Department of Economics, University College London

Information and identified sets

Jeff Rowley

7th September, 2014

Abstract

I explore the effect on the identified set of (a) enriching the support of an instrumental variable, and (b) permitting an exogenous variable to enter the structural equation for the outcome variable in a model with a binary endogenous variable. I show that enriching the support of an instrumental variable is advantageous. I show that omitting an exogenous variable induces a bias in the resulting identified set if that variable should be included. Further, I show that the average causal effect of the endogenous variable on the outcome variable is expressible as a weighted sum of another functional that I name the conditional average causal effect. I consider the average causal effect and the conditional average causal effect in the context of the effect of additional children on a mother's employment using United States census data.

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 study a non-parametric binary choice model with a binary endogenous variable. I explore the effect of incorporating information into the model on the identified set of the average causal effect of the endogenous variable on the outcome variable. 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 are relevant 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 the identified set of 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 important consequence is that the causal effect of the endogenous variable on the outcome variable depends upon the values of 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

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

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. In identifying models, if an omitted variable is a determinant of the endogenous variable then its effect on the outcome variable is instead attributable to the endogenous variable, and is a bias. A contribution that I make is in showing that incorporating information negates this bias, and is equivalent to a shift in the location of the estimate of the identified set. 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 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 additional children on a mother's employment using US census data. I extend Chesher and Rosen (2013) in a number of ways. First, I report statistical uncertainty in the estimate of the average causal effect of additional children on a mother's employment using a method that is outlined in Chernozhukov et al.

[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 $CATE(x)$.

(2013). Second, I enrich the support of the instrumental variable and explore the effect that this has on the estimate of the identified set of values for the average causal effect of additional children on a mother's employment, and on its accompanying confidence region. Third, I enrich individual response by permitting the structural equation for employment to depend upon 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 employment. 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 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 cardinality of a set S as $\text{card}(S)$. 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 additional children 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 that X is a function of mother's race or ethnicity (shortened to race), and that Z is a function of whether the oldest two children in the household share the same gender (shortened to child gender) and whether a mother experienced a multiple second birth.[3]

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*

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

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 M6.

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 $\boldsymbol{U}$. 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 dimension of U is a normalisation rather than a restriction since D has binary support.

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

Restriction M3 is a non-verifiable restriction. The restriction nests the restrictions $U \perp\!\!\!\perp Z|X$ (conditional independence) and $U \perp\!\!\!\perp X$, and nests the restrictions $U \perp\!\!\!\perp X$ and $U \perp\!\!\!\perp Z$ (marginal independence).[4] In the context of the application, the restriction implies that variables such as opportunity are independent of child gender conditional on race, and are independent of race.

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

[4]The restriction also implies the restrictions $U \perp\!\!\!\perp X|Z$ and $U \perp\!\!\!\perp Z$.

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 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 additional children 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 $d \in \mathcal{R}_D$ and for some $(z, z') \in \mathcal{R}_Z^2$.

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 associated with a variable that causes D.[5] 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 $(z, z') \in \mathcal{R}_Z^2$.

[5] This point is the subject of Figure 5, which studies causality.

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

I define credibility. I discuss the conditions under which a model is more credible than another. I discuss opposition to the assumption and use of partially identifying models.

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 (a) and (b) are motivated by two concerns. Namely, that characterising the estimate of the identified set of structures or the estimate of the identified 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 (a) and (b). In (a), 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. In (b), the feature that is identified is often less valuable than the original feature of interest. Nonetheless, it is promising that (b) 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.

First, the single equation model embeds the restriction that h is a non-parametric function (Restriction M5). 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 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.

Second, 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, the normalisation 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.

Third, the single equation model embeds the restriction that h excludes Z (Restriction M4). The restriction is an essential element of the single equation model if identification of causality is sought (and if Z is an instrumental variable). Further, the restriction is an essential element of any model of causality and not just an essential element of the single equation model. This point is the subject of Figure 5, which studies causality. In the context of the application, the restriction (together with Restriction M3) implies that child gender is conditionally independent of the employment of mothers. If a mother chooses to participate in the labour market only when her children share the same gender say, then this restriction is violated. It is plausible that child gender should not affect the decision of a mother to participate in the labour market and so the restriction is plausible.

Fourth, 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 the context of the application, whether there are three or more children in the household is endogenous. If mothers that incur large costs from employment also incur small costs from having children say, then whether there are three or more children in the household is correlated with a mother's employment. The single equation model does not restrict how strong this negative correlation should be, nor does it limit the reasons for or direction of endogenous variation.

Fifth, the single equation model embeds the restriction that $U \perp\!\!\!\perp (X, Z)$ (Restriction M3). Joint independence is a strong restriction on the joint distribution of (D, X, U). Joint independence is stronger than both conditional independence and marginal independence, and restricts the full distribution unlike mean independence or quantile independence restrictions. In general, joint independence is not a plausible restriction. In the context of the application, the restriction implies that variables such as opportunity are independent of child gender conditional on race, and are independent of race. It is plausible that opportunity is independent of child gender conditional on race since child gender is randomly assigned. However, there is no reason to suppose that opportunity is independent of race. Minority individuals need not have access to the same opportunities as otherwise equivalent white individuals. It is plausible that the distribution of opportunities for white individuals is more negatively skewed than the distribution of opportunities for minority individuals.

Sixth, the single equation model embeds the restriction that h is separable in U (Restriction M5). 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 additional children on a mother's employment is positive for all mothers or is negative for all mothers. There is no reason to suppose that individual response is monotonic. A mother that uses a paid-for childcare service might find the cost of such services prohibitive for additional children and so exit employment. In contrast, a mother that relies on family members for childcare might find need to return to employment to increase household income.

As a final remark regarding the credibility of the single equation model, 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. Restriction M6 permits the existence of compliers and defiers. In other words, it permits individual selection to be both positive and negative. In fact, the monotonicity restriction that is embedded in the LATE model is nested by Restriction M6. In the context of the application, it is plausible that whether there are three or more children in the household depends upon the gender of the oldest two children. If preferences are convex then it is plausible that whether there are three or more children in the household is more likely when the oldest two children share the same gender, but there may be other considerations which determine perceived child quality.

The single equation model is reasonably credible in general, and in the context of the application. Nonetheless, some of the restrictions that the single equation model embeds are unsatisfactory in general, and are implausible in the context of the application. Joint independence and monotonicity are two such restrictions.

Conditional independence is a more plausible restriction than joint independence. Recall that joint independence nests conditional independence and marginal independence, and that it is marginal independence that is implausible in the context of the application. It is possible that partial identification of the average causal effect of the endogenous variable on the outcome variable can be maintained under conditional independence (that joint independence is not an essential element of the model). I propose relaxing joint independence in favour of conditional independence as an extension.

Non-separability of h in U enriches individual response and is a more plausible restriction than monotonicity. To clarify, if h is non-separable in U then Restriction M1' is replaced by

$$Y = \mathbb{1}[q(D, X, U) < 0]$$

for some non-parametric function q. In general, non-separability of h in U permits the causal effect of (D, X) to vary with U. At the extreme, non-separability of h in U permits the causal effect of (D, X) to be positive for some values of U and negative for others. In the context of the application, it is plausible that additional children lead some mothers to enter the labour market and some mothers to exit the labour market, and that this is dependent upon the opportunity that a mother faces. This behaviour is consistent with non-separability of h in U. I propose relaxing monotonicity in favour of non-separability as an extension.

Finally, 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. Suppose that this restriction is replaced with the restriction that g is a non-parametric function that varies with (U, X, Z), and is non-separable in U. The resulting triangular model that embeds this restriction might have greater informational content than the single equation model with only a small loss in credibility. I propose strengthening the lack of constraints on g in favour of a non-parametric restriction as an extension.

2.2 Falsifiability

I show that the single equation model is falsifiable. I derive an instrumental inequality (Pearl, 1995) that is a sufficient condition for the single equation model to be observationally restrictive.

It is important that a model is falsifiable. If a model is not falsifiable then it cannot be rejected for any probability distribution of observable random variables that is consistent with the observable supports of these random variables. In other words, a model that is not falsifiable is always valid. The single equation model is falsifiable, and can be rejected if the following instrumental inequality is violated.

$$\max_{d\in\mathcal{R}_D}\max_{x\in\mathcal{R}_X}\sum_{y\in\mathcal{R}_Y}\max_{z\in\mathcal{R}_Z}\mathbb{P}(y,d|x,z)\leq 1 \tag{2.1}$$

Proof. I write $\mathbb{P}(y,d|x,z)$ as

$$\int_{u\in\mathcal{R}_U}\mathbb{P}(y,d,u|x,z)dU, \tag{2.2}$$

which is valid by the Law of Total Probability. Further, I write the integrand as the product decomposition

$$\mathbb{P}(y|d,x,z,u)\mathbb{P}(d|x,z,u)\mathbb{P}(u|x,z), \tag{2.3}$$

which is valid by Bayes' Theorem. I postulate that $\mathbb{P}(y,d|x,z)$ is generated by the single equation model. I write (2.3) as the product decomposition

$$\mathbb{P}(y|d,x,u)\mathbb{P}(d|x,z,u)\mathbb{P}(u), \tag{2.4}$$

which is valid by the restrictions that are embedded in the single equation model. Specifically, I use Restriction M3 and Restriction M4. I substitute (2.4) into (2.2) in place of the integrand.

$$\mathbb{P}(y,d|x,z)=\int_{u\in\mathcal{R}_U}\mathbb{P}(y|d,x,u)\mathbb{P}(d|x,z,u)\mathbb{P}(u)dU \tag{2.5}$$

I note that (2.5) holds for any z, and so holds for $z(y,d,x)$ that I define as $\text{argmax}_{z\in\mathcal{R}_X} \mathbb{P}(y,d|x,z)$.

$$\mathbb{P}(y,d|x,z(y,d,x)) = \int_{u\in\mathcal{R}_U} \mathbb{P}(y|d,x,u)\mathbb{P}(d|x,z(y,d,x),u)\mathbb{P}(u)dU \tag{2.6}$$

I sum both sides of (2.6) over $\mathcal{R}_Y$.

$$\sum_{y\in\mathcal{R}_Y} \mathbb{P}(y,d|x,z(y,d,x)) = \sum_{y\in\mathcal{R}_Y} \int_{u\in\mathcal{R}_U} \mathbb{P}(y|d,x,u)\mathbb{P}(d|x,z(y,d,x),u)\mathbb{P}(u)dU,$$

I note that $\mathbb{P}(d|x,z(y,d,x),u)$ is a probability, and so is bounded from above by unity.

$$\sum_{y\in\mathcal{R}_Y} \mathbb{P}(y,d|x,z(y,d,x)) \leq \sum_{y\in\mathcal{R}_Y} \int_{u\in\mathcal{R}_U} \mathbb{P}(y|d,x,u)\mathbb{P}(u)dU \tag{2.7}$$

I note that the right-hand side of (2.7) is an expectation, and so evaluates to a well-defined probability.

$$\sum_{y\in\mathcal{R}_Y} \mathbb{P}(y,d|x,z(y,d,x)) \leq \sum_{y\in\mathcal{R}_Y} \mathbb{P}(y|d,x) \tag{2.8}$$

I write

$$\sum_{y\in\mathcal{R}_Y} \mathbb{P}(y,d|x,z(y,d,x)) \leq 1,$$

which is valid by the Law of Total Probability. I write

$$\sum_{y\in\mathcal{R}_Y} \max_{z\in\mathcal{R}_Z} \mathbb{P}(y,d|x,z) \leq 1, \tag{2.9}$$

which is valid by the definition of $z(y,d,x)$. I note that the right-hand side of (2.9) is constant for any (d,x), and so holds for those values that maximise the left-hand side of the inequality. This completes the proof. □

I note that the instrumental inequality and its proof are adapted from Pearl (1995). I extend Pearl (1995) in that I permit the existence of an exogenous variable.

3 Identification

I introduce random set theory (Molchanov, 2005) as a tool for identification analysis. I discuss random set theory in the context of the single equation model. I exploit joint independence (Restriction M3) and monotonicity (Restriction M5), and derive sharp bounds on the distribution of U.

Artstein's Inequality (Artstein, 1983) is an important theorem of random set theory that is useful for deriving bounds on latent probability distributions. The usefulness of

Artstein's Inequality is its dual representation as a capacity functional and as a containment functional, and that it defines a sharp set. Together these properties determine that identification analysis that is conducted using Artstein's Inequality defines the identified set of a functional of a latent probability distribution, as opposed to a proper superset (of the identified set). For a selection ξ from a random closed set Ξ,

$$\mathbb{P}(\xi \in T|\mathcal{I}) \leq \mathbb{P}(\Xi \cap T \neq \varnothing|\mathcal{I}) \tag{3.1}$$

$$\mathbb{P}(\xi \in T|\mathcal{I}) \geq \mathbb{P}(\Xi \subseteq T|\mathcal{I}) \tag{3.2}$$

for all test sets T in the class of compact sets $\mathcal{T}$. Here, I write a conditional form (conditional on an arbitrary information set $\mathcal{I}$) of Artstein's Inequality since the identification analysis that is conducted exploits the independence relations that the single equation model embeds. (3.1) is the capacity functional representation of Artstein's Inequality, and (3.2) is its containment functional representation.

In the context of the single equation model, $(Y(0), Y(1))$ is a random closed set that has a well-defined probability distribution. I follow Chesher and Rosen (2013) in defining a level set $\mathcal{U}_h$.

$$\mathcal{U}_h(y, d, x) \equiv \{u : y = h(d, x, u)\}$$

The usefulness of $\mathcal{U}_h$ to the identification analysis is its synonymity with $(Y(0), Y(1))$. This point is the subject of Figure 12. I translate Artstein's Inequality as

$$\mathbb{P}(U \in T|x, z) \leq \mathbb{P}(\mathcal{U}_h(Y, D, x) \cap T \neq \varnothing|x, z) \tag{3.3}$$

$$\mathbb{P}(U \in T|x, z) \geq \mathbb{P}(\mathcal{U}_h(Y, D, x) \subseteq T|x, z) \tag{3.4}$$

for all T in the class of all compact sets on $[0,1]$, and for some $(x, z) \in \mathcal{R}_X \times \mathcal{R}_X$.[6] Further, I exploit joint independence and monotonicity, and write the left-hand side of (3.3) and of (3.4) as $\|T\|$.

The class of all compact sets on $[0,1]$ is a large class, and it is not feasible to compute Artstein's Inequality for all test sets in this class. Chesher et al. (2013) shows that it is sufficient to compute Artstein's Inequality for a smaller class of test sets, and names this class the class of core determining sets. For the single equation model, the class of core determining sets is the collection of

$$[0, p(d, x)] \text{ and } (p(d, x), 1]$$

[6]The restriction on the class of test sets is valid by joint independence and monotonicity

over $\mathcal{R}_D \times \mathcal{R}_X$.[7] The left-hand side of (3.3) and of (3.4) is then either $p(d,x)$ or $1 - p(d,x)$, depending upon the core determining set that is tested. Further, the class of core determining sets can be reduced to the collection of

$$[0, p(d,x)]$$

over $\mathcal{R}_D \times \mathcal{R}_X$. This refinement is valid since $[0, p(d,x)]$ is the complement of $(p(d,x), 1]$, and so the capacity functional representation of Artstein's Inequality for $[0, p(d,x)]$ is equivalent to the containment functional representation function of Artstein's Inequality for $(p(d,x), 1]$.

3.1 Identification analysis

I describe the identified set of the collection of $p(d,x)$ over $\mathcal{R}_D \times \mathcal{R}_X$. I employ random set theory as a tool for the identification analysis. I distinguish between the identified set of a functional of a latent probability distribution and the estimate of the identified set of that functional.

Assume that there is a particular ordering of the collection of $p(d,x)$ over $\mathcal{R}_D \times \mathcal{R}_X$. First, I define correspondences (set-valued functions) $\mathcal{A}_p$ and $\mathcal{B}_p$ as follows.

$$\mathcal{A}_p(\eta; d, x) \equiv \{a : p(a, \eta) \leq p(d,x)\}$$
$$\mathcal{B}_p(\eta; d, x) \equiv \{b : p(b, \eta) \geq p(d,x)\}$$

I write η so as to emphasise the distinction between η as a conditioning value, and x as a determinant of a test set. That is, for a given test set $[0, p(d,x)]$, there is a conditional form of Artstein's Inequality for each $\eta \in \mathcal{R}_X$. Second, I describe the identified set of $p(d,x)$ as

$$\sup_{\eta \in \mathcal{R}_X} \sup_{z \in \mathcal{R}_Z} \mathbb{P}(Y = 0, D \in \mathcal{A}_p(\eta; d, x) | \eta, z) \leq p(d,x) \leq 1 - \sup_{\eta \in \mathcal{R}_X} \sup_{z \in \mathcal{R}_Z} \mathbb{P}(Y = 1, D \in \mathcal{B}_p(\eta; d, x) | \eta, z), \tag{3.5}$$

which is valid by Artstein's Inequality. Note that alternative forms of Artstein's Inequality that rely only on marginal independence relations that are implied by joint independence

[7] Notice that $(p(d,x), 1]$ is an open set, but that Artstein's Inequality is defined for compact test sets. Strictly speaking, the class of core determining sets should include $\mathrm{cl}(p(d,x), 1]$ rather than $(p(d,x), 1]$. Since U is distributed continuously, there is zero mass at $p(d,x)$ and so the measure of $\mathrm{cl}(p(d,x), 1]$ is equal to the measure of $(p(d,x), 1]$.

also describe sets of $p(d,x)$, but that these sets constitute proper or improper supersets of (3.5). Further, (3.5) is defined for each possible ordering of the collection of $p(d,x)$ over $\mathcal{R}_D \times \mathcal{R}_X$. If (3.5) is empty for a particular ordering of the collection of $p(d,x)$ over $\mathcal{R}_D \times \mathcal{R}_D$, then that ordering can be rejected.

Generally, the sampling process does not identify $\mathbb{P}$. Instead, the sampling process identifies $\mathbb{P}_n$ that is representative of a proper subset of the population, and that is often distinct from $\mathbb{P}$. To estimate (3.5), it is natural to replace $\mathbb{P}$ with $\mathbb{P}_n$, which is valid by the assumption that the sampling process is informative of $\mathbb{P}$ as $n \to \infty$. This assumption is known as the analogue principle (Manski, 1988). It is somewhat unnatural to regard the sample analogue of (3.5) as an identified set since identification is a concept of the population. I am careful to the sample analogue of (3.5) as an estimate of the identified set to emphasise the distinction.

3.2 Dimensionality

I calculate the number of inequalities that describe the identified set, and how this number changes as $\mathrm{card}(\mathcal{R}_X)$ increases. I show that there is a curse of dimensionality.

Tractability is a legitimate concern: regardless of whether incorporating information is advantageous in the context of the single equation model, the computational cost of incorporating information may be prohibitive. Let

$$\mathrm{card}(\mathcal{R}_X) = K \text{ and } \mathrm{card}(\mathcal{R}_Z) = L,$$

and fix a particular ordering of the collection of $p(d,x)$ over $\mathcal{R}_D \times \mathcal{R}_X$. There are $2K$ parameters in the collection of $p(d,x)$ over $\mathcal{R}_D \times \mathcal{R}_X$, and there are $2K \times L$ inequality relations for each parameter. In sum, the number of inequalities for each ordering of the collection of $p(d,x)$ over $\mathcal{R}_D \times \mathcal{R}_X$ is

$$4K^2 \times L.$$

The total number of orderings of the collection of $p(d,x)$ over $\mathcal{R}_D \times \mathcal{R}_X$ is then

$$\sum_{j=1}^{2K} j! \left\{ {2K \atop j} \right\},$$

where $\left\{ {n \atop m} \right\}$ counts the number of ways that a set of cardinality n can be partitioned into m non-empty subsets (the Stirling Number of the Second Kind). Each $\left\{ {n \atop m} \right\}$ is multiplied by

$m!$, which accounts for the possible ordering over the m non-empty subsets. Combining the number of inequalities for a particular ordering of the collection of $p(d,x)$ over $\mathcal{R}_D \times \mathcal{R}_X$ with the number of possible orderings yields

$$4K^2 \times \sum_{j=1}^{2K} j! \left\{ {2K \atop j} \right\} \times L$$

as the number of inequalities that describe the identified set. There is a clear curse of dimensionality. It is sensible to automate the task of calculating the inequalities that describe the identified set.

3.3 Average causal effects

I state the definition of $ACE(D \to Y)$ when D is a binary variable. I discuss the relationship between $ACE(D \to Y)$ and its conditional counterparts. I show that $ACE(D \to Y)$ is expressible as a function of the parameters of the single equation model.

In economics, $ACE(D \to Y)$ is commonly referred to as the Average Treatment Effect. When D is a binary variable, this convention is intuitive since $ACE(D \to Y)$ is the effect of receiving treatment versus not receiving treatment; or,

$$\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]. \tag{3.6}$$

For the single equation model, (3.6) is equal to

$$\mathbb{E}[p(0,X)] - \mathbb{E}[p(1,X)],$$

and may be more conveniently expressed as

$$\sum_{x \in \mathcal{R}_X} \mathbb{P}(x)(p(0,x) - p(1,x)). \tag{3.7}$$

In fact, (3.7) reveals the relationship between $ACE(D \to Y)$ and its conditional counterparts: $p(0,x) - p(1,x)$ is $ACE(D \to Y|x)$, and so $ACE(D \to Y)$ is a weighted sum of its conditional counterparts.

$$ACE(D \to Y) = \sum_{x \in \mathcal{R}_X} \mathbb{P}(x) ACE(D \to Y|x)$$

I exploit this relationship to describe the identified set of $ACE(D \to Y)$. As the single equation model only partially identifies $ACE(D \to Y|X$, so the sum is a Minkowski sum.

Although p is a non-parametric function, it has a discrete domain and a discrete codomain. As such, p can be summarised by the collection of $p(d,x)$ over $\mathcal{R}_D \times \mathcal{R}_X$. Since knowledge of this collection is also sufficient to determine the distribution of $Y(d)$, the collection of $p(d,x)$ over $\mathcal{R}_D \times \mathcal{R}_X$ may be regarded as parameters of the single equation model. $ACE(D \to Y)$ and its conditional counterparts are then functions of the parameters of the single equation model.

4 Incorporating information

I explore the effect of incorporating information into the single equation model. I study a special case of the single equation model that embeds the restriction that h excludes X. That is, I study a special case of the single equation model that embeds the restriction that

$$Y = \mathbb{1}[r(D) < U]$$

for a non-parametric function r. I name this special case of the single equation model the simple single equation model, and note that this model is assumed in Chesher and Rosen (2013) for the purpose of estimating the average causal effect of additional children on a mother's employment. I study how the estimate of the identified set of $ACE(D \to Y)$, equal to $r(0) - r(1)$, changes as

(a.) I enrich $\mathcal{R}_Z$; and,

(b.) I permit X to enter h.

The model that I assume for (b) is then the single equation model, rather than the simple single equation model.

Chesher and Rosen (2013) does not report statistical uncertainty in the estimate of the average causal effect of additional children on a mother's employment. I extend Chesher and Rosen (2013) by constructing a confidence region for the estimate of the average causal effect of additional children on a mother's employment, and use this as a baseline for comparison in studying how the estimate of the identified set of the average causal effect of additional children on a mother's employment changes as information is incorporated into the simple single equation model. I use a method that is outlined in Chernozhukov et al. (2013) to compute statistical uncertainty. Although valid by

Bonferroni's inequality, the statistical uncertainty that I report is conservative. That is, for some statistical size α, the confidence region covers the average causal effect of additional children on a mother's employment in at least $1-\alpha$ samples. I discuss the inferential problem in Figure 6.

In the context of the application, I consider two definitions for Z in the framework of the simple single equation model. This approach is consistent with Chesher and Rosen (2013). In the first instance, I define Z as an indicator for the event that the oldest two children in the household share the same gender. I refer to this instance as Specification 1. In the second instance, I define Z as an indicator for the event that the second birth is a multiple birth. I refer to this instance as Specification 6.

$$\text{Specification 1} \quad Z = \begin{cases} 1 & \text{Male-Male} \cup \text{Female-Female}, \\ 0 & \text{Male-Female} \cup \text{Female-Male}. \end{cases}$$

$$\text{Specification 6} \quad Z = \begin{cases} 1 & \text{Multiple birth}, \\ 0 & \text{Single birth}. \end{cases}$$

Figure 9 and Figure 10 show the estimate of the identified set of $(\mathbb{E}[Y(0)], \mathbb{E}[Y(1)])$ for Specification 1 and Specification 6, respectively.[8]

In Specification 1, the estimate of the identified set of $(\mathbb{E}[Y(0)], \mathbb{E}[Y(1)])$ is large and disconnected. The gender of the oldest two children is not a good predictor of whether there are three or more children in the household.

$$\mathbb{P}_n(D = 1|Z = 1) = 0.414 \text{ versus } \mathbb{P}_n(D = 1|Z = 0) = 0.346$$

There is little statistical uncertainty in the estimate of the identified set; the estimate of the identified set is large relative to the confidence region for the estimate of the identified set. There is little statistical uncertainty in the estimate of the identified set as

(a) the event that the oldest two children in the household share the same gender is approximately as frequent as the event that the oldest two children in the household do not share the same gender; and,

(b) the gender of the oldest two children is not a good predictor of whether there are three or more children in the household.

[8] Note that the estimate of the identified set of $(\mathbb{E}[Y(0)], \mathbb{E}[Y(1)])$ is equivalent to the identified set of $(\mathbb{E}_n[Y(0)], \mathbb{E}_n[Y(1)])$. I prefer to use the former since it makes clear the uncertainty that is present in estimation.

Specifically, (a) means that each of the inequalities that describe the estimate of the identified set is a functional of a large number of data, and so there should be little statistical uncertainty in each of these inequalities. Further, (b) means that there is little difference between those inequalities that are binding constraints, and those that are not binding constraints.

In Specification 6, the estimate of the identified set of $(\mathbb{E}[Y(0)], \mathbb{E}n[Y(1)])$ collapses to a line, which covers a medium interval. The estimate of the identified set collapses due to the nature of the relationship between whether there are three or more children in the household and the event that the second birth is a multiple birth: clearly, if there is a multiple second birth then there must be three or more children in the household. The event that there is a multiple second birth is a special instrumental variable as it excludes some events in the probability space.

$$\mathbb{P}_n(D = 1|Z = 1) = 1 \text{ versus } \mathbb{P}_n(D = 1|Z = 0) = 0.375$$

An implication of its special status is that $\mathbb{E}_n[Y(1)]$ is identified. Essentially, if there is a multiple second birth then there is no selection and so $\mathbb{E}_n[Y(1)]$ is identified from the sub-population of mothers that experience a multiple second birth. There is considerable statistical uncertainty in the estimate of the identified set; the estimate of the identified set is small relative to the confidence region for the estimate of the identified set. In particular, there is considerable uncertainty in the $\mathbb{E}[Y(1)]$-direction of the estimate of the identified set of $(\mathbb{E}[Y(0)], \mathbb{E}[Y(1)])$ as

(a) the event that there is a multiple second birth is rare; and,

(b) the event that there is a multiple second birth is a good predictor of whether there are three or more children in the household.

In fact, whether the instrumental variable is a good predictor of the endogenous variable is of first-order importance only for estimation, and not for inference. The difference between those inequalities that are binding constraints, and those that are not binding constraints only becomes important if the statistical uncertainty in those inequalities that are binding constraints is considerable. In other words, if the instrumental variable is a rare event. In sum, the ideal instrumental variable is a good predictor of the endogenous variable and is approximately as frequent as not.

Figure 8 shows the estimate of the identified set of $ACE(D \to Y)$ for Specification 1 and for Specification 6 (and all other specifications that I conduct). In Specification 1, the estimate of the identified set is large and disconnected. The event that the oldest two children in the household share the same gender is insufficient to sign $ACE(D \to Y)$. Further, either side of zero the estimate of the identified set is not informative beyond what anecdotal evidence might suggest, and certainly cannot inform policy. The only information that Specification 1 conveys with respect to $ACE(D \to Y)$ is that there is a statistically significant effect (since the confidence region for the estimate of the identified set does not cover zero). In Specification 6, the estimate of the identified set is not disconnected. The event that there is a multiple second birth is sufficient to sign $ACE(D \to Y)$ as negative. Although Specification 6 is informative with respect to the sign of $ACE(D \to Y)$, it conveys little other information to inform policy.

4.1 Enriching the support of the instrumental variable

I explore the effect of enriching $\mathcal{R}_Z$ for the simple single equation model in the context of the application. Specifically, I disaggregate the event that the oldest two children in the household share the same gender into the possible permutations of gender for two children. Further, I combine the event that there is a multiple second birth with the possible permutations of gender for two children. I show that incorporating informational is advantageous.

In all, I consider eight specifications. Each specification is characterised by a different definition of Z. The results of these specifications are the subject of Table 1 through Table 4, and the tables also give the definition of Z in each specification. I make the following observations.

(a) Disaggregating the event that the oldest two children in the household share the same gender into either the event that the oldest two children in the household are male, or into the event that the oldest two children in the household are female has less informational content.

In Specification 2 I disaggregate the event that the oldest two children in the household share the same gender into the event that the oldest two children are male. In Specification 3 I disaggregate the event that the oldest two children in the household share the same gender into the event that the oldest two children are female. In both specifications, the estimate of the identified set of $ACE(D \to Y)$ is larger than the estimate of the identified set for Specification 1. Further, there is greater statistical uncertainty in the estimate of the identified set for Specification 2 and for Specification 3.

Either disaggregation maintains Z as a binary variable and, if preferences are convex, aggregates an event that the oldest two children in the household share the same gender with the event that the oldest two children do not share the same gender. For example, in Specification 2 the disaggregation aggregates the event that the oldest two children in the household are female with the event that the oldest two children do not share the same gender. Importantly, this dilutes the identifying power of Z (since the event that the oldest two children in the household share the same gender is associated with both values of Z). Further, each value of Z is no longer as frequent as not. The imbalance in the sub-populations leads to greater statistical uncertainty.

(b) Disaggregating the event that the oldest two children in the household share the same gender into the possible permutations of gender for two children has more informational content.

In Specification 4 I disaggregate the event that the oldest two children in the household share the same gender into the event that the oldest two children are male, the event that the oldest two children are female, and the event that the oldest two children do not share the same gender. In Specification 5 I disaggregate the event that the oldest two children in the household share the same gender into the possible permutations of gender for two children. In both specifications, the estimate of the identified set of $ACE(D \to Y)$ is smaller than the estimate of the identified set for Specification 1. There is greater statistical uncertainty in the estimate of the identified set for Specification 4 and for Specification 5, but the net effect of greater accuracy and less precision is to make the estimate of the identified set and the confidence region of the estimate of the identified set smaller in both cases. The estimate of the identified set of $ACE(D \to Y)$ is smaller than the estimate of the identified set for Specification 1.

Most importantly, either disaggregation increases $\text{card}(\mathcal{R}_Z)$. The advantage of disaggregating the event that the oldest two children in the household is that more variation is introduced that might be exploited. To illustrate, notice that

$$\mathbb{P}_n(A|\text{Male-Male} \cup \text{Female-Female}) \approx \frac{1}{2}\mathbb{P}_n(A|\text{Male-Male}) + \frac{1}{2}\mathbb{P}_n(A|\text{Female-Female})$$

for some event A. If

$$\mathbb{P}_n(A|\text{Male-Male}) \neq \mathbb{P}_n(A|\text{Female-Female})$$

(rather, the probabilities are not sufficiently close) then there is advantage to conditioning solely on one of these events. The reason is that one of the probabilities will constitute a more binding constraint. If preferences are convex but also exhibit gender bias then such an effect might be observed. For example, if

$$\mathbb{E}[\text{Male-Male-Third child}] > \mathbb{E}[\text{Female-Female-Third child}]$$

then the probability that there are three or more children in the household is different depending upon whether the oldest two children in the household are male, or are female. Such an effect might lead to an imbalance in the conditional probabilities. The disaggregation does not dilute the identifying power of Z.

(c) Combining the event that there is a multiple second birth with the possible permutations of gender for two children has more informational content.

In Specification 8 I combine the event that there is a multiple second birth with the possible permutations of gender for two children. There is greater statistical uncertainty in the estimate of the identified set for Specification 8, but the net effect of greater accuracy and less precision is to make the estimate of the identified set and the confidence region of the estimate of the identified set smaller than the estimate of the identified set for Specification 6.

Whether there are three or more children is dependent upon the quality margin (if preferences are convex then whether the oldest two children in the household share the same gender) and the quantity margin (whether there is a multiple second birth), and so combining information on both margins yields a more complete picture of the relationship. Some caution must be taken in combining information on both margins; namely, that the quantity margin has primacy over the quality margin. For example, if the oldest child in the household is male and there is a multiple second birth say, then whether the second oldest child is male or female is not relevant to whether there are three or more children in the household. A multiple second birth implies that there are three or more children in the household, regardless. The advantage of incorporating information on the quality margin is that quality is relevant when there is not a multiple second birth.

Incorporating information is advantageous, but (a) emphasises that simple redefinition of Z without enriching $\mathcal{R}_Z$ is insufficient for a model to have more informational context. To enrich $\mathcal{R}_Z$, it is essential that $\text{card}(\mathcal{R}_Z)$ increases such that there is greater variation to exploit. Although the gains from enriching $\mathcal{R}_Z$ in conducting estimation and inference are not substantial, the cost of incorporating information is low since the number of inequalities increases linearly in $\text{card}(\mathcal{R}_Z)$. If the task of calculating the inequalities that describe the estimate of the identified set is automated then there is little reason not to incorporate information. Further, the availability of information on so many events implies that the single equation model is overidentified, and so its credibility is testable (to an extent).

4.2 Enriching individual response

I explore the effect of permitting X to enter h for the simple single equation model. I discuss the implications for individual behaviour of assuming the simple single equation model versus assuming the single equation model, and discuss the misspecification issue that arises when individual behaviour is dependent upon X but the single equation model is assumed.

The principal difference between the simple single equation model and the single equation model is the treatment of X. The simple single equation model embeds the restriction that individual behaviour is not dependent upon X.[9] In the context of the application, the restriction implies that race does not affect a mother's employment. Cultural norms or discriminatory policies that reduce the incentive to seek employment for mothers of a particular race say, independently of ability, are excluded by the simple single equation model. Supposing that the restriction $U \perp\!\!\!\perp X$ is maintained (a reasonable restriction given the definition of X) then the interpretation of X for the simple single equation model is

[9] In fact, individual behaviour is permitted to depend upon X provided that this dependency is in the aggregate. In other words, if only the aggregate of X matters for individual behaviour, and so is constant across all economic agents.

(a) as an exogenous variable that is irrelevant to the determination of any component of the structural model; or,

(b) as an instrumental variable if X is figured to be a determinant of D, or is a good predictor of D.

In either case, if individual behaviour is dependent upon X but the simple single equation model is assumed, the restriction $U \perp\!\!\!\perp X$ is no longer valid. Note that, if the restrictions that single equation model embeds are maintained then

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

can always be written as

$$Y = \mathbb{1}[\zeta(D,X) < \lambda(D,V),] \tag{4.1}$$

where (ζ, λ) are arbitrary functions and V is an independent random variable. This transformation is valid by the quantile transform that is a monotone transformation of both sides of the inequality.[10] I define a function π such that

$$F_\upsilon(\pi(d)) \equiv r(d),$$

where F_υ is the distribution function of some as yet undefined variable.[11] Adding and subtracting $\pi(D)$ inside the indicator of (4.1) yields

$$Y = \mathbb{1}[\pi(D) < \lambda(D,V) + \pi(D) - \zeta(D,X)],$$

which upon defining

$$\upsilon \equiv \lambda(D,V) + \pi(D) - \zeta(D,X)$$

demonstrates why the restriction $U \perp\!\!\!\perp X$ is no longer valid (υ is a function of X).[12] The interpretation of the bias term $\zeta(D,X)$ is then as omitted variable bias in the case of

[10] For exposition, define

$$W \equiv \lambda(D,V)$$

and suppose that $W \sim F_W$, which is a well-defined distribution function. Applying F_W to both sides of the inequality yields

$$F_W(\zeta(D,X)) \equiv p(D,X) < U \equiv F_W(\lambda(D,V)).$$

To be precise, this final operation is the inverse quantile transform.

[11] I depart from the notation that is employed throughout the rest of the paper in order to emphasise the variable to which this distribution corresponds.

[12] Applying the inverse quantile transform to $\pi(D)$ and to υ yields

$$F_\upsilon(\pi(D)) \equiv r(D) < U \equiv F_\upsilon(\lambda(D,V) + \pi(D) - \zeta(D,X)).$$

(a), but more worryingly as a violation of the exclusion restriction if X is re-branded as an instrumental variable as in the case of (b).

4.2.1 Identification analysis

I describe the identified set of the collection of $r(d)$ over $\mathcal{R}_D$ for the simple single equation model. I employ random set theory as a tool for the identification analysis, and proceed as for the identification analysis of the single equation model.

Assume that there is a particular ordering of the collection of $r(d)$ over $\mathcal{R}_D$. First, I define correspondences $\mathcal{A}_r$ and $\mathcal{B}_r$ as follows.

$$\mathcal{A}_r(d) \equiv \{a : r(a) \leq r(d)\}$$
$$\mathcal{B}_r(d) \equiv \{b : r(b) \geq r(d)\}$$

Second, I describe the identified set of $r(d)$ as

$$\sup_{z\in\mathcal{R}_Z} \mathbb{P}(Y = 0, D \in \mathcal{A}_r(d)|z) \leq r(d) \leq 1 - \sup_{z\in\mathcal{R}_Z} \mathbb{P}(Y = 1, D \in \mathcal{B}_r(d)|z), \tag{4.2}$$

which is valid by Artstein's Inequality.

4.2.2 Model misspecification and bias

I discuss the bias that arises when individual behaviour is dependent upon X but the simple single equation model is assumed. I show that the simple single equation model does not necessarily identify a superset of the identified set if the restrictions that the single equation model embeds are maintained (model misspecification).

If the simple single equation model is misspecified then it does not (partially) identify $\mathbb{E}[Y(d)]$. Rather, the simple single equation model (partially) identifies the functional $r(d)$ that encompasses a bias. A natural question is then how does the identified set of $r(d)$ compare to the identified set of $\mathbb{E}[Y(d)]$, which is identified by the single equation model? I focus on the lower set of inequalities that describe the identified set in each case since the analysis is sufficient to address the question. I expand the left-hand inequality of (4.2) to

$$\sup_{z\in\mathcal{R}_Z} \sum_{x\in\mathcal{R}_X} \mathbb{P}(Y = 0, D \in \mathcal{A}_r(d)|x, z)\mathbb{P}(x|z), \tag{4.3}$$

which is valid by the Law of Total Expectation. Contrast this with the lower bound that is identified by the single equation model.

$$\sum_{x \in \mathcal{R}_X} \mathbb{P}(x) \sup_{\eta \in \mathcal{R}_X} \sup_{z \in \mathcal{R}_Z} \mathbb{P}(Y = 0, D \in \mathcal{A}_p(\eta; d, x) | \eta, z) \tag{4.4}$$

I study the differences between (4.3) and (4.4), emphasising each difference that I study in turn to facilitate comparison. I study the partial (*ceteris paribus*) effect of each difference.

(a) First, the supremum operator in (4.4) guarantees that the conditional probability $\mathbb{P}(\cdot|\eta, z)$ is at least as great as in (4.3).

$$\sum_{x \in \mathcal{R}_X} \mathbb{P}(x) \sup_{\eta \in \mathcal{R}_X} \sup_{z \in \mathcal{R}_Z} \mathbb{P}(Y = 0, D \in \mathcal{A}_p(\eta; d, x) | \eta, z)$$

This difference determines that the bias from model misspecification is a downwards bias.

(b) Second, the location of the supremum operator in (4.4) guarantees that the conditional probability $\mathbb{P}(\cdot|\eta, z)$ is at least as great as in (4.3).

$$\sum_{x \in \mathcal{R}_X} \mathbb{P}(x) \sup_{\eta \in \mathcal{R}_X} \sup_{z \in \mathcal{R}_Z} \mathbb{P}(Y = 0, D \in \mathcal{A}_p(\eta; d, x) | \eta, z)$$

It is somewhat misleading to study this difference independently of the weighting function but it is nonetheless enlightening to study the partial effect of moving the supremum operator inside the sum. If the difference between the probability weights is ignored then moving the supremum operator inside the sum determines that the bias from model misspecification is a downwards bias.

(c) Third, the probability weights in (4.4) are unconditional probabilities.

$$\sum_{x \in \mathcal{R}_X} \mathbb{P}(x) \sup_{\eta \in \mathcal{R}_X} \sup_{z \in \mathcal{R}_Z} \mathbb{P}(Y = 0, D \in \mathcal{A}_p(\eta; d, x) | \eta, z)$$

Together with moving the supremum operator inside the sum, the difference in probability weights determines that the bias from model misspecification is ambiguous.

(d) Fourth, $\mathcal{A}_r$ is not necessarily equal to $\mathcal{A}_p$.

$$\sum_{x \in \mathcal{R}_X} \mathbb{P}(x) \sup_{\eta \in \mathcal{R}_X} \sup_{z \in \mathcal{R}_Z} \mathbb{P}(Y = 0, D \in \mathcal{A}_p(\eta; d, x) | \eta, z)$$

Notice that the set $\mathcal{A}_r$ is constant across the summation, but that $\mathcal{A}_p$ can be different in each term of the summation. Further, it is unclear whether $\mathcal{A}_r$ and $\mathcal{A}_p$ are equal in each term of the summation. This difference determines that the bias from model misspecification is ambiguous.

Overall, the differences between (4.3) and (4.4) determine that the bias from model misspecification is ambiguous; the bias from (c) and (d) can outweigh the bias from (a) and (b). Extending this finding to the upper bound that is identified by the single equation model, the effect of permitting X to enter h is to alter the length of the identified set, to shift the location of the identified set, or a combination of both. With respect to location shift, the finding is consistent with the behaviour of identifying models.

4.2.3 Hispanic ethnicity

I estimate the identified set of the average causal effect of additional children on a mother's employment and the identified sets of its conditional counterparts, assuming that the single equation model is a valid representation of mothers' behaviour. I discuss the estimates. I discuss a practical issue regarding calculation of $ACE(D \to Y)$ from its conditional counterparts when there is more than one ordering that survives elimination.

I define X as the event that a mother is Hispanic. There is reason to suppose that the behaviour of Hispanic mothers is different to mothers of a different ethnic background due to cultural differences and immigration patterns.

$$\mathbb{P}_n(X = 0) = 0.926 \text{ versus } \mathbb{P}_n(X = 1) = 0.074$$

I define Z as the event that there is a multiple second birth. The choice of definition for Z reflects a desire to simplify the analysis, so as not to confound the analysis with issues of computation (recall that the event that there is a multiple second birth excludes many orderings).

When X and D are both binary variables there are four parameters to estimate, and 75 possible orderings over these parameters. Of these 75 possible orderings, 72 are not internally consistent given $\mathbb{P}_n$ and are eliminated. That is, given $\mathbb{P}_n$ and an ordering, the estimate of the identified set is empty for at least one of the parameters (in practice, an upper bound is less than a lower bound). The three orderings that survive elimination are

(a) $p(0,0) = p(0,1) < p(1,0) < p(1,1)$;

(b) $p(0,0) < p(0,1) < p(1,0) < p(1,1)$; and,

(c) $p(0,1) < p(0,0) < p(1,0) < p(1,1)$.

Even without assigning numerical values to the parameters, the three orderings are informative about the behaviour of mothers. First, the conditional average causal effect of childbirth given ethnicity is negative for both Hispanic mothers and for non-Hispanic mothers. On average, mothers exit employment in response to having additional children, and this is true for Hispanic mothers and for non-Hispanic mothers. The fact that this holds for mothers of Hispanic and non-Hispanic ethnicity means that the average conditional effect of childbirth is negative as well. Second, in households where there are three or more children, the percentage of Hispanic mothers that are in employment is lower than the percentage of non-Hispanic mothers that are in employment. This suggests that the hypothesis that there are cultural differences between ethnic groups that affects the employment status of mothers is valid, and supports the inclusion of X in h.[13]

Table 5 shows the estimates of the identified sets of the collection of $p(d,x)$ over $\mathcal{R}_D \times \mathcal{R}_X$. Table 6 shows the estimate of the identified set of the average causal effect of additional children on a mother's employment, and the estimates of the identified sets of its conditional counterparts. I make the following observations.

(a) In households where there are three or more children, the difference between the percentage of Hispanic mothers that are in employment and the percentage of non-Hispanic mothers that are in employment is 0.033 (a Hispanic mother is less likely to be employed by probability 0.033).

(b) The estimate of the identified set of the conditional average causal effect of additional children on a mother's employment given non-Hispanic ethnicity is constant across all orderings.

Although constancy of the estimate of the identified set across all orderings that survive elimination is a desirable property (since there is no ambiguity associated with the estimate), it is not in itself an informative property since the conditional average causal effect given a particular ethnicity is not independent of the conditional average causal effect given another ethnicity (in the sense that orderings must be preserved). Nonetheless, I suggest that it is a useful property in that the estimate of the identified set of the conditional average causal effect given Hispanic ethnicity and the estimate of the identified set

[13] Although without further statistical testing, the hypothesis that $p(1,0)$ and $p(1,1)$ are equal cannot be rejected.

of the average causal effect of additional children can both be expressed in terms of the conditional average causal effect given non-Hispanic ethnicity. This is perhaps a matter of presentation.

(c) The upper bound of the average causal effect of additional children on a mother's employment is more negative than when X is excluded from h (−0.053 versus −0.052).

Two conclusions can be drawn from this observation. First, that the average causal effect of additional children on a mother's employment is more negative than the simple single equation model would suggest. Second, that the bias when X is excluded from h is small relative to the length of the estimate of the identified set.

(d) The upper bound of the conditional average causal effect of additional children on a mother's employment given Hispanic ethnicity is more negative than given non-Hispanic ethnicity.

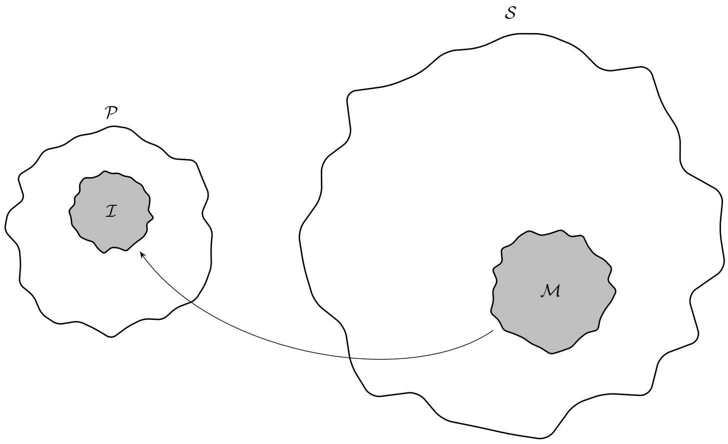
Although the effect of additional children may be larger for Hispanic mothers than for non-Hispanic mothers, this observation is not as informative as it might first appear. The orderings that survive elimination are insufficient to conclude that the conditional average causal effect of additional children on a mother's employment given Hispanic ethnicity is more negative than given non-Hispanic ethnicity. Nonetheless, the observation is informative as to the maximum and minimum values that the conditional average causal effect may take in each sub-population.

As a final remark regarding the calculation of the estimate of the identified set of $ACE(D \to Y)$ from the estimates of the identified sets of its conditional counterparts, note that it is sufficient to consider the Minkowski sum for each ordering. It is not necessary to combine the estimate of the identified set of $ACE(D \to Y|0)$ for one ordering with the estimate of the identified set of $ACE(D \to Y|1)$ for another ordering say.

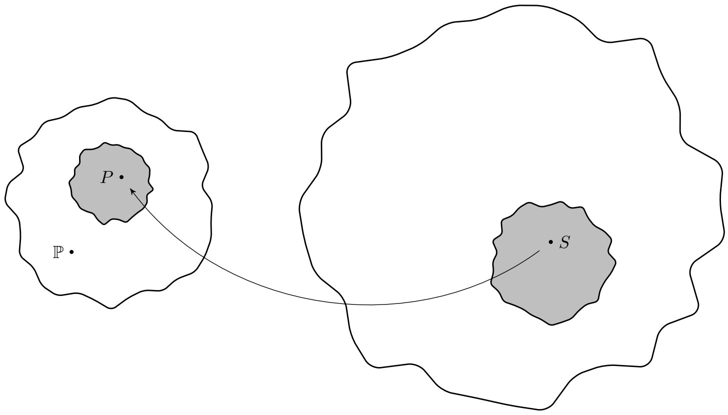
5 Conclusion

I have studied the effect of incorporating information into a single equation model. In the first instance, I studied the effect of enriching the support of an instrumental variable. I have shown that is advantageous to incorporate all available information into the model in this respect. In the context of the application that I studied, by enriching the support of the instrumental variable I was able to deliver a narrower estimate of the identified set of the average causal effect of additional children on a mother's employment. In the second instance, I studied the effect of permitting an exogenous variable to enter the structural equation for an outcome variable, and so determine its value. I have shown that the average causal effect is expressible as a function of its conditional counterparts, and that this expression takes the form of a Minkowski sum when the conditional average causal effect is partially identified. I have shown that omitting an exogenous variable from the structural equation when the exogenous variable does indeed determine the outcome variable induces a bias in the resulting estimate of the identified set. Further, I have shown that the direction of this bias is ambiguous. In the context of the application, I was able to determine that the estimate of the identified set of the average causal effect amongst Hispanic mothers is more negative than for non-Hispanic mothers, although I cannot rule out that the opposite is true for the conditional average causal effect itself. By incorporating information into the model, I am able to determine that the average causal effect of additional children on a mother's employment may be more negative than would otherwise be determined if this information was not incorporated. However, the bias from not incorporating this information is small relative to the length of the estimate of the identified set.

I suggest that there are numerous extensions that must be undertaken to improve upon this study. First, that the issue of computing statistical uncertainty for the average causal effect as a weighted sum (in which weights are themselves estimated) must be resolved. Second, the single equation model embeds some restrictions that are not necessarily plausible. I suggest that it would be useful to relax some of the more stringent restrictions that the model embeds.

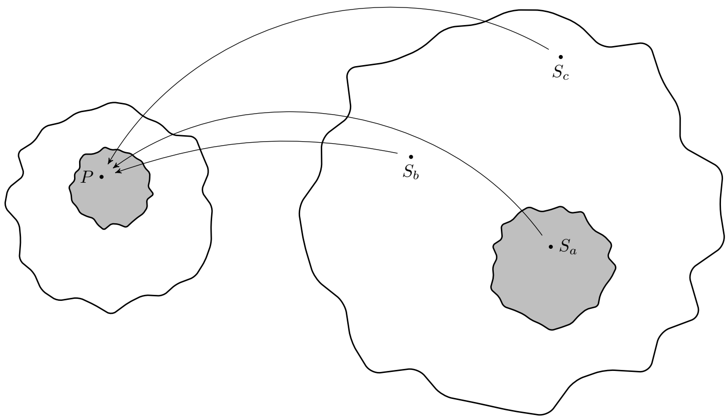


A model $\mathcal{M}$ is a set of structures that forms a proper subset of the class of all structures $\mathcal{S}$. Each structure in $\mathcal{M}$ generates a probability distribution in the class of all probability distributions (of observable variables) $\mathcal{P}$. Then the image $\mathcal{I}$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

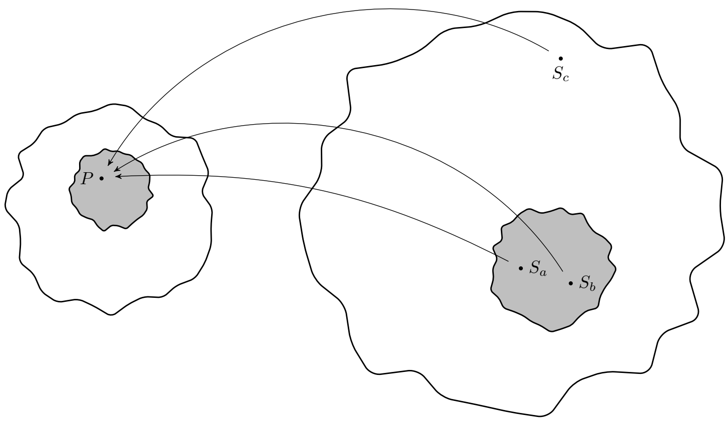


A structure S is incompatible with data if it generates a probability distribution (of observable variables) P that is distinct from a realised probability distribution $\mathbb{P}$. If all structures in $\mathcal{M}$ are incompatible with data then $\mathcal{M}$ is said to be observationally restrictive,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

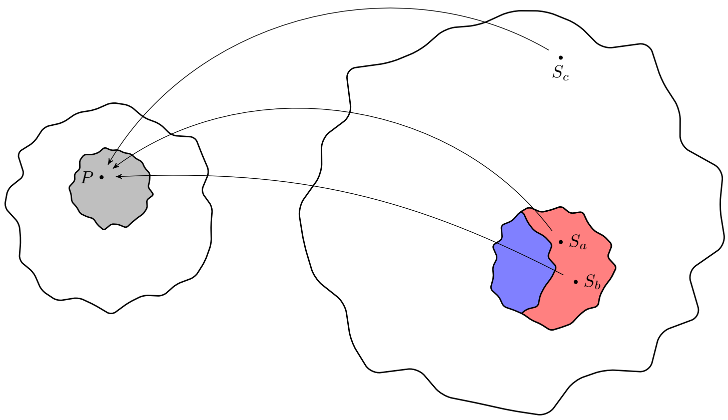


A model $\mathcal{M}$ is said to identify a structure S if the probability distribution (of observable variables) P that is generated by S is distinct from those generated by other structures in $\mathcal{M}$. The structures S_a, S_b and S_c are said to be observationally equivalent as they all generate 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.

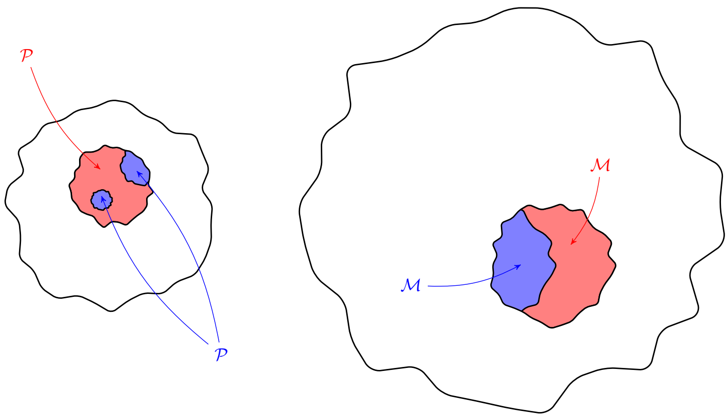


As S_a and S_b are observationally equivalent and are both admitted by $\mathcal{M}$ then $\mathcal{M}$ does not identify either S_a or S_b. Nonetheless, as $\mathcal{M}$ restricts the set of observationally equivalent structures that generate P to S_a and S_b then $\mathcal{M}$ partially identifies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



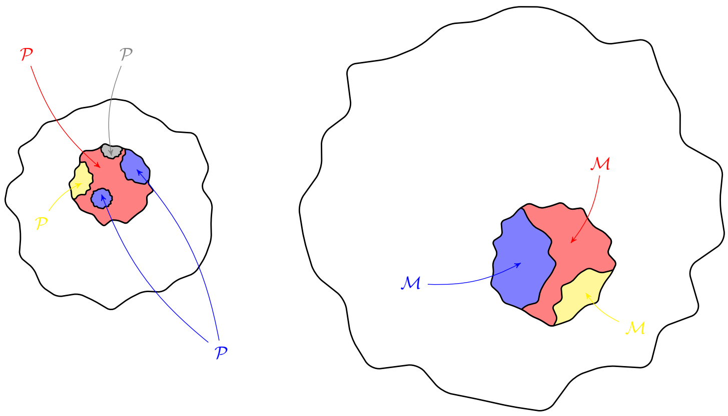
A structural characteristic χ is a function of a structure S. A model $\mathcal{M}$ can be partitioned such that structures in a partition deliver the same value for χ. Structures in the red partition $\mathcal{M}$ deliver the value a for χ, and structures in the blue partition $\mathcal{M}$ deliver 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.



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 red partition $\mathcal{P}$ and into the blue 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 3 $\mathcal{M}$ 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.

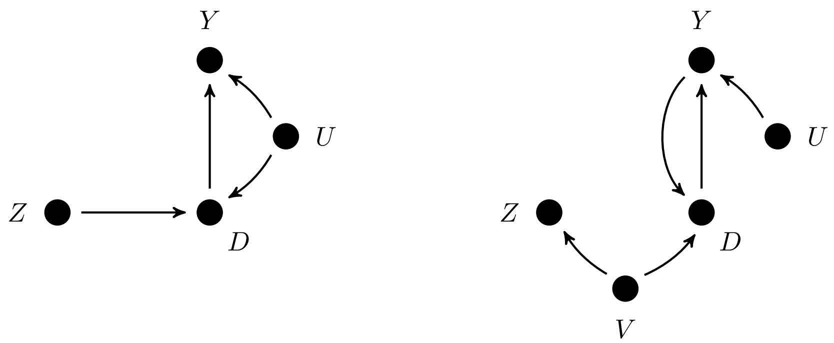
A structural characteristic χ is a function of a structure S. A model $\mathcal{M}$ can be partitioned such that structures in a partition deliver the same value for χ. Structures in the red partition $\mathcal{M}$ deliver the value a for χ, structures in the blue partition $\mathcal{M}$ deliver the 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 $\mathcal{P}$.



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 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 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.

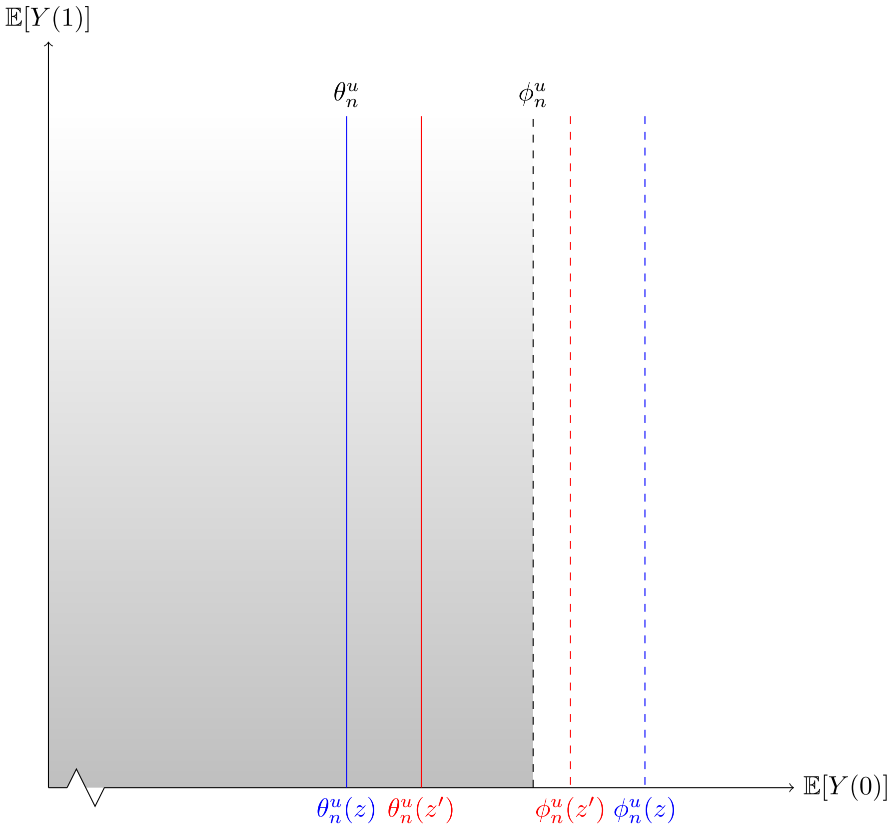
The left- and right-hand panels show directed acyclic graphs that are able to represent the same probability distribution of (Y, D). In the left-hand panel, U causes Y and D. The left-hand panel is a representation of selection in that particular values of D are strongly associated with particular values 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 converges to an equilibrium).



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.

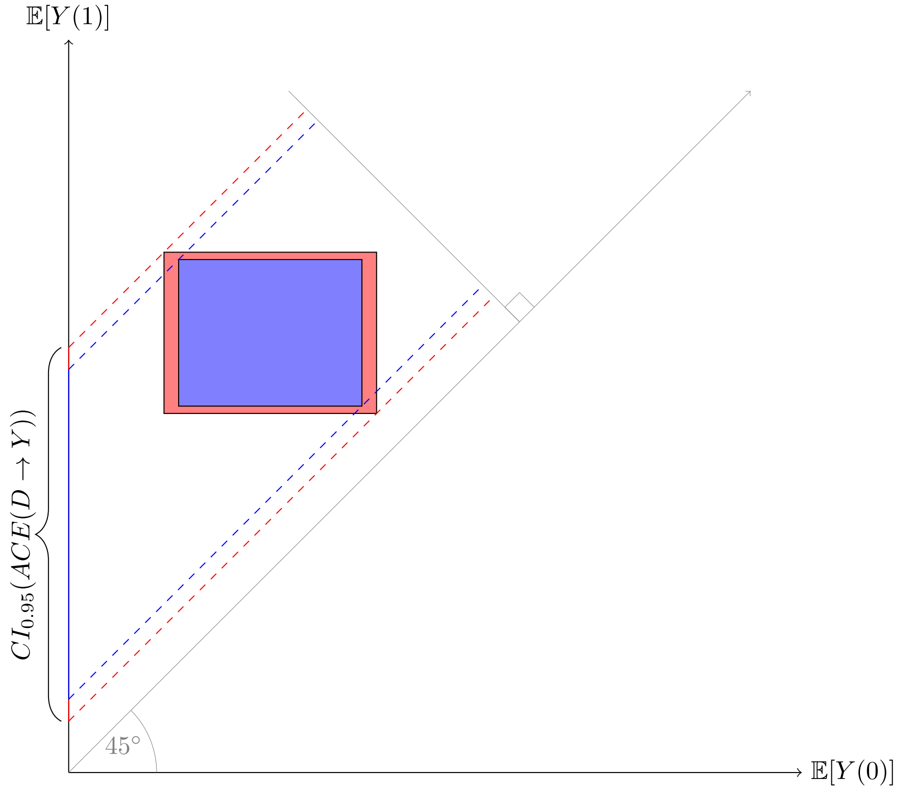
Suppose that $\theta_n^u(z)$ and $\theta_n^u(z')$ are estimates of upper bounds on the identified set of $\mathbb{E}[Y(0)]$. Similarly, suppose that $\phi_n^u(z)$ and $\phi_n^u(z')$ are one-sided $1-\alpha$ confidence regions for $\theta_n^u(z)$ and $\theta_n^u(z')$. $\phi_n^u(z') > \phi_n^u(z')$ if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



As $\theta_n^u(z)$ and $\theta_n^u(z')$ are both constraints on the identified set of $\mathbb{E}[Y(0)]$, 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 $\mathbb{P}(\mathbb{E}[Y(0)] > \phi_n^u) = \alpha$. 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 $\mathbb{P}(\mathbb{E}[Y(0)] > \theta_n^u(z)+k\sigma) = \alpha$. 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.

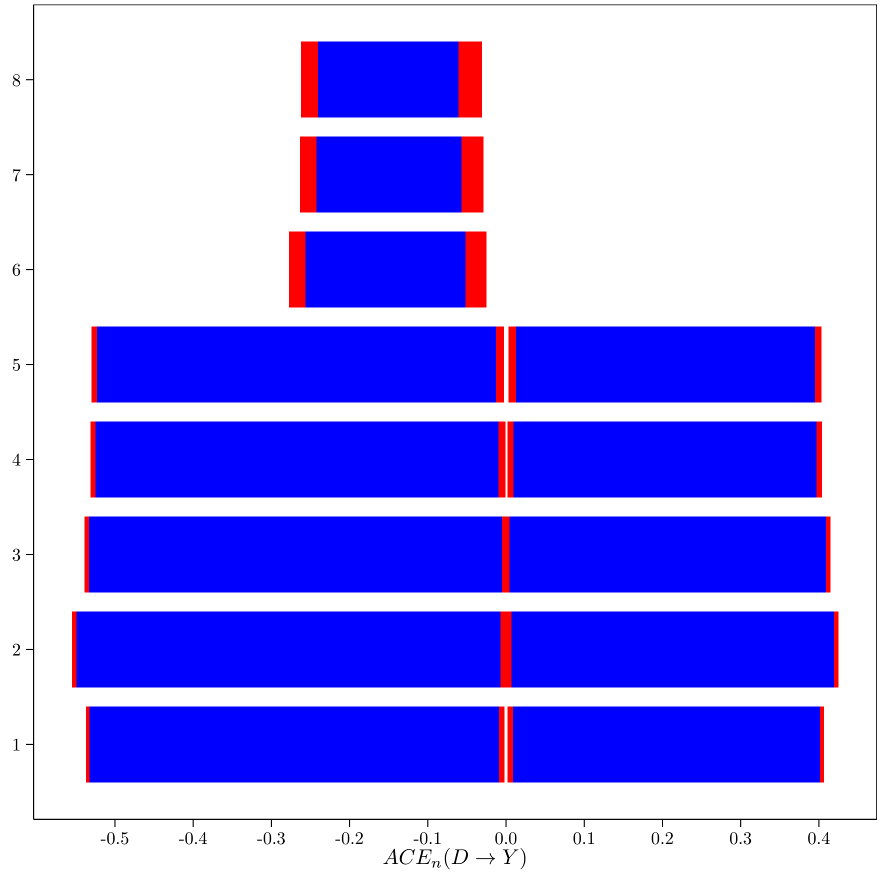
Suppose that the estimate of the identified set of $(\mathbb{E}[Y(0)], \mathbb{E}[Y(1)])$ is given by the blue rectangle, and that the $1-\alpha$-confidence region for this set is the union of the blue rectangle and the red polygon.



It is possible to recover the estimate of the identified set of $ACE(D \to Y)$ 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. Fourth, 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 identified value of $(\mathbb{E}_n[Y(0)], \mathbb{E}_n[Y(1)])$; 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 estimate of the identified set 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.

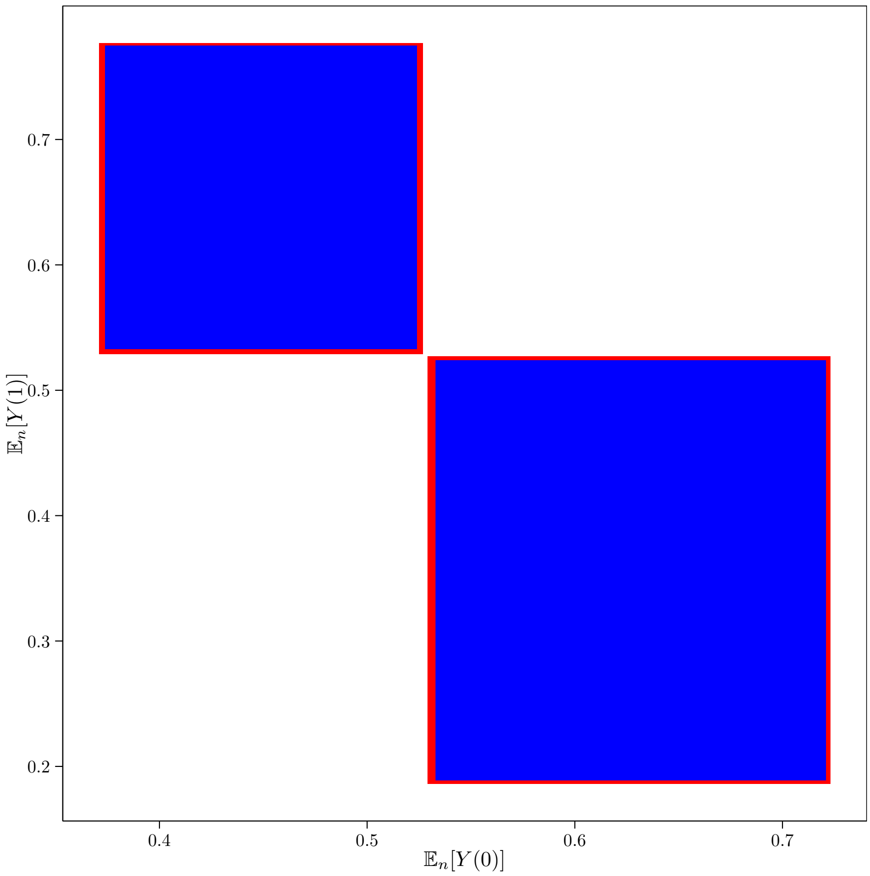
The plot shows how the estimate of the identified set of $ACE(D \to Y)$ changes as $\mathcal{R}_Z$ is varied. Each value on the y-axis corresponds to a different specification. In each specification, $\mathcal{R}_Z$ is varied and the estimate of the identified set of $ACE(D \to Y)$ is calculated. See Table 3 and Table 4 for the definition of $\mathcal{R}_Z$ in each specification (the column headed No. states the specification number in each table, and the column headed $\mathcal{R}_Z$ states the events that form the points of support for Z). The plot is a graphical representation of Table 3 and Table 4. Blue regions represent the estimate of the identified set of $ACE(D \to Y)$, and the union of blue and red regions represent 0.950 confidence regions for $ACE(D \to Y)$.



In Specification 1 through Specification 5, information relating to whether there is a multiple second birth is ignored. The estimate of the identified sets of values of $ACE(D \to Y)$ in these specifications are large and disconnected, which suggests that child gender is weakly associated with the number of children in a household. In other words, that child gender is a weak instrumental variable. Child gender is, by itself, insufficient to determine the sign of the average causal effect of the number of children in a household on a mother's employment. Specification 2 and Specification 3 are unable to rule out a null effect at the 0.950 significance level (the central red regions overlap at zero). In Specification 6 through Specification 8, information relating to whether there is multiple second birth is incorporated. The sign of the average causal effect of the number of children in a household on a mother's employment is negative, and significant. The length of the estimate of the identified set of $ACE(D \to Y)$ is decreasing as more information is incorporated, as is the 0.950 confidence region for $ACE(D \to Y)$.

Figure 8: Estimates of the identified set of $ACE(D \to Y)$ as $\mathcal{R}_Z$ is varied.

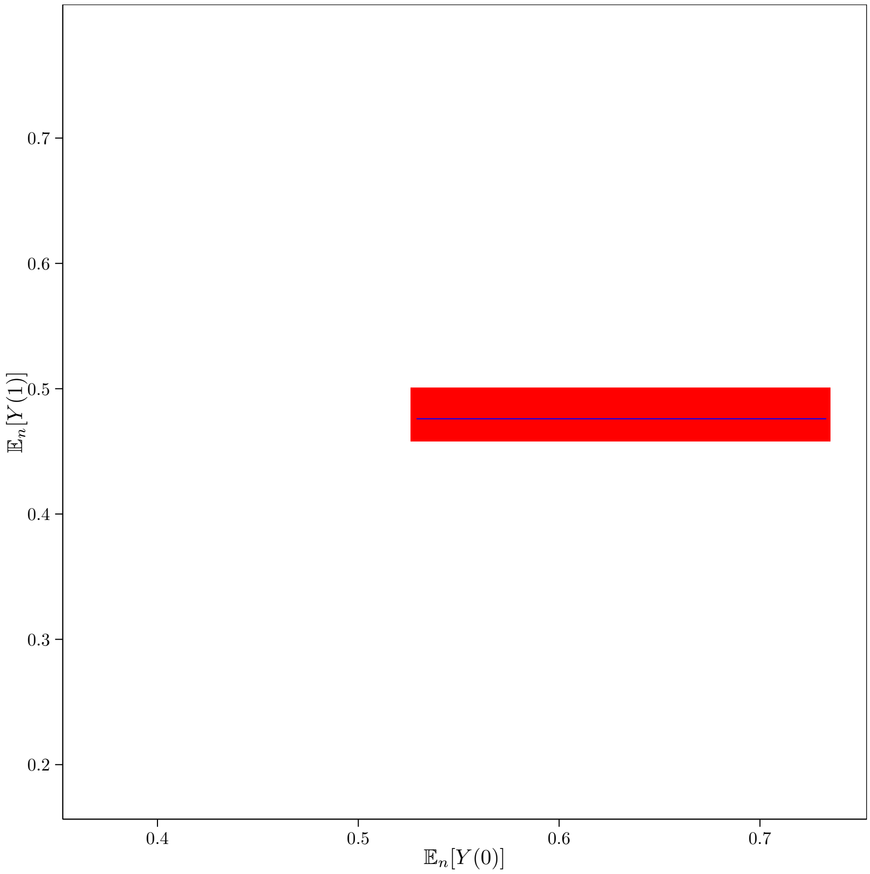
The plot shows the estimate of the identified set of $(\mathbb{E}[Y(0)], \mathbb{E}[Y(1)])$ when $\mathcal{R}_Z$ is formed of the events Male-Male $\cup$ Female-Female and Male-Female $\cup$ Female-Male. This is Specification 1 in Table 1. Blue regions represent the estimate of the identified set of $(\mathbb{E}[Y(0)], \mathbb{E}[Y(1)])$, and the union of blue and red regions represent 0.950 confidence regions for $(\mathbb{E}_n[Y(0)], \mathbb{E}_n[Y(1)])$. The confidence regions that are shown in the plot are different to those that are reported in Table 1. The distinction arises since Table 1 reports 0.950 confidence regions for $\mathbb{E}[Y(0)]$ and for $\mathbb{E}[Y(1)]$ whereas the plot shows the 0.950 confidence region for $(\mathbb{E}[Y(0)], \mathbb{E}[Y(1)])$.



The estimate of the identified set of $(\mathbb{E}[Y(0)], \mathbb{E}[Y(1)])$ is large and disconnected as child gender is a weak instrumental variable. There is little statistical uncertainty since the number of data is large, and as the event that the oldest two children in the household is frequent. A null effect can be ruled out but the estimate is otherwise relatively uninformative.

Figure 9: Estimate of the identified set of $(\mathbb{E}[Y(0)], \mathbb{E}[Y(1)])$ for Specification 1.

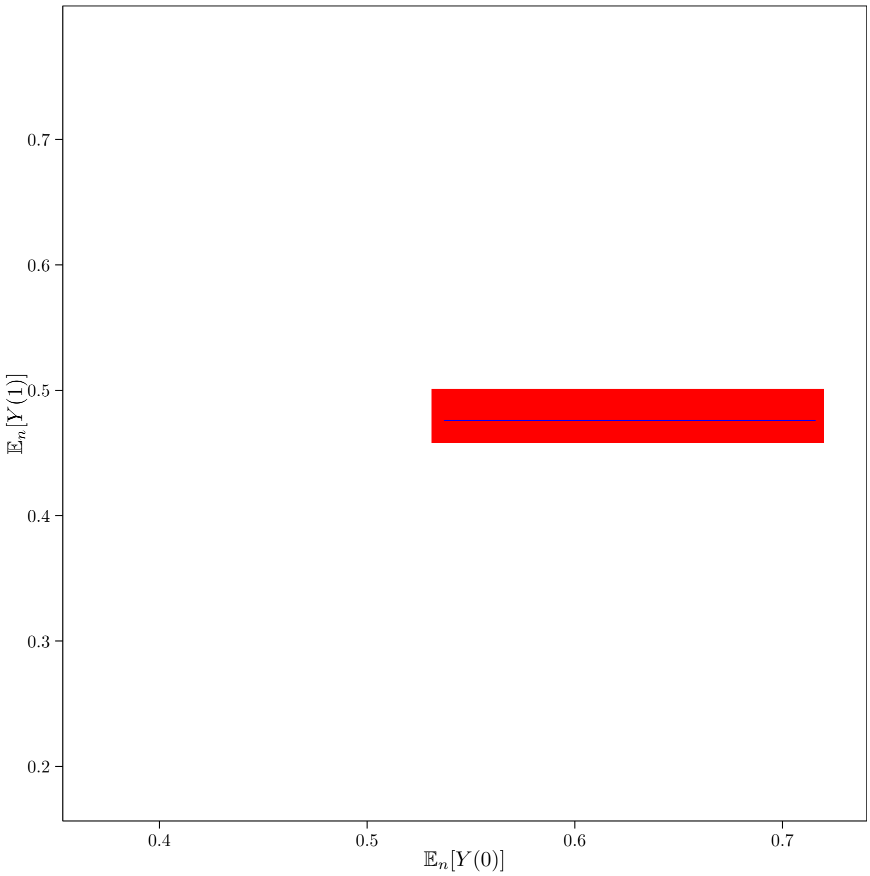
The plot shows the estimate of the identified set of $(\mathbb{E}[Y(0)], \mathbb{E}[Y(1)])$ when $\mathcal{R}_Z$ is formed of the events Multiple birth, Single birth $\cap$ (Male-Male $\cup$ Female-Female) and Single birth $\cap$ (Male-Female $\cup$ Male-Female). This is Specification 6 in Table 2. Blue regions represent the estimate of the identified set of $(\mathbb{E}[Y(0)], \mathbb{E}[Y(1)])$, and the union of blue and red regions represent 0.950 confidence regions for $(\mathbb{E}[Y(0)], \mathbb{E}[Y(1)])$. The confidence regions that are shown in the plot are different to those that are reported in Table 2. The distinction arises since Table 2 reports 0.950 confidence regions for $\mathbb{E}[Y(0)]$ and for $\mathbb{E}[Y(1)]$ whereas the plot shows the 0.950 confidence region for $(\mathbb{E}[Y(0)], \mathbb{E}[Y(1)])$.



The event that there is a multiple second birth is a special instrumental variable, and identifies $\mathbb{E}_n[Y(1)]$ in the sub-population of mothers that experience a multiple second birth. As such, the estimate of the identified set of $\mathbb{E}[Y(1)]$ is a point. The estimate of the identified set is small, although there is medium variation in the estimate of the identified set of $\mathbb{E}[Y(0)]$ that means that it is not overly informative about $ACE(D \to Y)$. The confidence region is large relative to the set due to the rarity of a multiple second birth. This is particularly apparent in the y-direction.

Figure 10: Estimate of the identified set of $(\mathbb{E}[Y(0)], \mathbb{E}[Y(1)])$ for Specification 6.

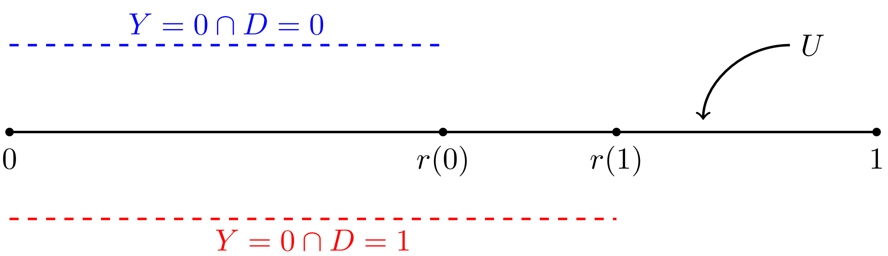
The plot shows the estimate of the identified set of $(\mathbb{E}_n[Y(0)], \mathbb{E}_n[Y(1)])$ when $\mathcal{R}_Z$ is formed of the events Multiple birth, Single birth $\cap$ Male-Male, Single birth $\cap$ Female-Female, Single birth $\cap$ Male-Female and Single birth $\cap$ Female-Male. This is Specification 8 in Table 2. Blue regions represent the estimate of the identified set of $(\mathbb{E}_n[Y(0)], \mathbb{E}_n[Y(1)])$, and the union of blue and red regions represent 0.950 confidence regions for $(\mathbb{E}_n[Y(0)], \mathbb{E}_n[Y(1)])$. The confidence regions that are shown in the plot are different to those that are reported in Table 2. The distinction arises since Table 2 reports 0.950 confidence regions for $\mathbb{E}_n[Y(0)]$ and for $\mathbb{E}_n[Y(1)]$ whereas the plot shows the 0.950 confidence region for $(\mathbb{E}_n[Y(0)], \mathbb{E}_n[Y(1)])$.



Specification 8 incorporates all available information into the special single equation model with respect to Z. As such, the estimate of the identified set of $(\mathbb{E}_n[Y(0)], \mathbb{E}_n[Y(1)])$ is the smallest possible set that is achievable without further restriction or additional data. There is advantage to Specification 8 relative to Specification 6 but it is clear that the gains are not substantial.

Figure 11: Estimate of the identified set of $(\mathbb{E}[Y(0)], \mathbb{E}[Y(1)])$ for Specification 8.

The goal of the identification analysis is to determine the location of $r(0)$ and of $r(1)$. The analysis is confounded by the fact that it is not known whether $r(0) > r(1)$, $r(0) < r(1)$ or $r(0) = r(1)$. The order of the collection of $r(d)$ is crucial since it is the order that determines the value of $\mathcal{A}_r$ and of $\mathcal{B}_r$. Suppose that $r(0) < r(1)$. What can be determined about the location of $p(0)$ from observable variables alone? Any economic agent that is characterised by $Y = 0 \cap D = 0$ must also be characterised by u somewhere in the blue dashed interval. Similarly, any economic agent that is characterised by $Y = 0 \cap D = 1$ must also be characterised by u somewhere in the red dashed interval. The location of $r(0)$ is confounded since it is not known what proportion of $Y = 0 \cap D = 1$ are located in $[0, r(0)]$ (the left partition of the red dashed interval), and what proportion of $Y = 0 \cap D = 1$ are located in $(r(0), r(1)]$ (the right partition of the red dashed interval). A lower bound on the location of $r(0)$ is then that all economic agents that are are characterised by $Y = 0 \cap D = 1$ are located in $(r(0), r(1)]$. The measure of economic agents that are located in $[0, r(0)]$ in this case is then $\mathbb{P}(Y = 0, D = 0)$. An upper bound on the location of $r(0)$ is then that all economic agents that are are characterised by $Y = 0 \cap D = 1$ are located in $[0, r(0)]$. The measure of economic agents that are located in $[0, r(0)]$ in this case is then $\mathbb{P}(Y = 0, D = 0 \cup 1)$. For this order, $\mathcal{A}_r(0) = \{0\}$ and $\mathcal{B}_r(0) = \{0, 1\}$.



Note that economic agents that are located in $[0, r(0)]$ are characterised by $Y(0) = 0$ and $Y(1) = 0$. In contrast, economic agents that are located in $(r(0), r(1)]$ are characterised by $Y(0) = 1$ and $Y(1) = 0$. The problem of determining the measure of economic agents that are characterised by each pair $(Y(0), Y(1))$ is then equivalent to locating $r(0)$ and $r(1)$. This point validates consideration of level sets of U.

Figure 12: A note on partial identification.

Table 1: $\mathcal{R}_Z$ and estimates of the identified set of $\mathbb{E}[Y(d)]$.

		Bound			
No.	$\mathcal{R}_Z$	$\mathbb{E}_n^-[Y(0)]$	$\mathbb{E}_n^-[Y(1)]$	$\mathbb{E}_n^+[Y(0)]$	$\mathbb{E}_n^+[Y(1)]$
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]
2	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]
3	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]
4	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				
5	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				

The column headed No. states the specification number. The column headed $\mathcal{R}_Z$ states the events that form the points of support of Z. For example, in Specification 1 Z is a random variable that takes the value z when the event Male-Male ∪ Female-Female occurs and the value z' when this does not occur. The event Male-Female is the event that the oldest two children in a household are male and female, respectively. The columns headed Bound are the estimates of the identified set of $\mathbb{E}[Y(D)]$. A superscript $-$ is written when the bounds are conditional on $\mathbb{E}[Y(0)] \geq \mathbb{E}[Y(1)]$, and a superscript $+$ is written when the bounds are conditional on $\mathbb{E}[Y(0)] \leq \mathbb{E}[Y(1)]$. 0.950 confidence regions for $\mathbb{E}[Y(D)]$ are in blue and are constructed from one-sided 0.975 confidence regions for the lower bound and the upper bound.

Table 2: $\mathcal{R}_Z$ and estimates of the identified set of $\mathbb{E}[Y(d)]$ (*continued*).

No.	$\mathcal{R}_Z$	Bound	
		$\mathbb{E}_n^-[Y(0)]$	$\mathbb{E}_n^-[Y(1)]$
6	Multiple birth	$[0.529, 0.733]$	0.476
	Single birth	$[0.526, 0.734]$	$[0.460, 0.498]$
7	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)		
8	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		

The column headed No. states the specification number. The column headed $\mathcal{R}_Z$ states the events that form the points of support of Z. For example, in Specification 1 Z is a random variable that takes the value z when the event Male-Male ∪ Female-Female occurs and the value z' when this does not occur. The event Male-Female is the event that the oldest two children in a household are male and female, respectively. The columns headed Bound are the estimates of the identified set of $\mathbb{E}[Y(D)]$. A superscript $-$ is written when the bounds are conditional on $\mathbb{E}[Y(0)] \geq \mathbb{E}[Y(1)]$, and a superscript $+$ is written when the bounds are conditional on $\mathbb{E}[Y(0)] \leq \mathbb{E}[Y(1)]$. 0.950 confidence regions for $\mathbb{E}[Y(D)]$ are in blue and are constructed from one-sided 0.975 confidence regions for the lower bound and the upper bound.

Table 3: $\mathcal{R}_Z$ and estimates of the identified set of $ACE(D \to Y)$.

No.	$\mathcal{R}_Z$	Bound	
		$ACE_n^-(D \to Y)$	$ACE_n^+(D \to Y)$
1	Male-Male ∪ Female-Female	$[-0.532, -0.009]$	$[0.009, 0.401]$
	Male-Female ∪ Female-Male	$[-0.537, -0.002]$	$[0.002, 0.406]$
2	Male-Male	$[-0.549, -0.007]$	$[0.007, 0.419]$
	Male-Female ∪ Female-Male ∪ Female-Female	$[-0.555, 0.001]$	$[-0.001, 0.425]$
3	Female-Female	$[-0.533, -0.005]$	$[0.005, 0.409]$
	Male-Male ∪ Male-Female ∪ Female-Male	$[-0.539, 0.003]$	$[-0.003, 0.415]$
4	Male-Male	$[-0.525, -0.010]$	$[0.010, 0.397]$
	Female-Female	$[-0.531, -0.001]$	$[0.002, 0.404]$
	Male-Female ∪ Female-Male		
5	Male-Male	$[-0.523, -0.013]$	$[0.013, 0.395]$
	Female-Female	$[-0.530, -0.003]$	$[0.003, 0.403]$
	Male-Female		
	Female-Male		

The column headed No. states the specification number. The column headed $\mathcal{R}_Z$ states the events that form the points of support of Z. For example, in Specification 1 Z is a random variable that takes the value z when the event Male-Male ∪ Female-Female occurs and the value z' when this does not occur. The event Male-Female is the event that the oldest two children in a household are male and female, respectively. The columns headed Bound are the estimates of the identified set of $ACE(D \to Y)$. A superscript − is written when the bounds are conditional on $ACE(D \to Y) \leq 0$, and a superscript + is written when the bounds are conditional on $ACE(D \to Y) \geq 0$. 0.950 confidence regions for $ACE(D \to Y)$ are in blue and are constructed from one-sided 0.975 confidence regions for the lower bound and the upper bound.

Table 4: $\mathcal{R}_Z$ and estimates of the identified set of $ACE(D \to Y)$ (*continued*).

No.	$\mathcal{R}_Z$	Bound $ACE_n^-(D \to Y)$
6	Multiple birth	$[-0.256, -0.052]$
	Single birth	$[-0.277, -0.025]$
7	Multiple birth	$[-0.242, -0.057]$
	Single birth $\cap$ (Male-Male $\cup$ Female-Female)	$[-0.263, -0.029]$
	Single birth $\cap$ (Male-Female $\cup$ Female-Male)	
8	Multiple birth	$[-0.240, -0.061]$
	Single birth $\cap$ Male-Male	$[-0.262, -0.031]$
	Single birth $\cap$ Female-Female	
	Single birth $\cap$ Male-Female	
	Single birth $\cap$ Female-Male	

The column headed No. states the specification number. The column headed $\mathcal{R}_Z$ states the events that form the points of support of Z. For example, in Specification 1 Z is a random variable that takes the value z when the event Male-Male $\cup$ Female-Female occurs and the value z' when this does not occur. The event Male-Female is the event that the oldest two children in a household are male and female, respectively. The columns headed Bound are the estimates of the identified set of $ACE(D \to Y)$. A superscript $-$ is written when the bounds are conditional on $ACE(D \to Y) \leq 0$, and a superscript $+$ is written when the bounds are conditional on $ACE(D \to Y) \geq 0$. 0.950 confidence regions for $ACE(D \to Y)$ are in blue and are constructed from one-sided 0.975 confidence regions for the lower bound and the upper bound.

Table 5: Non-eliminated orderings and estimates of the identified set of $p(D, X)$.

		Parameter			
No.	Ordering	$p(0,0)$	$p(0,1)$	$p(1,0)$	$p(1,1)$
(a)	$p(0,0) = p(0,1) < p(1,0) < p(1,1)$	$[0.272, 0.470]$	$[0.272, 0.470]$	0.521	0.555
(b)	$p(0,0) < p(0,1) < p(1,0) < p(1,1)$	$[0.272, 0.470]$	$[0.272, 0.494]$	0.521	0.555
(c)	$p(0,1) < p(0,0) < p(1,0) < p(1,1)$	$[0.272, 0.470]$	$[0.212, 0.470]$	0.521	0.555

The column headed No. references the ordering as listed in the main text. The column headed Ordering states the ordering of the collection of $p(d, x)$ over $\mathcal{R}_D \times \mathcal{R}_X$ that survives elimination. The columns headed Parameter are the estimates of the identified set of $p(D, X)$. The estimate of the identified set must preserve the ordering that generates it. For example, in (a), the estimate of the identified set of $p(0,1)$ could be written simply as $p(0,0)$ since the value of the parameters must be the same.

Table 6: Non-eliminated orderings and estimates of the identified set of $ACE(D \to Y|X)$ and of $ACE(D \to Y)$.

		Functional		
No.	Ordering	$ACE_n(D \to Y\|0)$	$ACE_n(D \to Y\|1)$	$ACE_n(D \to Y)$
(a)	$p(0,0) = p(0,1) < p(1,0) < p(1,1)$	$[-0.250, -0.052]$	$[-0.283, -0.085]$	$[-0.252, -0.054]$
(b)	$p(0,0) < p(0,1) < p(1,0) < p(1,1)$	$[-0.250, -0.052]$	$[-0.283, -0.061]$	$[-0.252, -0.053]$
(c)	$p(0,1) < p(0,0) < p(1,0) < p(1,1)$	$[-0.250, -0.052]$	$[-0.343, -0.085]$	$[-0.257, -0.054]$

The column headed No. references the ordering as listed in the main text. The column headed Ordering states the ordering of the collection of $p(d,x)$ over $\mathcal{R}_D \times \mathcal{R}_X$ that survives elimination. The columns headed Functional are the estimates of the identified set of $ACE(D \to Y)$ and its conditional counterparts. The estimate of the identified set must preserve the ordering that generates it. Therefore some caution must be taken in interpreting each interval (each interval represents the interval of maximum length for each functional; in practice, the value of another functional may truncate the interval of a functional).

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.

Artstein, Z. (1983). Distributions of random sets and random selections. *Israel Journal of Mathematics 46*(4), 313 – 324.

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.

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. (1988). *Analog Estimation Methods in Econometrics.* New York and London: Chapman and Hall.

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.

Pearl, J. (1995). On the testability of causal models with latent and instrumental variables. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 435 – 443. Morgan Kaufmann Publishers Inc.

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.