_ρ and so will be biased and inconsistent if mis-specified.

Rivers and Vuong (1988) further consider the properties of two simple estimators for probit models which are both analogous to 2SLS. These are 'pseudolikelihood' estimators because each involves replacing nuisance parameters with consistent estimators thereof. Identification depends crucially on (U, V) being normally distributed and on X given Z being linear.

First, consider the 'plug-in' estimator if $X = \alpha_0 + x\alpha_1 + V$. Stage one: fit the linear selection model and store $\widehat{X} = \widehat{\alpha}_0 + \widehat{\alpha}_1 Z$; stage two: fit $Y = \Phi(\beta_0^* + \widehat{X}\ \beta_1^*)$ to obtain consistent estimates of the scaled coefficients $\beta_j^* = \beta_j / \sqrt{1 + 2\beta_1\rho\sigma_v + \beta_1^2\sigma_v^2}$ for $j = 0, 1$. While the causal effect of X on the latent scale, β_1, cannot be identified, its direction (positive or negative) can through β_1^*. The plug-in estimator can be written as

$$\beta_1^* = \frac{\Phi^{-1}\{E(Y|Z = 1)\} - \Phi^{-1}\{E(Y|Z = 0\}}{E(X|Z = 1) - E(X|Z = 0)},$$

which has the same basic form as (2) and the Wald estimators considered in Section 7.

Second, consider the 'control function' estimator. Stage one: fit linear selection model as before, but instead of $\widehat{X}$ store the fitted residual $\widehat{V} = X - \widehat{\alpha}_0 - \widehat{\alpha}_1 Z$ and its estimated variance $\widehat{\sigma}_v^2$; stage two: fit $Y = \Phi(\beta_0^{**} + X\ \beta_1^{**} + \widehat{V}\lambda^{**})$ to obtain consistent estimates of $\beta_j^{**} = \beta_j / \sqrt{1 - \rho^2}$ and $\lambda^{**} = \lambda / \sqrt{1 - \rho^2}$, where $\lambda = \rho/\sigma_v$. Under this model, it follows that

$$\Phi^{-1}\{E(Y|X, V)\} - \Phi^{-1}\{E(Y|X = 0, V)\} = X\beta_1^{**},$$

the form of which we have already seen in the context of the probit SMM from Section 5. In contrast to the plug-in estimator, the coefficients are identified because $\widehat{\rho}$ is a function of $\widehat{\sigma}_v^2$ and $\widehat{\lambda}^{**}$; moreover, again in contrast to the plug-in estimator, the control function is consistent if selection is ignorable.

Compared to the probit SMM, the assumptions implicit in the control function approach are instantly seen to be stronger. First, the probit generating model satisfies $E\{Y(x)|Z\} = \Phi(\beta_0 + x\beta_1)$ and thus CMI $E\{Y(x)|Z\} = E\{Y(x)\}$ holds automatically and cannot be exploited for identification, which comes through functional form. On the other hand, the probit SMM does not use functional form but is required only to satisfy $E\{Y(0)|Z\} = E\{Y(0)\}$ and not $E\{Y(1)|Z\} = E\{Y(1)\}$. Second, the association model under the probit generating model is

$$E(Y|X,Z) = \Phi\left\{\beta_0^{**} - \alpha_0\lambda^{**} + (\beta_1^{**} + \lambda^{**})X - Z\alpha_1\lambda^{**}\right\},$$

which has the same form as the association model for the double-probit SMM estimator (namely, $E(Y|X,Z) = \Phi(\eta_0 + X\eta_1 + Z\eta_2)$). However, this similarity is superficial: the association model follows because the probit generating model constrains the exposure effect on the latent scale to be constant, the selection model to be linear, and U to be normally distributed and independent of Z; in contrast, the probit SMM under NEM $\Phi^{-1}\{E(Y|X,Z)\} - \Phi^{-1}[E\{Y(0)|X,Z\}] = X\psi_0$ is an *assumption*.

The operational simplicity of the two-stage estimators is an attractive feature for applications. Unfortunately, this simplicity hinges crucially on linearity of the selection model. If X is binary then the selection model is non-linear and neither estimator is consistent, even for the scaled coefficients. For example, stage one of the plug-in estimator becomes: fit $X = \Phi(\alpha_0 + \alpha_1 Z)$ and obtain $\widehat{E(X|Z)} = \Phi(\widehat{\alpha}_0 + \widehat{\alpha}_1 Z)$; stage two: fit $Y = \Phi\{\beta_0 + \widehat{E(X|Z)}\beta_1\}$. Essentially, the plug-in estimator relies on being able to construct a tractable expression for $E(Y|Z)$, which is easy under linear selection but falls down here because

$$E(Y|Z) = E_{U,V}[I\{\beta_0 + I(\alpha_0 + Z\alpha_1 + V > 0)\beta_1 + U > 0\}] \neq \Phi\{\beta_0 + E(X|Z)\beta_1\}.$$

Similarly for the control function method: stage one now involves calculating $\widehat{V} = X - \Phi(\widehat{\alpha}_0 + \widehat{\alpha}_1 Z)$ and using this residual in stage two, but the estimator is inconsistent because no λ can be found to ensure that $X \perp U - V\lambda$ if $V = X - \Phi(\alpha_0 + \alpha_1 Z)$.

6.2 Logistic and other models

In theory, the likelihood for any parametric model can be specified, but there are practical difficulties in specifying a suitable likelihood for non-normal (U, V). Despite this, pseudolikelihood estimators have been proposed for logistic models. Palmer et al. (2008) develop both plug-in and control variable approaches for logistic models and a linear selection model. To overcome the problem of non-normal latent variables, they assume that the outcome data come from the logistic mixed model (11) where $U_1 \sim N(0, \sigma_1^2)$. Thus expressions for $E(Y|Z)$ and $E(Y|X, V)$ can be constructed using the result that $E_A\{\text{expit}(B\omega + A)\} \backsimeq \text{expit}(B\omega^c)$ if $A \sim N(0, v)$, where $\omega^c = \omega/\sqrt{1 + cv}$ and $c = 16\sqrt{3}/15\pi$. It is shown that both the plug-in and control function approaches are consistent for scaled coefficients like those of the two-stage bivariate probit estimators, but that even the control function cannot produce consistent estimators for actual coefficients.

Nagelkerke et al. (2000) construct an IV estimator using arguments analogous to those for the control variable estimator above but for binary X. Their control variable approach is based on an additive error structure for the selection model $X = E(X|Z)+V$, which leads to an inconsistent estimator because $E(X|Z) + V \neq I\{f_X^*(Z, V) > 0\}$. Ten Have et al. (2003) found by simulation that this estimator can be badly biased if the association between X and U is not weak, that is, the true selection process is strongly non-ignorable.

Finally, we note that semi-parametric control function approaches for linear X based on non-parametric methods have been developed (e.g., Blundell and Powell, 2004).

7 Local Causal Effects

An alternative identification strategy is to focus not on population causal effects, but on so-called 'local' or 'complier-specific' causal effects. More generally, these effects are the parameters of principal strata (Frangakis and Rubin, 2002). The identification of

local effects requires only assumptions about the selection model for $X(z)$ and not for the scientific model, although semi-parametric assumptions are required if covariates are included.

To illustrate the nature of the identifying assumptions, consider the support of $X(0)$ and $X(1)$ in the simple example where X and Z are binary. Realisations from this distribution fall into one of four groups:

1. 'Compliers' $X(0) = 0$ and $X(1) = 1$.
2. 'Always-takers' $X(0) = 1$ and $X(1) = 1$.
3. 'Never-takers' $X(0) = 0$ and $X(1) = 0$.
4. 'Defiers' $X(0) = 1$ and $X(1) = 0$.

These groups are defined using what the study unit would have selected if its IV had taken another value, and so are defined in terms of unobservable counterfactuals. A monotonic selection mechanism requires that $X(z)$ is a non-decreasing function of z (or a non-increasing function, depending on the labelling). In this example, monotonic selection implies that the set of defiers is empty with probability one; more generally, monotonic selection is satisfied if $z > z'$ implies that $X(z) \geq X(z')$ for all pairs z, z' (Imbens and Angrist, 1994).

As we suggested earlier, the simple selection models (12) and (13) are special cases of monotonic selection mechanisms because both imply that either $X(1) \geq X(0)$ or $X(1) \leq X(0)$; for brevity, we herein assume that monotonic selection implies $X(1) \geq X(0)$ without loss of generality. However, some plausible selection models are not monotonic. For example, an extension of (12) to allow for heterogeneous exposure effects is

$$X(z) = \alpha_0 + z(\alpha_1 + V_1) + V_2,$$

but selection is not monotonic because $X(1) - X(0) = \alpha_1 + V_1 \not\geq 0$ if $\alpha_1 > 0$.

Imbens and Angrist (1994) give key results regarding local causal effect estimators under monotonic selection. These effects are local in the sense that the com-

plier group comprises *only* those whose exposure selection would be modified if they had (counterfactually) been characterised differently by the IV. An important result is that the classical IV estimator (2) is consistent for the 'local average treatment effect' LATE $= E\{Y(1) - Y(0)|X(1) > X(0)\}$ (Angrist et al., 1996); LATE is also known as the 'complier' average causal effect. If Z is discrete then the focus is on modelling LATE$(z, z') = E\{Y(1) - Y(0)|X(z) > X(z')\}$ for all $z > z'$. Two notable extensions to this result are: if both X and Z are (multi-valued) discrete random variables then the classical IV estimator is a weighted function of local causal effects (Imbens and Angrist, 1994); and Angrist et al. (2000) discuss an interpretation of the classical IV estimator (2) when X is continuous and $Y(x)$ is a smooth function of x. Crucially, all of these results hold no matter what the scientific model, and so are valid for binary outcomes.

A connection between local estimators and SMMs has already been alluded to. From Section 5, we know that the G-estimator for the additive SMM equals the classical IV estimator (2). Hence, if NEM fails but selection is monotonic then the additive SMM is consistent for the LATE; a similar relationship holds for the multiplicative SMM. Hernán and Robins (2006) shows that the G-estimator for the multiplicative SMM (21) is consistent for the local risk ratio LRR $= E\{Y(1)|X(1) > X(0)\}/E\{Y(0)|X(1) > X(0)\}$; Angrist (2001) also derived this estimator but without reference to SMMs. Hence, the multiplicative SMM is consistent for LRR if NEM fails. Conversely, the local odds ratio

$$\text{LOR} = \frac{E\left\{Y\left(1\right)|X\left(1\right) > X\left(0\right)\right\}}{E\left\{1 - Y\left(1\right)|X\left(1\right) > X\left(0\right)\right\}} \Big/ \frac{E\left\{Y\left(0\right)|X\left(1\right) > X\left(0\right)\right\}}{E\left\{1 - Y\left(0\right)|X\left(1\right) > X\left(0\right)\right\}},$$

is not the estimand of the double-logistic SMM if NEM fails unless $E\{Y(1)|X(1) = X(0) = 1\} = E\{Y(1)|X(1) > X(0)\}$ (Clarke and Windmeijer, 2009, app. 3), but a consistent estimator for LOR is given by Abadie (2003, eqs. 3-4).

The no contamination restrictions introduced in Section 5 constitute a stronger form of monotonic selection. For example, a placebo-control trial constrains $X(1) \geq X(0) = 0$, in which case causal effects among the exposed group are equivalent to local effects. Under

such restrictions, all the SMM parameters are identified and can thus be interpreted as local effects (c.f., Greenland, 2000; Robins and Rotnitzky, 2004). Clarke and Windmeijer (2010) review the connection between SMMs and local estimators in more detail.

Didelez et al. (2010) propose 'Wald' estimators for the risk ratio and odds ratio based on extending (2); these are written

$$\begin{aligned} \text{WaldRR} &= \exp\left[\frac{\log\{E(Y|Z=1)\} - \log\{E(Y|Z=0)\}}{E(X|Z=1) - E(X|Z=0)}\right], \\ \text{WaldOR} &= \exp\left[\frac{\text{logit}\{E(Y|Z=1)\} - \text{logit}\{E(Y|Z=0)\}}{E(X|Z=1) - E(X|Z=0)}\right], \end{aligned}$$

where $\text{logit}(p) = \log\{p/(1-p)\}$ and are approximately unbiased for CRR and COR, respectively, if X is symmetrically distributed, the true causal effect is small, and there is no effect modification by U. It is also straightforward to show that WaldRR and WaldOR are approximately consistent for LRR and LOR, respectively. For instance, a second-order Taylor series of $\text{logit}(p)$ around $p = 0.5$ is $2(2p-1)$, and applying this approximation to the numerator of WaldOR together with the usual arguments for local effects (e.g., Angrist et al., 1996) it follows that WaldOR $\simeq$ LOR.

If covariates are included in the model then estimators for covariate-conditional local treatment effects (i.e., where X is binary) can be applied. The earliest approaches to this problem are based on fully parametric specification of the generating models, with estimators based either on maximum likelihood using the EM algorithm or Markov chain Monte Carlo using data augmentation (e.g., Imbens and Rubin, 1997; Hirano et al., 2000; Yau and Little, 2001). Tan (2006) highlights that, as in Section 6, these approaches are sensitive to unverifiable parametric assumptions.

Abadie (2003) and Tan (2006) consider alternative, more robust approaches for estimating covariate-conditional local treatment effects. Abadie (2003) shows that the local expectation of any function $h(Y, X, C)$ can be identified by a weighted estimating equation, and that these weights are straightforward to calculate. In practice, this allows the specification of a weighted estimator based on a working model for the 'local average re-

sponse function' (LARF). The LARF is defined as $E\{Y|C, X = x, X(1) > X(0)\}$ and it identifies $E\{Y(x)|C, X(1) > X(0)\}$ under the IV core conditions and monotonic selection. In practice, a semi-parametric estimator for LARF can be based on weighted least-squares for the residual $h(Y, X, C) = Y - g(X, C)$, where $g(X, C)$ is a semi-parametric working model for LARF (for example, $g(X, C) = \Phi(\xi_1 X + \xi_2 C)$); similarly, a fully parametric specification of LARF can be used to derive a score function $h(Y, X, C)$ that can be weighted. In either case, if h is correctly specified then the weighted estimand is correct and the estimator is consitent. More realistically, working model h will be mis-specified, but either weighted estimator will be robust in the sense that its estimand corresponds to the working model that is closest to the truth: for the weighted least-squares estimator distance is measured in terms of mean-square error, whereas for the weighted maximum likelihood estimator it is measured by Kullback-Leibler distance.

Tan (2006) proposes two alternative approaches to the same problem. First, his 'regression estimator' hinges on the weak assumption that the selection model takes the form $X(z) = I\{\pi(Z, C) - V \geq 0\}$, where selection probability $\pi(Z, C) = \Pr(X = 1|Z, C)$ and V is uniform on $(0, 1)$ (Tan, 2006, prop. A.1). From this it follows that the association model can be written as $E(Y|X, Z, C) = \eta\{X, C, \pi(Z, C)\}$, that is, a function of the selection probabilities $\pi(X, C)$ as well as exposure and covariates. Under monotonic selection, CMI ($E\{Y(x)|Z, C\} = E\{Y(x)|C\}$) leads to an expression for the required local expectations in terms of these two models; for example,

$$E\{Y(1)|X(1) > X(0), C\} = \frac{\pi(1, C)\eta\{1, C, \pi(1, C)\} - \pi(0, C)\eta\{1, C, \pi(0, C)\}}{\pi(1, C) - \pi(0, C)}.$$

In practice, generalised linear models for π and η are specified, where the association model includes the selection probabilities in the linear predictor. A two-stage regression estimator (first-stage: fit the selection model π; second-stage: fit association model $\eta(X, C, \widehat{\pi})$) is consistent provided that both models are correctly specified. The key distinction between the regression estimator and previous likelihood-based approaches is

that both sets of semi-parametric assumptions relate to distributions of observed quantities, which in principle can be empirically verified. Tan (2006) also develops the 'weighting method' estimator for $E\{Y(x)|X(1) > X(0)\}$, which is based on models for $E(X|Z)$, $E(Y|X)$ and the conditional distribution of Z given C. Combining both approaches, a double-robust, semi-parametric efficient estimator can be obtained with similar properties to the G-estimators and double-SMM estimators in Section 5 (Tan, 2006, p. 1612).

8 Discussion

Dawid (2000) highlights that causal inference generally requires assumptions that are 'metaphysical', in that the observed data alone contain insufficient information to identify the causal parameters. The core conditions which must be satisfied by an IV are inherently metaphysical. Throughout this paper, we have side-stepped this issue and assumed that the analyst has available a valid IV satisfying all three core conditions, but even in randomised experiments, where the IV corresponds to the randomisation indicator, only condition P1 will automatically be satisfied: condition P3 can be verified empirically but the exclusion restriction (P2) depends on the study units' compliance decisions being independent of their outcomes. For example, in unblinded clinical trials where patients can make their own informal prognoses and judgements, condition P2 will not always hold. Similarly, in Mendelian randomisation studies, genetic IVs are assessed with respect to conditions P2 and P3 on the basis of continually developing scientific understanding (e.g., Lawlor et al., 2008).

Chesher (2010) proves that it is only possible to set-identify, or bound, causal effects for non-linear scientific models under the IV core conditions. Two approaches to calculating bounds were considered in Section 3, but in practice are difficult to calculate in general scenarios where there are covariates and X and Z are non-binary. Bounds can be wide if the non-compliance rate is high and the IV is weakly associated with the

exposure. However, this is a true reflection of how informative the data are about the causal effect, and not a drawback with the method itself.

From the analyst's perspective, point estimates of population causal parameters are desirable but require further assumptions. From Chesher's analysis, it is seen that identification can generally be obtained in two ways: via parametric assumptions about the latent variables (that is, (U, V) in our generating model framework); or via direct assumptions linking observed and counterfactual potential outcomes that exploit the conditional mean independence (CMI) assumption (4). The rule of thumb is that better estimators have more empirically verifiable identification assumptions (i.e., ones which could be tested if the sample size was infinite) and fewer unverifiable metaphysical assumptions.

Of th estimators we consider, those longest established are generalised method of moments (GMM) and maximum likelihood (ML). GMM estimators (Section 4) do not fit into Chesher's framework but are instead non-linear analogues of two-stage least-squares (2SLS). While these estimators cannot be consistent for population causal parameters, the bias can be small in scenarios where rich covariate information is included in the analysis, thus making plausible the assumption that selection is approximately ignorable; for example, the marginal estimator is shown in simulations to have small bias for scenarios with approximately ignorable selection (Ten Have et al., 2003). Conversely, ML estimators (Section 6) can be consistent but are identified through a wholly metaphysical parametric specification of the latent variable distribution.

Structural mean model (SMM) estimators for causal effects in the exposed group have a number of advantages over GMM and ML. SMMs are essentially semi-parametric assumptions about how the causal effect varies with X and Z (and possibly C). Identification is obtained by additionally assuming that the causal effect among the exposed does not vary with Z, that is, there is no effect modification by Z (NEM). Clarke and Windmeijer (2010) highlight that NEM implicitly places strong restrictions on the un-

known (and unknowable) generating model, and so is inherently metaphysical in nature. Using a simulation study, they show that even minor failures of NEM for SMMs can lead to disproportionately large bias. In one example, data are generated in such a way that the parameters of the multiplicative SMM are $\exp(\psi_0) = 1.063$ and $\exp(\psi_0 + \psi_1) = 1.068$ (a 0.5 percent difference) so that the risk ratio among the exposed equals 1.066, but the relative bias of its G-estimator is $1.122 - 1.066$ or 5.3 percent (Clarke and Windmeijer, 2010, sec. 6.3). The double-logistic and double-probit SMM estimators further require semi-parametric assumptions about the association model for the conditional distribution of Y given X, Z and C, but these are empirically verifiable.

Encouragement designs which satisfy 'no contamination' restrictions on exposure following assignment are a powerful aid for identification. Somer and Zeger (1991) originally proposed an IV estimator for an encouragement design wherein all subjects assigned to the control group are excluded from any sort of treatment, be it the active treatment or a placebo. In their case, an estimator for CRR follows from a multiplicative model that incorporates plausible assumptions about the compliance behaviour and outcomes in the control group, in the counterfactual scenario where the controls can select treatment. More generally, no contamination restrictions such as those in randomised placebo-controlled trials enforce constraints like $\Pr(X = 0|Z = 0) = 1$. Greenland (2000) develops an IV estimator for CRR among the exposed group, and Robins and Rotnitzky (2004) explicate the important role these play in identifying the parameters of SMMs. In fact, the power of these constraints for identifying causal parameters is a limiting case of the behaviour of identified sets for discrete-outcome structural models, in which set-identification tends to point-identification as the strength of association between the IV and the exposure increases (Chesher, 2010, sec. 2).

Estimators of local effects require weaker assumptions than those for population causal effects. We have highlighted the close correspondence between saturated SMMs and local

estimators of additive causal effects and risk ratios (e.g., Hernán and Robins, 2006), and also more general identification results for parameters of the complier group (Abadie, 2003). Dawid (2000) criticises the potential outcomes approach for, among other things, its focus on causal effects for a subgroup whose individual members cannot be identified. Moreover, local estimators require that the selection model is monotonic. Monotonicity is an inherently metaphysical assumption that can be violated by plausible selection models in which the effect of the IV on selection varies between subjects. An exception to this rule is for study designs which constrain selection to be monotonic; for example, encouragement designs with no contamination restrictions effectively force a strong type of monotonicity because constraints like $\Pr(X = 0|Z = 0) = 1$ imply that $X(1) \geq X(0) = 0$ with probability 1.

Tan (2006) develops two alternative estimators for covariate-conditional local expectations. In doing so, he makes a distinction between 'modern' IV frameworks like his and those based on a fully parametric specification (e.g., Imbens and Rubin, 1997). Chesher (2010) does not explicitly consider the identification of local causal effects, but CMI plays a crucial role here too. For example, Tan's 'regression estimator' is based on an expression of the covariate-conditional local expectation in terms of two empirically verifiable semi-parametric models, and this expression follows only if CMI holds. While SMMs require stronger assumptions than models for local parameters, these estimators still fall within a modern framework: G-estimators for additive and multiplicative SMMs (Robins, 1994), and the double estimators for logistic and probit SMMs (Vansteelandt and Goetghebeur, 2003), generally rely on the metaphysical NEM assumption, but the double-estimators' association models are empirically verifiable, and all of these estimators are locally robust in the sense of being consistent if the SMM is mis-specified under the 'sharp' null hypothesis that there is no causal effect.

To conclude, we note that extensive simulation studies comparing different binary IV

estimators have been conducted by Didelez et al. (2010) and Vansteelandt et al. (2010). These simulations are based on specific choices of generating model, and in the latter case explicitly constrain NEM to hold, but in applications the form of this model cannot be known; indeed, there is no evidence that the logistic and logistic SMM models used to simulate data in these studies approximate realistic generating models at all. Instead, we have highlighted the inherently metaphysical nature of the identification assumptions; how the strength of these assumptions increases with the target parameter; and how different estimators exploit parametric functional form or CMI to obtain identification. The most sophisticated estimators like SMMs seek to limit assumptions where possible to be semi-parametric and empirically verifiable; moreover, robustness to model mis-specification can be incorporated using recent developments in estimating equation theory. In practice, the fundamentally metaphysical nature of the binary IV problem means that researchers should ideally look to assess robustness of inferences using sensitivity analyses involving a number of alternative estimators (e.g., Rassen et al., 2008; Clarke and Windmeijer, 2010; Vansteelandt et al., 2010), but with added weight given to those estimators which do not rely on fully parametric assumptions.

References

[1] Abadie, A. (2003), Semiparametric instrumental variable estimation of treatment response models, Journal of Econometrics 113, 231-263.

[2] Angrist, J.D. (2001), Estimation of limited dependent variable models with dummy endogenous regressors: simple strategies for empirical practice (with discussion), Journal of Business and Economic Statistics 19, 2-28.

[3] Angrist, J.D., Graddy, K. and Imbens, G.W. (2000), The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish, Review of Economic Studies 67, 499-527.

[4] Angrist, J.D., Imbens, G.W. and Rubin, D.B. (1996), Identification of causal effects using instrumental variables, Journal of the American Statistical Association 91, 444-455.

[5] Balke, A. and Pearl, J. (1997), Bounds on treatment effects from studies with imperfect compliance, Journal of the American Statistical Association 92, 1171-1176.

[6] Blundell, R.W. and Powell, J.L. (2004), Endogeneity in semiparametric binary response models, Review of Economic Studies 71, 655-679.

[7] Carroll, R.J., Ruppert, D., Stefanski, L. and Crainiceanu, C. (2006), Measurement Error in Nonlinear Models: A Modern Perspective (2nd. edition), London: Chapman & Hall/CRC Press.

[8] Clarke, P. and Windmeijer, F. (2009), Instumental variable estimators for binary outcomes, CMPO Working Paper 09/209, Centre for Market and Public Organisation, University of Bristol, UK.

[9] Clarke, P.S. and Windmeijer, F. (2010), Identification of causal effects on binary outcomes using structural mean models, Biostatistics (in press).

[10] Chesher, A. (2010), Instrumental variable models for discrete outcomes, Econometrica 78, 575-601.

[11] Dawid, A.P. (2000), Causal inference without counterfactuals (with discussion), Journal of the American Statistical Association 95, 407-448.

[12] Dawid, A.P. (2003), Causal inference using influence diagrams: the problem of partial compliance (with discussion), in: Green, P.J., Hjort, N.L. and Richardson, S., Highly Structured Stochastic Systems, Oxford: Oxford University Press.

[13] Didelez, V., Meng, S. and Sheehan, N. (2010), On the bias of IV estimators for Mendelian Randomisation, Statistical Science (in press).

[14] Didelez, V. and Sheehan, N. (2007), Mendelian randomization as an instrumental variable approach to causal inference, Statistical Methods in Medical Research 16, 309-330.

[15] Frangakis, C.E. and Rubin, D.B. (2002), Principal stratification in causal inference, Biometrics 58, 21-29.

[16] Goetghebeur, E. and Vansteelandt, S. (2005), Structural mean models for compliance analysis in randomized clinical trials and the impact on measures of exposure, Statistical Methods in Medical Research 14, 397-415.

[17] Goldberger, A.S. (1972), Structural equation methods in social sciences, Econometrica 40, 979-1001.

[18] Greenland, S. (2000), An introduction to instrumental variables for epidemiologists, International Journal of Epidemiology 29, 722-729.

[19] Greenland, S., Pearl, J. and Robins, J.M. (1999), Confounding and collapsibility in causal inference, Statistical Science 14, 29-46.

[20] Heckman, J. (1979), Dummy endogenous variables in a simultaneous equation system, Econometrica 46, 931-959.

[21] Hernán, M.A. and Robins, J.M. (2006), Instruments for causal inference: an epidemiologist's dream?, Epidemiology 17, 360-372.

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

[23] Hogan, J.W. and Lancaster, T. (2004), Instrumental variables and inverse probability weighting for causal inference from longitudinal studies, Statistical Methods in Medical Research 13, 17-48.

[24] Imbens, G.W. (2001), Comment on: Angrist, J.D., Estimation of limited dependent variable models with dummy endogenous regressors: simple strategies for empirical practice, Journal of Business and Economic Statistics 19, 17-20.

[25] Imbens, G.W. and Angrist,. J. (1994), Identification and estimation of local average treatment effects, Econometrica 62, 467-476.

[26] Imbens, G.W. and Lemieux, T. (2008), Regression discontinuity designs: a guide to practice, Journal of Econometrics 142, 615-635.

[27] Imbens, G.W. and Rubin, D.B. (1997), Bayesian inference for causal effects in randomized experiments with noncompliance, Annals of Statistics 25, 305-327.

[28] Johnston, K.M., Gustafson, P., Levy, A.R., and Grootendorst, P. (2008), Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research, Statistics in Medicine 27, 1539-1556.

[29] Lawlor, D.A., Harbord, R.M., Sterne, J.A.C., Timpson, N. and Davey Smith, G. (2008), Mendelian randomization: using genes as instruments for making causal inferences in epidemiology, Statistics in Medicine 27, 1133-1163.

[30] Lee, L.F. (1981), Simultaneous equation models with discrete and censored dependent variables, in: Manski, C. and McFadden, D. (eds.), Structural Analysis of Discrete Data with Economic Applications, Cambridge, MA: MIT Press.

[31] Manski, C.F. (1990), Nonparametric bounds on treatment effects, American Economic Review 80, 319-324.

[32] Mullahy, J. (1997), Instrumental-variable estimation of count data models: applications to models of cigarette smoking behavior, Review of Economics and Statistics 79, 586-593.

[33] Nagelkerke, N., Fidler, V., Bernsen, R. and Borgdorff, M. (2000), Estimating treatment effects in randomized clinical trials in the presence of non-compliance, Statistics in Medicine 19, 1849-1864.

[34] Palmer, T.M., Thompson, J.R., Tobin, M.D., Sheehan, N.A. and Burton, P.R. (2008), Adjusting for bias and unmeasured confounding in Mendelian ranomization studies with binary responses, International Journal of Epidemiology 37, 1161-1168.

[35] Rassen, J.A., Schneeweiss, S., Glynn, R.J., Mittleman, M.A. and Brookhart, M.A. (2008), Instrumental variable analysis for estimators of treatment effects with dichotomous outcomes, American Journal of Epidemiology 169, 273-284.

[36] Rivers, D. and Vuong, Q.H. (1988), Limited information estimators and exogeneity tests for simultaneous probit models, Journal of Econometrics 39, 347-366.

[37] Robins, J.M. (1989), The analysis of randomised and non-randomised AIDS treatment trials using a new approach to causal inference in longitudinal studies, in: Sechrest, L., Freeman, H. and Mulley, A. (eds.), Health Service Research Methodology: A Focus on AIDS, 113-159, Washington, DC: US Public Health Service, National Center for Health Services Research.

[38] Robins, J.M. (1994), Correcting for non-compliance in randomized trials using structural nested mean models, Communications in Statistics - Theory and Methods 23, 2379-2412.

[39] Robins, J.M., Hernán, M.A. and Brumback, A. (2000), Marginal structural models and causal inference in epidemiology, Epidemiology 11, 550-560.

[40] Robins, J.M. and Ritov, Y. (1997), Towards a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models, Statistics in Medicine 16, 285-319.

[41] Robins, J.M. and Rotnitzky, A. (2004), Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models, Biometrika 91, 763-783.

[42] Rosenbaum, P.R. and Rubin, D.B. (1983), The central role of the propensity score in observational studies for causal effects, Biometrika 70, 41-55.

[43] Rubin, D.B. (2008), For objective causal inference, design trumps analysis, Annals of Applied Statistics 2, 808-840.

[44] Tan, Z. (2006), Regression and weighting methods for causal inference using instrumental variables, Journal of the American Statistical Association 101, 1607-1618.

[45] Ten Have, T.R., Joffe, M. and Cary, M. (2003), Causal logistic models for non-compliance under randomized trials with non-compliance and a dichotomous outcome, Statistics in Medicine 24, 191-210.

[46] van der Laan, M.J., Hubbard, A. and Jewell, N.P. (2007), Estimation of treatment effects in randomized trials with non-compliance and a dichotomous outcome, Journal of the Royal Statistical Society, Series B 69, 463-482.

[47] Vansteelandt, S., Babanezhad, M. and Goetghebeur, E. (2008), Correcting instrumental variables estimators for systematic measurement error, Statistica Sinica 19, 1223-1246.

[48] Vansteelandt, S., Bowden, J., Babanezhad, M. and Goetghebeur, E. (2010), On instrumental variable estimation of the causal odds ratio. Technical Report, Center for Statistics, Ghent University, Ghent, Belgium.

[49] Vansteelandt, S. and Goetghebeur, E. (2003), Causal inference with generalized structural mean models, Journal of the Royal Statistical Society, Series B 65, 817-835.

[50] Vansteelandt, S. and Goetghebeur, E. (2005), Sense and sensitivity when correcting for observed exposures in randomized clinical trials, Statistics in Medicine 24, 191-210.

[51] Wooldridge, J.M. (2002), Econometric Analysis of Cross-sectional and Panel Data, MA: MIT Press.

[52] Yau, L.H.Y. and Little, R.J.A. (2001), Inference for the complier-average causal effect from longitudinal data subject to noncompliance and missing data, with an application to a job training assessment for the unemployed, Journal of the American Statistical Association 96, 1232-1244.

Figure 1: A directed acyclic graph representing conditional independence relationships implied by a structural model for Y given X and U and a non-ignorable selection mechanism, along with the core conditions that must be satisfied by instrumental variable Z. Each node represents a variable (square nodes are observed and circular nodes are unobserved variables) with edges between variables denoting pairs that are not conditionally independent. Directed edges with arrows indicate causal direction, and undirected edges indicate an association about which no causal direction is assumed.