The Identification Zoo - Meanings of Identification in Econometrics

Arthur Lewbel
Boston College

First version January 2016, Final preprint version October 2019,
Published version: Journal of Economic Literature, December 2019, 57(4).

Abstract

Over two dozen different terms for identification appear in the econometrics literature, including set identification, causal identification, local identification, generic identification, weak identification, identification at infinity, and many more. This survey: 1. gives a new framework unifying existing definitions of point identification, 2. summarizes and compares the zooful of different terms associated with identification that appear in the literature, and 3. discusses concepts closely related to identification, such as normalizations and the differences in identification between structural models and causal, reduced form models.

JEL codes: C10, B16
Keywords: Identification, Econometrics, Coherence, Completeness, Randomization, Causal inference, Reduced Form Models, Instrumental Variables, Structural Models, Observational Equivalence, Normalizations, Nonparametrics, Semiparametrics.
I would like to thank Steven Durlauf, Jim Heckman, Judea Pearl, Krishna Pendakur, Frederic Vermeulen, Daniel Ben-Moshe, Xun Tang, Juan-Carlos Escanciano, Jeremy Fox, Eric Renault, Yingying Dong, Laurens Cherchye, Matthew Gentzkow, Fabio Schiantarelli, Andrew Pua, Ping Yu, and five anonymous referees for many helpful suggestions. All errors are my own.

Corresponding address: Arthur Lewbel, Dept of Economics, Boston College, 140 Commonwealth Ave., Chestnut Hill, MA, 02467, USA. (617)-552-3678, lewbel@bc.edu, https://www2.bc.edu/arthur-lewbel/

1 Introduction

Econometric identification really means just one thing: model parameters or features being uniquely determined from the observable population that generates the data.[1] Yet well over two dozen different terms for identification now appear in the econometrics literature. The goal of this survey is to summarize (identify) and categorize this zooful of different terms associated with identification. This includes providing a new, more general definition of identification that unifies and encompasses previously existing definitions.

This survey then discusses the differences between identification in traditional structural models vs. the so-called reduced form (or causal inference, or treatment effects, or program evaluation) literature. Other topics include set vs. point identification, limited forms of identification such as local and generic identification, and identification concepts that relate to statistical inference, such as weak identification, irregular identification, and identification at infinity. Concepts that are closely related to identification, including normalizations, coherence, and completeness are also discussed.

The mathematics in this survey is kept relatively simple, with a little more formality provided in the Appendix. Each section can be read largely independently of the others, with only a handful of concepts carried over from one section of the zoo to the next.

The many terms for identification that appear in the econometrics literature include (in alphabetical order): Bayesian identification, causal identification, essential identification, eventual identification, exact identification, first order identification, frequentist identification, generic identification, global identification, identification arrangement, identification at infinity, identification by construction, identification of bounds, ill-posed identification, irregular identification, local identification, nearly-weak identification, nonparametric identification, non-robust identification, nonstandard weak identification, overidentification, parametric identification, partial identification, point identification, sampling identification, semiparametric identification, semi-strong identification, set identification, strong identification, structural identification, thin-set identification, underidentification, and weak identification. This survey gives the

[1] The first two sections of this survey use identification in the traditional sense of what would now be more precisely called "point identification." See section 3 for details.

meaning of each, and shows how they relate to each other.

Let denote an unknown parameter, or a set of unknown parameters (vectors and/or functions) that we would like to learn about, and ideally, estimate. Examples of what could include are objects like regressor coefficients, or average treatment effects, or error distributions. Identification deals with characterizing what could potentially or conceivably be learned about parameters from observable data. Roughly, identification asks, if we knew the population that data are drawn from, would be known? And if not, what could be learned about ?

The study of identification logically precedes estimation, inference, and testing. For to be identified, alternative values of must imply different distributions of the observable data (see, e.g., Matzkin 2013). This implies that if is not identified, then we cannot hope to find a consistent estimator for . More generally, identification failures complicate statistical analyses of models, so recognizing lack of identification, and searching for restrictions that suffice to attain identification, are fundamentally important problems in econometric modeling.

The next section, Section 2, begins by providing some historical background. The basic notion of identification (uniquely recovering model parameters from the observable population), is now known as "point identification." Section 3 summarizes the basic idea of point identification. A few somewhat different characterizations of point identification appear in the literature, varying in what is assumed to be observable and in the nature of the parameters to be identified. In Section 3 (and in an Appendix), this survey proposes a new definition of point identification (and of related concepts like structures and observational equivalence) that encompasses these alternative characterizations or classes of point identified models that currently appear in the literature.

Section 3 then provides examples of, and methods for obtaining, point identification. This section

need the notion of "ceteris paribus," that is, holding other things equal. The formal application of this concept to economic analysis is generally attributed to Alfred Marshall (1890). However, Persky (1990) points out that usage of the term ceteris paribus in an economic context goes back to William Petty (1662).[2]

The textbook example of an identification problem in economics, that of separating supply and demand curves, appears to have been first recognized by Philip Wright (1915), who pointed out that what appeared to be an upward sloping demand curve for pig iron was actually a supply curve, traced out by a moving demand curve. Philip's son, Sewall, invented the use of causal path diagrams in statistics.

A different identification problem is that of identifying the true coefficient in a linear regression when regressors are measured with error. Robert J. Adcock (1877, 1878), and Charles H. Kummell (1879) considered measurement errors in a Deming regression, as popularized in W. Edwards Deming (1943)[5]. This is a regression that minimizes the sum of squares of errors measured perpendicular to the fitted line. Corrado Gini (1921) gave an example of an estimator that deals with measurement errors in standard linear regression, but Ragnar A. K. Frisch (1934) was the first to discuss the issue in a way that would now be recognized as identification. Other early papers looking at measurement errors in regression include Neyman (1937), Wald (1940), Koopmans (1937), Reiersøl (1945, 1950), Roy C. Geary (1948), and James Durbin (1954). Tamer (2010) credits Frisch (1934) as also being the first in the literature to describe an example of set identification.

3 Point Identification

In modern terminology, the standard notion of identification is formally called point identification. Depending on context, point identification may also be called global identification or frequentist identification. When one simply says that a parameter or a function is identified, what is usually meant is that it is point identified.

Early formal definitions of (point) identification were provided by Koopmans and Reiersøl (1950), Hurwicz (1950), Fisher (1966) and Rothenberg (1971). These include the related concepts of a structure and of observational equivalence. See Chesher (2008) for additional historical details on these classical identification concepts.

In this survey I provide a new general definition of identification. This generalization maintains the intuition of existing classical definitions while encompassing a larger class of models than previous definitions. The discussion in the text below will be somewhat informal for ease of reading. More rigorous definitions are given in the Appendix.

[5]Adcock's publications give his name as R. J. Adcock. I only have circumstantial evidence that his name was actually Robert.

3.1 Introduction to Point Identification

Recall that is the parameter (which could include vectors and functions) that we want to identify and ultimately estimate. We start by assuming there is some information, call it , that we either already know or could learn from data. Think of as everything that could be learned about the population that data are drawn from. Usually, would either be a distribution function, or some features of distributions like conditional means, quantiles, autocovariances, or regression coefficients. In short, is what would be knowable from unlimited amounts of whatever type of data we have. The key difference between the definition of identification given in this survey and previous definitions in the literature is that previous definitions generally started with a particular assumption (sometimes only implicit) of what constitutes (examples are the Wright-Cowles identification and Distribution Based identification discussed in Section 3.3).

Assume also that we have a model, which typically imposes some restrictions on the possible values could take on. A simple definition of (point) identification is then that a parameter is point identified if, given the model, is uniquely determined from .

For example, suppose for scalars Y, X, and , our model is that $Y = X + e$ where $E[X^2] > 0$ and $E(eX) = 0$, and suppose that , what we can learn from data, includes the second moments of the vector (Y, X). Then we can conclude that is point identified, because it is uniquely determined in the usual linear regression way by $= E(XY)/E[X^2]$, which is a function of second moments of (Y, X).

Another example is to let the model be that a binary treatment indicator X is assigned to individuals by a coin flip, and Y is each individual's outcome. Suppose we can observe realizations of (X, Y) that are independent across individuals. We might therefore assume that , what we can learn from data, includes $E(Y \mid X)$. It then follows that the average treatment effect is identified because, when treatment is randomly assigned, $= E(Y \mid X = 1) - E(Y \mid X = 0)$, that is, the difference between the mean of Y among people who have $X = 1$ (the treated) and the mean of Y among people who have $X = 0$ (the untreated).

Both of the above examples assume that expectations of observed variables are knowable, and so can

But in this case $F(Y \mid X)$ can only be known for the values of X that can be chosen in the experiment (e.g., it may be impossible to run the experiment at a temperature X

regarding the collection of data, e.g., selection, measurement errors, and survey attrition. The other is assumptions regarding the generation of data, e.g., randomization or statistical and behavioral assumptions. Arellano (2003) refers to a set of behavioral assumptions that suffice for identification as an *identification arrangement*. Ultimately, both types of assumptions determine what we know about the model and the DGP, and hence determine what identification is possible.

3.2 Defining Point Identification

Here we define point identification and some related terms, including structure and observational equivalence. The definitions provided here generalize and encompass most previous definitions provided in the literature. The framework here most closely corresponds to Matzkin (2007, 2012). Her framework is essentially the special case of the definitions provided here in which is a distribution function. In contrast, the traditional textbook discussion of identification of linear supply and demand curves corresponds to the special case where is a set of limiting values of linear regression coefficients. The relationship of the definitions provided here to other definitions in the literature, such as those given by the Cowles foundation work, or in Rothenberg (1971), Sargan (1983), Hsiao (1983), or Newey and McFadden (1994), are discussed below. In this section, the provided definitions will still be somewhat informal, stressing the underlying ideas and intuition. More formal and detailed definitions are provided in the Appendix.

Define a *model* M to be a set of functions or constants that satisfy some given restrictions. Examples of what might be included in a model are regression functions, error distribution functions, utility functions, game payoff matrices, and coefficient vectors. Examples of restrictions could include assuming regression functions are linear or monotonic or differentiable, or that errors are normal or fat tailed, or that parameters are bounded.

Define a model value m to be one particular possible value of the functions or constants that comprise M. Each m implies a particular DGP (data generating process). An exception is incoherent models (see Section 4), which may have model values that do not correspond to any possible DGP.

Define to be a set of constants and/or functions about the DGP that we assume are known, or

knowable from data. Common examples of might be data distribution functions, conditional mean functions, linear regression coefficients, or time series autocovariances.

Define a set of *parameters* to be a set of unknown constants and/or functions that characterize or summarize relevant features of a model. Essentially, can be anything we might want to estimate. Parameters could include what we usually think of as model parameters, such as regression coefficients, but could also be, e.g., the sign of an elasticity, or an average treatment effect.

The set of parameters may also include *nuisance* parameters, which are defined as parameters that are not of direct economic interest, but may be required for identification and estimation of other objects that are of interest. For example, in a linear regression model might include not only the regression coefficients, but also the marginal distribution function of identically distributed errors. Depending on context, this distribution might not be of direct interest and would then be considered a nuisance parameter. It is not necessary that nuisance parameters, if present, be included in , but they could be.

We assume that each particular value of *m* implies a particular value of and of (violations of this assumption can lead to incoherence or incompleteness, as discussed in a later section). However, there could be many values of *m* that imply the same or the same . Define the *structure* $s(\cdot,\cdot)$ to be the set of all model values *m* that yield both the given values of and of .

Two parameter values and θ are defined to be *observationally equivalent* if there exists a such that both $s(\cdot,\cdot)$ and $s(\cdot,\theta)$ are not empty. Roughly, and θ observationally equivalent means there exists a value such that, if is true, then either the value or θ could also be true. Equivalently, and θ being observationally equivalent means that there exists a and model values *m* and m_θ such that model value *m* yields the values and , and model value m_θ yields the values and θ.

We're now ready to define identification. The parameter is defined to be *point identified* (often just called *identified*) if there do not exist any pairs of possible values and θ that are different but observationally equivalent.

Let Θ denote the set of all possible values that the model says could be. One of these values is the unknown true value of , which we denote as θ_0. We say that the particular value θ_0 is point identified

if $_0$ not observationally equivalent to any other in $\mathbf{2}$. However, we don't know which of the possible values of (that is, which of the elements of $\mathbf{2}$) is the true $_0$. This is why, to ensure point identification, we generally require that no two elements and e in the set $\mathbf{2}$ having $\mathbf{6}\ e$ be observationally equivalent. Sometimes this condition is called *global identification* rather than point identification, to explicitly say that $_0$ is point identified no matter what value in $\mathbf{2}$ turns out to be $_0$.

We have now defined what it means to have parameters be point identified. We say that the *model is point identified* when no pairs of model values m and m in M are observationally equivalent (treating m and m as if they were parameters). Since every model value is associated with at most one value of , having the model be identified is sufficient, but stronger than necessary, to also have any possible set of parameters be identified.

The economist or econometrician defines the model M, so we could in theory enumerate every $m \in M$, list every value of and that is implied by each m, and thereby check every pair $s\ ;\ /$ and $s\ ;e$ to see if is point identified or not. The difficulty of proving identification in practice is in finding tractable ways to accomplish this enumeration. Note that since we do not know which value of is the true one, proving identification in practice requires showing that the definition holds for any possible , not just the true value.

We conclude this section by defining some identification concepts closely related to point identification. Later sections will explore these identification concepts in more detail.

The concepts of local and [(.)25hone,

fo6ill values

those that are observationally equivalent to some other element of Θ, and those that are not. If θ_0 is in the second group, then it's identified, otherwise it's not. Since θ_0 could be any value in Θ, and we don't know which one, to prove point identification in general we would need to show that the first group is empty. The parameter is defined to be *generically identified* if the first group is extremely small (formally, if

.

How does this example fit the general definition of identification? Here each value of is a particular continuous, monotonically increasing distribution function F. In this example, each model value m happens to correspond to a unique value of , because each possible distribution of W has a unique distribution function. In this example, for any given candidate value of and , the structure $s(\ ;\ )$ is either an empty set or it has one element. For a given value of and , if $= F$ and $F(\ ) = 1/2$ (the definition of a median) then set $s(\ ;\ )$ contains one element. That element m is the distribution that has distribution function F. Otherwise, if $= F$ where $F(\ ) \neq 1/2$, the set $s(\ ;\ )$ is empty. In this example, it's not possible to have two different parameter values and e be observationally equivalent, because $F(\ ) = 1/2$ and $F^{e} = 1/2$ implies $= e$ for any continuous, monotonically increasing function F. Therefore is point identified, because its true value $_0$ cannot be observationally equivalent to any other value .

Example 2: Linear regression. Consider a DGP consisting of observations of $Y; X$ where Y is a scalar, X is a K vector. The observations of Y and X might not be independent or identically distributed. Assume the first and second moments of X and Y are constant across observations, and let be the set of first and second moments of X and Y. Let the model M be the set of joint distributions of $e; X$ that satisfy $Y = X^{0} + e$, where is some K vector of parameters, e is an error term, $E(Xe) = 0$ for an error term e, and where $e; X$

define and . For example, suppose we had IID observations of $Y; X$. We could then have defined to be the joint distribution function of $Y; X$, and defined to include both the coefficients of X and the distribution function of the error term e. Given the same model M, including the restriction that $E\ XX^{0}$ is nonsingular, we would then have semiparametric identification of .

Example 3: treatment. Suppose the DGP consists of individuals who are assigned a treatment of $T = 0$ or $T = 1$, and each individual generates an observed outcome Y. Assume $Y; T$ are independent across individuals. In the Rubin (1974) causal notation, define the random variable $Y(t)$ to be the outcome an individual would have generated if he or she were assigned $T = t$. The observed Y satisfies $Y = Y(T)$. Let the parameter of interest be the average treatment effect (ATE), defined by $= E(Y(1) - Y(0))$. The model M is the set of all possible joint distributions of $Y(1)$, $Y(0)$, and T. One possible restriction on the model is Rosenbaum and Rubin's (1983) assumption that $(Y(1); Y(0))$ is independent of T. This assumption, equivalent to random assignment of treatment, is what Rubin (1990) calls unconfoundedness. Imposing unconfoundedness means that M only contains model values m (i.e., joint distributions) where $(Y(1); Y(0))$ is independent of T.

The knowable function from this DGP is the joint distribution of Y and T. Given unconfoundedness, is identified because unconfoundedness implies that $= E(Y \mid T = 1) - E(Y \mid T = 0)$, which is uniquely determined from . Heckman, Ichimura, and Todd (1998) note that a weaker sufficient condition for identification of by this formula is the mean unconfoundedness assumption that $E(Y(t) \mid T) = E(Y(t))$. If we had not assumed some form of unconfoundedness, then might not equal $E(Y \mid T = 1) - E(Y \mid T = 0)$. More relevantly for identification, without unconfoundedness, there could exist different joint distributions of $Y(1)$, $Y(0)$, and T (i.e., different model values m) that yield the same joint distribution , but have different values for . Those different values would then be observationally equivalent to each other, and so we would not have point identification.

The key point for identification is not whether we can write a closed form expression like $E(Y \mid T = 1) - E(Y \mid T = 0)$ for , but whether there exists a unique value of corresponding to every possible .

These constructions can all be generalized to 0.00(ut).9552 T]TJ/Fr 7.iTd [o0

generally be defined as the assumption that $(Y(1), Y(0))$ is independent of T conditional on a set of observed covariates X. This also corresponds to Heckman and Robb's (1985) selection on observables assumption. In this case identification requires an additional condition, called the overlap condition, which says that for every value x that X can take on, we can observe individuals who have $T = 1$ and $X = x$, and other individuals who have $T = 0$ and $X = x$. This implies that we can identify both $E(Y \mid T = 1, X = x)$ and $E(Y \mid T = 1, X = x)$, and the average treatment effect is then identified by $= E(E(Y \mid T = 1, X) - E(Y \mid T = 0, X))$, where the outer expectation is over all the values that X can take on.

Example 4: linear supply and demand. Consider the textbook example of linear supply and demand curves. Assume we have, for each time period, a demand equation $Y = bX + cZ + U$ and a supply equation $Y = aX + \varepsilon$, where Y is quantity, X is price, Z is income, and U and ε are mean zero errors, independent of Z. Each model value m could consist of a particular joint distribution of Z, U, and ε in every time period. Note that these distributions could change over time. Different values of coefficients a, b, and c, and different distributions of Z, U, and ε for all time periods correspond to different model values m. The model M is the set of all possible model values m that satisfy the assumptions. Here could be defined as the vector (π_1, π_2) of reduced form coefficients $Y = \pi_1 Z + V_1$ and $X = \pi_2 Z + V_2$ where V_1 and V_2 are mean zero, independent of Z. Suppose $= a$, meaning that what we want to identify is the coefficient of price in the supply equation. Solving for the reduced form coefficients we have that $\pi_1 = ac/(a - b)$ and $\pi_2 = c/(a - b)$.

In this example, what model values m comprise a given structure $s(\ ,\ )$? Along with distributions, each m includes a particular value of a, b, and c. So the model values m that are in a given structure $s(\ ,\ )$ are the ones that satisfy the equations $= a$, $\pi_1 = ac/(a - b)$, and $\pi_2 = c/(a - b)$. Note in particular that, if $c \neq 0$, then $\pi_1/\pi_2 = a$, so $s(\ ,\ )$ is empty if $c \neq 0$ and $\pi_1/\pi_2 \neq$. Whenever $s(\ ,\ )$ is not empty, it contains many elements m, because there are many different possible distributions of Z, U, and ε for the given value of and .

Without additional assumptions, is not identified in this example. This is because any two values

and e will be observationally equivalent when $\quad D .0; 0/$. The more familiar way of saying the same thing is that, for identification of the supply equation, we need the instrument Z to appear in the demand equation, and therefore we need $c \neq 0$, which implies that $\quad_2 \neq 0$. If we include in the definition of the model that $c \neq 0$, then $\quad$ is identified.

Example 5: latent error distribution. Suppose the DGP is IID observations of scalar random variables $Y; X$, so $\quad$ is the joint distribution of $Y; X$. The model M is the set of joint distributions of $X; U$ satisfying the restrictions that X is continuously distributed, $U \perp X$ (meaning U is independent of X), and that Y

regression coefficients. These could, e.g., be the probability limits of regression coefficients estimated using time series, cross section, or panel data.

Many identification arguments in econometrics begin with one of three cases: Either is a set of reduced form regression coefficients, or is a data distribution, or is the maximizer of some function. These starting points are sufficiently common that they deserve names, so I will call these classes Wright-Cowles identification, distribution based identification, and extremum based identification.

Wright-Cowles Identification: The notion of identification most closely associated with the Cowles foundation concerns the simultaneity of linear systems of equations like supply and demand equations. This is the same problem considered earlier by Philip and Sewall Wright, as discussed in the previous section, so call this concept Wright-Cowles identification. Let Y be a vector of endogenous variables, and let X be a vector of exogenous variables (regressors and instruments). Define to be a matrix of population reduced form linear regression coefficients, that is, denotes a possible value of the set of

here because there could be more than one reduced form matrix value that is consistent with the true $_0$. For example, if $_0$ is just the coefficient of price in the supply equation, there could be many possible reduced form coefficient matrices , corresponding to different possible values of all the other coefficients in the structural model.

A convenient feature of Wright-Cowles identification is that it can be applied to time series, panel, or other DGP's with dependence across observations, as long as the reduced form linear regression coefficients have some well defined limiting value .

Identification of linear models can sometimes be attained by combining exclusions or other restrictions on the matrix of structural coefficients with restrictions on the covariance matrix of the errors. In this case we could expand the definition of to include both the matrix of reduced form coefficients and the covariance matrix of the reduced form error terms, conditional on covariates. Now we're assuming more information (specifically, the error covariance matrix) is knowable, and so the structure s can now include restrictions not just on coefficients, but also on error covariances. More generally the structure could have all kinds of restrictions on the first and second moments of Y given X. In models like these, identification is sometimes possible even without instruments of the type provided by exclusion restrictions. Examples include the LISREL model of Jöreskog (1970) and the heteroskedasticity based identification of Lewbel (2012, 2018).

Distribution Based Identification: Distribution based identification is equivalent to the general definition of identification given by Matzkin (200703 Td6s

cieHurwicz66(E7)01970iFis E7 E7en8(g15(E7)019709istTes)662(Te,)-15(E7)auutE7 of E7E7 ervations, 21uivalent

metric restrictions on the knowable distribution function, like continuity or existence of moments. Distribution based identification is suitable for IID data, where would be knowable by the Glivenko-Cantelli theorem, or may apply to some non-IID DGP's where the distribution is sufficiently parameterized.

Here could be parameters of a parameterized distribution function, or features of the distribution like moments or quantiles, including possibly functions like conditional moments. Alternatively, could consist of constants or functions describing some behavioral or treatment model that is assumed to generate data drawn from the distribution . The structure *s . ; /* will be an empty set if the given distribution function doesn't have the features or parameter values . Two vectors and e are observationally equivalent if there's a distribution function that can imply values or e. So is point identified if it's uniquely determined from knowing the distribution function .

Note a key difference between Wright-Cowles and distribution based identification and is that the latter assumes an entire distribution function is knowable, while the former is based on just having features of the first and second moments of data be knowable.

Extremum Based Identification

density or mass function of W_i, then this could be a maximum likelihood estimator. Now define G by $G(\theta) = E[g(W_i, \theta)]$. More generally, $G(\theta)$ could be the probability limit of the objective function of a given extremum estimator. The parameter θ is point identified if, for every value of G allowed by the model, there's only a single value of θ that corresponds to any of the values of θ that maximize $G(\theta)$.

Suppose G is, as above, the probability limit of the objective function of a given extremum estimator. A standard assumption for proving consistency of extremum estimators is to assume $G(\theta)$ has a unique maximum θ_0, and that θ_0 equals a known function of (or subset of) θ_0. See, e.g., Section 2 of Newey and McFadden (1994). This is a sufficient condition for extremum based identification.

For linear models, Wright-Cowles identification can generally be rewritten as an example of, or a special case of, extremum based identification, by defining G to be an appropriate least squares objective function. In parametric models, distribution based identification can also often be recast as extremum based identification, by defining the objective function G to be a likelihood function.

Extremum based identification can be particularly convenient for contexts like time series or panel data models, where the distribution of data may change with every observation, or for social interactions models where there is complicated dependence across observations. When the DGP is complicated, it may be difficult to define everything that is knowable about the DGP, but feasible to show that the maximizing values of a given objective function G

model consists of G functions defined by $G(\cdot/\mathsf{D} \quad E \cdot W \quad \mathrm{j}\ \mathrm{j}/^2$. The set of arguments that maximize this G

tion, as in example 5 where we had $\mathsf{D}\ F_U$.

be known, and then ask whether there exists a unique value of vectors or functions that satisfy the restrictions defined by the structure $s(\cdot;\theta)$. Some results in economics do take this form, for example, the revealed preference theory of Samuelson (1938, 1948), Houthakker (1950), and Mas-Colell (1978) provides conditions under which indifference curves are point identified from demand functions. Here the model is the set of restrictions on demand functions (e.g., homogeneity and Slutsky symmetry) that arise from maximization of a regular utility function under a linear budget constraint. This identification theorem makes no direct reference to data, though it is empirically relevant because we believe we can estimate (and hence can identify) demand functions from observable data.

This example illustrates the sense mentioned earlier in which definitions of identification are somewhat circular or recursive. We start by assuming that a set of vectors or functions are 'knowable,' which essentially means we assume is identified. Then given the assumed identification of , we define identification of a set of other parameters . Equivalently, identification of parameters can only be proven conditional on already knowing that something else, , is identified. For example, the Samuelson, Houthakker, and Mas-Colell theorems say that, given revealed preference assumptions, if demand functions are identified, then indifference curves are identified. A separate question would then be when or whether demand functions themselves can be identified.

3.5 Common Reasons for Failure of Point identification

Parameters often fail to be point identified for one of six somewhat overlapping reasons: model incompleteness, perfect collinearity, nonlinearity, simultaneity, endogeneity, or unobservability.

Incompleteness arises in models where the relationships among variables are not fully specified. An example is games having multiple equilibria, where the equilibrium selection rule is not known or specified. Incompleteness can also arise in structures characterized at least in part by inequality constraints. It is sometimes possible for parameters to be point identified in incomplete models, but often incompleteness causes, or at least contributes to, failure of identification. Incompleteness is discussed further in Section 4.

Perfect collinearity is the familiar problem in linear regression that one cannot separately identify

Randomization is a useful source of identification, primarily because it prevents simultaneity. It can't be the case that Y and X are determined jointly if X is determined by a random process that is independent of Y.

Endogeneity is the general problem of regressors being correlated with errors. Simultaneity is one source of endogeneity, but endogeneity can arise in other ways as well. Sampling issues such as measurement errors and selection can cause endogeneity. Even when a regressor X is determined by a coin flip, if some people are not observed, or are observed with error, in ways that correlate with X, then we could end up having an endogeneity problem. Endogeneity can also arise when errors that correspond to unobserved covariates may correlate with observables, just as observables often correlate with each other. For example, the error in a production function may correspond to an unobserved factor of production such as entrepreneurship, and may therefore correlate with other factors of production. Or the error in a wage equation may correspond to an individual's ability or drive, and so correlate with other factors that determine wages, like education. In the causal diagrams literature, colliders are endogenous covariates.

The last common source of nonidentification is unobservability. Many models contain unobserved heterogeneity, which typically take the form of nonadditive or nonseparable error terms. Examples are unobserved random utility parameters in consumer demand models, unobserved state variables in dynamic optimization models, and unobserved production efficiency as in stochastic frontier models. The causal or reduced form literature is often concerned with unobservable counterfactuals, e.g., what an untreated individual's outcome would have been had they been treated. In structural models, many of the concepts we would like to estimate, such as an individual's utility level, are unobservable.

Still other concepts may in theory be observable, but are difficult to measure, and so in practice are

reduced form, causal analysis like LATE or difference-in-difference.[6]

The second, more important reason is that Z itself could be endogenous, and the problems resulting from adding an endogenous Z regressor to the model could be worse than the confounding issue. For example, consider a regression of wages Y on a gender dummy X and other covariates to uncover a causal effect of gender on wages (as might result from wage discrimination). The endogenous

3.7 Identification by Functional Form

instruments, in this model the parameters in both the supply and demand equations can be identified!

Substituting out

Historically, identification by functional form assumed completely parameterized models with no unknown functions. However, that is often much stronger than needed for identification. For example, suppose, as in the previous supply and demand example, we have demand given by $Y = bX + cZ + U$ where X is endogenous, and so is correlated with U. In that example, if Z is independent of U, then any nonlinear function of Z would be a valid instrument for X in the sense of being uncorrelated with U.

a and b are identified without any instruments or other outside information as long as either some error or the true X is not normal. So the standard assumption of normality turns out to be the worst possible functional form for identification with measurement error. Lewbel (1997b) shows that, in this model, if the measurement error is symmetrically distributed and the true X is asymmetric, then $X \quad \overline{X}^{\,2}$ is a valid instrument for X. Schennach and Hu (2013) show that the Reiersøl result can be extended to obtain identification of $Y \ \mathsf{D} \ g$.

functional form restrictions can be used to test validity of a potential "true" instrument. For example, in the linear model $Y = a + bX + cZ + U$, where Z is exogenous, we may have some outside variable W that we think is a valid instrument for X. We could estimate the model by two stage least squares, using a constant, W, Z, and Lewbel's (2012) heteroskedasticity based constructed variable R defined above as instruments. With both W and R as instruments for X, the model is overidentified (see the next section for details on ove-ridentification), so one can test jointly for validity of all the instruments, using e.g., a Sargan (1958) and Hansen (1982) J-test. If validity is rejected, then either the model is misspecified or at least one of these instruments is invalid. If validity is not rejected, it is still possible that the model is wrong or the instruments are invalid, but one would at least have increased confidence in both the outside instrument

and

der conditions arising from extremum estimators, such as the score functions associated with maximum likelihood estimation.

Suppose the model consists mainly of a set of equalities like these. We then say that parameters are *exactly identified* if removing any one these equalities causes to no longer be point identified. The parameters are *overidentified* when can still be point identified after removing one or more of the equalities, and they are *underidentified* when we do not have enough equalities to point identify .

If is a J vector, then it will typically take J equations of the form $E\ g.W; / \ \mathsf{D}\ 0$ to exactly identify . Having the number of equations equal or exceed the number of unknowns is called the *order condition* for identification. The order condition is typically necessary ody for point ident-r ficatios, tnoughien mney ppli2catioes, the order(condition)-205(m(y)-205sufo)25ifier(for)-205gendercr idenH en7n.

bleg tovor a vector

4 Coherence, Completeness, and Reduced Forms

Although often ignored in practice, consideration of coherence and completeness of models should logically precede the study of identification. Indeed, most proofs of point identification either implicitly or explicitly assume the model has a unique reduced form, and therefore (as discussed below) assume both coherence and completeness. For example, the models considered in Matzkin's (2005, 2007, 2012) identification surveys are coherent and complete. In contrast, incompleteness often results in parameters being set identified but not point identified.

Let Y be a vector of endogenous variables, and let V be a set of observables and unobservables that determine Y. Here V could contain unknown parameters, exogenous observed covariates and error terms. Let $\bullet_{\boldsymbol{v}}$ and $\bullet_y$ be the sets of all values that V and Y can take on, respectively. Consider a proposed model M of the form $Y = H(Y, V)$. By saying this equation is the model M, what is meant that each model value $m \in M$ implies a DGP in which V and Y satisfy this equation.

This model is defined to be *coherent* if for each $\boldsymbol{v} \in \bullet_{\boldsymbol{v}}$ there exists a $y \in \bullet_y$ that satisfies the equation $y = H(y, \boldsymbol{v})$. The model is defined to be *complete* if for each $\boldsymbol{v} \in \bullet_{\boldsymbol{v}}$ there exists at most one value of $y \in \bullet_y$ that satisfies the equation $y = H(y, \boldsymbol{v})$. A *reduced form* of the model is defined as a function (or mapping) G such that $y = G(\boldsymbol{v})$, so a reduced form expresses the models' endogenous variables Y in terms of V. Having both coherence and completeness means that for each $\boldsymbol{v} \in \bullet_{\boldsymbol{v}}$ there exists a unique $y \in \bullet_y$ that satisfies $y = H(y, \boldsymbol{v})$. Having a model be both coherent and complete therefore guarantees the existence of a unique reduced form $y = G(\boldsymbol{v})$ for the model, because then G can be uniquely defined by $G(\boldsymbol{v}) = H(G(\boldsymbol{v}), \boldsymbol{v})$.

This definition of completeness and coherence is used by Tamer (2003). Completeness as defined here is an extension of the concept of statistical completeness. Statistical completeness is discussed in Newey and Powell (2003) for identification of nonparametric IV models, and in parametric models is associated with sufficient statistics. Gourieroux, Laffont, and Monfort (1980) defined a model to be coherent if, in Tamer's terminology, the model is both coherent and complete. Heckman (1978) referred to this combination of both coherence and completeness as the "principal assumption" and as "conditions for existence

of the model."

Incoherent or incomplete models arise in some simultaneous games, e.g., based on Tamer (2003), the industry entry game discussed by Bresnahan and Reiss (1991) can be incoherent if the game has no Nash equilibrium, or incomplete if there are multiple equilibria. Aradillas-Lopez (2010) removes the incompleteness in these games by showing how a unique Nash equilibrium exists when players each possess some private information.

Entry games are an example of a system of equations involving discrete endogenous variables. More generally, issues of incoherency and incompleteness can readily arise in simultaneous systems of equations involving limited dependent variables. Examples are analyzed by Blundell and Smith (1994), Dagenais (1997), and Lewbel (2007a). To illustrate, consider the simple model

$$Y_1 = I(Y_2 + U_1 \geq 0)$$

$$Y_2 = Y_1 + U_2$$

where is a coefficient, U_1 and U_2 are unobserved error terms, $V = (U_1, U_2)$, $Y = (Y_1, Y_2)$, and I is the indicator function that equals one if its argument is true and zero otherwise. These equations could for example be the reaction functions of two players in some game, where player one makes a binary choice Y_1 (such as whether to enter a market or not), and player two makes some continuous decision Y_2 (such as the quantity to produce of a good).

It is not obvious that this simple model could suffer from incoherence or incompleteness, and so a researcher who is not familar with these issues could easily make the mistake of attempting to estimate this model by standard methods (e.g., maximum likelihood assuming U_1 and U_2 are normal).

Subtstituting the second equation into the first gives $Y_1 = I(Y_1 + U_1 + U_2 \geq 0)$. Using this equation one can readily check that if $U_1 + U_2 < 0$ then both $Y_1 = 0$ and $Y_1 = 1$ satisfy the model, and therefore the model is incomplete if the errors can satisfy this inequality. Also, if $0 \leq U_1 + U_2 <$ then neither $Y_1 = 0$ nor $Y_1 = 1$ will satisfy the model, making the model incoherent. This model is both coherent and complete if and only if $= 0$ or $U_1 + U_2$ is constrained to not lie between zero and .

The above system of equations is simultaneous, in that Y_1 is a function of Y_2 and Y_2 is a function

of Y_1. A pair of equations is said to be triangular if either Y_1 is a function of Y_2 or Y_2 is a function of Y_1, but not both. For example, the model above is triangular if $\delta \neq 0$ (since in that case Y_2 depends on Y_1, but not vice versa), and in that case the model is also coherent and complete. In fact, if U_1 and U_2 can take on any value (e.g., if they were normal), then the model is coherent and complete *only* if $\delta \neq 0$. Lewbel (2007a) shows this type of result is generic, i.e., that simultaneous systems of equations containing a dummy endogenous variable and separable errors generally need to either be triangular or to restrict the supports of the errors to be coherent and complete. It is possible to overcome this generic problem, constructing complete, coherent simultaneous systems containing dummy endogenous variables,

needs to be changed, while a finding of incompleteness means that the model may need to be completed. Determining an equilibrium selection rule is an example of completing a model that is otherwise incomplete due to having multiple equilibria.

Even without changing or completing the model, parameters of incoherent or incomplete models can sometimes be point identified and estimated. See Tamer (2003) for examples. However, incomplete models usually have parameters that are set rather than point identified, as in Manski and Tamer (2002). This is because, when multiple values of Y can correspond to each V, it will often be the case that the different values of Y will correspond to different values of .

Incompleteness or incoherency can arise in models with multiple decision makers, such as strategically interacting players in a game. Models of a single optimizing agent will typically be coherent though sometimes incomplete, such as when the same utility or profit level can be attained in more than one way. Incoherency or incompleteness can also arise in such models when the decision making process is either incorrectly or incompletely specified, or is not characterized by optimizing behavior. Equilibrium selection mechanisms or rules for tie breaking in optimization models can be interpreted as techniques for resolving incompleteness. Another common source of incompleteness is behavioral restrictions on structures that take the form of inequality rather than equality constraints, yielding multiple possible values of Y for the same V.

5 Causal Reduced Form vs. Structural Model Identification

Among economists doing empirical work, recent years have seen a rapid rise in the application of so-called reduced form or causal inference methods, usually based on randomization. This so-called "credibility revolution," as exemplified by, e.g., Angrist and Pischke (2008), Levitt and List (2009), and Banerjee and Duflo (2009), arose in economics long after the standard theory of identification was developed in the context of structural modeling. As a result, most surveys of identification in econometrics, such as Hsiao (1983) or Matzkin (2007, 2012), do not touch on identification as it is used in this literature.

Proponents of these methods often refer to their approach as a reduced form methodology. Other

commonly used terms for these methods include causal modeling, causal inference, treatment effects modeling, program evaluation, or mostly harmless econometrics.[7]

To distinguish them from structural model based methods, I will simply refer to these types of analyses as causal, or causal reduced form methods. Two key characteristics of causal methods are 1. A focus on identification and estimation of treatment effects rather than deep parameters, 2. An emphasis on natural or experimental randomization (rather than restrictions on how treatment may affect outcomes) as a key source of identification. However, many exceptions to these characterizations exist. For example, numerous structural analyses, like the famous Roy (1951) model, also seek to identify treatment effects. Some reduced form methods, like difference in difference estimation (see section 3.6), are not based on random assignment. Some literatures, such as Pearl (2000, 2009), focus on using minimal structural type assumptions (like causal diagrams) to aid in identifying causal effects. And a growing number of empirical structural analyses make use of data obtained from randomized control trial (RCT) experiments.

Despite these many exceptions, causal methods generally focus on identification and estimation of treatment effects based on random assignment, either of treatment itself from a RCT, or of some variable that correlates with treatment (i.e., a randomly assigned instrument). In the causal literature, instruments are defined as variables that one can plausibly argue are randomly determined, and that correlate with treatments of interest. A large part of the causal literature is devoted to designing and interpreting RCTs. These are particularly popular in, e.g., development economics. Much of the rest of the causal literature entails searching for and exploiting instruments (as from natural experiments) for identification.

Causal methods largely forego attempts to identify so-called structural or deep parameters, that is, parameters of models based on equations representing the behavior of various economic agents (such parameters are assumed to be unaffected by the treatment). Instead, causal analyses focus on identifying treatment effects. These are the average (across the population or across some subpopulation) of the

[7]Terms like reduced form modeling or causal modeling are potentially confusing, since "reduced form" has a specific meaning discussed earlier in the structural modeling context, and structural methods are also often employed to identify causal or treatment effects. Mostly Harmless Econometrics is the title of Angrist and Pischke's (2008) book promoting these approaches. The name is in turn based on a satirical science fiction novel, that humorously also features the phrase, "infinite improbability."

1. Causal analyses based on randomization can be augmented with structural econometric methods to deal with identification problems caused by data issues such as attrition, sample selection, measurement error, and contamination bias. For example, Conlon and Mortimer (2016) use a field experiment to estimate the causal effects of temporarily removing a popular brand from vending machines. They combine observed experimental outcomes with a simple structural model of purchase timing, to deal with the fact that purchase outcomes are only observed when the machines are serviced.

2. It is not just reduced form methods that require instrument independence. Identification in structural models also often depends on independence assumptions, and the use of randomization can increase confidence that required structural assumptions are satisfied. In short, good reduced form instruments are generally also good structural model instruments. An example is Ahlfeldt, Redding, Sturm, and Wolf (2015), which uses a natural experiment (the partition of Berlin) to identify a structural model of the economic gains associated with people living and working near each other in cities.[8]

3. Identifiable causal effects can provide useful benchmarks for structural models. For example, suppose we have a structural model with parameters that are identified by assumed behavioral restrictions. One might estimate these behavioral model parameters using data from large surveys, and then check whether treatment effects implied by the estimated structural parameters equal treatment effects that are identified and estimated using small randomized trial data sets drawn from the same underlying population. Another example is Andrews, Gentzkow and Shapiro (2017, 2018), who construct summary statistics based on the estimated joint distribution of reduced form parameters (like the moments used to estimate LATE) and structural model parameters. They use these statistics to assess the extent to which structural results depend on intuitively transparent identifying information.

4. Economic theory and structure can provide guidance regarding the external validity of causal parameters. For example, in a causal analysis one can't say how even a small change in treatment policy would change the resulting effects of treatment. Weak structural assumptions can overcome this limita-

[8]In awarding this paper the 2018 Frisch Medal, the Econometric Society's medal committee wrote that this paper, "provides an outstanding example of how to credibly and transparently use a quasi-experimental approach to structurally estimate model parameters."

tion. For example, in regression discontinuity designs the cutoff, i.e., the threshold discontinuity point, is often a relevant policy variable (such as the grade at which one qualifies for a scholarship). Dong and Lewbel (2015) show that, with a mild structural assumption called local policy invariance, one can identify how treatment effects estimated by regression discontinuity designs would change if the threshold were raised or lowered, even when no such change in the threshold is observed. Their estimator also provides a direct measure of the stability of regression discontinuity treatment effects (see Cerulli, Dong, Lewbel, and Poulsen 2017). Frölich and Huber (2017) use structural assumptions regarding a second instrument to separate direct from indirect effects of treatment on outcomes. Yet another example is Rosenzweig and Udry (2016), who use structure to model how average treatment effects (returns from policy interventions) estimated from randomized control trials, vary with macro shocks such as weather.

5. One can use causal methods to link randomized treatments to observable variables, and use structure to relate these observables to more policy relevant treatments and outcomes. For example, it has been documented that middle aged and older women in India have much higher mortality rates than would be expected, based on household income levels and the mortality rates of their spouses. Calvi (2016) uses a causal analysis to link changes in women's household bargaining power (stemming from a change in inheritance laws) to their health outcomes. One might then speculate that this established causal link between household power and health could explain the excess mortality rates of older women. But such speculation is nothing more or less than crude structural modeling. Instead of speculating, Calvi then constructs estimates of women's relative poverty rates based on structural models of their bargaining power, as defined by their consumption and control of household resources. She finds that these structurally estimated relative poverty rates can explain more than 90% of the women's higher than expected observed mortality rates by age. Most causal analyses include informal speculation regarding the wider implications of estimated treatment effects. More convincing than such informal discussions is formally establishing those connections and correlations with the rigor imposed by structural model identification and estimation.

Another, related example is Calvi, Lewbel, and Tommasi (2017). This paper estimates a LATE where treatment is defined as women's control over most resources within a household, and as above the out-

comes are family health measures, and the instrument is changes in inheritance laws. However, in this case the relevant treatment indicator cannot be directly observed, and so is estimated using a structural model of household behavior. Since structural models can be misspecified and have estimation errors, the estimated treatment indicator will be mismeasured for some households. The paper therefore proposes and applies an alternative estimator, called MR-LATE (mismeasurement robust LATE), that accounts for the potential measurement errors in observed treatment that may arise from misspecification or estimation errors in the structural model. In this example, the use of structure allows the application of LATE to identify a more policy relevant treatment effect than would otherwise be possible.

6. Big data analyses on large data sets can uncover promising correlations. Structural analyses of such data could then be used to uncover possible economic and behavioral mechanisms that underlie these correlations, while randomization might be used to verify the causal direction of these correlations. It is sometimes claimed that machine learning, natural experiments, and randomized controlled trials are replacing structural economic modeling. This is, if anything, backwards: as machine learning and experiments uncover ever more previously unknown correlations and connections, the desire to understand these newfound relationships will rise, leading to an increase, not a decrease, in the demand for structural economic theory and models.

7. Structural type assumptions can be used to clarify when and how causal effects may be identified. Examples are the structural causal models and causal diagrams, like directed acyclic graphs, summarized in Pearl (2000, 2009) and the more accessible Pearl and Mackenzie (2018). Another line of research that formally unifies structural and randomization based approaches to causal modeling is Vytlacil (2002), Heckman, Urzua and Vytlacil (2006), Heckman and Vytlacil (2007), and Heckman (2008, 2010).

5.2 Randomized Causal vs. Structural Identification: An Example

An obstacle to comparing causal vs. structural analyses is that these methods are usually described using different notations. So, to facilitate the comparison, a causal model's assumptions will here be rewritten completely in terms of the corresponding restrictions on a structural model, and vice versa. Both models

will be described in both notations.

Let Y be an observed outcome, let T be a binary endogenous regressor (think of T as indicating whether one receives a treatment or not), and let Z be a binary variable that is correlated with T, which we will use as an instrument. Assume includes the first and second moments of (Y, T, Z). In practice the DGP is such that these moments can be consistently estimated by sample averages.

The example structural model considered here will be the linear regression model $Y = a + bT + e$ for some error term e and constants a and b, under the standard instrumental variables identifying assumption that $E(eZ) = 0$. The corresponding causal model will be the local average treatment effect (LATE) model of Imbens and Angrist (1994). The key difference between these specific models is that, in the structural model, any heterogeneity of the impact of T on Y is assumed to be included in the error term e and hence is assumed to be uncorrelated with Z. This is a behavioral assumption, since it restricts the distribution of responses to treatment (i.e., behavior) in the population. The LATE model drops this behavioral restriction, replacing it with a randomized Z and a "no defiers" assumption (defined later), and instead identifies the average effect of T on Y for a subpopulation called compliers. While comparison of these models is not new (see, e.g., Imbens and Angrist 1994, Angrist, Imbens, and Rubin 1996, Imbens and Rubin 1997, Vytlacil 2002 and Heckman 1997, 2008, 2010), the goal here is to use the models to illustrate differences between identification in a popular structural and a popular causal model, both in terms of their assumptions and their notation.

What makes one model or analysis structural and another causal? As discussed earlier, structural models are generally assumed to have fixed deep or policy invariant parameters (like the coefficient b) and identifying restrictions (like $E(eZ) = 0$) that together summarize and place assumed limits on behavior. In short, structural models are generally models of economic behavior, ideally derived from and identified by economic theory.

In contrast, while it is impossible to avoid making some assumptions regarding behavior, causal models attempt to make as few such assumptions as possible. Causal models instead usually exploit randomization as the primary source of parameter identification (though some models that don't involve explicit

(1978). [10] Notwithstanding this long history, reduced form causal analyses often start with the counterfactual notation of Rubin (1974). In this notation, $Y(t)$ is defined as the random variable denoting the outcome Y that would occur if the treatment T equals t. Since $Y = U_0 + U_1 T$, it follows that $Y(0) = U_0$ (since that's what you get if you set $T = 0$) and $Y(1) = U_0 + U_1$. So both $Y(1)$ and $Y(0)$ are random variables. Note that observed Y satisfies $Y = Y(T)$.

In the same way, $T(z)$ denotes the random variable describing what one's treatment T would be if $Z = z$, so $T(0) = V_0$ and $T(1) = V_0 + V_1$. Note that since T and Z are both binary, we can without loss of generality say that V_0 and $V_0 + V_1$ are binary, that is, both V_0 and $V_0 + V_1$ can only equal zero or one.

Consistent with the above structural random coefficients model, this potential outcome notation assumes that Z

or SUTVA, which is the assumption any one person's outcome is unaffected by the treatment that other people receive. The term SUTVA was coined by Rubin (1980), but the concept goes back at least to Cox (1958), and indeed may be implicit in Splawa-Neyman (1923), Neyman, Iwaszkiewicz, and Kolodziejczyk (1935), and Fisher (1935). SUTVA is a strong behavioral assumption that essentially rules out social interactions, peer effects, network effects, and many kinds of general equilibrium effects. Although a goal of causal modeling is to make as few behavioral assumptions as possible, the behavioral SUTVA assumption is generally accepted in this literature, in part because it can be enforced in many purely experimental settings (by, e.g., physically separating experimental subjects until the experiment is over).

In contrast to laboratory settings, many natural or field experiments may (despite randomization) still violate SUTVA, due to the effects of people interacting with each other, either directly or via markets. When SUTVA is violated, most causal inference estimators become invalid, and point identification of causal effects becomes far more difficult to obtain. When SUTVA is violated one must typically make behavioral assumptions (i.e., the types of assumptions more commonly associated with structural models) to gain point identification, or construct more complicated experiments aimed at identifying the magnitude of spillover effects of some people's treatments to other's outcomes, or settle for set identification of causal effects. See Manski (2013), Lazzati (2015), Angelucci and Di Maro (2016), and Laffers and Mellace (2016) for examples of dealing with SUTVA violations by each of these methods. See also Rosenbaum (2007), who discusses inference when SUTVA is violated.

SUTVA can be interpreted as another type of exclusion restriction, in which $Y_i(0)$ and $Y_i(1)$, the potential outcomes for any given person i, are assumed to be independent of T_j for any other person j. In the structural model, SUTVA corresponds to a restriction on the causal correlations of $U_{1i}; U_{0i}$ (the outcome model random coefficients of person $; U_{0i}$)]TJ/F97 11.9552 Tf 5.758 0 Td [(U)]TJ/F58

We have now assumed some exclusions and some independence. Next, observe that the regressor T is, by construction, related to V_0 and V_1. The regressor T is also potentially endogenous, and so could be correlated with U_0 and U_1 as well. However, Z is supposed to be an instrument, so now let us add the causal assumption that $\{Y(1), Y(0), T(1), T(0)\}$ is independent of Z. This is equivalent in the structural notation to assuming that $\{U_1, U_0, V_0, V_1\}$ is independent of Z. This assumption would be justified by assuming random assignment of Z. This assumption, while corresponding to standard unconfoundedness in the causal literature, is stronger than would typically be assumed in the structural linear regression model. However, let us maintain this independence to facilitate comparison between the two approaches. One final assumption we maintain for both the structural and causal frameworks is that $E(V_1) \neq 0$, or equivalently in causal notation, that $E(T(1) - T(0)) \neq 0$. This assumption ensures that the instrument Z is relevant.

Now define the parameter c by $c = cov(Z, Y)/cov(Z, T)$, which is identified by construction. This c would be the limiting value of the estimated coefficient of T in a linear instrumental variables regression of Y on a constant and on T, using a constant and Z as instruments. We are going to consider the interpretation of this parameter under the assumptions of the causal LATE model and under the assumptions of our structural linear regression model. With just the assumptions we have so far the parameter c satisfies

$$c = \frac{cov(Z, Y)}{cov(Z, T)} = \frac{cov(Z, U_0 + U_1 T)}{cov(Z, (V_0 + V_1 Z))} = \frac{cov(Z, U_0 + U_1 (V_0 + V_1 Z))}{cov(Z, (V_0 + V_1 Z))}$$
$$= \frac{cov(Z, U_1 V_1 Z)}{cov(Z, V_1 Z)} = \frac{E(U_1 V_1) var(Z)}{E(V_1) var(Z)} = \frac{E(U_1 V_1)}{E(V_1)}.$$

The difference between our particular causal and structural models will consist only of different assumptions regarding the equation $c = E(U_1 V_1)/E(V_1)$.

structural instrument validity, i.e., $cov(e, Z) = 0$, actually mean? Under the above assumption that Z is independent of $\{U_1, U_0, V_0, V_1\}$, we have that $cov(e, Z) = cov(U_1, V_1)var(Z)$, so the instrument Z is valid in the structural sense if $cov(U_1, V_1) = 0$. Note from above that

$$c = \frac{E(U_1 V_1)}{E(V_1)} = \frac{E(U_1)E(V_1) + cov(U_1, V_1)}{E(V_1)} = E(U_1) + \frac{cov(U_1, V_1)}{E(V_1)}$$

so the structural restriction $cov(U_1, V_1) = 0$ makes $c = b = E(U_1)$. What does this structural coefficient b correspond to in causal notation? Recall the average treatment effect (ATE) is defined by $E[Y(1) - Y(0)]$. Now $E[Y(t)] = E[(U_0 + U_1 t)] = E(U_0) + E(U_1)t$ for $t = 0$ and for $t = 1$. Therefore $E[Y(1) - Y(0)] = E(U_1) = b$. So the structural coefficient b is precisely the causal ATE.

To summarize, first, the structural instrument validity assumption that $cov(e, Z) = 0$ is equivalent to $cov(U_1, V_1) = 0$. Second, under this assumption, the instrumental variables estimand c equals the structural linear regression coefficient b, which in turn equals the ATE. Again it should be emphasized that what has been provided here is just one possible set of structural modeling assumptions. For example, under milder assumptions, Heckman (1997) interprets the structural instrumental variable's regression coefficient as the ATT (the average treatment effect on the treated).

Now consider a causal identification argument. Define compliers to be individuals for whom T and Z are the same random variable, that is, a complier is an individual i who has $T_i = z$ when $Z_i = z$ for z being either zero or one. Similarly, let defiers denote people for whom T and $1 - Z$ are the same random variable. Recalling that $T = V_0 + V_1 Z$, anyone who has $V_1 = 1$ must also have $V_0 = 0$ (otherwise T would equal two which is impossible). It follows that compliers are exactly the people who have $V_1 = 1$. Imbens and Angrist (1994) define the local average treatment effect (LATE) to be the average treatment effect among compliers. In our notation, LATE is defined to equal $E[Y(1) - Y(0) \mid V_1 = 1]$. Note that these authors use the word 'local' in LATE to mean identification of the ATE for just a subset of people (the compliers). This is quite different from how the word local was used in the previous identification literature (see Section 6 below).

With these definitions in mind, consider again the equation for c. From the equation $T = V_0 + V_1 Z$, we noted that compliers are individuals who have $V_1 = 1$, and these people must also have $V_0 = 0$,

structural and causal approaches require assumptions regarding unobservables (not just on V_1 or U_1, but also assumptions like SUTVA), and a priori neither method's identifying assumptions are more or less plausible or restrictive than the other.

It is important to recall that these particular assumptions and models are not universal or required features of causal vs. structural methods. For example, there exist structural analyses that don't assume $cov(e, Z) = 0$ and identify, or partially identify, objects similar to LATE, and there exist causal treatment effects that can be identified even if defiers are present. The point of this example is just to illustrate the different types of assumptions, and associated estimands, that are typical in the two methodologies.

One way to interpret the difference in assumptions in this example is that the structural assumption $cov(e, Z) = 0$, which reduces to $cov(U_1, V_1) = 0$, is a restriction on the heterogeneity of the treatment effect U_1. Essentially this restriction says that an individual's type V_1 (complier if $V_1 = 1$, defier if $V_1 = -1$, always or never taker if $V_1 = 0$) is on average unrelated to the magnitude of their personal treatment effect U_1. This is a behavioral restriction regarding the outcome Y (or more precisely on how treatment affects Y), and behavioral restrictions on how covariates can affect outcomes are typical structural type assumptions.

In contrast, the causal assumption that nobody has $V_1 = -1$ (no defiers) is a restriction on the heterogeneity of types of individuals. This is still a behavioral restriction, but it only restricts behavior regarding the determination of treatment T relative to the instrument Z. This no defiers assumption does not restrict how the outcome Y might depend on treatment. This illustrates a general feature of causal methods, which is to assume as little as possible about how outcomes depend on treatment, preferring instead to make relatively stronger assumptions regarding the determination of treatment. In short, causal methods generally make fewer or weaker assumptions regarding the Y equation and stronger assumptions regarding the T equation.

In the causal model, we pay a price for dropping restrictions on the outcome equation. First, we can't know precisely who the compliers are, because the definition of a complier involves unobservables (or, equivalently, counterfactuals). The LATE is the average treatment effect for a subset of the population,

but we don't know who is in that subset, and we don't know how treatment may affect anyone else. This disadvantage is mitigated by the fact that we can estimate the probability that anyone is a complier, as a function of their observable characteristics (see Angrist and Pischke 2008). Also, if Z is a policy variable, then compliers might be a subpopulation of interest. One more mitigating factor is that, given an instrument or other covariates, one can often calculate bounds on ATE (an example of set identification). See, e.g., Manski (1990) and Balke and Pearl (1997).

On the other hand, treatment might be something individuals actively seek out, typically for reasons that relate to the outcome. An example is the Roy (1951) model, where people choose a treatment (like moving to a new location, accepting a job offer, or taking a drug) because of the outcome they expect from that treatment. Anyone who chooses their own treatment in this way will generally not be a complier (since compliers have treatment given by the randomly determined Z). This is a reappearance of the point made earlier that structure is needed to identify causal relationships in which treatment correlates with outcomes. For more on this point see Heckman (2008), who illustrates limitations of reduced form estimators like LATE in comparison to more general treatment effect models that allow for self selection. By focusing on compliers, LATE essentially only looks at the subset of people for whom treatment was randomly assigned. To the extent that compliers are not representative of the population as a whole, LATE may be unreliable for policy analyses. Of course, a similar objection might be made to the structural interpretation of c; it too could be unreliable for policy analyses if the population does not, at least approximately, satisfy the assumed behavioral restrictions.

Another limitation of the reduced form methodology is how it extends to more general treatments. When the treatment is many valued or even continuous, the number of types (compliers, deniers, etc.) that one needs to define and restrict becomes large and complicated in the causal framework. In contrast, the structural restriction $cov(U_1, V_1) = 0$, or equivalently, $cov(e, Z) = 0$, remains unchanged regardless of how many values T or Z can take on.

A related limitation of LATE is that the definition of a complier depends on the definition of the instrument Z. Suppose we saw a different instrument $\hat{Z}$ instead of Z, and we let $\hat{e} = cov(\hat{Z}, Y)/cov(\hat{Z}, T)$. If

$\hat{z}$ is a valid instrument in the structural sense, then we will have $c \le e \le b$, meaning that any structurally valid instrument will identify the same population ATE b

As before, let us again analyze the meaning of $c = cov(Z,Y)/cov(Z,X)$. For this simultaneous system, the structural analysis is exactly the same as before: We can rewrite the Y equation as $Y = a + bX + e$ where $b = E(U_1)$ and $e = U_0 + (U_1 - b)X$. If the structural assumption $cov(e,X) = 0$ holds then $c = b$ and both equal $E(U_1)$, the average marginal effect of X on Y.

In contrast, a causal analysis of this system is possible, but is much more complex. Angrist, Graddy and Imbens (2000) provide conditions, similar to those required for LATE, under which c will equal a complicated weighted average of conditional expectations of U_1, with weights that depend on Z. And even this limited result, unlike the simple structural restriction, requires X to be binary.

Another limitation of applying causal methods to simultaneous systems is that the counterfactual notation itself rules out some types of structural models. For example, consider an incomplete model similar to that of Section 4, where endogenous variables Y and X are determined by $Y = I(X + U \geq 0)$, $X = Y + Z + V$, and $1 - U + Z + V < 0$. As above, Y, X, and Z are observables and U and V are unobserved error terms. This model could represent a game between two players, where the equations are the reaction functions of one player who chooses Y and the other who chooses X. In this model, reduced form equations for Y and X as functions of Z, U, and V do not exist. Even if you knew the joint distribution of (Z,U,V), the probability distribution of potential outcomes $Y(x)$ would not be defined. When we use the potential outcome notation, we assume that the potential outcomes $Y(x)$ are random variables having well defined (albeit unknown) distributions. This structural model cannot be represented by the potential outcome notation. Use of the potential outcome notation in this model imposes additional assumptions or restrictions that are not part of the underlying structural model, like assuming the existence of an equilibrium selection rule for the game. More generally, use of counterfactual notation in a model implicitly assumes that reduced forms exist in that model (indeed, causal models are often called reduced form models).

A final limitation in applying causal analyses to simultaneous systems is the SUTVA restriction discussed earlier. Many kinds of structural models involving simultaneous systems exist in which treatment of one person may causally affect

cial interactions, network effects, and general equilibrium models. For example, the Progresa program in Mexico (and its successor Oportunidades) is a widely cited example of randomized treatment assignment, but people may choose to move to communities that have the program, or change behavior either from interacting with treated individuals, or in expectation that the program will expand to their own community. Behrman and Todd (1999) discuss these and other potential SUTVA violations associated with Progresa. Similarly, macroeconomic treatments often either cannot be randomized, or would violate SUTVA if they were. As noted earlier, some work does exist on causal model identification when SUTVA is violated, but typically such models require behavioral, structural type assumptions for point identification of treatment effects. See, e.g., Manski (2013).

As the above examples illustrate, in many settings involving interactions or simultaneous systems, causal estimands can be difficult to identify or interpret without structural, behavioral type model restrictions. This may explain why causal inference is more popular in fields that traditionally focused on partial equilibrium analyses (e.g., labor economics and micro development), but has made fewer inroads in fields where general equilibrium models requiring simultaneous systems are the norm (e.g., industrial organization and macroeconomics).

5.4 Randomized Causal vs. Structural Identification: Conclusions

This subsection provides a short summary of the relative advantages and disadvantages of causal vs. structural approaches to identification, though as noted earlier, best practice will often be to combine features of both methodologies.

One great advantage of causal based methods is their long history of success in the hard sciences.

mands. As long as the framework is coherent and complete, so that potential outcomes are well defined,

in the calibrated value of the rate of time preference (some other structural parameters, like the degree of risk aversion, have less consensus). Most of these calibrated values are obtained from multiple structural models, though some have been additionally investigated by laboratory and field experiments.

More generally, experience drawn from a variety of structural models has lead to a consensus among economists regarding ranges of values for parameters, such as price and income elasticities, that are widely recognized as reasonable. Empirical structural analyses have in addition revealed many behavioral relationships, going back at least to Engel's (1857) law, that appear to hold up almost universally.

The main disadvantage of imposing behavioral restrictions for identification is that reality is complicated, so every structural model we propose is likely to be oversimplified and hence misspecified. As Box (1979) famously observed, "all models are wrong, but some are useful." Moreover, one generally does not know the extent to which misspecification can corrupt the interpretation and applicability of structural model estimates.

Causal models can of course also be misspecified, e.g., in the Les

tures to identify parameter vectors and functions.

The third subsection below describes the role of normalizations in identification. Normalizations are prominently used in the literature on nonparametric, semiparametric, and set identification, but are rarely discussed. The fourth subsection below uses special regressors to provide examples of nonparametric, semiparametric, and set identification, and the use of normalizations.

6.1 Nonparametric and Semiparametric Identification

In Section 3.2, we defined *nonparametric identification* as the case where consists of functions or infinite sets. As discussed earlier in Section 2.4, the Glivenko–Cantelli theorem proves that, with IID observations of a vector W, the distribution function of W is consistently estimated by the empirical distribution function. It follows that the distribution function of W, $F(W)$, is nonparametrically identified by construction, where the construction is to take the probability limit of the empirical distribution function. If W is continuously distributed, then its probability density function $f(W)$ is also nonparametrically identified, using the construction $f(W) = \partial F(W)/\partial W$ with $F(W)$ identified. For another example, assume IID observations of continuous Y, X. Suppose we have the nonparametric regression model $Y = m(X) + e$ with $E(e \mid X) = 0$. Then $m(X) = E(Y \mid X)$ as long as this expectation exists. The conditional expectation function $m(X)$ can be constructed from the joint distribution of Y and X, which is itself identified, so we have by construction that $m(X)$ is nonparametrically identified.

Recall that *parametric identification* was defined as the case where is a finite set of constants, and all the different possible values of also correspond to different values of a finite set of constants. Identification that is neither parametric nor nonparametric is called *semiparametric identification*. For example, given we have IID observations of random variables Y, X, Z, a partially linear model is defined as the model $Y = m(Z) + X' + e$ where m is an unknown function,

is identification of the function $m(X)$ in the model $Y = m(X) + e$, but instead of the nonparametric regression assumption that $E(e \mid X) = 0$, it is assumed that $E(e \mid Z) = 0$ for some vector of instruments Z. The parameter to identify is the function $m(X)$, what is knowable, , is $F(Y, X \mid Z)$ (the joint distribution of Y, X given Z), and the restrictions that define the model are the equation $E(e \mid Z) = 0$ along with some regularity conditions. This equation can be written as $\int_{supp(Y,X \mid Z = z)} (Y - m(X)) \, dF(Y, X \mid z) = 0$ for all $z \in supp(Z)$. We have identification if this integral equation can be uniquely solved for $m(X)$. So here identification corresponds to uniqueness of the solution of an integral equation. Newey and Powell (2003) show that identification here is equivalent to an example of statistical completeness.

In contrast, if Y, X, and Z were all discrete, then identification of would be parametric identification. In that case, the integral equation would reduce to a matrix equation, and identification would only require simple nonsingularity of a moment matrix, as in linear instrumental variables regression models. When identification of vectors of parameters depends on inverting nonsingular matrices, it is sometimes possible to extend these same arguments to the identification of functions through the use of so-called operator methods. These methods roughly correspond to inverting the integrals (like expectations of continuous random variables) that express in terms of , in the same way that the sums (like expectations of discrete random variables) may be solved for vectors by matrix inversion. Schennach (2007) is a prominent example, identifying a nonparametric regression function with a mismeasured regressor. In applications of operator methods, concepts like completeness and injectivity are crucial to identification, as infinite dimensional analogues to the invertibility of matrices.

More generally, econometric models often involve moments, which take the form of integrals. As a result models requiring nonparametric or semiparametric identification frequently require integral equations to have unique solutions. The above nonparametric instrumental variables model is one example. Other

(1951) on recovering joint distributions from marginals, and Peterson (1976) on competing risks models. Much of the systematic study of set identification is attributed to Manski (e.g., Manski 1990, 1995, 2003).

While not typically associated with the set identification literature, an early example of combining theoretical restrictions with data to obtain inequalities and hence bounds on behavior are the revealed preference inequalities derived by Afriat (1967) and Varian (1982, 1983).

Interest in set identification grew when methods for doing inference on estimators of set identified parameters began to be developed, as in Manski and Tamer (2002). In addition to inference, much of the modern literature on set identification deals with the derivation of sharp bounds, with verifying that one has obtained the smallest possible identified set given the available information, and with finding situations where the identified set is very small relative to $\boldsymbol{\mathcal{Z}}$.

One reason why parameters may be set rather than point identified is incompleteness of the underlying model. It can be difficult to uniquely pin down parameters when more than one value of the endogenous variables Y can be associated with any given value of covariates and errors. See Tamer (2003) and Manski (2007) for examples. Set rather than point identification is also common when data are incompletely observed, e.g., when regressors are censored, discretized, mismeasured, or observations are missing not at random. See, e.g., Manski and Tamer (2002) and references therein. We also often get sets rather than points when economic theory only provides inequalities rather than equalities on behavior, as in Pakes, Porter, Ho, and Ishii (2015).

As noted earlier, parameters may fail to be point identified when they are at least partly defined in terms of unobserved variables, such as counterfactuals in treatment effects models (see, e.g., Manski (1990) and Balke and Pearl (1997)). Different parameter values may be associated with different values these unobservables may take on, leading to set identification. Schennach (2014) provides a general technique for deriving a collection of observable moments that implicitly characterize the identified set in models that are defined by moments over both observables and unobservables.

Some proponents of set identification methods, such as Chesher and Rosen (2017), argue that economic theory rarely provides enough restrictions to point identify model parameters of interest, resulting in a

great deal of econometric literature devoted to complicated or poorly motivated tricks to obtain point identification. They essentially argue that set identification should be treated as the usual situation. A less extreme view is that we should first see what assumptions are needed to obtain point identification. Then, examine what happens to the identified set when the strongest or least defensible point identifying assumptions are dropped. For example, Lewbel (2012) uses a strong heteroskedasticity restriction to obtain identification in models where ordinary instruments would usually be used for estimation, but are unavailable. That paper includes construction of identified sets when this strong point identifying restriction is relaxed.

Khan and Tamer (2010) define *non-robust identification* as the situation where an otherwise point identified parameter loses even set identification when an identifying assumption is relaxed. For example, suppose we wish to estimate $\mu = E(Y)$ where the scalar random variable Y can take on any value, but what is knowable, μ is the distribution of Y, defined by $Y = \min(b, Y)$ for some positive constant b. For example, our DGP may consist of IID observations of Y, which is a censored version of the true Y. Here μ is not point identified unless there's no censoring, meaning $b = \infty$. The identifying assumption $b = \infty$ is non-robust, because if it does not hold then, whatever the distribution of Y is, μ could take on any value. For example, even if Y has only a 1% chance of being larger than b, it could take on an arbitrarily large value with that .01 probability, resulting in μ being arbitrarily large. A non-robust identifying assumption is one that is crucial in the sense that, without it, the data do not limit the range of values μ could take on.[13]

Set identification is an area of active research in econometric theory, but it is not yet frequently used in empirical work. Perhaps the main obstacle to its application is that existing methods for estimation

estimators and inference is an ongoing area of research.

Modern econometrics is often criticized for being too complicated. This is a theme that appears in, e.g., Angrist and Pischke (2008) and in Freedman (2005). The essence of these critiques is that model complexity makes it difficult to discern or assess the plausibility of underlying identifying assumptions, and too difficult to implement modern estimators. It is therefore perhaps ironic that removing complicated identifying assumptions often leads to set rather than point identification of parameters, which then typically requires even more rather than less mathematically complicated econometrics for identification, estimation, and inference.

6.3 Normalizations in Identification

Nonparametric or semiparametric identification results often require so-called normalizations, but I do not know of any previous survey that has reviewed the general issues associated with normalizations for identification and estimation. To see what is meant by a normalization, consider the linear index model $E(Y \mid X) = g(X'\beta)$, where g is some strictly monotonically increasing function and β is an unknown vector. Many common econometric models are special cases of linear index models. For example, linear regression is a linear index model where g is the identity function, and the binary probit model is a linear index model where g is the cumulative standard normal distribution function. Binary logit models and many censored and truncated regression models are also special cases of linear index models.

Assume F is the joint distribution of Y and X and $E(XX')$ is nonsingular. If g is known (like the logit or probit model), then β is point identified. The proof is by construction: $\beta = [E(XX')]^{-1}E(Xg^{-1}(E(Y \mid X)))$.

Is β still point identified when the function g is unknown? In general, the answer is no. Intuitively, one could double β and suitable redifine g, leaving $g(X'\beta)$ unchanged. so β and 2β are observationally equivalent. Formally, Let $\theta = (g, \beta)$, so both the function g and the vector β are unknown. For any given positive constant c, define $\tilde{\theta} = (c, \tilde{g}(z) = g(cz))$, and $\tilde{\theta} = (\tilde{g}, \tilde{\beta})$. Then for any X we have $E(Y \mid X) = g(X'\beta) = \tilde{g}(X'\tilde{\beta})$, which shows that θ and $\tilde{\theta}$ are observationally equivalent. So unless our model contains some other information about g or β, the vector β is not identified. At best β might

be set identified, where the identified set includes all vectors that are proportional to the true . Suppose that all of the elements of the identified set for are proportional to the true , so all have the form of $e\ D = c$. Then we would say that is *identified up to scale*. An early example in the semiparametric

given any θ we can define an observationally equivalent $\tilde{g}$ such that, $\tilde{g}(X'\theta) = g(X'\theta + \delta)$. As with scale restrictions, location restrictions may or may not be free normalizations, depending on the use of the model.

Consider a threshold crossing binary choice model, that is, $Y = I(\beta + X'\theta + e \geq 0)$ where e is an unobserved error that is independent of X. This is a special case of the linear index model, since it implies that $E(Y \mid X) = g(\beta + X'\theta)$ where g is the distribution function of $-(\beta + e)$. Here identification requires both a location and a scale normalization. In parametric models, these normalizations are usually imposed on e. For example, the probit model is the special case of the threshold crossing binary choice model where e has a standard normal distribution. This includes the restriction that e has mean zero and variance one, which uniquely determines the location and scale of $\beta + X'\theta + e$. We could instead have imposed the location and scale restrictions that $\beta = 0$, $\theta'\theta = 1$, and assumed that e has an arbitrary mean and variance. Both ways of expressing the model are observationally equivalent.

Unlike parametric models, in semiparametric models it is more common to impose location and scale normalizations on model parameters like β and θ instead of on error distribution parameters like $E(e)$ and $var(e)$.

for an unknown strictly monotonic, invertible function h, where the function $G(X;v)$ is defined by9552 Tf 11

equivalent to saying that utility is ordinally but not cardinally identified, or that utility is identified up to an arbitrary monotonic transformation, which we may call a normalization. Similarly, the threshold crossing model $Y = I(X'\beta + e \geq 0)$ is observationally equivalent to $Y = I(g(X'\beta + e) \geq g(0))$ where g is any strictly monotonically increasing function. Without more information, we could therefore never tell if one's actual utility level were $X'\beta + e$ or $g(X'\beta + e)$ for any strictly monotonically increasing function g. As before, whether choice of g corresponds to a free normalization or to a behavioral restriction depends on context.

Some final notes on normalizations are these. First, parametric and semiparametric models often use different normalizations. The location and scale normalizations on coefficients vs on error moments discussed above are an example. When comparing parametric and semiparametric estimates, one should either recast in the same normalization, or compare summary measures like marginal effects that are independent of the choice of normalization. Relatedly, choice of what normalization to make, even if

is identified given . Note that $F_{U|Z}(u \mid z)$ is only identified for values of u that X can equal.

This is an example of semiparametric identification of a function. The intuition behind this identification, which is the basis of special regressor estimation (see Lewbel 1997a, 2000, 2014), is that the distribution of the unobserved latent error U can be identified because the model contains $U + X$ for a covariate X, and variation in X moves the dependent variable in the same way that variation in U does. Let us now consider examples of models that exploit this idea.

Example: Set Identification of the Latent Mean. In this example we let Z be empty, and consider identification of $= E(U)$. Let us also assume that U can take on any value (its support is the whole real line). We have from above that $F_U(u)$ is identified, but only for value of u that are in the support of X. This means that if X has support on the whole real line, then $F_U(u)$ is identified for all values of u, and therefore we can identify $E(U) = \int_{-1}^{1} u dF_U(u)$. This so-called large support assumption on X is needed to identify $E(U)$, because calculating the mean of a random variable depends on the entire distribution function of that variable. In contrast, other features of the distribution of U can be identified even if X has very limited support. For example, if we wanted to identify the median rather than the mean of X, then we would only require that the support of X includes the point x that makes $E(Y \mid X = x) = 1/2$.

Suppose now that the support of X equals the interval $[a, b]$ for some finite constants a and b. Then $F_U(u)$ is only identified for values of u in the range $b - a \leq u \leq a - b$. In this case, as noted by Khan and Tamer (2010), $E(U)$ is not even set identified, so identification of $E(U)$ is non-robust. This is because the distribution of U could have mass arbitrarily far below b or arbitrarily far above a allowing $E(U)$ to take on any value. On the other hand, if either $a = -1$ or $b = 1$, then we could place bounds on $E(U)$, giving us set identification, and if both $a = -1$ and $b = 1$, then from above $E(U)$ is point identified. Point or set identification of $E(U)$ is also possible when a and b are both bounded if we are given some additional information about or restrictions on the tails of U. An example is that $E(U)$ will be point identified with bounded a and b if a condition Magnac and Maurin (2007) call tail symmetry holds. Even mild restrictions on the tails of U, such as having the variance of U be finite, may suffice to yield at least set identification of $E(U)$ when a and b are bounded.

Example: General Binary Choice. Suppose we continue to have IID observations of $Y; X; Z$ with $Y = I(X + U > 0)$ where $U \perp X \mid Z$, but now in addition assume that $U = g(Z) + e$ with $g(Z) = E(U \mid Z)$, so $Y = I(X + g(Z) + e > 0)$. If $g(Z)$ were linear and e was normal, this would be a probit model. Instead we have a very general binary choice model where the latent variable contains an unknown function $g(Z)$ and the distribution of the latent error term e is also unknown and could be heteroskedastic. Note that the assumption that the coefficient of

construction.

Example: Binary Choice With Random Coefficients. Before considering binary choice, consider first the simpler linear random coefficients model. Suppose for the moment that we had IID observations of continuous U_i and Z_i, so the distribution function $F_{U|Z}$ is identified. Suppose further that U and Z satisfy the linear random coefficients model $U = Z'e$, where e is a vector of random coefficients having an unknown distribution, and e is independent of the vector Z. Then it can be shown that, under some standard conditions, the distribution of e is nonparametrically identified. See, e.g., Beran and Millar (1994) or Beran, Feuerverger, and Hall (1996). The proof is based on the conditional characteristic function of U given Z, but some intuition for why identification is possible can be obtained just by looking at simple moments. From $E(U|Z) = Z'E(e)$ we can identify $E(e)$, From$Z$

regressors.

Identification theorems for binary choice models that predate special regressors, but which can be reinterpreted as special cases of special regressor based identification methods, include Cosslett (1983), Manski (1985), Horowitz (1992), and Lewbel (1997a). For more on the construction and use of special regressors for identification, see Lewbel, Dong, and Yang (2012), Lewbel (2014), and Dong and Lewbel (2015).

7 Limited Forms of Identification

For many models it is difficult or impossible to lay out conditions that formally ensure parameters are point identified. One possible response to this problem is to use estimators that only require set identification, though these are often difficult or intractable to apply. At the other extreme, one might simply ignore the problem and just assume identification, though any resulting estimator could be poorly behaved. A middle way is to establish conditions that make identification likely in some sense. Examples are local identification and generic identification. These are conditions that are weaker than point identification, but are often easier to prove. Given local or generic identification, it is then less of a leap of faith to assume point identification holds.

7.1 Local and Global Identification

For a given true value of $_0$, recall that point identification of $_0$ means that there is no other 2 $\boldsymbol{2}$ (the set of all possible values of , according to the model) that is observationally equivalent to $_0$. Since the true value of $_0$ is unknown, proving point identification requires that no distinct pairs of values and e in $\boldsymbol{2}$ be observationally equivalent to each other. As noted earlier, this condition is sometimes called *global identification*, emphasizing how point identification must hold whatever the true value of $_0$ turns out to be. A recent study that focuses on conditions for global identification is Komunjer (2012).

A necessary condition for global identification, and one that is often easier to verify in practice, is local

observationally equivalent to any other value in that interval.

This example generalizes in some ways. For example, suppose in some model that θ is an interval. If is set identified, and the set has a finite number of elements, then is locally identified. Similarly, consider an extremum identification problem where the objective function is complicated. In such cases it may be difficult to rule out the possibility of a finite number of local optima, in which case one might show local but not necessarily global identification. More generally, in nonlinear models it is often easier to provide conditions that ensure local rather than global identification.

Local identification may be sufficient in practice if we have enough economic intuition about the estimand to know that the correct should lie in a particular region. Lewbel (2012) gives an example of a model with a parameter and associated estimator that is set identified. The parameter is a coefficient in a simultaneous system of equations, and the identified set has two elements, one positive and one negative. So in this case we only have local identification, but if economic theory is sufficient to tell us the sign of the parameter a priori, then that local identification may suffice for estimation. Note that in this particular example, we could have used economic theory to restrict the model, redefining θ to only include values of with the correct sign, and then declaring the parameter to be globally identified.

The notion of local identification is described by Fisher (1966) in the context of linear models, and is

For parametric models that can (if identified) be estimated by maximum likelihood, this first order condition is equivalent to the condition that the information matrix evaluated at the true be nonsingular. Newey and McFadden (1994) and Chernozhukov, Imbens, and Newey (2007) give semiparametric extensions of the Sargan rank result. Chen, Chernozhukov, Lee, and Newey (2014) provide a general rank condition for local identification of a finite parameter vector in models defined by conditional moment restrictions. Chen and Santos (2015) provide a concept of local overidentification that can be applied to a large class of semiparametric models.

7.2 Generic Identification

Like local identification, generic identification is a weaker condition than point identification, is a necessary condition for point identification, and is often easier to prove than point identification. Also like local identification, one may be more comfortable assuming point identification for estimation purposes, if one can show that at least generic identification holds.

Let $\mathcal{B}$ be a subset of $\mathcal{B}$, defined as follows: Consider every $\in \mathcal{B}$. If is observationally equivalent to any other $\neq \in \mathcal{B}$, then include in $\mathcal{B}$. This construction means that if $_0$ takes on a value that is in $\mathcal{B}$ then $_0$ is not point identified, otherwise $_0$ is point identified. Proving point identification for any value $_0$ might take on requires that $\mathcal{B}$ be empty. Following McManus (1992), the parameter is defined to be *generically identified* if $\mathcal{B}$ is a measure zero subset of $\mathcal{B}$.

To interpret what generic identification means, imagine that nature chooses a value $_0$ by randomly picking an element of $\mathcal{B}$. Assume all elements of $\mathcal{B}$

nonsingular. If we drew J^2 random numbers from some continuous distribution and put them in a matrix, the probability that the matrix would be singular is zero, so in this example the order condition implies generic identification of the model. Similarly, the coefficients in a linear regression model $Y = X'\beta + e$ with $E(e \mid X) = 0$ are generically identified if the probability is zero that nature chooses a distribution function for X with the property that $E(XX')$ is singular.

Another example is the regression model with measurement error. Assume the DGP is IID observations of (Y, X). Suppose the model is $X = X^* + U$ and $Y = X^*\beta + e$, where the unobserved model error e, the unobserved measurement error U, and the unobserved true covariate X^* are all mutually independent with mean zero. An early result in the identification literature is Reiersøl (1950), who showed that in this model, despite not having instruments, the coefficient β is identified when (Y, X) has any joint distribution except a bivariate normal. We could then say that β is generically identified if the set of possible joint distributions that (Y, X) might be drawn from is sufficiently large, as would be true if, e.g., e could have been drawn from any continuous distribution. Similarly, Schennach and Hu (2013) show that under the same mutual independence of e, U, and X^*, the function m in the nonparametric regression model $Y = m(X^*) + e$ is nonparametrically identified as long as m and the distribution of e are not members of a certain parametric class of functions. So again one could claim that mutual independence of e, U, and X^* leads to generic identification of m, as long as m could have been any smooth function or if e could have been drawn from any smooth distribution.

Generic identification is sometimes seen in social interactions models. In many such models, showing point identification is intractable, but one can establish generic identification. See, e.g., Blume, Brock, Durlauf, and Ioannides (2011).

The term generic identification is sometimes used more informally, to describe situations in which identification holds except in special or pathological cases, but where it might be difficult to explicitly describe all such cases. An example is the generic identification results in Chiappori and Ekeland (2009). These formal and informal definitions of generic identification coincide if we can interpret the special or pathological situations as arising with probability zero.

(1984). Bound, Jaeger, and Baker (1995) specifically raised the issue of weak instruments in an empirical context. An early paper dealing with the problem econometrically is Staiger and Stock (1997). A survey of the weak instruments problem is Stock, Wright, and Yogo (2002).

The usual source of weak identification is low correlations among variables used to attain identification. A typical example is when the correlation between an instrument Z and the covariate X it is instrumenting is close to zero. Associated parameters would not be identified if the correlation was actually zero, and so identification is weak (usually stated as saying the instrument Z is weak) when this correlation is close to zero. Given a vector of regressors X and a vector of instruments Z in a linear regression model, the first stage of two stage least squares is to regress X on Z to get fitted values $\hat{X}$, and some or all of the model coefficients may be weakly identified if the matrix $E\left[\hat{X}\hat{X}'\right]$ is ill conditioned, i.e., close to singular. More generally, in a GMM model weak identification may occur if the moments used for estimation yield noisy or generally uninformative estimates of the underlying parameters.

The key feature of weakly identified parameters is not that they are imprecisely estimated with large standard errors (though they do typically have that feature). Rather, weakly identified parameters have the property that standard asymptotic theory provides a poor approximation to the true precision of estimation. Moreover, higher order asymptotics don't help, since they too depend on precise parameter estimates. In contrast, strongly identified parameters are defined as parameters for which standard estimated asymptotic distributions provide good approximations to their actual finite sample distributions.

Nonparametric regressions are also typically imprecisely estimated, with slower than parametric convergence rates and associated large standard errors. But nonparametric regressions are not said to be weakly identified, because standard asymptotic theory adequately approximates the true precision with which those parameters are estimated. Similarly, parameters that suffer from irregular or thin set identification, such as those based on identification at infinity, are also not called weakly identified, since standard asymptotic theory, again at slower than parametric rates, can still typically be applied.

To illustrate, consider a parameter vector that is identified, and could be estimated at parametric rates using an extremum estimator (one that maximizes an objective function) like least squares or GMM or

maximum likelihood. Elements of this parameter vector will be weakly identified if any objective function we might use for estimation is relatively flat in one or more directions involving those parameters. This flatness of the objective function leads to imprecision in estimation. But more relevantly, flatness also means that standard errors and t-statistics calculated in the usual ways (either analytically or by bootstrapping) will be poorly estimated, because they depend on the inverse of a matrix of derivatives of the objective function, and that matrix will be close to singular.

Weak identification resembles multicollinearity, which in a linear regression would correspond to $E[XX']$ instead of $E[\hat{X}\hat{X}']$ being ill-conditioned. Like multicollinearity, it is not the case that a parameter either "is or "is not" weakly identified. Rather, relative weakness of identification depends on the sample size. A model that suffers from multicollinearity when the sample size is $n = 100$ may be fine when $n = 1000$. Similarly, A parameter that is weakly identified (meaning that standard asymptotics provide a poor finite sample approximation to the actual distribution of the estimator) when $n = 100$ may be strongly identified when $n = 1000$. This is why weakness of identification is generally judged by rules of thumb rather than formal tests. For example, Staiger and Stock (1997) suggest the rule of thumb for linear two stage least squares models that instruments are potentially weak if the F-statistic on the excluded regressors in the first stage of two stage least squares is less than 10. See also Inoue and Rossi (2011) for who provide a test for (extremum based) strong identification, where alternatives include weak identification and a lack of extremum based point identification.

It is important to make a distinction between parameters that are weakly identified, and the models that econometricians use to deal with weak identification. In real data, weak identification is purely a finite sample problem that disappears when n gets sufficiently large. This makes it difficult to provide asymptotic theory to deal with the problem. Econometricians have therefore devised a trick, i.e., an alternative asymptotic theory, to provide better approximations to true finite sample distributions than are obtained with standard asymptotics.

To understand this trick, consider the simple two equation system $Y = \beta X + U$ and $X = Z + V$ where the DGP consists of IID observations of the mean zero random scalars $Y, X,$ and Z, while U

and V are unobserved mean zero errors that are uncorrelated with Z. In this case, as long as $\neq 0$, the parameter is identified by $= E(ZY)/E(ZX)$. A corresponding estimator would replace these expectations with sample averages, yielding the standard linear instrumental variables estimator. However, since $E(ZX) = E(Z^2)$, if is close to zero then $E(ZX)$ will be close to zero, making weakly identified. But how close is close? Small errors in the estimation of $E(ZX)$ will yield large errors in the estimate of . The bigger the sample size, the more accurately $E(ZX)$ can be estimated, and hence the closer can be to zero without causing trouble.

To capture this idea asymptotically, econometricians pretend that the true value of is not a constant, but instead takes a value that drifts closer to zero as the sample size grows. That is, we imagine that the true model is $X = {}_n Z + U_x$, where ${}_n = bn^{-1/2}$ for some constant b. The larger n gets, the smaller the coefficient ${}_n$ becomes. This gives us a model where suffers from the weak identification problem at all sample sizes, and so can be analyzed using asymptotic methods. Typically, in a drifting parameter model like this, the constant b and hence the parameter ${}_n$ is not identified, so tests and confidence regions for have been developed that are robust to weak instruments, that is, they do not depend on consistent estimation of ${}_n$. See, e.g., Andrews, Moreira, and Stock (2006) for an overview of such methods.

In the econometrics literature, saying that a parameter in a model "is1(that)62 Tf 12.g-261(miified,)-226(sayn)-

In$_n$ D 4550200-309(455e)-297 0.ts,ntur9(455t5(w)-26e)-a

(2015), and references therein. These refer to models where parameters, or their impact on , drift to zero at rates other than $n^{1=2}$, or more generally where the model may contain a mix of drifting, nondrifting, and purely unidentified parameters.

One final note is that weak instruments are often discussed in the context of models that also have many instruments. However, the econometric difficulties associated with many instruments are distinct from those associated with weak instruments, and some separate theory exists for dealing with many instruments, weak instruments, or the combination of the two.

8.2 Identification at Infinity or Zero; Irregular and Thin set identification

Based on Chamberlain (1986) and Heckman (1990), *identification at infinity* refers to the situation in which identification is based only on the joint distribution of data at points where one or more variables go to infinity. For example, suppose our DGP is IID observations of scalar random variables $Y; D; Z$. Assume $Y = D \cdot Y$ where D is a binary variable equal to zero or one, Y is a latent unobserved variable that is independent of Z, and $\lim_{z \to 1} E(D \mid Z = z) = 1$. The goal is identification and estimation of $= E(Y)$. This is a selection model, where Y is selected (observed) only when $D = 1$. For example D could be a treatment indicator, Y is the outcome if one is treated, is what the average outcome would equal if everyone in the population were treated, and Z is an observed variable (an instrument) that affects the probability of treatment, with the probability of treatment going to one as Z goes to infinity. Here is identified by $= \lim_{z \to 1} E(Y \mid Z = z)$. The problem is that Y and D may be correlated, so looking at the unconditional mean of Y confounds the two. But everyone who has Z equal to infinity is treated, so looking at the mean of Y just among people having arbitrarily large values of Z eliminates the problem. In real data we would estimate by $b = \sum_{i=1}^{n} w(n; Z_i) Y_i = \sum_{i=1}^{n}$

mators based on such identification will typically converge slowly (slower than parametric root n rates). The same estimation problems can also arise whenever identification is based on Z taking on a value or range of values that has probability zero. Khan and Tamer (2010) call this general idea *thin set identification*. For example, Manski's (1985) maximum score estimator for binary choice models is based on the assumption that the conditional median of a latent error equals zero. This assumption is another example of thin set identification, because it gets identifying power only from information at a single point (the median) of a continuously distributed variable.

Khan and Tamer (2010) and Graham and Powell (2012) use the term *irregular identification* to describe cases where thin set identification leads to slower than root-n rates of estimation. Not all parameters that are thin set identified or identified at infinity are irregular. For example, estimates of $\mathsf{D}\,E.Y\,/$ above can converge at parametric rates if Z has a strictly positive probability of equaling infinity. More subtly, the 'impossibility' theorems of both Chamberlain and Khan and Tamer showing that some thin set identified models cannot converge at rate root n assume that the variables in the DGP have finite variances. So, e.g., Lewbel's (2000) binary choice special regressor estimator with endogenous regressors is thin set identified. But this estimator can converge at rate root n, avoiding irregular identification and overcoming Khan and Tamer's impossibility theorem, either by having a special regressor with strictly larger support than the model's latent index, or by having a special regressor with infinite variance. Similarly, Khan and Tamer point out that the average treatment effect model of Hahn (1998) and Hirano, Imbens and Ridder (2003) is also generally irregularly identified, and so will not attain the parametric root n rates derived by those authors unless a latent index has extremely thick tails, just as Lewbel (2000) requires a special regressor with thick tails (or larger support than the latent variable) to avoid being irregular.

It is easy to confuse irregular identification with weak identification, but they are not the same. Both types of parameters are point identified by the usual definitions, and both refer to properties of the underlying data and model that cause problems with estimation and inference regardless of the choice of estimator.

The difference is that asymptotic theory for weakly identified parameters is based on models where true

However, if g is discontinuous, then $\hat{b}$ will generally not be consistent. This discontinuity is the problem of ill-posedness.[14] Ill-posedness is an identification concept like weak identification or identification at infinity, because it is a feature of the underlying model, , and .

When identification is ill-posed, construction of a consistent estimator requires "regularization," that is, some way to smooth out the discontinuity in g. However, regularization generally introduces bias, and obtaining consistency then requires some method of shrinking this bias as the sample size grows. This in turn generally results in slower convergence rates.

Nonparametric estimation of a probability density function is an example of an ill-posedness problem. Consider estimation of the density function $f(w)$, defined by $f(w) = dF(w)/dw$. There does not generally exist a continuous g such that $f = g(F)$. Correspondingly, one cannot just take a derivative of the empirical distribution function $\hat{F}(w)$ with respect to w to estimate $f(w)$. This problem is ill-posed, and so regularization is needed to consistently estimate f. The standard Rosenblatt-Parzen kernel density estimator is an example of regularization. This estimator, using a uniform kernel, is equivalent to letting $\hat{f}(w) = \hat{F}(w + bng) - \hat{F}(w - bng)$

Other common situations in econometrics where ill-posedness arises are in models containing mismeasured variables where the measurement error distribution needs to be estimated, and in random coefficient models where the distribution of the random coefficients is unknown. Although the term "Ill-Posed Identification" does not actually appear in the literature (that name is therefore being proposed here), the general problem of ill-posedness in econometrics is well recognized. See, e.g., Horowitz (2014) for a survey. The concept of well-posedness, the opposite of ill-posedness, is originally due to Hadamard (1923).

8.4 Bayesian and Essential Identification

We have already seen that the usual notion of identification can be called point identification or global identification. Two more names for the same concept that appear in the literature are *frequentist identification* and *sampling identification*. These terms are used to contrast the role of identification in frequentist statistics from its role in Bayesian statistics. In a Bayesian model, a parameter is a random variable rather than a constant, having both a prior and a posterior distribution. There is a sense in which point identification is irrelevant for Bayesian models, since one can specify a prior distribution for , and obtain a posterior distribution, regardless of whether is point identified or not. See Lindley (1971) and Poirer (1998) for examples and discussions of the implications for Bayes estimation when parameters are not point identified.

Still, there are notions of identification that are relevant for Bayesians. Gustafson (2005) defines parameters [illegible]

distribution. That is,

Unlike statistical inference, there is not a large body of general tools or techniques that exist for proving identification. As a result, identification proofs are often highly model specific and idiosyncratic. Some general techniques for obtaining or proving identification in a variety of settings do exist. These include control function methods as generalized in Blundell and Powell (2004), special regressors as in Lewbel (2000), contraction mappings to obtain unique fixed points as applied in the Berry, Levinsohn, and Pakes (1995) model, classes of integral equations corresponding to moment conditions that have unique solutions, such as completeness as applied by Newey and Powell (2003), the observational equivalence characterization theorem of Matzkin (2005), and the moments characterization theorem of Schennach (2014). Development of more such general techniques and principles would be a valuable area for future research.

Finally, one might draw a connection between identification and big data. Varian (2014) says, "In this period of “big data,” it seems strange to focus on sampling uncertainty, which tends to be small with large datasets, while completely ignoring model uncertainty, which may be quite large." In big data, the observed sample is so large that it can be treated as if it were the population. Identification deals precisely with what can be learned about the relationships among variables given the population, i.e., given big data. A valuable area for future research would be to explore more fully the potential linkages between methods used to establish identification and techniques used to analyze big data.

This paper has considered over two dozen different identification related concepts, as listed in the introduction. Given the increasing recognition of its importance in econometrics, the identification zoo is likely to keep expanding.

10 Appendix: Point Identification Details

This Appendix presents the definition of point identification and related concepts with somewhat more mathematical rigor and detail than in Section 3. These derivations are very similar to those of Matzkin (2007, 2012), though Matzkin only considers the case in which is a data distribution function.

Define a *model* M to be a set of functions or sets that satisfy some given restrictions. These could

model value m_0 must satisfy $_0 = \mathcal{S}(m_0)$. If more than one value of $m \in M$ satisfies $_0 = \mathcal{S}(m)$, then we cannot tell which of these values of m is the true one.

This definition sidesteps the deeper question of what is actually meant by truth of a model, since models are assumed to only approximate the real world. All we are saying here about the true model value m_0 is that it doesn't conflict with what we can observe or know, which is $_0$.

Define a set of *parameters* to be a set of unknown constants and/or functions that characterize or summarize relevant features of a model. Essentially, can be anything we might want to estimate (more precisely, will generally be estimands, i.e., population values of estimators of objects that we want to learn about). Parameters could include what we usually think of as model parameters, e.g. regression coefficients, but could also be, e.g., the sign of an elasticity, or an average treatment effect.

Assume that there is a unique value of associated with each model value m (violation of this assumption relates to the coherence and completeness conditions; see Section 4 for details). Let $= \mathcal{I}(m)$ be the function or mapping that defines the particular parameter value that corresponds to the given model value m. The true parameter value $_0$ satisfies $_0 = \mathcal{I}(m_0)$.

Define $\Theta = \{ \mid = \mathcal{I}(m) \text{ where } m \in M\}$. So Θ is the set of all values of that are possible given the model M. Any $\notin \Theta$ is ruled out by the model. We can therefore think of Θ as embodying all of the restrictions on that are implied by the model.

Similar to Θ, define $\mathcal{B} = \{ \mid = \mathcal{S}(m) \text{ where } m \in M\}$. So $\mathcal{B}$ is the set of all that are possible given the model M. Any that is not in $\mathcal{B}$ is ruled out by the model. While the set Θ embodies or describes all the restrictions on the parameters that are implied by the model, the set $\mathcal{B}$ embodies all of the observable restrictions that are implied by the model (assuming that what we can observe is). The functions $\mathcal{S}$ and $\mathcal{I}$ are $\mathcal{S}: M \to \mathcal{B}$ and $\mathcal{I}: M \to \Theta$.

Define the *structure* $s(,)$ to be the set of all model values m that can yield both the given values and , that is, $s(,) = \{m \mid = \mathcal{S}(m), = \mathcal{I}(m), \text{ and } m \in M\}$. We can think of the structure as embodying the relationship between the parameters and what we could learn from data, which is .

Two sets of parameter values and θ are defined to be *observationally equivalent* in the model M if

there exists a $2 \in \mathcal{S}$ such that $s(\cdot, \cdot, \theta) \neq \varnothing$ and $s(\cdot, \cdot, \theta^e) \neq \varnothing$. Equivalently, θ and θ^e are observationally equivalent if there exist model values m and m^e in M such that $\theta = \lambda_1(m)$, $\theta^e = \lambda_1(m^e)$, and $\lambda_5(m) = \lambda_5(m^e)$. Roughly, θ and θ^e observationally equivalent means there exists a value λ_5 such that, if λ_5 is true, then either the value θ or θ^e could also be true.

Given observational equivalence, we have what we need to define identification. The parameter θ is defined to be *point identified* (also sometimes called *globally identified* and often just called *identified*) in the model M if, for any $\theta \in \Theta$ and $\theta^e \in \Theta$, having θ and θ^e be *observationally equivalent* implies $\theta = \theta^e$. Let $\theta_0 \in \Theta$ denote the unknown true value of θ. We can say that the particular value θ_0 is point identified if θ_0 is not observationally equivalent to any other value of $\theta \in \Theta$. The key point is that all we can know is θ_0, and $\theta_0 = \lambda_5(m_0)$. We therefore can't distinguish between m_0 and any other m for which $\lambda_5(m) = \lambda_5(m_0)$, and so we can't distinguish between $\theta = \lambda_1(m)$ and $\theta_0 = \lambda_1(m_0)$ if any such m exists. And, since we don't know before hand which of the possible values of θ will be the θ_0 that we see, and we don't know which of the possible values of θ is the true θ_0, to ensure point identification we require that no pairs of values θ and θ^e be observationally equivalent.

In practice, ensuring point identification may require that the definition of the model rules out some model values m, specifically, those for which $\lambda_1(m)$ is observationally equivalent to some $\lambda_1(m^e)$. Equivalently, the set Θ may be limited by ruling out values that can't be point identified.

We have now defined what it means to have parameters θ be point identified. We say that the *model is point identified* when no pairs of model values m and m^e in*be points(ensure)-297(point)-2963tionally equivalent.*

53-54.

Afriat, S. N. (1967), “The construction of utility functions from expenditure data,” International economic review, 8(1), 67-77.

Ahlfeldt, G., S. Redding, D.Sturm, and N. Wolf (2015) "The Economics of Density: Evidence from the Berlin Wall,” Econometrica, 83(6), 2127–2189.

Amemiya, T., (1985), Advanced econometrics. Harvard University Press, Cambridge, MA.

Anderson, T. W., Rubin, H. (1949), “Estimation of the parameters of a single equation in a complete system of stochastic equations,” The Annals of Mathematical Statistics, 46-63.

Princeton university press.

Berry, S. and P. Haile (2014), "Identification in Differentiated Products Markets Using Market Level Data," Econometrica 82(5) 1749-1797.

Conditional Moment Restrictions" in Badi H. Baltagi, R. Carter Hill, Whitney K. Newey, Halbert L. White (ed.) Essays in Honor of Jerry Hausman (Advances in Econometrics, Volume 29) Emerald Group Publishing Limited, 455 - 477.

Calvi, R. (2016), "Why Are Older Women Missing in India? The Age Profile of Bargaining Power and Poverty" Unpublished Manuscript, Rice University.

Calvi, R., A. Lewbel, and D. Tommasi (2017), “Women’s Empowerment and Family Health: Estimating LATE with Mismeasured Treatment,” Unpublished Manuscript, Boston College.

Cerulli, G., Y. Dong, Lewbel, A., and Poulsen, A. (2017), “Testing Stability of Regression Discontinuity Models,” forthcoming, Advances in Econometrics, vol. 38, Regression Discontinuity Designs: Theory and Applications, M. D. Cattaneo and J. C. Escanciano, editors.

Chamberlain, G. (1986), “Asymptotic efficiency in semi-parametric models with censoring,” Journal of Econometrics, 32(2), 189-218.

Chen S., A. Khan S., and X. Tang (2016), "Informational content of special regressors in heteroskedastic binary response models," Journal of Econometrics, 193(1), 162-182.

Chen, X., Chernozhukov, V., Lee, S., Newey, W. K. (2014), “Local identification of nonparametric and semiparametric models,” Econometrica, 82(2), 785-809.

Chen, X. and A. Santos (2015), "Overidentification in Regular Models," Cowles Foundation Discussion Paper 1999, Yale University.

Cheng, X. (2015), "Robust Inference in Nonlinear Models With Mixed Identification," Journal of Econometrics, 189(1) 207-228.

Chernozhukov, V., Hong, H. and Tamer, E. (2007), Estimation and Confidence Regions for Parameter Sets in Econometric Models. Econometrica, 75: 1243–1284.

Chernozhukov, V., Imbens, G. W., & Newey, W. K. (2007), Instrumental variable estimation of nonseparable models. Journal of Econometrics, 139(1), 4-14.

Chesher, A. (2008), Lectures on Identification, available at: http://economics.yale.edu/sites/default/files/files/Work
Seminars/Econometrics/chesher1-080416.pdf

as an appendix to Engel E. (1895) “Die Lebenskosten Belgischer Arbeiter Familien frfther und jetzt," Bulletin de l’institut international de statistique, tome IX, premiere livraison, Rome.

Escanciano, J-C, D. Jacho-Chávez, and A. Lewbel (2016), “Identification and Estimation of Semiparametric Two Step Models,” Quantitative Economics, 7, 561-589.

Fisher, R.A. (1935), The Design of Experiments. Edinburgh: Oliver and Boyd.

Fisher, F. (1959), “Generalization of the Rank and Order Conditions for Identifiability," Econometrica, 27, 431-447.

Fisher, F. (1966), “The identification problem in econometrics,” McGraw-Hill, New York. 1966

Florens, J.P., Mouchart, M. and J.M., Rolin (1990), Elements of Bayesian statistics. Dekker: New York.

Florens, J.P., and A. Simoni (2011), Bayesian Identification and Partial Identification, unpublished manuscript.

Choice Models," Econometrica, 81: 581–607.

Geary, R. C. (1948), "Studies in the relations between economic time series," Journal of the Royal Statistical Society, series B, 10, 140-158.

Gini, C. (1921), “Sull’interpoliazone di una retta quando i valori della variabile indipendente sono affetti da errori accidentali,” Metroeconomica, 1, 63–82.

Graham, B. S., and Powell, J. L. (2012), "Identification and Estimation of Average Partial Effects in “Irregular” Correlated Random Coefficient Panel Data Models," Econometrica, 80(5), 2105-2152.

Granger, C. W. J. (1969) “Investigating Causal Relations by Econometric Models and Cross-Spectral Methods,” Econometrica, 37(3), 424-438

Gustafson, P. (2005), "On Model Expansion, Model Contraction, Identifiability and Prior Information: Two Illustrative Scenarios Involving Mismeasured Variables," Statistical Science, 20(2), 111-140.

Gustafson, P. (2015), Bayesian Inference for Partially Identified Models: Exploring the Limits of Limited Data, CRC press: Boca Raton.

Haavelmo, T. (1943), “The statistical implications of a system of simultaneous equations,” Econometrica, 11, 1-12.

Haberman, S. J. (1974), The Analysis of Frequency Data, University of Chicago Press.

Hadamard, J. (1923), Lectures on Cauchy’s Problem in Linear Partial Differential Equations. New Haven: Yale University Press.

Hahn, J. (1998): “On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects,” Econometrica, 66, 315–331.

Hansen, L. P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators". Econometrica 50(4): 1029–1054.

Härdle, W. and T. M. Stoker (1989), "Investigating Smooth Multiple Regression by the Method of Average Derivatives," Journal of the American Statistical Association, 84(408), 986-995.

Heckman, J. J. (1978), “Dummy Endogenous Variables in a Simultaneous Equation System,” Econometrica, 46, 931-59.

Heckman, J. J. (1990), “Varieties of selection bias,” The American Economic Review, 80, 313-318.

Heckman, J. J. (1997), "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations," The Journal of Human Resources, 32(3), 441-462.

Heckman, J. J. (2008), “Econometric Causality,” International Statistical Review, 76, 1-27.

Heckman, J. J. (2010), "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy," Journal of Economic Literature, 48: 356-98.

Heckman, J. J. and B. E. Honoré (1989), "The identifiability of the competing risks model," Biometrika, 76, 325-30.

Heckman, J. J. and B. E. Honoré (1990), "The empirical content of the Roy model," Econometrica, 58, 1121-1149.

Heckman, J. J., H. Ichimura, and P. Todd. (1998), “Matching as an econometric evaluation estimator,” The Review of Economic Studies, 65(2), 261-294.

Heckman, J. J. and R. Pinto (2015), "Causal Analysis After Haavelmo," Econometric Theory, 31(1), 115-151.

Heckman, J. J., Robb Jr, R. (1985), “Alternative methods for evaluating the impact of interventions: An overview,” Journal of Econometrics, 30(1), 239-267.

Heckman, J. J., S. Urzua and E. Vytlacil, (2006) "Understanding Instrumental Variables in Models with Essential Heteroa309(e)25b[(with)-2"Td [(with)-25cTItm in Models

Economics, Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave, Macmillan.

Horowitz, J. L. (1992) “A Smoothed Maximum Score Estimator for the Binary Response Model,” Econometrica, 60(3), 505–531.

Horowitz, J. L. (2014), "Ill-Posed Inverse Problems in Economics," Annual Review of Economics, 6: 21-51.

Houthakker, H. S. (1950), “Revealed preference and the utility function,” Economica, 159-174.

Hsiao, C., (1983), Identification, in: Griliches, Z. and M.D. Intriligator, (Eds.), Handbook of econometrics, Vol. 1. 223-283, North Holland, Amsterdam, The Netherlands.

Hume, D. (1739) A Treatise of Human Nature. Edited by L.A. Selby-Bigge. Oxford: Clarendon Press, 1888.

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

Ichimura, H. and T. S. Thompson (1998), “Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution," Journal of Econometrics, 86(2): 269-295.

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

Imbens, G. W., and D. B. Rubin (1997), "Estimating Outcome Distributions for Compliers in Instrumental Variables Models," The Review of Economic Studies, 64(4), 555-574.

Inoue, A. and B. Rossi (2011), "Testing for weak identification in possibly nonlinear models" Journal of Econometrics 161(2), 246-261.

Jöreskog, K. G. (1970), "A general method for analysis of covariance structures," Biometrika 57, 239–251.

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

Kitagawa, T., (2015)“A Test for Instrument Validity,” Econometrica, 83, 2043-2063.

Klein, R. and F. Vella, (2010), "Estimating a class of triangular simultaneous equations models without

exclusion restrictions," Journal of Econometrics, 154(2), 154-164.

Koopmans, T. C. (1937), Linear regression analysis of economic time series. DeErven F. Bohn, Haarlem:Netherlands.

Koopmans, T. C. (1949), “Identification problems in economic model construction,” Econometrica, 17, 125-144.

Koopmans, T. C., Reiersøl, O. (1950), “The identification of structural characteristics,” The Annals of Mathematical Statistics, 165-181.

Koopmans, T. C., Rubin, H., Leipnik, R. B. (1950), “Measuring the equation systems of dynamic economics,” Statistical inference in dynamic economic models, 10.

Komunjer, I. (2012), “Global Identification in Nonlinear Models with Moment Restrictions,” Econometric Theory, 28 (4), 719-729.

Kummell, C. H. (1879), "Reduction of observation equations which contain more than one observed quantity". The Analyst (Annals of Mathematics) 6(4), 97–105.

Laffers, L. and G. Mellace (2016), "Identification of the Average Treatment Effect when SUTVA is violated," unpublished manuscript.

Lazzati, N. (2015), "Treatment Response with Social Interactions: Partial Identification via Monotone Comparative Statics," Quantitative Economics, 6, 49-83

Lee, L. F. and A. Chesher, (1986), “Specification Testing when Score Test Statistics are Identically Zero,” Journal of Econometrics, 31, 121-149.

Lee, S., and A. Lewbel (2013), "Nonparametric identification of accelerated failure time competing risks models," Econometric Theory, 29(5), 905-919.

Lee, Y.-Y. and H.-H. Li, (2018), "Partial effects in binary response models using a special regressor," Economics Letters, 169, 15-19.

Levitt, S. D. and J. A. List (2009), "Field experiments in economics: The past, the present, and the future," European Economic Review, 53(1), 1–18.

Lewbel, A. (1997a), "Semiparametric Estimation of Location and Other Discrete Choice Moments,"

Lindley, D.V. (1971), Bayesian statistics: a review. SIAM:Philadelphia.

Lucas, R. (1976), “Econometric Policy Evaluation: A Critique,” in: K. Brunner and A. Meltzer (eds.), The Phillips Curve and Labor Markets, Carnegie-Rochester Conference Series on Public Policy, 1, 19–46.

Magnac, T., Maurin, E. (2007), “Identification and information in monotone binary models,” Journal of Econometrics, 139(1), 76-104.

Magnac, T. and Thesmar, D. (2002), "Identifying Dynamic Discrete Decision Processes",Econometrica, 70, 801-816.

Manski, C. F. (1975), "Maximum Score Estimation of the Stochastic Utility Model of Choice," Journal of Econometrics, 3, 205-228.

Manski, C. F. (1985), “Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator,” Journal of Econometrics, 27(3), 313-333.

Manski, C. F. (1990), “Nonparametric bounds on treatment effects,” The American Economic Review, 319-323.

Manski, C. F., (1993), "Identification of Endogenous Social Effects: The Reflection Problem," The Review of Economic Studies, 60, 531-542.

Manski, C. F. (1995), “Identification problems in the social sciences,” Harvard University Press.

Manski, C. F. (2003), “Partial identification of probability distributions,” New York: Springer.

Manski, C. F. (2007), “Partial identification of counterfactual choice probabilities,” International Economic Review, 48(4), 1393-1410.

Manski, C. F. (2013), "Identification of treatment response with social interactions," The Econometrics Journal, 16, S1-S23.

Manski, C. F., Tamer, E. (2002), “Inference on regressions with interval data on a regressor or outcome,” Econometrica, 70(2), 519-546.

Marschak, J., Andrews, W. H. (1944), “Random simultaneous equations and the theory of production,” Econometrica, 12, 143-205.

Marshall, A. (1890). Principles of Economics. 1 (First ed.). London: Macmillan.

Mas-Colell, A. (1978), "On Revealed Preference Analysis," Review of Economic Studies, 45, 121-131.

Matzkin, R.L. (2005),“Identification of consumers’ preferences when individuals’ choices are unobservable,” Economic Theory, 26, 423-443.

Matzkin R.L. (2007),“Nonparametric identification,” In Handbook of Econometrics, Vol 6B, ed JJ Heckman, EE Leamer, pp 5307–5368 Amsterdam: Elsevier

Matzkin RL. (2008), Identification in nonparametric simultaneous equations models. Econometrica 76 945f

Experimentation," Journal Of the Royal Statistical Society, II, 2, 154-180.

Pakes, A., J. Porter, K. Ho, and J. Ishii (2015), "Moment Inequalities and Their Application," Econometrica, 83, 315-334.

Pastorello, S., V. Patilea, and E. Renault (2003), Iterative and Recursive Estimation in Structural Non-adaptive Models," Journal of Business & Economic Statistics, 21:4, 449-509.

Pearl, J. (1988), Probabilistic Reasoning in Intelligent Systems, San Mateo, California: Morgan Kaufmann.

Pearl, J. (2000), Causality: Models, Reasoning, and Inference, Cambridge: Cambridge University Press.

Pearl, J. (2009), "Causal Inference in Statistics: An Overview," Statistics Surveys 2, 96-146.

Pearl, J. (2015), "Trygve Haavelmo and the Emergence of Causal Calculus," Econometric Theory 31, 152–179.

Pearl, J. and D. Mackenzie (2018), The Book of Why: The New Science of Cause and Effect, New York: Basic Books.

Persky, J. (1990), "Retrospectives: Ceteris Paribus," Journal of Economic Perspectives, 4(2), 187–193.

Peterson, A. V. (1976), Bounds for a joint distribution function with fixed sub-distribution functions: Application to competing risks. Proceedings of the National Academy of Sciences, 73(1), 11-13.

Petty, W. (1662), "A Treatise of Taxes and Contributions. Reprinted in Kelley, Augustus M.,and Charles Hull, eds., The Economic Writings of Sir William Petty. London: N. Brooke, 1936

Phillips, P. C. (1983), “Exact small sample theory in the simultaneous equations model,” Handbook of econometrics, 1, 449-516.

Poirier, D. J. (1998), "Revising Beliefs in Nonidentified Models," Econometric Theory, 14, 483-509.

Reiersøl, O. (1941), “Confluence analysis by means of lag moments and other methods of confluence analysis,” Econometrica, 9, 1-24.

Rubin, D. B. (1990), “Formal mode of statistical inference for causal effects,” Journal of Statistical Planning and Inference, 25(3), 279-292.

Rust, J. (1994), "Structural Estimation of Markov Decision Processes", in Handbook of Econometrics, Vol. 4, ed. by R. Engle and D. McFadden, Amsterdam: North Holland, 3081-3143

Sargan, J. D. (1958), "The Estimation of Economic Relationships Using Instrumental Variables". Econometrica 26 (3): 393–415.

Sargan, J. D. (1959), “The estimation of relationships with autocorrelated residuals by the use of instrumental variables,” Journal of the Royal Statistical Society. Series B (Methodological), 91-105.

Sargan, J. D. (1983), “Identification and lack of identification,” Econometrica, 51, 1605-1633.

Samuelson, P. A. (1938), “A note on the pure theory of consumer’s behaviour,” Economica, 61-71.

Samuelson, P. A. (1948), “Consumption theory in terms of revealed preference,” Economica, 243-253.

Scheiber, N. (2007), "Freaks and Geeks; How Freakonomics is Ruining the Dismal Science," New Republic, April 2, 27-31.

Schennach, S. M. (2007), “Instrumental Variable Estimation of Nonlinear Errors in Variables Models,” Econometrica, 75(1), 201-239.

Schennach, S. M., and Hu, Y. (2013), "Nonparametric identification and semiparametric estimation of classical measurement error models without side information," Journal of the American Statistical Association, 108(501), 177-186

Schennach, S. M. (2014), “Entropic Latent Variable Integration via Simulation," Econometrica, 82(1), 345-385.

Sims, C. (1972), "Money, income and causality," American Economic Review 62, 540–52.

Simpson, E. H. (1951), "The Interpretation of Interaction in Contingency Tables," Journal of the Royal Statistical Society, Series B. 13, 238–241.

Splawa-Neyman, J. (192238–241.Samtical1), ocalExp3,mod2(ofTd [(Econometrica)-250(26)ssay(yal)]ingenc)1P

Staiger, D. O., and J. Stock, (1997), “Instrumental Variables Regression With Weak Instruments,” Econometrica, 65, 557-586.

Stock, J. H., Trebbi, F. (2003), “Retrospectives Who Invented Instrumental Variable Regression?,” Journal of Economic Perspectives, 177-194.

Stock, J. H., Wright, J. H., & Yogo, M. (2002), A survey of weak instruments and weak identification in generalized method of moments. Journal of Business & Economic Statistics, 20(4) 518–529.

Tamer, E. (2003), “Incomplete simultaneous discrete response model with multiple equilibria,” The Review of Economic Studies, 70(1), 147-165.

Tamer, E. (2010), “Partial identification in econometrics,” Annual Review of Economics, 2(1), 167-195.

Theil, H. (1953), “Repeated least squares applied to complete equation systems,” The Hague: central planning bureau.

Tinbergen, Jan (1930), “Bestimmung und Deutung von Angebotskurven: Ein Beispiel,” Zietschrift fur Nationalokonomie. 1, pp. 669-79.

Varian, H. R. (1982), “The nonparametric approach to demand analysis,” Econometrica, 50, 945-973.

Varian, H. R. (1983), “Non-parametric tests of consumer behaviour,” The Review of Economic Studies, 50(1), 99-110.

Varian, H. R. (2014), “Big data: New tricks for econometrics,” The Journal of Economic Perspectives, 28(2), 3-27.

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

Wald, A. (1940), "The Fitting of Straight Lines if Both Variables Are Subject to Error". Annals of Mathematical Statistics 11 (3): 284–300.

Wald, A. (1943), "A Method of Estimating Plane Vulnerability Based on Damage of Survivors," Statistical Research Group, Columbia University. CRC 432, reprint from July 1980, Center for Naval Analyses.

Wald, A. (1950), “Statistical decision functions,” John Wiley and Sons, New York; Chapman and Hall,

London.

Working, H. (1925), “The statistical determination of demand curves,” The Quarterly Journal of Economics, 503-543.

Working, Elmer J. (1927), “What Do Statistical Demand Curves Show?” Quarterly Journal of Economics. 41:1, pp. 212-35.

Wright, J. (2003), "Detecting Lack of Identification in GMM," Econometric Theory, 19, 322–330.

Wright, Philip G. (1915), “Moore’s Economic Cycles,” Quarterly Journal of Economics. 29:4, pp.631-641.

Wright, Philip G. (1928), “The Tariff on Animal and Vegetable Oils,” New York: Macmillan.

Wright, Sewall. (1925), “Corn and Hog Correlations,” U.S. Department of Agriculture Bulletin, 1300, pp. 1-60.