Department of Economics, University College London

Information and admissible sets

Jeff Rowley

14th August, 2014

Abstract

«Abstract here»

Acknowledgements. *I acknowledge the* R *and* LaTeX *communities, and the wealth of knowledge that they have made freely available to all. I thank Andrew Chesher and Toru Kitagawa for their supervision and support. I further thank Adam Rosen for helpful discussion. I gratefully acknowledge financial support from the Economic and Social Research Council (ESRC).*

I explore the effect of incorporating information for a non-parametric binary choice model. The model permits endogenous variation in a scalar random variable that is due to non-random selection, and it is the average causal effect of this endogenous variable on the outcome variable that is of interest. The model embeds an exclusion restriction and an independence restriction that together define an instrumental variable but is silent as to the relationship between the endogenous variable and the instrumental variable. I restrict the relationship between the outcome variable and the endogenous variable up to a non-parametric threshold crossing function. The model is credible (Manski, 2013) in that it embeds restrictions that impose weaker constraints on assumed behaviour, but does not identify the average causal effect of the endogenous variable on the outcome variable.[1] Rather, the model partially identifies the average causal effect of the endogenous variable on the outcome variable.

I define information to be those additional characteristics of economic agents that are observable with the caveat that these characteristics are exogenous and relevant are to the latent structure. It is convenient to think of such characteristics as being predetermined and immutable; characteristics that result from choices that are made jointly with the outcome variable are excluded by the definition. Accordingly, exogenous variables and instrumental variables are each regarded as information, and I distinguish between these classes of information. I study how the admissible set of values for the average causal effect of the endogenous variable on the outcome variable changes as each class of information is incorporated into the model separately.

It is useful to distinguish between classes of information since each class enters the latent structure in a different way. Exogenous variables are permitted to enter the structural equation for the outcome variable and to determine the endogenous variable. As such, exogenous variables can be seen to enrich both individual response and individual selection, respectively. An

[1] Assumptions that cannot be tested using data. The model does embed some non-trivial non-verifiable restrictions that might be relaxed.

important consequence is that the causal effect of the endogenous variable on the outcome variable depends upon the value of the exogenous variables when individual response is enriched. In contrast, instrumental variables are excluded from the structural equation for the outcome variable by definition and so only enrich individual selection. Given this, the effect of incorporating information is different depending upon the class of information that is being incorporated into the model.

Incorporating information of either class is generally sensible for a number of reasons. Firstly, incorporating information is known to be efficient; variation that is attributable to an observable variable is instead attributable to unobservable heterogeneity when that variable is omitted. Secondly, the effect of incorporating information for partially identifying models is not well-documented; one hypothesis is that incorporating information narrows bounds on admissible sets. Such an effect is not documented in identifying models precisely because such models deliver a point estimate (a set of length zero), but point estimates may shift as information is incorporated. A contribution that I make is in showing that incorporating information leads to narrower bounds on the admissible sets that are delivered by the model. A further reason to particularly favour incorporating exogenous variables is that the average causal effect of the endogenous variable on the outcome variable in identifiable sub-populations can be recovered. I name this structural characteristic the conditional average causal effect of the endogenous variable on the outcome variable, and index it by the conditioning value.[2] Understanding the effect of an intervention in sub-populations can be interesting if the intervention can be targeted or if the intervention is to be applied elsewhere in a population that differs according to its observable characteristics.

A relevant question is how to relate conditional causal effects to (unconditional) causal effects. More precisely, how does the average causal effect of the endogenous variable on the outcome variable relate to its conditional counterparts? I show that the average causal effect of the endogenous variable on the outcome variable can be expressed as a Minkowski summation of its conditional counterparts when the non-parametric binary choice model is augmented. I derive sharp bounds on the conditional average causal effect by applying random set theory. I employ the capacity (or containment) functional as in Chesher et al. (2013) as a matter of choice, rather than the Aumann expectation as in Beresteanu et al. (2012). As I show that the average causal effect of the endogenous variable on the outcome variable can be expressed as a Minkowski summation of its conditional counterparts, I derive sharp bounds on the average causal effect of the endogenous variable on the outcome variable.[3] I establish the conditions under which bounds on conditional causal effects can be informative about bounds on (unconditional) causal effects. That is, I establish the conditions under which bounds on conditional causal effects can be used to narrow bounds on (unconditional) causal effects, exploiting the mapping from one to the other.

I demonstrate application of the non-parametric binary choice model, elucidating the practical difficulties that arise when estimating set identifying models (focusing on those issues that arise from incorporating information). As in Chesher and Rosen (2013), I estimate the average causal effect of childbirth on a mother's labour force participation using US census data. I extend Chesher and Rosen (2013) in a number of ways. Firstly, I report statistical uncertainty in the estimate of the average causal effect of childbirth on a mother's labour force participation using a method that is outlined in Chernozhukov et al. (2013). Secondly, I enrich the support of the instrumental variable and explore the effect that this has on the admissible set of values for the average causal effect of childbirth on a mother's labour force participation, and on its accompanying confidence region. Thirdly, I enrich individual response by permitting the structural equation for labour force participation to depend upon the age and other such predetermined and immutable characteristics of mothers. I discuss the complication of calculating statistical uncertainty when exogenous variables are permitted to enter the structural equation for labour force participation. With respect to the second and third extensions, it is necessary that I augment the model by embedding additional restrictions. In fact, Chesher and Rosen (2013) describe the augmented non-parametric binary choice model that I assume but simplify this model for application (by

[2] The conditioning value is specifically the value of the exogenous variables. Heckman and Vytlacil (2005) defines a parameter $ATE(x)$ that is equivalent to the conditional average causal effect of the endogenous variable on the outcome variable at the conditioning value x. Khan and Tamer (2010) and Abrevaya et al. (2013) instead refer to this parameter as the conditional average treatment effect and abbreviate this to $CATE(x)$.

[3] Molchanov (2005) is a useful companion in the study of random sets.

excluding exogenous variables from the structural equation for the outcome variable). I discuss how the augmented model relates to the simplified model in each case and the credibility of the additional restrictions that are embedded in the augmented model.

Related research

Other notable non-parametric binary choice models are described in Balke and Pearl (1997) and Shaikh and Vytlacil (2011), and general non-parametric models of choice are described in Chesher (2005), Kitagawa (2009) and Chesher (2010).

Balke and Pearl (1997) assumes a triangular model (the model embeds a structural equation for the outcome variable and a structural equation for the endogenous variable; see Strotz and Wold (1960) for a detailed discussion of triangular models) that relaxes separability of unobservable heterogeneity in the structural equation for the outcome variable. The cost is that the model is no longer silent as to the relationship between the endogenous variable and the instrumental variable. The model does not permit exogenous variables to enter the structural equation for the outcome variable. I discuss the credibility of separability of unobservable heterogeneity in the main text. Shaikh and Vytlacil (2011) assumes a triangular model but maintains separability of unobservable heterogeneity in the structural equation for the outcome variable. The model permits exogenous variables to enter the structural equation for the outcome variable.

Chesher (2005) and Kitagawa (2009) describe non-parametric models that permit continuous variation in the outcome variable. Chesher (2005) assumes a triangular model that relaxes separability of unobservable heterogeneity in the structural equation for the outcome variable. The model permits exogenous variables to enter the structural equation for the outcome variable, although local invariance of the structural equation for the outcome variable to variation in the exogenous variables is embedded. The model is uninformative when there is binary variation in the endogenous variable but is informative when there is discrete variation. Kitagawa (2009) extends Balke and Pearl (1997) to permit discrete and continuous variation in the outcome variable, and studies commonly invoked restrictions on covariation of the instrumental variable and unobservable heterogeneity.

Chesher (2010) describes an ordered choice model that permits discrete variation in the outcome variable. Chesher (2010) assumes a single equation model that relaxes separability of unobservable heterogeneity in the structural equation for the outcome variable, although monotonicity of the structural equation for the outcome variable in unobservable heterogeneity is embedded. The model permits binary or discrete variation in the endogenous variable.

Notation

I study a probability space $(\Omega, \Sigma, \mathbb{P})$. I define random variables on this probability space. I write random variables as upper case Latin letters, and I write realisations (or specific values) of random variables as lower case Latin letters. I write the support of A as $\mathcal{R}_A$. I write the counterfactual value of A when B has a causal interpretation and is externally fixed as $A(b)$. I write the average causal effect of B on A as $ACE(B \to A)$, and the conditional average causal effect of B on A given C as $ACE(B \to A|c)$.

I refer to Y as the outcome variable, to D as the endogenous variable, to X as the endogenous variable, to Z as the instrumental variable, and to U as unobservable heterogeneity. Despite the use of *the*, I permit (X, Z) to be vectors. I write the structural equation for the outcome variable as h, and the structural equation for the endogenous variable as g.

I write the expectation operator as $\mathbb{E}$, and the indicator function as $\mathbb{1}$. I write A is independent of B as $A \perp\!\!\!\perp B$. To distinguish between population and sample quantities, I subscript sample quantities by n.

I introduce further terminology and notation in Figure 1 through Figure 4. This specifically relates to models and structures, and is consistent with the approach that is formally laid out in Hurwicz (1950) and in Koopmans and Reiersøl (1950).

Application

I estimate the average causal effect of childbirth on a mother's employment using United States census data. The data are obtainable from Angrist (2014), and are described in Angrist and Evans (1998). To summarise, the dataset consists of 254,654 households that were recorded as part of the 1980 United States census. The dataset specifically contains observations of married households with at least two children under the age of 18 years and where the mother is aged between 21 years and 35 years. For clarity, I translate each variable in the data into the mathematical notation that I employ.

$$Y \equiv \mathbb{1}[\text{Mother is employed in 1979}]$$
$$D \equiv \mathbb{1}[\text{Three or more children in the household}]$$

As (X, Z) are continually redefined in the main text, I do not define these variables as I do (Y, D). Instead, I note those variables that at some point or other form part of the definition of (X, Z). X is a function of mother's race or ethnicity (shortened to race).[4] Z is a function of whether the oldest two children in the household share the same gender (shortened to child gender), and whether the second pregnancy was a multiple pregnancy.

I refer to the application throughout the main text so as to illustrate how technical conditions on variables and on the relationship between variables restrict the behaviour of economic agents, in this case mothers. For brevity, I simply refer to race when discussing X in the context of the application, and child gender when discussing Z in the context of the application.

1 A non-parametric model of binary choice

Axiom. *Economic agents are utility maximising, selecting between alternatives in a choice set according to the utility that they attach to that choice. Utility is perfectly observable by economic agents and is determined by a well-defined utility function for each choice. Each agent is permitted to value each choice differently.*

I introduce the non-parametric binary choice model that is described in Chesher and Rosen (2013) (hereafter, the single equation model). The single equation model constitutes the set of structures that are consistent with Restriction M1 through Restriction M5.

M1. Discrete support. (Y, D, X, Z) are observable and have discrete supports (with at least two points of support). Further, (Y, D) have binary supports and are normalised such that

(a) $\mathcal{R}_Y = \{0, 1\}$ and

(b) $\mathcal{R}_D = \{0, 1\}$,

respectively.

Restriction M1 is a verifiable restriction. (Y, D, X, Z) are observable and so it is trivial to verify that each variable satisfies its support restriction. (Y, D) are normalised to be consistent with the application, but any other supports $\{y_0, y_1\}$ and $\{d_0, d_1\}$ can be generated by an affine transformation of h and of g.

M2. Scalar U. U is an unobservable scalar such that $\mathcal{R}_U$ is an open subset of $\mathbb{R}$ with strictly positive Lebesgue measure.

Restriction M2 is a non-verifiable restriction. The restriction is not overly restrictive in that it is equivalent to those determinants of utility that are unobservable being defined on a set of cardinality no greater than $2^{\aleph_0}$. As most economic variables are defined on $\mathbb{R}$ or on a subset of $\mathbb{R}$ it is uncontroversial to assume that this restriction is satisfied. In the context of the application, the restriction implies that variables such as job application ability and taste for leisure are quantifiable and can be defined on $\mathbb{R}$.

[4] Angrist and Evans (1998) also treats mother's age, and mother's age at the time of her first birth (shortened to birthing age) as exogenous variables.

M3. Joint independence. $U \perp\!\!\!\perp (X, Z)$.

Restriction M3 is a non-verifiable restriction. The restriction nests the restrictions $U \perp\!\!\!\perp X$ and $U \perp\!\!\!\perp Z$ (marginal independence), and is the strongest possible restriction that might be imposed upon the joint distribution of (X, Z, U). The restriction excludes some correlation structures of (X, Z) that are permitted by marginal independence. In the context of the application, the restriction implies that variables such as job application ability and taste for leisure are independent of combinations of race and child gender, and are independent of race and of child gender.

M4. Exclusion. $Y = h(D, X, U)$.

Restriction M4 is a non-verifiable restriction. The restriction excludes Z from h and so excludes Z from having a causal effect on Y. The restriction is equivalent to an order condition. In the context of the application, the restriction implies that child gender does not have a causal effect on the employment rate of mothers.

M5. Monotonicity. h is a non-parametric threshold crossing function that is separable in U. h is normalised to be increasing in U, and U is normalised to be distributed uniformly on the unit interval.

Restriction M5 is a non-verifiable restriction. The restriction implies that individual response is monotonic. In the context of the application, the restriction implies that the causal effect of childbirth on a mother's employment is positive for all mothers or is negative for all mothers. The restriction permits the threshold to be a non-parametric function of (D, X) and implies that the distribution of U can be relatively unrestricted beyond Restriction M2.

M6. Relevance. There exist values (z, z') such that $\mathbb{P}(d|z) \neq \mathbb{P}(d|z')$ for all $d \in \mathcal{R}_D$.

Restriction M6 is a verifiable restriction. The restriction states that Z covaries with D. A simple interpretation is that Z causes D, but the restriction itself is weaker than this in that it permits Z to be correlated with a variable that causes D. Despite this, I opt to refer to Z as an instrumental variable. The restriction is equivalent to a rank condition. In the context of the application, the restriction implies that the probability of having three or more children varies with child gender. For example, if the probability of having three or more children is greater when the oldest two children in the household share the same gender.

The single equation model partially identifies $ACE(D \to Y)$. The single equation model also partially identifies $ACE(D \to Y|x)$ for all $x \in \mathcal{R}_X$. Restriction M1 through Restriction M6 can be written more compactly as Restriction M1' through Restriction M6'.

M1'. $Y = \mathbb{1}[p(D, X) > U]$.

M2'. $U|(X, Z) \sim \text{unif}(0, 1)$.

M3'. $\mathbb{P}(d|z) \neq \mathbb{P}(d|z)$ for all $d \in \mathcal{R}_D$ and for some $(z, z') \in \mathcal{R}_Z$.

M4'. $\mathcal{R}_D = \{0, 1\}$.

M5'. $\mathcal{R}_X = \{x_1, ..., x_K\}$ and $K < \infty$.

M6'. $\mathcal{R}_Z = \{z_1, ..., z_L\}$ and $L < \infty$.

2 Credibility in economic modelling

Credibility is a statement of the validity and the plausibility of the restrictions that a model embeds, and is a desirable property. The need to discuss both validity and plausibility arises because restrictions can be either verifiable or non-verifiable. The distinction between verifiable and non-verifiable restrictions is that verifiable restrictions are testable using data while non-verifiable restrictions cannot be tested even if data is collected for the population. As verifiable restrictions can be rejected or not rejected on the basis of observed behaviour, it makes sense to talk about such restrictions as being valid or invalid. In contrast, the validity of non-verifiable

restrictions is indeterminable. Whether to accept a set of non-verifiable restrictions as an accurate representation of how economic agents behave is subjective and depends upon how plausible the restrictions seem. Restrictions that are founded in economic theory, or that impose weaker constraints on assumed behaviour are more plausible (a view that is consistent with Occam's razor, a widely accepted principle of parsimony).

I regard a model as incredible if the verifiable restrictions that it embeds are invalid. I regard a model as more credible relative to another if the verifiable restrictions that it embeds are valid, and if the sum of the non-verifiable restrictions that it embeds are more plausible. Manski (2013) adopts an equivalent stance, formalised as The Law of Decreasing Credibility.

Models that embed restrictions that impose weaker constraints on assumed behaviour are typically not uniformly identifying. Instead, such models are typically partially identifying. More commonly,

(a) a more restrictive model is assumed that identifies a feature of interest; or,

(b) identification of a different feature is sought and a model that embeds restrictions that impose weak constraints on assumed behaviour is assumed.

I suggest that these responses to partial identification are motivated by two concerns. Namely, that characterising the admissible set of structures or the admissible set of values for a structural characteristic of interest can be complex and computationally difficult, and that partially identifying models do not produce unique conclusions. Although tractability is a legitimate concern, there is an inherent and widespread misunderstanding that models that do not produce unique conclusions are inferior regardless of the restrictions that they embed. Conclusions that are produced by more credible models should always be preferred, even if these conclusions display ambiguity.

I caution against both responses. In the first response, a more restrictive model is assumed specifically for the purpose of achieving identification. Koopmans and Reiersøl (1950) remarks that a model should be constructed purely from prior knowledge of the studied behaviour, and to do otherwise violates scientific honesty. Similarly, Manski (2013) remarks that this response displays incredible certitude. In the second response, the feature that is identified is often less valuable than the original feature of interest. Nonetheless, it is promising that the second response should implicitly recognise the importance of credibility.

2.1 The credibility of the single equation model

I discuss the credibility of the single equation model, generally and in the context of the application. I focus on the non-verifiable restrictions that the model embeds since it is these restrictions that are of principal interest when selecting from competing models.

Firstly, the single equation model embeds the restriction that h is a non-parametric function (Restriction M4). In general, non-parametric restrictions are plausible since they permit the output of a function to depend arbitrarily on the value of its arguments. Non-parametric functions are flexible and are able to capture arbitrary variation that could otherwise only be captured using high-order polynomial functions or indicator functions. In particular, non-parametric functions are well-suited to capturing interaction between the arguments of a function. For example, if the difference in the employment rate of mothers between the counterfactual environments of two children in the household versus three or more children in the household varies systematically with race. Non-parametric functions are also well-suited to settings in which an argument is a categorical (and discrete) variable with no natural ordering. For example, race is a categorical variable with no natural ordering.

Secondly, the single equation model embeds the normalisation that U is distributed uniformly on the unit interval (Restriction M5). Note that, as a normalisation and not as a restriction, it imposes no constraints on the distribution of U. In general, restrictions that impose constraints on the distribution of U are implausible. This follows from the definition of U as a projection of unobservable determinants of utility onto an ordered set: there is no reason to suppose that economic agents should be distributed on this set according to some well-behaved distribution. Further, the normalisation is a normalisation and not a restriction because the single equation model embeds the restriction that h is a non-parametric function. The restriction that h is a

non-parametric function implies not only that h depends arbitrarily on the value of its argument, but that it permits U to be distributed non-parametrically.

Thirdly, the single equation model embeds the restriction that h excludes Z. Establishing that the relationship between Y and D is causal necessitates the existence of an external and measurable factor that causes variation in D. This is the subject of Figure 5. The restriction that h excludes Z is then an essential element of the single equation model if identification of causality is desired and if Z is associated with the external and measurable factor.[5] Further, this is true of any model of causality and not just the single equation model. In the context of the application, the restriction implies that child gender does not have a causal effect on the employment rate of mothers. If child gender affects the cost of employment by decreasing the value of leisure say, then this condition is violated. Even if the condition does not hold exactly, the effect of child gender on the employment rate of mothers is likely to be small relative to the effect of additional children in the household and so the restriction is plausible.

Fourthly, the single equation model is silent as to the relationship between D and Z. More importantly, it is silent as to the determination of D. That is, the single equation model embeds the restriction that the codomain of g is $\{0,1\}$ but otherwise is silent as to the arguments and functional form of g. The lack of constraints on g is important since the single equation model then permits any endogenous relationship between Y and D, and does not restrict the relationship between D and Z to be causal. In general, the lack of constraints on g is a plausible restriction. In the context of the application, it is plausible that there should be no feedback from the employment of a mother to whether there are three or more children in the household contemporaneously. If such a feedback is present then it is likely from lagged employment and not from contemporaneous employment. Further, it is plausible that child gender should have a causal effect on whether there are three or more children in the household since child gender is likely to enter the preferences of mothers over children. This raises a further point that the single equation model might be strengthened to embed some restrictions that impose constraints on g. Such restrictions need only be weak but might permit other restrictions that are embedded in the single equation that are not necessarily as plausible to be weakened.

Fifthly, the single equation model embeds the restriction that $U \perp\!\!\!\perp (X, Z)$.

Sixthly, the single equation model embeds the restriction that h is separable in U. In general, this is an implausible restriction since it is implies that individual response is monotonic. In the context of the application, the restriction implies that the causal effect of an additional child in the household on a mother's employment is positive for all mothers or is negative for all mothers. There is no reason to suppose that this should be the case. A mother that uses a paid-for childcare service might find the cost of such services prohibitive with the addition of a child in the household and so exit employment. In contrast, a mother that relies on family members for childcare might find need to return to employment to feed the additional mouth. Non-separability of U enriches individual response and is a more plausible restriction. If h were non-separable in U then Restriction M1' would be replaced by the equation

$$Y = \mathbb{1}[p(D, X, U) > 0].$$

In general, non-separability of U permits the causal effect of (D, X) to vary with U.

As a final remark, Restriction M6 is a weaker restriction than the monotonicity restriction that is embedded in the model that is described in Imbens and Angrist (1994) (hereafter, the late model), which identifies the local average treatment effect. The restriction permits the existence of compliers and defiers, and is satisfied whenever the monotonicity restriction that is embedded in the late model is satisfied. In the context of the application, it is plausible that mothers should prefer to have additional children if the oldest two children in the household share the same gender. Both the monotonicity restriction that is embedded in the late model and Restriction M6 that is embedded in the single equation model are plausible in this case.

[5] I reiterate that it is not necessary for Z to be the external and measurable factor that causes D, only that it is associated with the external and measurable factor. This point is made in Figure 5.

3 Incorporating information

3.1 Enriching the support of the instrumental variable

3.2 Enriching individual response

First of all what is the parameter that Chesher and Rosen (2013) identifies? They are after $ACE(D \to Y)$ which is a function of $Y(1)$ and $Y(0)$. In the model that doesn't include the Xs, the parameter that they identify is $p(d)$. What is this?

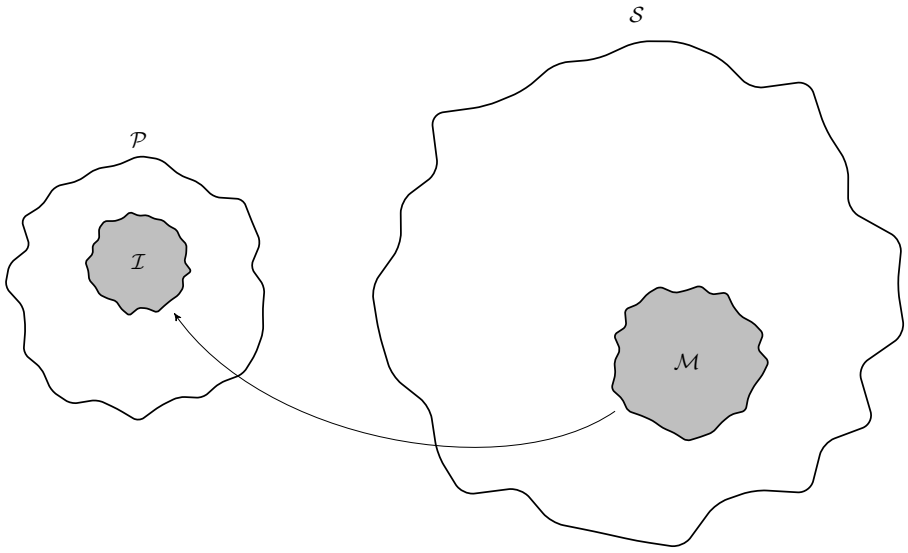
$$\begin{aligned} Y &= \mathbb{1}[p(D) > U] \\ \mathbb{E}[Y(d)] &= \mathbb{E}[\mathbb{1}[p(d) > U]] \\ &= \mathbb{P}(U < p(d)) \\ &= p(d) \end{aligned}$$

So $p(d)$ is equal to $\mathbb{E}[Y(d)]$. So what is this in the augmented model?

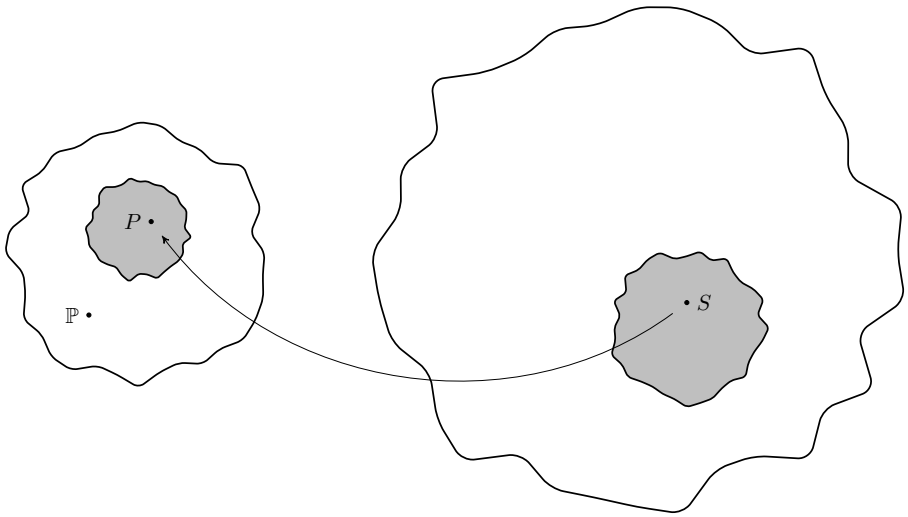
$$\begin{aligned} Y &= \mathbb{1}[p(D, X) > U] \\ \mathbb{E}[Y(d)] &= \mathbb{E}[\mathbb{1}[p(d, X) > U]] \\ &= \mathbb{E}[\mathbb{E}[\mathbb{1}[p(d, X) > U] | X = x]] \\ &= \mathbb{E}[\mathbb{E}[\mathbb{1}[p(d, x) > U]]] \\ &= \sum_{x \in \mathcal{R}_X} \mathbb{P}(x) \mathbb{E}[\mathbb{1}[p(d, x) > U]] \\ &= \sum_{x \in \mathcal{R}_X} \mathbb{P}(x) \mathbb{P}(U < p(d, x)) \\ &= \sum_{x \in \mathcal{R}_X} \mathbb{P}(x) p(d, x) \end{aligned}$$

Hence, I have that

$$p(d) = \sum_{x \in \mathcal{R}_X} \mathbb{P}(x) p(d, x)$$

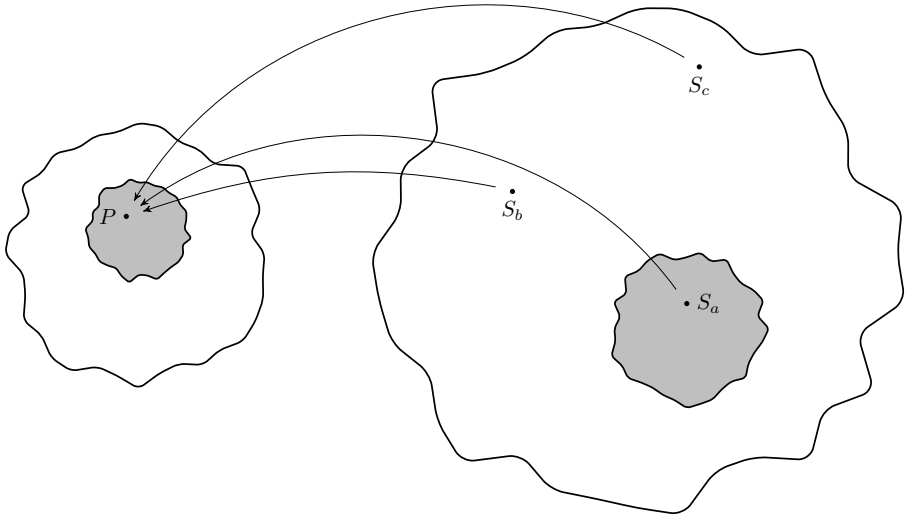


(a) A model $\mathcal{M}$ is a set of structures that forms a proper subset of the class of all structures $\mathcal{S}$. Each structure in $\mathcal{M}$ generates a probability distribution in the class of all probability distributions (of observable variables) $\mathcal{P}$. Then the image $\mathcal{I}$ is the set of all probability distributions that are generated by structures in $\mathcal{M}$.

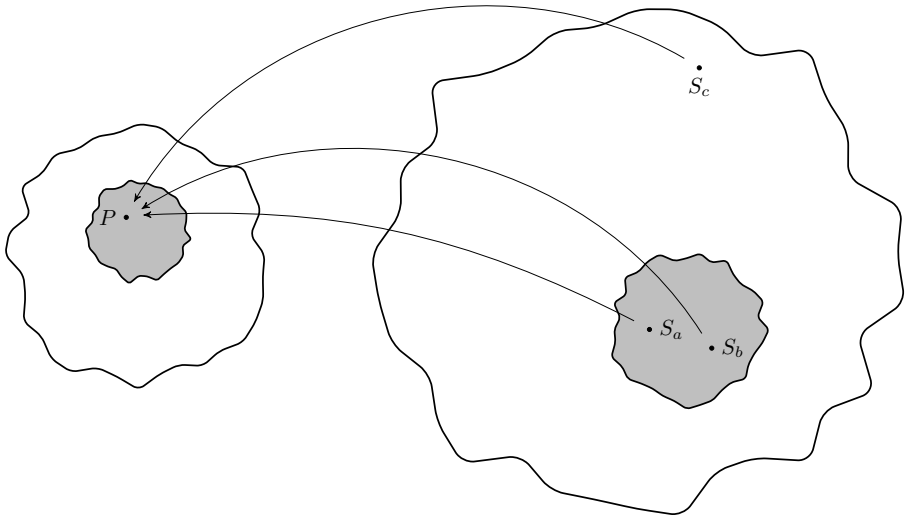


(b) A structure S is incompatible with data if it generates a probability distribution (of observable variables) P that is distinct from a realised probability distribution $\mathbb{P}$. If all structures in $\mathcal{M}$ are incompatible with data then $\mathcal{M}$ is said to be observationally restrictive, and is falsified. This condition is equivalent to $\mathbb{P} \in \mathcal{P} \setminus \mathcal{I}$.

Figure 1: Structures, models, probability distributions (of observable variables), and falsifiability.

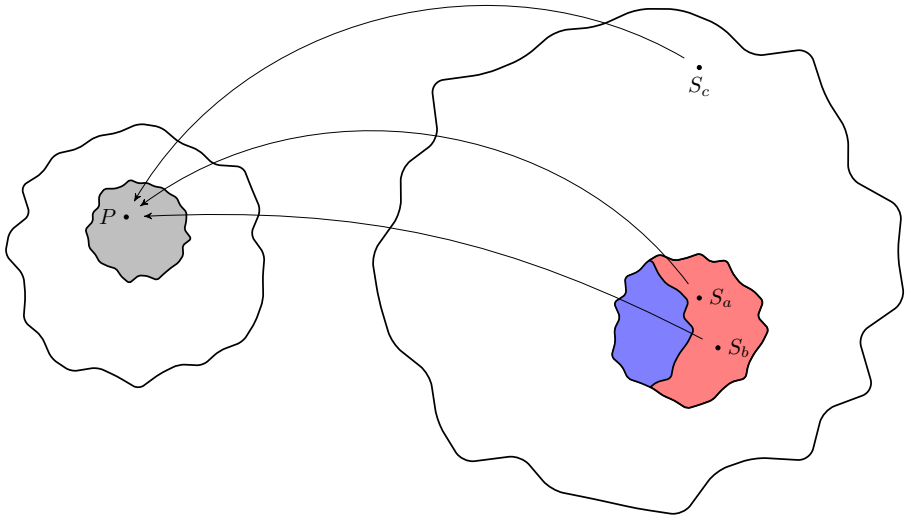


(a) A model $\mathcal{M}$ is said to identify a structure S if the probability distribution (of observable variables) P that is generated by S is distinct from those generated by other structures in $\mathcal{M}$. The structures S_a, S_b and S_c are said to be observationally equivalent as they all generate P but S_b and S_c are not admitted by $\mathcal{M}$. As S_a is the only structure that is admitted by $\mathcal{M}$ and that generates P, S_a is identified by $\mathcal{M}$. For completeness, $\mathcal{M}$ is said to be uniformly identifying if it identifies each structure that it admits.

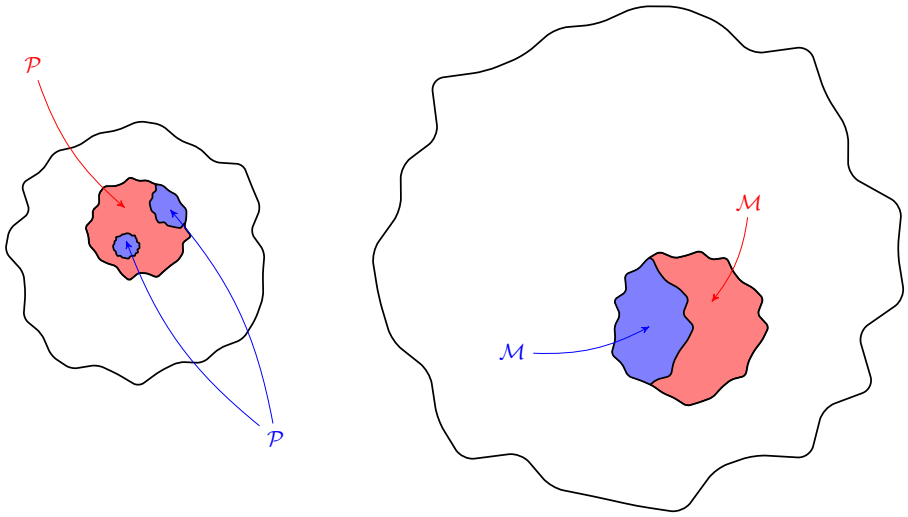


(b) As S_a and S_b are observationally equivalent and are both admitted by $\mathcal{M}$ then $\mathcal{M}$ does not identify either S_a or S_b. Nonetheless, as $\mathcal{M}$ restricts the set of observationally equivalent structures that generate P to S_a and S_b then $\mathcal{M}$ partially identifies S_a (and S_b to within $\{S_a, S_b\}$).

Figure 2: Identification and non-identification of a structure, and partial identification of a structure.



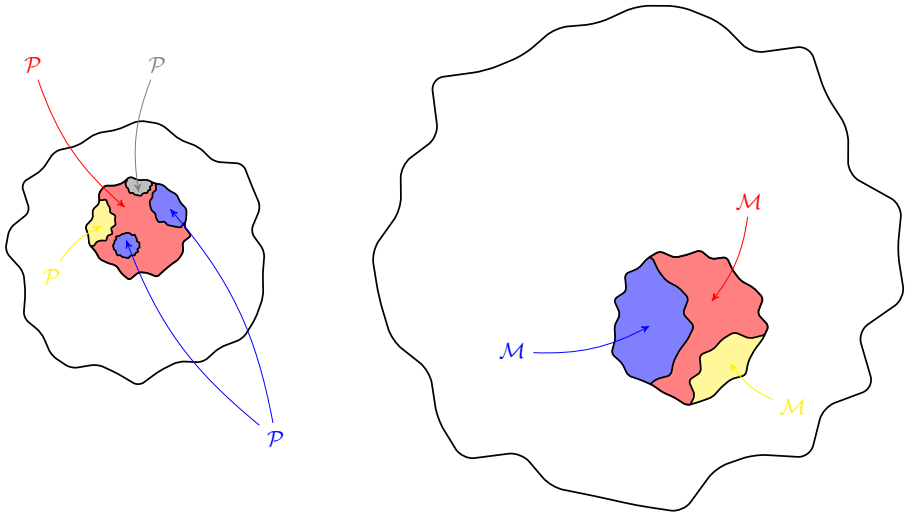
(a) A structural characteristic χ is a function of a structure S. A model $\mathcal{M}$ can be partitioned such that structures in a partition deliver the same value for χ. Structures in the red partition $\mathcal{M}$ deliver the value a for χ, and structures in the red partition $\mathcal{M}$ deliver the value b for χ. If χ is constant across all observationally equivalent structures that $\mathcal{M}$ admits then $\mathcal{M}$ is said to identify χ. As $\chi(S_a)$ is equal to $\chi(S_b)$ (is equal to a) $\mathcal{M}$ identifies χ.



(b) If $\mathcal{M}$ identifies χ for all structures in $\mathcal{M}$ then $\mathcal{M}$ is said to uniformly identify χ. The class of all probability distributions (of observable variables) is partitioned into the blue partition $\mathcal{P}$ and into the red partition $\mathcal{P}$. Probability distributions in $\mathcal{P}$ are generated by (potentially many) structures in $\mathcal{M}$, and probability distributions in $\mathcal{P}$ are generated by (potentially many) structures in $\mathcal{M}$. It is important that the number of partitions in $\mathcal{M}$ and in $\mathcal{P}$ are equal, although that number can be countably infinite. In the context of Figure 3b $\mathcal{M}$ uniformly identifies χ since observationally equivalent structures that $\mathcal{M}$ admits are in the same colour of $\mathcal{M}$. More conveniently, whether $\mathcal{M}$ uniformly identifies χ can be determined by the existence of an identifying correspondence G, a functional. P is a probability distribution in $\mathcal{P}$, and P is a probability distribution in $\mathcal{P}$. Then $\mathcal{M}$ uniformly identifies χ if the value of $G(P)$ is a and if the value of $G(P)$ is b, holding for any such P and P. Notice that if $\mathcal{M}$ uniformly identifies all χ then $\mathcal{M}$ also uniformly identifies structures.

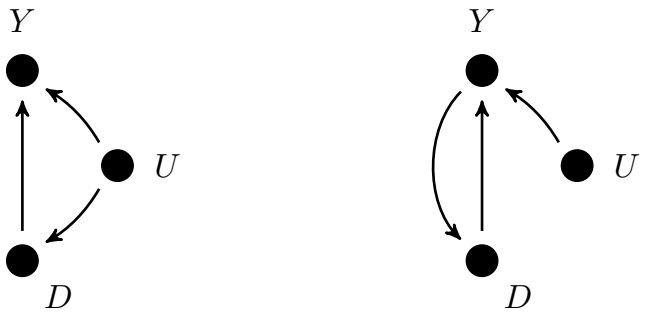
Figure 3: The identification of structural characteristics, and identifying correspondences.

(a) A structural characteristic χ is a function of a structure S. A model $\mathcal{M}$ can be partitioned such that structures in a partition deliver the same value for χ. Structures in the red partition $\mathcal{M}$ deliver the value a for χ, structures in the blue partition $\mathcal{M}$ deliver the value b for χ, and structures in the yellow partition $\mathcal{M}$ deliver the value c for χ. The class of all probability distributions (of observable variables) $\mathcal{P}$ is partitioned into the red partition $\mathcal{P}$, into the blue partition $\mathcal{P}$, into the yellow partition $\mathcal{P}$ and into the grey partition $\mathcal{P}$. Probability distributions in a colour of $\mathcal{P}$ are generated by (potentially many) structures in the same colour of $\mathcal{M}$; the exception is probability distributions in $\mathcal{P}$ which are generated by (potentially many) structures in $\mathcal{M}$ and in $\mathcal{M}$. P is a probability distribution in $\mathcal{P}$ with probability distributions defined similarly for each colour in $\mathcal{P}$.

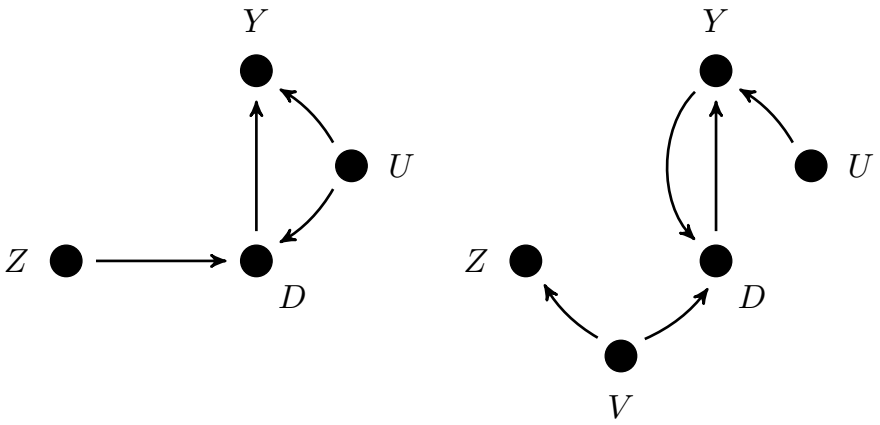


(b) That probability distributions in $\mathcal{P}$ are generated by structures in $\mathcal{M}$ and in $\mathcal{M}$ creates a complication; the value of χ is not constant across observationally equivalent structures that $\mathcal{M}$ admits and that generate a probability distribution in $\mathcal{P}$. So $\mathcal{M}$ does not uniformly identify χ. Consideration of the identifying correspondence G determines that this corresponds to there being structures in $\mathcal{M}$ for which G does not deliver the value of χ when applied to the probability distributions that these structures generate. Nonetheless, if $\mathcal{M}$ restricts the set of values of χ for any probability distribution in $\mathcal{P}$ then $\mathcal{M}$ does have some non-trivial identifying power for χ. Then $\mathcal{M}$ is said to uniformly partially identify χ if $\mathcal{M}$ and $\mathcal{P}$ can each be partitioned into countably many disjoint subsets and that a probability distribution in a partition of $\mathcal{P}$ is not generated by a structure in at least one partition of $\mathcal{M}$, holding for any such partition of $\mathcal{P}$. In the context of Figure 4 $\mathcal{M}$ identifies χ up to $\{a, c\}$, $\mathcal{M}$ identifies χ uniquely to b, and $\mathcal{M}$ identifies χ up to $\{a, c\}$. Each partition of $\mathcal{P}$ includes probability distributions that are generated by structures in at least one partition of $\mathcal{M}$. Equivalently, if G is permitted to be a multivalued functional (or one-to-many) then $\mathcal{M}$ uniformly partially identifies χ if G exists and if $G(P)$ contains the set of values of χ that are delivered by structures that generate P, holding for all such P. A caveat must be applied here; G cannot be trivial in the sense that it is constant across all such P. Clearly this definition of G does not exclude the possibility that there is multiplicity of identifying correspondences that satisfy this property. Sharpness is a desirable property in such circumstances; a functional G that can be shown to deliver smaller sets according to some well-defined distance measure across all possible P (and that satisfies the properties above) should be preferred to any alternative identifying correspondence.

Figure 4: Partial identification of a structural characteristic.



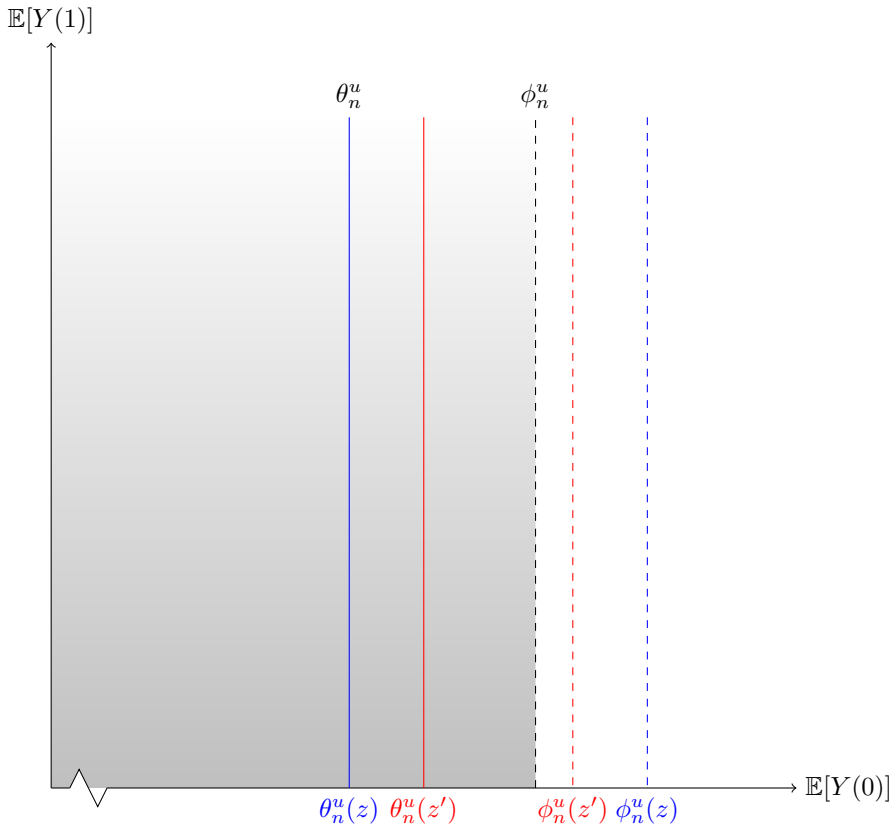
(a) The left- and right-hand panels show directed acyclic graphs that are able to represent the same probability distribution of (Y, D). In the left-hand panel, U causes Y and D. The left-hand panel is a representation of selection in that particular values of D are strongly associated with particular values of Y independently of the causal effect of D. In the right-hand panel, U causes Y but has an indirect effect on D through Y. The right-hand panel is a representation of simultaneity in that Y is both a cause of and an effect of D. The equivalence between the two directed acyclic graphs is that Y can always be written as $\xi(D, U)$ for $\xi : \mathcal{R}_D \times \mathcal{R}_U \to \mathcal{R}_Y$ (provided that the right-hand panel is convergent).



(b) To recover the causal effect of D on Y, it is necessary that there exists an external and measurable factor that causes variation in D. This external and measurable factor is known as an instrumental variable. In the left-hand panel, Z causes D. It is convenient to think of Z as a switch that forces D to take particular values. The difference between the value of Y when Z is `on` versus when Z is `off` is the causal effect of D on Y. In the right-hand panel, it is V that causes D (V is unobservable). As V causes D and Z, it may be sufficient to look at Z to measure exogenous variation in D (although it is not always). As such, the causal effect of D on Y is recoverable using variation in Z. This is an important point about the nature of an instrumental variable; namely, that the relationship between D and Z need not be causal.

Figure 5: A note on causality.

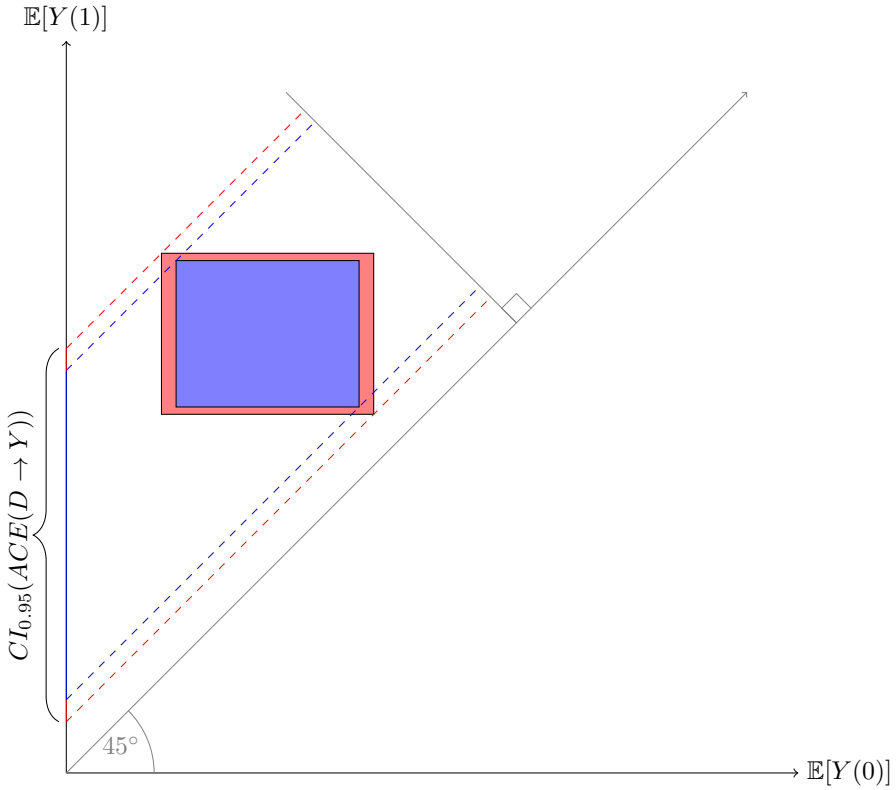
(a) Suppose that $\theta_n^u(z)$ and $\theta_n^u(z')$ are estimates of upper bounds on $\mathbb{E}[Y(0)]$. Similarly, suppose that $\phi_n^u(z)$ and $\phi_n^u(z')$ are one-sided $1-\alpha$ confidence regions for $\theta_n^u(z)$ and $\theta_n^u(z')$. $\phi_n^u(z') > \phi_n^u(z')$ if there is greater variation in the estimate of $\theta_n^u(z)$ (if there are fewer observations of z than z', say).



(b) As $\theta_n^u(z)$ and $\theta_n^u(z')$ are both binding constraints on $\mathbb{E}[Y(0)]$, the minimum of these upper bounds must be binding. The minimum upper bound is written as θ_n^u and is $\theta_n^u(z)$. The inferential problem is to determine ϕ_n^u, which satisfies $\mathbb{P}(\theta_n > \phi_n^u) = \alpha$. A naïve approach would be to look only at the one-sided $1-\alpha$ confidence region for $\theta_n^u(z)$, which is $\phi_n^u(z)$. This approach ignores variation in $\theta_n^u(z')$. An equally naïve approach would be to simply add aggregate variation to $\theta_n^u(z)$. In other words, to add the weighted average of $\phi_n^u(z) - \theta_n^u(z)$ and $\phi_n^u(z') - \theta_n^u(z')$ to $\theta_n^u(z)$. This approach is standard but fails in this case because it does not account for the fact that $\theta_n^u(z) < \theta_n^u(z')$. Inference must account for the fact that upward variation in $\theta_n^u(z')$ does not matter so long as $\theta_n^u(z) < \theta_n^u(z')$; equivalently, that $\theta_n^u(z')$ is a one-sided $1-\gamma$ confidence region for $\theta_n^u(z)$ for $\gamma > \alpha$. Chernozhukov et al. (2013) solves the inferential problem by adjusting the critical value that is associated with the one-sided $1-\alpha$-confidence region. That is, Chernozhukov et al. (2013) adjusts k such that k that solves $\mathbb{P}(\theta_n > \theta_n^u(z) + k\sigma) = \alpha$. The solution for k yields ϕ_n^u with the one-sided $1-\alpha$ confidence region for θ_n^u given by the grey area. The distribution of θ_n^u over repeated samples is non-standard in this case and the bootstrap is not necessarily consistent (Bugni, 2010).

Figure 6: A note on the inferential problem.

(a) Suppose that the admissible set of values of $(\mathbb{E}[Y(0)], \mathbb{E}[Y(1)])$ is given by the blue rectangle, and that the $1-\alpha$-confidence region for this set is the union of the blue rectangle and the red polygon.



(b) It is possible to recover $ACE(D \to Y)$ from the plot. First, note that $ACE(D \to Y)$ is increasing in the y-direction and is decreasing in the x-direction. Second, note that $ACE(D \to Y)$ is constant along any line with unit gradient. Third, note that the value of $ACE(D \to Y)$ along any line with unit gradient is dependent upon the value of the intercept of this line. Fourthly, note that a projection from the normal of a line with unit gradient is a line that has unit gradient. For example, the blue dashed line is a projection from the normal of the 45° line; notice that the blue dashed line is parallel to the 45° line. Fifth, note that any projection from the normal of the 45° line that passes through the blue rectangle is an admissible value of $(\mathbb{E}[Y(0)], \mathbb{E}[Y(1)])$; equivalently, that any projection from the normal of the 45° line that passes through the union of the blue rectangle and the red polygon is in the $1-\alpha$ confidence region of $(\mathbb{E}[Y(0)], \mathbb{E}[Y(1)])$. Together these five facts suggest that the admissible set of values of $ACE(D \to Y)$ and its $1-\alpha$ confidence region can be recovered from the projection from the normal of the 45° line onto the y-axis. This method gives a geometric interpretation to $ACE(D \to Y)$.

Figure 7: A note on recovering the average causal effect.

$\mathcal{R}_Z$	Bound			
	θ_n^{0-}	θ_n^{1-}	θ_n^{0+}	θ_n^{1+}
Male-Male ∪ Female-Female	$[0.533, 0.721]$	$[0.189, 0.524]$	$[0.374, 0.524]$	$[0.533, 0.775]$
Male-Female ∪ Female-Male	$[0.530, 0.723]$	$[0.187, 0.527]$	$[0.371, 0.527]$	$[0.530, 0.777]$
Male-Male	$[0.530, 0.731]$	$[0.182, 0.523]$	$[0.359, 0.523]$	$[0.530, 0.778]$
Male-Female ∪ Female-Male ∪ Female-Female	$[0.528, 0.733]$	$[0.179, 0.527]$	$[0.357, 0.527]$	$[0.528, 0.781]$
Female-Female	$[0.529, 0.729]$	$[0.196, 0.524]$	$[0.362, 0.524]$	$[0.529, 0.771]$
Male-Male ∪ Male-Female ∪ Female-Male	$[0.527, 0.731]$	$[0.193, 0.529]$	$[0.360, 0.529]$	$[0.527, 0.774]$
Male-Male	$[0.533, 0.721]$	$[0.196, 0.523]$	$[0.374, 0.523]$	$[0.533, 0.771]$
Female-Female	$[0.530, 0.723]$	$[0.192, 0.528]$	$[0.371, 0.528]$	$[0.530, 0.775]$
Male-Female ∪ Female-Male				
Male-Male	$[0.536, 0.718]$	$[0.196, 0.523]$	$[0.376, 0.523]$	$[0.536, 0.771]$
Female-Female	$[0.531, 0.722]$	$[0.193, 0.528]$	$[0.372, 0.528]$	$[0.531, 0.775]$
Male-Female				
Female-Male				

Table 1: 95% confidence regions for parameters $(1-p)$ constructed using 97.5% one-sided confidence.

	Bound	
$\mathcal{R}_Z$	θ_n^{0-}	θ_n^{1-}
Multiple birth	$[0.529, 0.733]$	0.476
Single birth	$[0.526, 0.734]$	$[0.460, 0.498]$
Multiple birth	$[0.534, 0.718]$	0.476
Single birth ∩ (Male-Male ∪ Female-Female)	$[0.530, 0.721]$	$[0.460, 0.498]$
Single birth ∩ (Male-Female ∪ Female-Male)		
Multiple birth	$[0.537, 0.716]$	0.476
Single birth ∩ Male-Male	$[0.532, 0.720]$	$[0.460, 0.498]$
Single birth ∩ Female-Female		
Single birth ∩ Male-Female		
Single birth ∩ Female-Male		

Table 2

	Bound	
$\mathcal{R}_Z$	$ACE_n(D \to Y)^-$	$ACE_n(D \to Y)^+$
Male-Male ∪ Female-Female	$[-0.532, -0.009]$	$[0.009, 0.401]$
Male-Female ∪ Female-Male	$[-0.537, -0.002]$	$[0.002, 0.406]$
Male-Male	$[-0.549, -0.007]$	$[0.007, 0.419]$
Male-Female ∪ Female-Male ∪ Female-Female	$[-0.555, 0.001]$	$[-0.001, 0.425]$
Female-Female	$[-0.533, -0.005]$	$[0.005, 0.409]$
Male-Male ∪ Male-Female ∪ Female-Male	$[-0.539, 0.003]$	$[-0.003, 0.415]$
Male-Male	$[-0.525, -0.010]$	$[0.010, 0.397]$
Female-Female	$[-0.531, -0.001]$	$[0.002, 0.404]$
Male-Female ∪ Female-Male		
Male-Male	$[-0.523, -0.013]$	$[0.013, 0.395]$
Female-Female	$[-0.530, -0.003]$	$[0.003, 0.403]$
Male-Female		
Female-Male		

Table 3: 95% confidence regions for parameters ACE constructed using 98.75% one-sided confidence for each parameter.

	Bound
$\mathcal{R}_Z$	θ_n^{0-}
Multiple birth	$[-0.256, -0.052]$
Single birth	$[-0.277, -0.025]$
Multiple birth	$[-0.242, -0.057]$
Single birth ∩ (Male-Male ∪ Female-Female)	$[-0.263, -0.029]$
Single birth ∩ (Male-Female ∪ Female-Male)	
Multiple birth	$[-0.240, -0.061]$
Single birth ∩ Male-Male	$[-0.262, -0.031]$
Single birth ∩ Female-Female	
Single birth ∩ Male-Female	
Single birth ∩ Female-Male	

Table 4

References

Abrevaya, J., Y.-C. Hsu, and R. P. Lieli (2013). Estimating conditional average treatment effects. Technical report, Working paper.

Angrist, J. D. (2014). Angrist data archive. [Online; accessed 25-July-2014].

Angrist, J. D. and W. N. Evans (1998). Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review 88*(3), 450 – 477.

Balke, A. and J. Pearl (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association 92*(439), 1171 – 1176.

Beresteanu, A., I. Molchanov, and F. Molinari (2012). Partial identification using random set theory. *Journal of Econometrics 166*(1), 17 – 32.

Bugni, F. A. (2010). Bootstrap inference in partially identified models defined by moment inequalities: Coverage of the identified set. *Econometrica 78*(2), 735 – 753.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013). Intersection bounds: Estimation and inference. *Econometrica 81*(2), 667 – 737.

Chesher, A. (2005). Nonparametric identification under discrete variation. *Econometrica 73*(5), 1525 – 1550.

Chesher, A. (2010). Instrumental variable models for discrete outcomes. *Econometrica 78*(2), 575 – 601.

Chesher, A. and A. M. Rosen (2013). What do instrumental variable models deliver with discrete dependent variables?. *American Economic Review 103*(3), 557 – 562.

Chesher, A., A. M. Rosen, and K. Smolinski (2013). An instrumental variable model of multiple discrete choice. *Quantitative Economics 4*(2), 157 – 196.

Heckman, J. J. and E. Vytlacil (2005). Structural equations, treatment effects, and econometric policy evaluation1. *Econometrica 73*(3), 669 – 738.

Hurwicz, L. (1950). Generalization of the concept of identification. *Statistical Inference in Dynamic Economic Models (T. Koopmans, ed.). Cowles Commission, Monograph 10*, 245 – 257.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467 – 475.

Khan, S. and E. Tamer (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica 78*(6), 2021 – 2042.

Kitagawa, T. (2009). Identification region of the potential outcome distributions under instrument independence. *Econometrica (Revise and resubmit)*.

Koopmans, T. C. and O. Reiersøl (1950). The identification of structural characteristics. *Annals of Mathematical Statistics 21*(2), 165 – 181.

Manski, C. F. (2013). *Public Policy in an Uncertain World: Analysis and Decisions.* Cambridge and London: Harvard University Press.

Molchanov, I. (2005). *Theory of Random Sets.* Springer.

Shaikh, A. M. and E. J. Vytlacil (2011). Partial identification in triangular systems of equations with binary dependent variables. *Econometrica 79*(3), 949 – 955.

Strotz, R. H. and H. O. Wold (1960). Recursive vs. nonrecursive systems: An attempt at synthesis (part i of a triptych on causal chain systems). *Econometrica: Journal of the Econometric Society*, 417 – 427.