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/

Table of Contents

1. Introduction 3

2. Historical Roots of Identification 5

3. Point Identification 8
   3.1 Introduction to Point Identification 9
   3.2 Defining Point Identification 12
   3.3 Examples and Classes of Point Identification 15
   3.4 Proving Point Identification 24
   3.5 Common Reasons for Failure of Point Identification 27
   3.6 Control Variables 30
   3.7 Identification by Functional Form 32
   3.8 Over, Under, and Exact Identification, Rank and Order conditions 36

4. Coherence, Completeness and Reduced Forms 38

5. Causal Reduced Form vs. Structural Model Identification 41
   5.1 Causal or Structural Modeling? Do Both 43
   5.2 Causal vs. Structural Identification: An Example 46
   5.3 Causal vs. Structural Simultaneous Systems 56
   5.4 Causal vs. Structural Conclusions 58

6. Identification of Functions and Sets 61
   6.1 Nonparametric and Semiparametric Identification 62
   6.2 Set Identification 65
   6.3 Normalizations in Identification 68
   6.4 Examples: Some Special Regressor Models 73

7. Limited Forms of Identification 77
   7.1 Local and Global Identification 77
   7.2 Generic Identification 80

8. Identification Concepts that Affect Inference 82
   8.1 Weak vs. Strong Identification 82
   8.2 Identification at Infinity or Zero; Irregular and Thin set identification 86
   8.3 Ill-Posed Identification 88
   8.4 Bayesian and Essential Identification 90

9. Conclusions 91

10. Appendix: Point Identification Details 92

11. References 95

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 θ 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 θ 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 θ from observable data. Roughly, identification asks, if we knew the population that data are drawn from, would θ be known? And if not, what could be learned about θ?

The study of identification logically precedes estimation, inference, and testing. For θ to be identified, alternative values of θ must imply different distributions of the observable data (see, e.g., Matzkin 2013). This implies that if θ is not identified, then we cannot hope to find a consistent estimator for θ. 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 also includes a discussion of typical sources of non-identification, and of some traditional identification related concepts like overidentification, exact identification, and rank and order conditions. Identification by functional form is described, and examples are provided, including constructed instruments based on second and higher moment assumptions. Appropriate use of such methods is discussed.

Next is Section 4, which defines and discusses the concepts of coherence and completeness of models. These are closely associated with existence of a reduced form, which in turn is often used as a starting point for proving identification. This is followed by Section 5, which is devoted to discussing identification concepts in what is variously known as the reduced form, or program evaluation, or treatment effects, or causal inference literature. This literature places a particular emphasis on randomization, and is devoted to the identification of parameters that can be given a causal interpretation. Typical methods and assumptions used to obtain identification in this literature are compared to identification of more traditional structural models. To facilitate this comparison, the assumptions of the popular local average treatment effect (LATE) causal model, which are usually described in potential outcome notation, are here rewritten using a traditional structural notation. The relative advantages and disadvantages of randomization based causal inference methods vs. structural modeling are laid out, and a case is made for combining both approaches in practice.

Section 6 describes nonparametric identification, semiparametric identification, and set identification. This section also discusses the related role of normalizations in identification analyses, which has not been analyzed in previous surveys. Special regressor methods are then described, mainly to provide examples of these concepts.

Section 7 describes limited forms of identification, in particular, local identification and generic identification. Section 8 considers forms of identification that have implications for, or are related to, statistical inference. These include weak identification, identification at infinity, ill-posed identification, and Bayesian identification. Section 9 then concludes, and an Appendix provides some additional mathematical details.

2 Historical Roots of Identification

Before discussing identification in detail, consider some historical context. I include first names of early authors in this section to promote greater knowledge of the early leaders in this field.

Before we can think about isolating, and thereby identifying, the effect of one variable on another, we

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.[3] Sewall Wright (1925) applied those methods to construct an instrumental variables estimator, but in a model of exogenous regressors that could have been identified and estimated by ordinary least squares. The idea of using instrumental variables to solve the identification problem arising from simultaneous systems of equations first appears in Appendix B of Philip Wright (1928). Stock and Trebbi (2003) claim that this is the earliest known solution to an identification problem in econometrics. They apply a stylometric analysis (the statistical analysis of literary styles) to conclude that Philip Wright was the one who actually wrote Appendix B, using his son's estimator to solve their identification problem.

In addition to two different Wrights, two different Workings also published early papers relating to the subject: Holbrook Working (1925) and, more relevantly, Elmer J. Working (1927). Both wrote about statistical demand curves, though Holbrook is the one for whom the Working-Leser Engel curve is named.

Jan Tinbergen (1930) proposed indirect least squares estimation (numerically recovering structural parameters from linear regression reduced form estimates), but does not appear to have recognized its usefulness for solving the identification problem.

[2]Petty's (1662) use of the term ceteris paribus gives what could be construed as an early identification argument, identifying a determinant of prices. On page 50 of his treatise he writes, "If a man can bring to London an ounce of Silver out of the Earth in Peru, in the same time that he can produce a bushel of Corn, then one is the natural price of the other; now if by reason of new and more easie Mines a man can get two ounces of Silver as easily as formerly he did one, then Corn will be as cheap at ten shillings the bushel, as it was before at five shillings caeteris paribus."

[3]Sewall's first application of causal paths was establishing the extent to which fur color in guinea pigs was determined by developmental vs genetic factors. See, e.g., Pearl (2018). So while the father Philip considered pig iron, the son Sewall studied actual pigs.

The above examples, along with the later analyses of Trygve Haavelmo (1943), Tjalling Koopmans (1949), Theodore W. Anderson and Herman Rubin (1949), Koopmans and Olav Reiersøl (1950), Leonid Hurwicz (1950), Koopmans, Rubin, and Roy B. Leipnik (1950), and the work of the Cowles Foundation more generally, are concerned with identification arising from simultaneity in supply and demand. Other important early work on this problem includes Abraham Wald (1950), Henri Theil (1953), J. Denis Sargan (1958), and results summarized and extended in Franklin Fisher's (1966) book. Most of this work emphasizes exclusion restrictions for solving identification in simultaneous systems, but identification could also come from restrictions on the covariance matrix of error terms, or combinations of the two, as in Karl G. Jöreskog (1970). Milton Friedman's (1953) essay on positive economics includes a critique of the Cowles foundation work, essentially warning against using different criteria to select models versus criteria to identify them.

A standard identification problem in the statistics literature is that of recovering a treatment effect. Derived from earlier probability theory, identification based on randomization was developed in this literature by Jerzy Splawa-Neyman (1923)[4], David R. Cox (1958), and Donald B. Rubin (1974), among many others. Pearl (2015) and Heckman and Pinto (2015) credit Haavelmo (1943) as the first rigorous treatment of causality in the context of structural econometric models. Unlike the results in the statistics literature, econometricians historically focused more on cases where selection (determining who is treated or observed) and outcomes may be correlated. These correlations could come from a variety of sources, such as simultaneity as in Haavelmo (1943), or optimizing self selection as in Andrew D. Roy (1951). Another example is Wald's (1943) survivorship bias analysis (regarding airplanes in world war II), which recognizes that even when treatment assignment (where a plane was hit) is random, sample attrition that is correlated with outcomes (only planes that survived attack could be observed) drastically affects the correct analysis. General models where selection and outcomes are correlated follow from James J. Heckman (1978). Causal diagrams (invented by Sewall Wright as discussed above) were promoted by Judea Pearl (1988) to model the connections between treatments and outcomes.

[4]Neyman's birth name was Splawa-Neyman, and he published a few of his early papers under than name, including this one.

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 θ 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 ϕ, that we either already know or could learn from data. Think of ϕ as everything that could be learned about the population that data are drawn from. Usually, ϕ would either be a distribution function, or some features of distributions like conditional means, quantiles, autocovariances, or regression coefficients. In short, ϕ 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 ϕ (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 ϕ could take on. A simple definition of (point) identification is then that a parameter θ is point identified if, given the model, θ is uniquely determined from ϕ.

For example, suppose for scalars Y, X, and θ, our model is that $Y = X\theta + e$ where $E\left(X^2\right) \neq 0$ and $E\left(eX\right) = 0$, and suppose that ϕ, what we can learn from data, includes the second moments of the vector (Y, X). Then we can conclude that θ is point identified, because it is uniquely determined in the usual linear regression way by $\theta = E\left(XY\right)/E\left(X^2\right)$, 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 ϕ, what we can learn from data, includes $E\left(Y \mid X\right)$. It then follows that the average treatment effect θ is identified because, when treatment is randomly assigned, $\theta = E\left(Y \mid X = 1\right) - E\left(Y \mid X = 0\right)$, 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

be included in ϕ. Since sample averages can be observed, to justify this assumption we might appeal to the consistency of sample averages, given conditions for a weak law of large numbers.

When discussing empirical work, a common question is, "what is the source of the identification?" that is, what feature of the data is providing the information needed to determine θ? This is essentially asking, what needs to be in ϕ?

Note that the definition of identification is somewhat circular or recursive. We start by assuming some information ϕ is knowable. Essentially, this means that to define identification of something, θ, we start by assuming something else, ϕ, is itself identified. Assuming ϕ is knowable, or identified, to begin with can itself only be justified by some deeper assumptions regarding the underlying DGP (Data Generating Process).

We usually think of a model as a set of equations describing behavior. But more generally, a model is whatever set of assumptions we make about, and restrictions we place on, the DGP. This includes both assumptions about the behavior that generates the data, and about how the data are collected and measured. These assumptions in turn imply restrictions on ϕ and θ. In this sense, identification (even in purely experimental settings) *always* requires a model.

A common starting assumption is that the DGP consists of n IID (Independently, Identically Distributed) observations of a vector W, where the sample size n goes to infinity. We know (by the Glivenko–Cantelli theorem, see Section 3.4 below) that with this kind of data we could consistently estimate the distribution of W. It is therefore reasonable with IID data in mind to start by assuming that what is knowable to begin with, ϕ, is the distribution function of W.

Another common DGP is where each data point consists of a value of X chosen from its support, and conditional upon that value of X, we randomly draw an observation of Y, independent from the other draws of Y given X. For example, X could be the temperature at which you choose to run an experiment, and Y is the outcome of the experiment. As $n \to \infty$ this DGP allows us to consistently estimate and thereby learn about $F(Y \mid X)$, the conditional distribution function of Y given X. So if we have this kind of DGP in mind, we could start an identification proof for some θ by assuming that $F(Y \mid X)$ is knowable.

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 of a million degrees).

With more complicated DGPs (e.g., time series data, or cross section data containing social interactions or common shocks), part of the challenge in establishing identification is characterizing what information ϕ is knowable, and hence is appropriate to use as the starting point for proving identification. For example, in a time series analyses we might start by supposing that the mean, variance, and autocovariances of a time series are knowable, but not assume information about higher moments is available. Why not? Either because higher moments might not be needed for identification (as in vector autoregression models), or because higher moments may not be stable over time.

Other possible examples are that ϕ could equal reduced form linear regression coefficients, or, if observations of W follow a martingale process, ϕ could consist of transition probabilities.

What to include in ϕ depends on the model. For example, in dynamic panel data models, the Arellano and Bond (1991) estimator is based on a set of moments that are assumed to be knowable (since they can be estimated from data) and equal zero in the population. The parameters of the model are identified if they are uniquely determined by the equations that set those moments equal to zero. The Blundell and Bond (1998) estimator provides additional moments (assuming functional form information about the initial time period zero distribution of data) that we could include in ϕ. We may therefore have model parameters that are not identified with Arellano and Bond moments, but become identified if we are willing to assume the model contains the additional information needed for Blundell and Bond moments.

Even in the most seemingly straightforward situations, such as experimental design with completely random assignment into treatment and control groups, additional assumptions regarding the DGP (and hence regarding the model and ϕ) are required for identification of treatment effects. Typical assumptions that are routinely made (and may often be violated) in this literature are assumptions that rule out certain types of measurement errors, sample attrition, censoring, social interactions, and general equilibrium effects.

In practice, it is often useful to distinguish between two types of DGP assumptions. One is assumptions

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 ϕ is a distribution function. In contrast, the traditional textbook discussion of identification of linear supply and demand curves corresponds to the special case where ϕ 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 ϕ to be a set of constants and/or functions about the DGP that we assume are known, or

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

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

The set of parameters θ 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 θ 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 θ, but they could be.

We assume that each particular value of m implies a particular value of ϕ and of θ (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 ϕ or the same θ. Define the *structure* $s(\phi, \theta)$ to be the set of all model values m that yield both the given values of ϕ and of θ.

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

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

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

if θ_0 not observationally equivalent to any other θ in Θ. However, we don't know which of the possible values of θ (that is, which of the elements of Θ) is the true θ_0. This is why, to ensure point identification, we generally require that no two elements θ and $\widetilde{\theta}$ in the set Θ having $\theta \neq \widetilde{\theta}$ be observationally equivalent. Sometimes this condition is called *global identification* rather than point identification, to explicitly say that θ_0 is point identified no matter what value in Θ turns out to be θ_0.

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 $\widetilde{m}$ in M are observationally equivalent (treating m and $\widetilde{m}$ as if they were parameters). Since every model value is associated with at most one value of θ, having the model be identified is sufficient, but stronger than necessary, to also have any possible set of parameters θ be identified.

The economist or econometrician defines the model M, so we could in theory enumerate every $m \in M$, list every value of ϕ and θ that is implied by each m, and thereby check every pair $s(\phi, \theta)$ and $s\left(\phi, \widetilde{\theta}\right)$ to see if θ 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 θ is the true one, proving identification in practice requires showing that the definition holds for any possible θ, 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 generic identification deal with cases where we can't establish point identification for all θ in Θ. *Local identification* of θ_0 means that there exists a neighborhood of θ_0 such that, for all values $\theta \neq \theta_0$ in this neighborhood, θ is not observationally equivalent to θ_0. As with point identification, since we don't know ahead of time which value of θ is θ_0, to prove that local identification holds we would need to show that for any $\widetilde{\theta} \in \Theta$, there exists a neighborhood of $\widetilde{\theta}$ such that, for any $\theta \neq \widetilde{\theta}$ in this neighborhood, θ is not observationally equivalent to $\widetilde{\theta}$.

Generic identification roughly means that the set of values of θ in Θ that cannot be point identified is a very small subset of Θ. Suppose we took all the values of θ in Θ, and divided them into two groups:

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 θ is defined to be *generically identified* if the first group is extremely small (formally, if the first group is a measure zero subset of Θ). Both local and generic identification are discussed in more detail later.

The true parameter value θ_0 is said to be *set identified* (sometimes also called *partially identified*) if there exist some values of $\theta \in \Theta$ that are not observationally equivalent to θ_0. So the only time a parameter θ is not set identified is when all $\theta \in \Theta$ are observationally equivalent. For set identified parameters, the *identified set* is defined to be the set of all values of $\theta \in \Theta$ that are observationally equivalent to θ_0. Point identification of θ_0 is therefore the special case of set identification in which the identified set contains only one element, which is θ_0.

Parametric identification is where θ is a finite set of constants, and all the different possible values of ϕ also correspond to different values of a finite set of constants. *Nonparametric identification* is where θ consists of functions or infinite sets. Other cases are called *semiparametric identification*, which includes situations where, e.g., θ includes both a vector of constants and nuisance parameters that are functions. As we will see in Section 6, sometimes the differences between parametric, semiparametric, and nonparametric identification can be somewhat arbitrary (see Powell 1994 for further discussion of this point).

3.3 Examples and Classes of Point Identification

Consider some examples to illustrate the basic idea of point identification.

Example 1: a median. Let the model M be the set of all possible distributions of a random variable W having a strictly monotonically increasing distribution function. Our DGP (data generating process) consists of IID (independent, identically distributed) draws of W. From this DGP, what is knowable is $F(w)$, the distribution function of W. Let our parameter θ be the median of W. In this simple example, we know θ is identified because it's the unique solution to $F(\theta) = 1/2$. By knowing F, we can determine

θ.

How does this example fit the general definition of identification? Here each value of ϕ is a particular continuous, monotonically increasing distribution function F. In this example, each model value m happens to correspond to a unique value of ϕ, because each possible distribution of W has a unique distribution function. In this example, for any given candidate value of ϕ and θ, the structure $s(\phi, \theta)$ is either an empty set or it has one element. For a given value of ϕ and θ, if $\phi = F$ and $F(\theta) = 1/2$ (the definition of a median) then set $s(\phi, \theta)$ contains one element. That element m is the distribution that has distribution function F. Otherwise, if $\phi = F$ where $F(\theta) \neq 1/2$, the set $s(\phi, \theta)$ is empty. In this example, it's not possible to have two different parameter values θ and $\widetilde{\theta}$ be observationally equivalent, because $F(\theta) = 1/2$ and $F\left(\widetilde{\theta}\right) = 1/2$ implies $\theta = \widetilde{\theta}$ for any continuous, monotonically increasing function F. Therefore θ is point identified, because its true value θ_0 cannot be observationally equivalent to any other value θ.

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 ϕ 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 $Y = X'\theta + e$, where θ is some $K-$vector of parameters, e is an error term, $E(Xe) = 0$ for an error term e, and where e, X has finite first and second moments. The structure $s(\phi, \theta)$ is nonempty when the moments comprising ϕ satisfy $E\left[X\left(Y - X'\theta\right)\right] = 0$ for the given θ. To ensure point identification, we could add the additional restriction on M that $E\left(XX'\right)$ is nonsingular, because then θ would be uniquely determined in the usual way by $\theta = E\left(XX'\right)^{-1} E(XY)$. However, if we do not add this additional restriction, then we can find values $\widetilde{\theta}$ that are observationally equivalent to θ by letting $\widetilde{\theta} = E\left(XX'\right)^{-} E(XY)$ for different pseudoinverses $E\left(XX'\right)^{-}$.

As described here, identification of θ is parametric, because θ is a vector, and ϕ can be written as a vector of moments. However, some authors claim linear regression as being semiparametric, because it includes errors e that have an unknown distribution function. This distinction depends on how we

define ϕ and θ. For example, suppose we had IID observations of Y, X. We could then have defined ϕ to be the joint distribution function of Y, X, and defined θ 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 $E\left(XX'\right)$ is nonsingular, we would then have semiparametric identification of θ.

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 θ be the average treatment effect (ATE), defined by $\theta = E(Y(1) - Y(0))$. 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 ϕ from this DGP is the joint distribution of Y and T. Given unconfoundedness, θ is identified because unconfoundedness implies that $\theta = E(Y \mid T = 1) - E(Y \mid T = 0)$, which is uniquely determined from ϕ. Heckman, Ichimura, and Todd (1998) note that a weaker sufficient condition for identification of θ 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 θ 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 ϕ, but have different values for θ. 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 θ, but whether there exists a unique value of θ corresponding to every possible ϕ.

These constructions can all be generalized to allow for covariates, e.g., unconfoundedness can more

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 θ is then identified by $\theta = E(E(Y \mid T = 1, X) - E(Y \mid T = 0, X))$, 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 ϕ could be defined as the vector (ϕ_1, ϕ_2) of reduced form coefficients $Y = \phi_1 Z + V_1$ and $X = \phi_2 Z + V_2$ where V_1 and V_2 are mean zero, independent of Z. Suppose $\theta = a$, 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 $\phi_1 = ac/(a - b)$ and $\phi_2 = c/(a - b)$.

In this example, what model values m comprise a given structure $s(\phi, \theta)$? 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 $s(\phi, \theta)$ are the ones that satisfy the equations $\theta = a$, $\phi_1 = ac/(a - b)$, and $\phi_2 = c/(a - b)$. Note in particular that, if $c \neq 0$, then $\phi_1/\phi_2 = a$, so $s(\phi, \theta)$ is empty if $c \neq 0$ and $\phi_1/\phi_2 \neq \theta$. Whenever $s(\phi, \theta)$ 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 ϕ and θ.

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

θ and $\widetilde{\theta}$ will be observationally equivalent when $\phi = (0, 0)$. 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 $\phi_2 \neq 0$. If we include in the definition of the model that $c \neq 0$, then θ is identified.

Example 5: latent error distribution. Suppose the DGP is IID observations of scalar random variables Y, X, so ϕ 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 $Y = I(X + U > 0)$, which means that Y is one if $X + U$ is positive and zero otherwise. Consider identification of the function $F_U(u)$, the distribution function of U, so in this example θ is a function, and the identification is nonparametric.

For any value x that the random variable X can take on, we have $E(Y \mid X = x) = \Pr(X + U > 0 \mid X = x) = \Pr(x + U > 0) = 1 - \Pr(U \leq -x) = 1 - F_U(-x)$. This shows that the function $F_U(u)$ is identified for $u = -x$, since it can be recovered from the function $E(Y \mid X = x)$, and that function is known given ϕ. This is the logic behind the identification of Lewbel's (2000) special regressor estimator, which is discussed further in Section 4. Note however that the function $F_U(u)$ might not be identified everywhere. It is only identified for values of u that are in the support of $-X$.

In examples 1, 2, and 5 above, data are assumed to be IID observations of some vector we can call W, and therefore what we start by assuming is knowable, ϕ, is the distribution function of W. But in other contexts like time series data, or in social interactions models, it is generally not reasonable to assume the distribution function is known, because the dimension of the joint distribution function grows at the same rate that the sample size grows. In such contexts it is more plausible to start by assuming that only some stable features of the joint distribution of data are knowable. For example, if the data consists of observations of a vector W in different time periods, with a distribution that potentially changes every period, it might be reasonable to assume W follows a martingale process, and then ϕ might consist of transition probabilities. Or for structures that correspond to linear stationary times series models, ϕ might be autocovariances at all lag lengths. In examples 3 and 4 above, ϕ consists of the true values of linear

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 ϕ is a set of reduced form regression coefficients, or ϕ is a data distribution, or ϕ 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 ϕ to be a matrix of population reduced form linear regression coefficients, that is, ϕ denotes a possible value of the set of coefficients that are obtained from a linear projection of Y onto X.

The model M is a set of linear structural equations. The restrictions that define the model include exclusion assumptions, e.g., an element of X that is known to be in the demand equation, but is excluded from the supply equation, and therefore serves as an instrument for price in the demand equation.

Let θ be a set of structural model coefficients we wish to identify. For example, θ could be the coefficients of one equation (say, the demand equation in a supply and demand system), or θ could be all the coefficients in the structural model, or θ could just be a single price coefficient. More generally, θ could be some function of coefficients, like an elasticity with respect to X in a model that is linear in functions of X.

For a given value of θ and ϕ, the structure $s(\phi, \theta)$ is the set of all model values having structural coefficients (or functions of structural coefficients) θ and reduced form coefficients ϕ. Recall that point identification of θ_0 requires that there not exist any ϕ, θ, and model value m such that m is in both $s(\phi, \theta_0)$ and $s(\phi, \theta)$ where $\theta \neq \theta_0$. This means there can't be any $\theta \neq \theta_0$ that satisfies the restrictions of the model and has the same matrix of reduced form coefficients ϕ that the true θ_0 might have. We say "might have"

here because there could be more than one reduced form matrix value that is consistent with the true θ_0. For example, if θ_0 is just the coefficient of price in the supply equation, there could be many possible reduced form coefficient matrices ϕ, 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 ϕ.

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 ϕ 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 (2007, 2012). It is derived from Koopmans and Reiersøl (1950), Hurwicz (1950), Fisher (1966), and Rothenberg (1971). These authors defined notions of a structure, observational equivalence, and identification with varying degrees of explicitness. In distribution based identification, what is assumed to be knowable, ϕ, is the distribution function of an observable random vector Y (or the conditional distribution function of Y given a vector X). In these earlier references, the restrictions of the model implied that ϕ was some known parametric family of distribution functions, and so had a known functional form. In that case model parameters could typically be estimated by maximum likelihood estimation. However, one might more generally define the model by only imposing nonpara-

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

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

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: Following Sargan (1959, 1983) and Amemiya (1985), Newey and McFadden (1994) define a notion of identification that, by analogy with extremum estimators, we will call extremum based identification. Extremum estimators are estimators that maximize an objective function, such as GMM (generalized method of moments) or least squares estimation. For extremum based identification, each model value m is associated with the value of a function G. What is assumed to be knowable, ϕ, is the value (or set of values) of vectors or functions ζ that maximize $G(\zeta)$. The parameters θ would typically be subsets of or functions of elements of ζ, but could more generally be features of G or of the underlying DGP that generates values of G.

To see the connection between extremum based identification and estimation, consider the example of extremum estimators that maximize an average with IID data. Let ζ be a vector of parameters. Assume $\widehat{\phi}$ equals the set of all ζ that maximize $\sum_{i=1}^{n} g(W_i, \zeta)/n$ where g is a known function, and IID W_i are observations of an observable data vector. If $-g(W_i, \zeta)$ is a squared error term for observation i, then this would be a linear or nonlinear least squares estimator, while if g is the log of the probability

density or mass function of W_i, then this could be a maximum likelihood estimator. Now define G by $G(\zeta) = E(g(W_i, \zeta))$. More generally, $G(\zeta)$ 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 ζ that maximize $G(\zeta)$.

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 $G(\zeta)$ has a unique maximum ζ_0, and that θ_0 equals a known function of (or subset of) ζ_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 are knowable, which is all we need for extremum based identification.

A key feature of extremum based identification is that, for the purpose of identification, nothing about the DGP is considered to be known except for maximizing values of the objective function G. This is an advantage in that extremum based identification is what is needed to establish consistency and other asymptotic properties of any given extremum based estimator. But this key feature also has the drawback of not saying anything about whether θ could have been identified from other features of the underlying DGP that might be knowable. For example, suppose the DGP is IID observations w_i of a bounded scalar random variable W and $\theta_0 = E(W)$. Then, applying distribution based identification, we would conclude that that θ_0 is identified. However, suppose we had considered extremum based identification where the

model consists of G functions defined by $G(\zeta) = -E\left[(W - |\zeta|)^2\right]$. The set of arguments that maximize this G is the set $\phi = \{\theta_0, -\theta_0\}$. As a result, θ_0 and $-\theta_0$ are observationally equivalent, and therefore θ is not identified using this G. Of course, in this case we failed to identify θ only because the objective function failed to account for other information that would have been knowable given IID data. This is precisely the limitation of extremum based identification. Failure of extremum based identification can be due either to some fundamental source of nonidentification in the DGP, or due to the particular choice of objective function.

This problem typically does not apply to parametric distribution based identification, because, under regularity conditions, the likelihood function contains all of the information about parameters that is available in the population. However, this issue can arise in Wright-Cowles identification. By defining ϕ just in terms of features of first and second moments, Wright-Cowles has the extremum based identification problem of ignoring additional information in the DGP that might sometimes be both useful and available for identification. In particular, third and higher moments are generally available in linear models, and can sometimes be used to obtain identification even when Wright-Cowles type exclusion restrictions are not available. One simple example is Lewbel (1997b), which uses some information in third moments to obtain identification in models containing mismeasured covariates without instruments. Other examples are given later in section 3.7, on identification by functional form.

Wright-Cowles, distribution, and extremum based identification are all examples of point identification. They differ only in what they regard as the knowable information ϕ in the DGP.

3.4 Demonstrating Point Identification

How can we show that a given set of parameters θ are point identified? In each of the five specific examples given in the previous section, we used "identification by construction." Identification by construction means that we can write a closed form expression for θ as a function of ϕ. In example 1 we had $\theta = F^{-1}(1/2)$, in example 2 we had $\theta = E\left(XX'\right)^{-1} E(XY)$, and in 3 we had $\theta = E(Y \mid X = 1) - E(Y \mid X = 0)$. Even when θ is a function rather than a vector, identification is often proved by construc-

tion, as in example 5 where we had $\theta = F_U(u) = 1 - E(Y \mid X = -u)$. While identification by construction is the commonest way to prove that parameters are identified, it is not necessary for.identification that such closed form expressions exist.

In the statistics literature, it is common to not consider identification separately from estimation, but to instead just directly prove that a proposed estimator is consistent. Suppose you can find an estimator $\widehat{\theta}$ and prove that, under the assumed DGP, $\widehat{\theta}$ is consistent. Then that also proves θ is identified by construction, where the construction is that θ equals the probability limit of $\widehat{\theta}$. For example, suppose $\theta = E(W)$ and you want to prove that θ is identified. One way you could do so would be to let $\widehat{\theta}$ be the sample mean of W, and show that a weak law of large numbers applies to your given DGP. Then θ would be identified by the construction $\theta = plim\left(\widehat{\theta}\right)$. One caution in applying this method is that some proofs of consistency either implicitly or explicitly assume identification. For example, Theorem 2.1 of Newey and McFadden (1994), which proves the consistency of extremum estimators, includes extremum based identification as one of its assumptions. This theorem could therefore not be used to prove identification, because even though it shows that $\theta = plim\left(\widehat{\theta}\right)$, it assumes identification to do so.

An important example of a direct consistency proof is the Glivenko–Cantelli theorem. Assume the DGP is IID observations of a vector W. Let $F(w)$ be the distribution function of W, evaluated at the value w. The empirical distribution function $\widehat{F}$ is the estimator of the function F defined as $\widehat{F}(w) = \sum_{i=1}^{n} I(W_i \leq w)/n$. This is nothing more than estimating the probability that $W \leq w$ by counting how many observations of W_i in the sample are less than or equal to w, and dividing by the sample size. The Glivenko–Cantelli theorem says that as long as our observations are IID, $\widehat{F}(w)$ is a uniformly consistent estimator of $F(w)$. This is the formal justification for our frequent starting assumption that, with IID data, what is knowable, ϕ, is a distribution function. Similarly, with non-IID data, the weak law of large numbers can justify assuming that means are knowable.

Although it is sometimes possible, as in the previous paragraph, to prove identification and estimator consistency simultaneously, analysis of identification logically precedes estimation, and identification does not by itself imply that estimators with any particular desired properties exist. For example, with IID

observations of a random variable X that has a finite mean, we might desire an estimator for $\theta = E(X)$ that converges in mean square, but if X has sufficiently thick tails then no such estimator may exist. Ill-conditioned identification and non-robust identification (discussed later) are two other situations where, despite being point identified, any estimator of θ will have some undesirable properties.

Another general method of showing that a parameter is identified is to prove that the true θ_0 is the unique solution to some maximization problem defined by the model. For example, if one can show that a likelihood function is globally concave, then maximum likelihood will have a unique maximizing value, and identification can then be established by showing that the unique maximizer in the population equals the true θ_0. Identification of the parameters in a standard probit or logit model can be established in this way (see Haberman 1974).

Based on earlier work by Roehrig (1988), Matzkin (2008) provides a theorem that, for a large class of regular simultaneous equations models, characterizes when two structures are observationally equivalent, and so provides a useful tool for establishing identification. Yet another technique for proving identification is to show that θ_0 is the unique fixed point in a contraction mapping based on the model. This method is employed, e.g., by Berry, Levinsohn, and Pakes (1995) in what is known as the BLP model. They do not formally prove identification of the BLP model, but they do establish that an important necessary condition for identification holds by showing that their error term inversion is a contraction mapping. Pastorello, Patilea, and Renault (2003) use a fixed point extremum based identification assumption for their latent backfitting estimator.

Examples of identification proofs that apply many of the above techniques can be found in Matzkin (2005, 2007, 2012). In some cases, it is possible to empirically test for identification. These are generally tests of extremum based identification, in that they are based on the behavior of associated extremum based estimators. Examples include Cragg and Donald (1993), Wright (2003), Inoue and Rossi (2011), Bravo, Escanciano, and Otsu (2012), and Arellano, Hansen, and Sentana (2012).

An interesting property of point identification is that it can be applied without reference to any data at all. One could for example propose a theoretical model in which some functions ϕ are assumed to

be known, and then ask whether there exists a unique value of vectors or functions θ that satisfy the restrictions defined by the structure $s(\phi, \theta)$. 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 θ are point identified from demand functions ϕ. 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 ϕ are 'knowable,' which essentially means we assume ϕ is identified. Then given the assumed identification of ϕ, we define identification of a set of other parameters θ. Equivalently, identification of parameters θ can only be proven conditional on already knowing that something else, ϕ, is identified. For example, the Samuelson, Houthakker, and Mas-Colell theorems say that, given revealed preference assumptions, if demand functions ϕ are identified, then indifference curves θ 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 θ 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

the coefficients in a linear regression like $Y_i = a + bX_i + cZ_i + e_i$ when X_i is itself linear in Z_i. A similar problem arises in nonparametric regression, in that one usually cannot nonparametrically identify the function $g(X, Z) = E(Y \mid X, Z)$ when the observed X is itself a deterministic function of Z, i.e., when $X = h(Z)$ for some function h.

Nonlinearity can cause nonidentification by allowing equations to have multiple solutions. For example, suppose we have the model $Y = (X - \theta)^2 + e$ with $E(e) = 0$. Then all we may know about the true value of θ is that it satisfies $E\left(Y - (X - \theta_0)^2\right) = 0$ and therefore the true θ equals one of the two roots of the equation $E\left(Y - X^2\right) + 2E(X)\theta - \theta^2 = 0$. Without more information, no amount of data can tell us which root is the true θ, and therefore θ is not point identified (it is, however, set identified, with the identified set consisting of the two roots). An example of more information that could yield point identification might be knowing the sign of θ. See also Brown (1983).

Simultaneity is the familiar source of nonidentification that arises from X and Y being determined jointly or simultaneously, as in the case of price and quantity in a market. This is the previously discussed issue in Wright-Cowles identification. Recall example 4 in section 3.3, where we had the supply curve $Y = aX + \varepsilon$ and a demand curve $Y = bX + cZ + U$ where Y is log quantity, X is log price, Z is log income, and ε and U are error terms with $E[(\varepsilon, U) \mid Z] = 0$. For simplicity, assume all these variables have mean zero. This allows us to simplify expressions by leaving out constant terms (nothing essential would change here by including constants). Multiplying the supply and demand equations by Z and taking expectations gives $E(ZY) = aE(ZX)$ and $E(ZY) = bE(ZX) + cE\left(Z^2\right)$. These show that a is identified by $a = E(ZY)/E(ZX)$ as long as $E(ZX) \neq 0$, and that $E(ZX) = E\left(Z^2\right)c/(a - b)$.

The supply equation is identified using the instrument Z, but without more information, b and c are not identified. The demand curve is observationally equivalent to $Y = \widetilde{b}X + \widetilde{c}Z + \widetilde{U}$ where, for any constant λ, $\widetilde{b} = \lambda b + (1 - \lambda)a$, $\widetilde{c} = \lambda c$, and $\widetilde{U} = \lambda U + (1 - \lambda)\varepsilon$. The standard graphical interpretation is that variation in Z moves the demand curve, thereby causing the intersection of the two curves at various values of Z to trace out the supply curve, but no information is provided about the slope of the demand curve. Essentially, only one point on the demand curve is observed.

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 treated as unobserved. Examples are an individual's information set in a game or in a dynamic optimization problem, and individual's bargaining power within a household. Many identification theorems employ combinations of assumptions and observable data to identify functions of unobservable (or unobserved) variables. One example is the use of identifiable compensating variation and equivalent variation to bound (set identify) unobserved true consumer surplus. Another is the use of unconfoundedness to overcome

unobservability of counterfactuals in the identification of average treatment effects.

3.6 Control Variables

"I controlled for that." This is perhaps the commonest response to a potential identification question in econometric modeling (particularly in simple regressions and in Difference-in-Difference analyses). Consider a model where θ is (or includes) the effect of one variable X on another variable Y. The concern is that the effect of X on Y identified by the model does not equal the desired estimand θ, because of the presence of some other so-called "confounding" connection between X and Y.

The solution of adding a control variable refers to the inclusion of another variable Z in the model to fix this problem. For example, in a study of how physical exercise X affects weight gain Y, one would want to include age of the participants as a control Z. Another example is the "parallel trends" assumption that underlies difference-in-difference models. Parallel trends is equivalent to assuming that inclusion of time and group dummies Z in the model completely controls for all confounding relationships between X and Y other than the desired causal treatment effect.

The idea of a control variable is that Z explains the confounding relationship between X and Y, so by putting Z in the model we can statistically hold Z fixed, and thereby "control" for the alternative, unintended connection between X and Y. Fixing Z is assumed to enforce the ceteris paribus condition mentioned in section 2. Alternatives to adding control variables include various forms of matching, randomization, and data stratification.

There are two reasons why simply including covariates intended to act as controls may not fix these identification problems, and indeed can potentially make them worse. The first reason is functional form. Unless we have a structural model of how Z affects Y, one should include Z in the model in a highly flexible (ideally nonparametric) way. The alternative is accepting that the model might be misspecified. In particular, the common practice of including controls additively and linearly (i.e., as additional regressors in a linear regression) is a strong structural modeling assumption, even if the model is intended to be a

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). One might think to include occupation Z in the regression, to control for the fact that women may choose different occupations from men, and that occupation affects wages. However Z is endogenous in this case, in that the wages that are offered to men and to women affect their occupation choice. So unless we can properly instrument for Z in the regression, including Z in the model will still not yield a causal effect of X on Y.

In the causal diagram literature (see, e.g., Pearl 2000, 2009), a distinction is made between "confounders" and "colliders." A confounder is an exogenous variable that, if observed, can be included in the model as a control. In contrast, occupation Z in the above example is a collider, and, because of the nature of its endogeneity, it cannot be simply included in the model as a control.

A similar argument applies to difference-in-difference models. The dummies and other covariates included in these models are intended to act as controls for all sorts of potentially endogenous group and time related effects. But if any of these dummies or covariates are colliders (or highly correlate with colliders), the causal interpretation of the difference-in-difference estimand may be lost.

These issues with potential controls are closely related to the Berkson (1946) and Simpson (1951) paradoxes in statistics. The bottom line is that either implicit or explicit consideration of underlying structure is needed to convincingly argue that covariates included in the model as controls will actually function as they are intended. This is true even in reduced form analyses like difference-in-difference models.

[6] An exception is so-called "saturated" models. If the controls are binary, and one includes all possible interactions of the controls with each other and with X, then the resulting model is equivalent to being nonparametric in the controls.

3.7 Identification by Functional Form

Identification by functional form is when identification holds when we assume some functions in the model have specific parametric or semiparametric forms, but where identification may fail to hold without these parametric or semiparametric restrictions. An example of identification by functional form is given by the Heckman (1978) selection model. That model consists of the two equation system $Y = \left(b'X + U\right) D$ and $D = I\left(a'X + \varepsilon \geq 0\right)$ where the data are observations of Y, D, and a vector X. This model says that the dummy variable D is one when $a'X + \varepsilon$ is positive, and Y is observed and equal to $b'X + U$ when D is one, otherwise Y is unobserved and so we just set it equal to zero. Assume errors U and ε are independent of X. In this model, the coefficient vector b can be identified if the coefficient (an element of the vector b) of a continuous element of X is known to equal zero. This is an exclusion restriction, like the kind used to get instruments in Wright-Cowles identification of supply and demand equations. However, even if we have no such exclusion restriction, b can still be identified by assuming that the errors U and ε are jointly normal. Here it is the functional form of the error distribution that provides identification by functional form.

Actually, in this selection model joint normality is much stronger than necessary for identification without an exclusion restriction. Escanciano, Jacho-Chávez, and Lewbel (2016) show that little more than nonlinearity in $E\left(D \mid X\right)$ will suffice. A similar result is given in Dong (2012).

As noted in section 3.5, models that are nonlinear in parameters may fail to be identified because nonlinear equations can have multiple solutions. However, nonlinearity can also sometimes help provide identification by functional form, as the following surprising example shows. Suppose we continue with the classical Cowles model considered in example 4 in sections 3.3 and 3.5, except that now, while the demand curve is still $Y = bX + cZ + U$, we let the supply curve be $Y = dX^2 + aX + \varepsilon$. Assume U and ε are still independent of Z and have mean zero. Constant terms are omitted for simplicity. We still have no exogenous variable in the supply equation to serve as an instrument for X in the demand equation. We also only have the single exogenous Z in the demand equation, and two endogenous regressors (X and X^2) that have coefficients we need to identify in the supply equation. Despite this apparent shortage of

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

Substituting out Y in the two equations gives $dX^2 + (a-b)X - cZ + \varepsilon - U = 0$, and solving for X yields the reduced form equation $X = \left(b - a \pm \left((b-a)^2 + 4(cZ - \varepsilon + U)d\right)^{1/2}\right)/2d$. This shows that X is linear in $(Z+\gamma)^{1/2}$ for some γ, and squaring this shows X^2 is linear in Z and $(Z+\gamma)^{1/2}$. So Z is a valid instrument for X^2, and a function of Z that is correlated with $(Z+\gamma)^{1/2}$ can serve as an instrument for X. Generally $Z^{1/2}$ would be correlated with $(Z+\gamma)^{1/2}$, so we can identify and estimate b and c by an instrumental variables regression of Y on X and Z, using $Z^{1/2}$ and Z as instruments. Similarly, we can identify and estimate d and a by an instrumental variables regression of Y on X and X^2, again using $Z^{1/2}$ and Z as instruments.

Formally proving identification entails showing that the equations $E\left(Y - dX^2 - aX \mid Z = z\right) = 0$ and $E\left(Y - bX - cZ \mid Z = z\right) = 0$ for all z on the support of Z can be uniquely solved for a, b, c, and d. This requires that the model contain a few mild additional assumptions. For example, identification would fail if Z only took the values zero and one. For some rough graphical intuition on why identification is possible here, observe now that since the supply curve is nonlinear, as Z shifts the demand curve, one sees intersections of supply at different points along the demand curve, instead of always at the same point. Unlike the linear case, we now get information about more than one point on the demand curve as Z moves supply, allowing us to trace out and thereby identify the demand curve.

This idea of identifying linear coefficients by exploiting nonlinearity elsewhere in the system can be greatly generalized. For example, consider the model $Y = h\left(Z'b, g(Z)\right) + \varepsilon$ and $X = g(Z) + U$ where the functions h and g are unknown, and the joint distribution of ε and U is unknown but independent of Z. Models like these can arise with endogenous regressors or with sample selection (the Heckman selection model discussed at the start of this subsection is an example). A model like this would generally be identified by making an exclusion assumption, e.g., assuming that some element of b equals zero. Escanciano, Jacho-Chávez, and Lewbel (2016) show that the coefficients b and the functions h and g can generally be identified in this model without exclusion assumptions, with a key requirement being that g is nonlinear.

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. Identification wouldn't require knowing that the supply equation was specifically a quadratic functional form. Most forms of nonlinearity in supply would suffice. As these examples show, identification by functional form does not necessarily require specific parametric models like a known error distribution or a completely parameterized equation. Often what suffices is just some general functional restrictions on the model, such as linearity in one equation and nonlinearity in another.

Consider again the model $Y = a + bX + c'Z + U$ with endogenous X, so X is correlated with U, and suppose $X = \alpha + \beta'Z + e$. Suppose we have no exclusion assumption, meaning no element of c is known to equal zero, and therefore we have no outside source of instruments. This model can sometimes be identified by exploiting heteroskedasticity instead of nonlinearity in the X equation to identify the model. Linearly regress X on a constant and on Z, then take the residuals $\widehat{e}$ from that regression and let $R = \left(Z - \overline{Z}\right)\widehat{e}$ where $\overline{Z}$ is the sample average of Z. Under some assumptions regarding heteroskedasticity, Lewbel (2012) shows that R is a valid vector of instruments for X in the Y equation. See also Lewbel (2018) and Baum and Lewbel (2019). One example of assumptions that make this identification work is if X suffers from classical measurement error, so U contains both model and measurement error. If the true model error in the Y equation is homoskedastic and e is heteroskedastic, then R is a valid instrument for X. This procedure is also valid, resulting in consistent estimates, under some standard factor structure assumptions. Other examples of heteroskedasticity based identification include Rigoban (2003) and Klein and Vella (2010).

Still another example of identification by functional form is the model $Y = a + bX + U$ where X is assumed to suffer from classical measurement error. Assuming that the true model error and measurement error are independent of the true, nonmismeasured X, Reiersøl (1950) obtained the surprising result that

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 $\left(X-\overline{X}\right)^2$ is a valid instrument for X. Schennach and Hu (2013) show that the Reiersøl result can be extended to obtain identification of $Y = g(X) + U$ with mismeasured X for almost any function g. This model is not identified for only a few specific functional forms of g and a few specific functional forms of U. In these models the assumption that the true regression error is not just uncorrelated with X but is independent of X has strong identifying power.

Identification based on functional form (e.g., using constructed instruments as in the above examples), generally depends on relatively strong modeling assumptions. So, when they are available, it is usually better to instead use 'true' outside instruments, that is, instruments that are known or believed to be excluded and exogenously determined based on randomization or on strong economic theory. Some proponents of causal inference methods (discussed in Section 5 below) only accept randomization as a valid source of exogenous variation.

In practice one is often not sure if a candidate outside instrument is a valid instrument. An instrument might be invalid because the economic theory leading to its exclusion restriction is wrong. Even with randomized instruments in a causal or experimental setting, assumptions like SUTVA or no defiers (see Section 5 below) could be violated, making the instrument invalid despite randomization. Or identification with a randomized instrument could fail due to problems such as measurement errors or attrition correlated with treatment (that is, observations that are missing not at random). It can therefore often be useful to combine identification by outside instruments or randomization with identification based on functional form.

In particular, identification based on functional form, such as constructed instruments, can be used to provide overidentifying information for model tests and for robustness checks (see the next section for the definition of overidentification). The overidentification provided by constructed instruments or by

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 W and the constructed instrument R. Both might then be used in estimation to maximize efficiency.

One could also just estimate the model separately using W or R as an instrument, and compare the resulting estimates of b and c. If the estimates are similar across these different sets of identifying assumptions, then that provides support for the model and evidence that the results are not just artifacts of one particular identifying assumption. More generally, identification based on functional form or constructed instruments is preferably not used in isolation, but rather is ideally employed in conjunction with other means of obtaining identification, both as a way to check robustness of results to alternative identifying assumptions and to increase efficiency of estimation. We should have more confidence that estimated effects are reliable if different sources of identification that may be available (randomization, exclusions, functional form, constructed instruments) all yield similar estimates.

3.8 Over, Under, and Exact Identification, Rank and Order conditions

Models often contain collections of equalities involving θ. Common examples are conditional or unconditional moments, i.e., equations of the form $E\left[g\left(W,\theta\right)\right] = 0$ or $E\left[g\left(W,\theta\right) \mid Z\right] = 0$ where W and Z are observable variables and g is a known function. Most parametric and many semiparametric estimators are based on information of this form, including linear and nonlinear least squares, two stage least squares, generalized method of moments (GMM) estimators, quantile regressions, and more generally the first or-

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 θ are *exactly identified* if removing any one these equalities causes θ to no longer be point identified. The parameters are *overidentified* when θ 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 θ.

If θ is a $J-$vector, then it will typically take J equations of the form $E\left[g\left(W,\theta\right)\right]=0$ to exactly identify θ. 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 but not sufficient for point identification, though in many applications, satisfying the order condition may suffice for generic identification as defined later in Section 7.

Being able to uniquely recover a vector θ from a set of linear equations of the form $B\theta=0$ requires the *rank condition* on the matrix B that the rank of B equal the number of elements of θ. More general rank conditions for nonlinear models exist based on the rank of relevant Jacobian matrices. See Fisher (1959, 1966), Rothenberg (1971), Sargan (1983) and Section 8.1 below on local identification. Bekker and Wansbeek (2001) show that the rank conditions for identification can be evaluated from data in a variety of settings.

Generally, when parameters are overidentified, it is possible to test validity of the moments used for identification. Intuitively, given overidentification one could estimate the J vector θ using different combinations of J moments, and test if the resulting estimates of θ all equal each other. In practice, more powerful tests exist that work with all the moments simultaneously, in particular, the Sargan (1958) and Hansen (1982) J-test. Arellano, Hansen, and Sentana (2012) discuss testing for underidentification.

The terminology discussed in this section is generally from the Cowles foundation era, e.g., the term 'order condition' dates back at least to Koopmans (1949). This terminology comes from the era where θ would be a vector, not a function. An extension is Chen and Santos (2015), who define a notion of local overidentification for semiparametric models.

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

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

$$\begin{aligned} Y_1 &= I\left(Y_2 + U_1 \geq 0\right) \\ Y_2 &= \theta Y_1 + U_2 \end{aligned}$$

where θ is a coefficient, U_1 and U_2 are unobserved error terms, $V = (\theta, U_1, U_2)$, $Y = (Y_1, Y_2)$, 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 $Y_1 = I\left(\theta Y_1 + U_1 + U_2 \geq 0\right)$. Using this equation one can readily check that if $-\theta \leq U_1 + U_2 < 0$ 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 $0 \leq U_1 + U_2 < -\theta$ 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 $\theta = 0$ or $U_1 + U_2$ is constrained to not lie between zero and $-\theta$.

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 $\theta = 0$ (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 $\theta = 0$. 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, by writing the system as a triangular one where the direction of triangularity is itself endogenous. This means constructing a model where sometimes Y_2 depends on Y_1 and at other times Y_1 depends on Y_2. For example, if we replace the above model with

$$\begin{aligned} Y_1 &= I\left(DY_2 + U_1 \geq 0\right) \\ Y_2 &= (1-D)\,\theta Y_1 + U_2 \end{aligned}$$

Where D is a binary random variable, then the model becomes complete and coherent. In the game context, D could be an indicator of which player moves first, which could vary either deterministically or randomly. Lewbel (2007a) shows that this is often the only way to attain coherence and completeness without restricting the support of the errors.

Incoherence can be interpreted as a form of model misspecification, since it implies that for some observable values of V there does not exist a corresponding value of Y, whereas in data when we observe V we would also observe some value of Y. Incoherent models can therefore be invalidated by a single data point (an observation of V and Y when the model says no Y value corresponds to that V value). This also means that incoherent models cannot used to make predictions about Y over the space of all values of V.

In contrast to incoherent models, we may think of an incomplete model as a model that is not fully specified, since for some feasible values of V, the model does not determine the corresponding unique value for Y. So, in many applications, a finding of incoherence can mean that the model is wrong and

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 θ.

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."

change in outcome that results from a change in a covariate (the treatment). Examples of such treatment effect parameters are average treatment effects (ATE), average treatment effects on the treated (ATT), marginal treatment effects (MTE), local average treatments effects (LATE), and quantile treatment effects (QTE).

A counterfactual is defined as what would have occurred under some different treatment than the one that was actually administered. More generally, a counterfactual can be what an outcome would have equaled if a covariate had taken a different value from the value it actually had. As discussed earlier, typical obstacles to attaining identification in structural models include simultaneity and endogeneity. In contrast, in the causal framework, as emphasized by Rubin (1974) and Angrist and Pischke (2008), the main obstacle to identification is that parameters of interest are defined in terms of unobservable counterfactuals. Note, however, that treatment effect parameters can also be expressed in structural terms, and their identification issues can then be recast as endogeneity issues. See, e.g., Pearl (2000, 2009, 2015) and Heckman (2008, 2010).

5.1 Randomized Causal or Structural Modeling? Do Both

Heated debates exist on the relative merits of causal vs. structural methods. An implication of Angrist and Pischke's (2008) book title is that structural modeling includes harmful econometrics. Proponents of structural methods have in turn referred to reduced form studies demeaningly as cuteonomics, and their practitioners as randomistas (implying an emphasis on randomization based methodology over economic importance). Some prominent economics journals are known for being either friendly or hostile to structural methods. These debates have even spilled over into the popular press, e.g., Scheiber (2007).

Before getting into details regarding the two methodologies, it should be pointed out that the perceived conflict between proponents of causal, reduced form methods vs. structural modeling approaches is somewhat artificial. Researchers do not need to choose between building structural models vs. estimating treatment effects. For both identification and estimation, the strengths of both approaches may be combined in many ways, including these:

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 ϕ 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

randomization, like difference in difference designs, are also associated with the causal literature). Also, instead of identifying deep behavioral parameters, causal models instead focus on just identifying treatment effects, i.e, the average across some subpopulation of the change in an outcome that results from a change in a covariate (the treatment).

It must be emphasized that what we are here calling structural vs. causal restrictions are just common examples in the literature. They do not define what makes a model structural or causal. Not all structural models are linear regressions, not all linear regressions are structural, and not all causal analyses are LATEs. Rather, these are just typical examples of the kinds of models each literature uses, and the kinds of restrictions that each literature imposes. Structural models generally include behavioral restrictions regarding outcomes, such as assumptions on how Y may depend on T and Z. Causal models generally minimize such behavioral assumptions regarding outcomes, and instead assume more about how Z and T are determined.

We begin with a general triangular model where Y is determined by T along with error terms, and T is determined by Z and errors. Because both T and Z are binary, we can without loss of generality write this model as a linear random coefficients model, where

$$Y = U_0 + U_1 T \quad \text{and} \quad T = V_0 + V_1 Z$$

for some error terms U_0, U_1, V_0, and V_1.[9]

As noted earlier, consideration of identification involving outcomes based on treatment goes back to Splawa-Neyman (1923), Wright (1925), Haavelmo (1943), Wald (1943), Roy (1951), and Heckman

[9]To see that this construction is without loss of generality, suppose we had $Y = G\left(T, \widetilde{U}\right)$ and $T = R\left(Z, \widetilde{V}\right)$ where $\widetilde{U}$ and $\widetilde{V}$ are vectors of unobservable errors and both G and R are completely arbitrary, unknown functions. This structure assumes that Z does not directly affect Y, that is, we could have equivalently started from the more general model $Y = G\left(T, Z, \widetilde{U}\right)$ and then assumed that $G\left(T, z, \widetilde{U}\right)$ takes the same value whether $z = 1$ or $z = 0$. Similarly, this structure assumes that Y does not directly affect T. Now define $U_0 = G\left(0, \widetilde{U}\right)$, $U_1 = G\left(1, \widetilde{U}\right) - G\left(0, \widetilde{U}\right)$, $V_0 = R\left(0, \widetilde{V}\right)$, and $V_1 = R\left(1, \widetilde{V}\right) - R\left(0, \widetilde{V}\right)$. We then without loss of generality have $Y = U_0 + U_1 T$ and $T = V_0 + V_1 Z$. This derivation depended crucially on having T and Z be binary. Generally, having a nonparametric model be the same as a linear model is a very special property of binary regressors.

(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 does not directly affect Y. We could have started (see Angrist, Imbens, and Rubin 1996) from a more general counterfactual $Y(t, z)$ denoting the value of Y if both $T = t$ and $Z = z$, and the exclusion assumption would then be that $Y(t, 1) = Y(t, 0)$. Similarly the $T(z)$ notation assumes that Y does not directly affect T. See Heckman (2008) for further generalizations allowing for explicit sources of nonrandomness in treatment assignment.

The treatment effect for an individual is defined as $Y(1) - Y(0)$, or equivalently as U_1, which is the difference between the outcome one would have if treated vs. not treated. This treatment effect is random and generally varies across individuals. So we can think of the general model as being one of heterogeneous treatment effects. The average treatment effect, ATE, would then be defined as $\theta = E(Y(1) - Y(0))$.

So far, no assumptions have been made apart from these exclusion restrictions, so at this level of generality there is no difference between the causal and structural framework. Now we start to make assumptions regarding the random variables U_1, U_0, V_1, and V_0, or equivalently regarding the random variables $Y(1)$, $Y(0)$, $T(1)$ and $T(0)$.

A common assumption in the causal inference literature is the stable unit treatment value assumption,

[10] Notions of causality that are relevant to the economics and econometrics literature have an even longer history, going back to Hume (1739) and Mill (1851). There are also the temporal based definitions of causality due to Granger (1969) and Sims (1972). See Hoover (2006) for a survey of alternative notions of causality in economics and econometrics.

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 i) with V_{0j}, V_{1j}, and Z_j. For simplicity, let us assume that the set of random variables $\{U_{1i}, U_{0i}, V_{0i}, V_{1i}, Z_i\}$ is independent across individuals i, which is sufficient and stronger than necessary to ensure that SUTVA holds. A structural model alternative to SUTVA might be to model the dependence of these variables across individuals, e.g., by specifying a social interactions model as in Blume, Brock, Durlauf, and Ioannides (2011).

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 $c = E(U_1 V_1)/E(V_1)$.

Consider the structural model first. In the structural model, we may rewrite the $Y = U_0 + U_1 T$ equation in the usual linear regression form $Y = a + bT + e$, where a is the constant term, $b = E(U_1)$ is the coefficient of T, and the error is $e = (U_1 - b) T + U_0 - a$. Since the instrumental variables estimator converges to c, the question we ask is, when does instrumental variables estimation yield the structural coefficient b, or equivalently, when does $c = b$? The answer is that $c = b$ when $cov(e, Z) = 0$. In the structural linear regression model, this is the definition of validity of an instrument Z. But what does

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 $V_0 = 0$,

making $T = Z$. It follows similarly that defiers are individuals who have $V_1 = -1$ and $V_0 = 1$, to make $T = 1 - Z$. The only remaining possibilities are called always takers, who have $V_1 = 0$ and $V_0 = 1$, and never takers, who have $V_1 = 0$ and $V_0 = 0$. Let P_υ denote the probability that $V_1 = \upsilon$. Then, by definition of expectations,

$$c = \frac{E(U_1 V_1)}{E(V_1)} = \frac{E(U_1 V_1 \mid V_1 = 1) P_1 + E(U_1 V_1 \mid V_1 = 0) P_0 + E(U_1 V_1 \mid V_1 = -1) P_{-1}}{E(V_1 \mid V_1 = 1) P_1 + E(V_1 \mid V_1 = 0) P_0 + E(V_1 \mid V_1 = -1) P_{-1}}$$
$$= \frac{E(U_1 \mid V_1 = 1) P_1 - E(U_1 \mid V_1 = -1) P_{-1}}{P_1 - P_{-1}}$$

Imbens and Angrist (1994) assume that there are no defiers in the population. This rules out the $V_1 = -1$ case, making $P_{-1} = 0$. This in turn makes the above equation simplify to $c = E(U_1 \mid V_1 = 1)$, which is the LATE. So, under the no defiers assumption, the instrumental variables estimand c equals, and therefore identifies, the LATE, which is the ATE for compliers.

Interestingly, if one imposes both our structural assumption $cov(U_1, V_1) = 0$ and our reduced form "no defiers" restriction $P_{-1} = 0$, then it can be shown that

$$0 = [E(U_1 \mid V_1 = 0) - E(U_1 \mid V_1 = 1)](1 - P_1).$$

This means that, with the no defiers assumption, the only way LATE would equal ATE is either if $E(U_1 \mid V_1)$ is the same for $V_1 = 0$ and $V_1 = 1$, or if everyone in the population is a complier.

Even with binary treatment, it is possible to relax the above listed assumptions in both the structural and the causal framework. For example, one might identify alternative LATEs (e.g., by allowing some defiers to exist whose effects are offset by other compliers), and it is possible to have structural identification under weaker or alternative assumptions (e.g., relaxing the independence assumptions and/or allowing for selection effects). But, for simplicity, let us stick with the above described canonical assumptions on binary treatment, and assess how the two approaches compare.

Given our independence and exclusion assumptions, the only difference between the structural and causal assumptions we have made here is the following: The causal analysis assumes nobody has $V_1 = -1$, identifies LATE (which is ATE for compliers), and identifies nothing about people who are not compliers. In contrast the structural analysis assumes $cov(U_1, V_1) = 0$ and identifies the population ATE. Both the

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 $cov\left(U_1, V_1\right) = 0$, or equivalently, $cov\left(e, Z\right) = 0$, 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 $\widetilde{Z}$ instead of Z, and we let $\widetilde{c} = cov\left(\widetilde{Z}, Y\right) / cov\left(\widetilde{Z}, T\right)$. If

$\widetilde{Z}$ is a valid instrument in the structural sense, then we will have $c = \widetilde{c} = b$, meaning that any structurally valid instrument will identify the same population ATE b. This also provides potential testability or falsifiability of the structural model, since one could estimate and thereby test if $c = \widetilde{c}$. Rejecting the null that $c = \widetilde{c}$ corresponds to rejecting the structural identifying assumption, or more formally, rejecting the joint hypothesis that both instruments are valid in the structural sense. Whether this test has power is a separate question, e.g., it is possible that structural assumptions are violated but still $c = \widetilde{c}$ just by chance.

In contrast, even if both Z and $\widetilde{Z}$ are valid instruments from the causal point of view, then $c \neq \widetilde{c}$ in general. Both estimates are LATEs, but since Z and $\widetilde{Z}$ are different, the definition of who is a complier will also be different in the two cases, so c and $\widetilde{c}$ will differ in value, meaning, and interpretation. Here c and $\widetilde{c}$ could have different signs or magnitudes, and the only causal conclusion one could draw is that different complier populations have different responses to treatment. This difficulty of testing or falsifying LATE is a possible drawback of the model.[11]

More simply, under the structural identifying assumptions, if different instruments yield different parameter values, then by definition at least one of the instruments must not be valid. In the causal version of the model, different instruments yield different parameter values because they correspond to different estimands, each of which is deemed to be causally meaningful, because each equals the average treatment effect for some, albeit unknown, set of people in the population.

5.3 Randomized Causal vs. Structural Simultaneous Systems

Suppose that instead of a treatment affecting an outcome, where the direction of causality is assumed, we had a simultaneous system of equations, say $Y = U_0 + U_1 X$ and $X = H(Y, Z, V)$, where (U_0, U_1) and V are unobserved error vectors, Y, X, and Z are observed variables, and H is some unknown function. We now use X instead of T because it may no longer be a treatment and need not be binary. But the main difference is that now Y appears in the H equation, so the model is no longer triangular.

[11] There are a few weak inequalities one can test that are implied by causal instrument validity in the LATE context. For example, with a valid instrument the number of implied compliers in any subgroup of people can't be negative. See, e.g., Kitagawa (2015) for associated tests.

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 $-1 \leq U + Z + V < 0$. 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 the outcome of others. Examples include models of peer effects, so-

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. Randomized controlled trials are often referred to as the gold standard for empirical work in other fields (despite recent alarms being raised regarding publication bias and failures of replication in some experimental sciences). The goal of exploiting randomization in field and natural experiments in economics is to get as close as possible to replicating that gold standard in economics.

Another virtue of causal methods is the fundamental nature of treatment effects as interpretable esti-

mands. As long as the framework is coherent and complete, so that potential outcomes are well defined, causal estimands like ATE, ATT, etc., are meaningful. In contrast, when a structural model is misspecified, the deep policy invariant parameters it attempts to identify can become meaningless, or at best difficult to interpret.

One advantage of structural models over causal models is that, as the name implies, they contain information about underlying behavioral structure. Structural models can incorporate, or be used to test, restrictions implied by economic theory, and can likewise gain identification from incorporating such restrictions, even when external sources of randomization are not present.

Structural models can also cope with many data issues that cause difficulties for causal analyses. For example, we can build structures that account for self selection into treatment, or account for measurement problems such as sample attrition (observations that are missing not at random), or deal with SUTVA violations like network effects and social interactions.

As Heckman (2008) emphasizes, random assignment makes it impossible to point identify some features of the joint distribution of potential outcomes, like what fraction of the population would benefit from treatment. But, e.g., in the Roy (1951) model and closely related competing risks models, the fact that selection is based on maximization over potential outcomes provides information about the joint distribution of potential outcomes that could not be identified if assignment were random. See Heckman and Honoré (1989, 1990) and Lee and Lewbel (2013) for nonparametric identification of these types of models. Another example of self selection providing information (which is unavailable under random assignment) is given in Brock and Durlauf (2001), who show that the presence of self selection can overcome the Manski (1993) reflection problem, which is an obstacle to identification in many social interactions models.

A common objection to the structural approach is the question of its success. Where are the deep parameters that structural models have uncovered? What are their widely agreed upon values? One answer is to look at the use of calibration methods. Calibrated models, widely used in macroeconomics, include a large number of parameters that users treat as known. Many of these parameters have values that proponents of these models largely agree upon, e.g., there is very little variation across applications

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 LATE model even if an instrument Z is randomly assigned, the population may contain defiers or SUTVA may be violated. As discussed earlier, the relative plausibility of structural vs. causal identifying assumptions depends on context. Methods like those of Pearl (2000, 2009), using minimal structural assumptions to identify causal parameters, may be particularly useful in this regard,

Another source of potential misspecification, in both structural and causal models, is the role of conditioning covariates. Conditioning requires functional forms that may either be parametrically misspecified (the popularity of linear probability models in reduced form methods is a case in point), or are nonparametrically specified and thereby suffer the curse of dimensionality upon estimation.

A big issue for both structural and causal models is external validity, that is, if the environment changes even slightly, how would an identified parameter or treatment effect change? The deep parameters in structural models are supposed to be constants that do not vary across environments, but behavioral restrictions that are valid in one context may not hold in another, and correct specifications may change due to factors like the well known Lucas (1976) critique. External validity is an even larger problem with causal

models, since these have no underlying economic or behavioral restrictions that one can assess in a new environment. For example, Rosenzweig and Udry (2016) document how macro conditions that one cannot control for, like weather, can dramatically affect estimated treatment effects obtained in randomized controlled trials. The search for empirical methods of assessing external validity of both causal and structural models is an active area of research.

Another limitation of causal methods is that economic policy is often concerned with characteristics that cannot be directly observed, like utility, risk aversion, noncognitive skills, bargaining power, expectations, or social welfare. In general, structural, behavioral assumptions are required to link observables to these elusive concepts, and hence to evaluate the impacts of treatment on them.

The rise of identification based on randomization in economics has been referred to as a "credibility revolution." Applied economic research in recent years has certainly benefited from this focus on credibility. But as the previous sections show, in practice both causal and structural methodologies in economics depend on a host of assumptions, only some of which are testable. Either can lead to invalid (perhaps we should say incredible) inference when their identifying assumptions are violated. Both sources of identification have their advantages and disadvantages. Ideally, future work will creatively combine the best features of both.

6 Identification of Functions and Sets

The first two subsections below discuss two modern literatures: nonparametric or semiparametric identification, and set identification. These two subsections are relatively brief in comparison to the size of these literatures, in part because good summaries of these literatures already exist. See Tamer (2010) for a survey of the set identification literature, and Chesher and Rosen (2017) for more current definitions and references on the subject. Surveys of semiparametric and nonparametric identification include Powell (1994), Chesher (2007), and Matzkin (2007, 2012). Powell (1994) focuses on semiparametrics, Chesher (2007) deals with nonadditive models with nonseparable errors, and Matzkin (2007, 2012) provides many examples of using restrictions such as additivity, monotonicity, conditional independence, and index struc-

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 θ 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 θ is a finite set of constants, and all the different possible values of ϕ 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 $Y = m(Z) + X'\beta + e$ where m is an unknown function, β is a finite vector of unknown parameters, and $E(e \mid X, Z) = 0$. Here the unknown parameters include both a vector of constants β and a function m. Identification of θ is semiparametric, since it contains both a parametric component (the parameter vector β) and a nonparametric component $m(Z)$. Following Robinson (1988), Observe that $E(Y \mid Z)$

and $E(X \mid Z)$ are nonparametrically identified, and $Y - E(Y \mid Z) = (X - E(X \mid Z))'\beta$. It follows that β is identified by construction, regressing $Y - E(Y \mid Z)$ on $X - E(X \mid Z)$, if $var(X \mid Z)$ is nonsingular. Finally, given β, m is identified by $m(Z) = E\left(Y - X'\beta \mid Z\right)$. Other relatively early examples of semiparametric identification include Manski (1975, 1985) and Cosslett (1983).

Semiparametric identification can also used to refer to identification of a vector of constants that are of interest in a nonparametric model. For example, in the nonparametric regression model $Y = m(X) + e$, since $m(X)$ is nonparametrically identified, it follows that the average derivative vector $\beta = E\left[\partial m(X)/\partial X\right]$ is identified. Identification and estimation of average derivatives and related estimators are generally called semiparametric. See, e.g. Härdle and Stoker (1989).

The difference between parametric and semiparametric identification can be somewhat arbitrary. For example, consider the ordinary linear regression model where observations of Y, X are independent, $Y = X'\beta + e$ with $E(e) = 0$, and e is independent of X. If we let θ be β and let ϕ be the first and second moments of Y, X (like in Wright-Cowles identification), then this example is parametric identification. But if we let θ be both β and the distribution of e, and let ϕ be the joint distribution of Y, X, then this is semiparametric identification. Similarly, the distinction between semiparametric and nonparametric identification can be somewhat arbitrary. See Powell (1994) for further discussion of the differences between parametric, semiparametric, and nonparametric frameworks. These types of distinctions can be traced back at least to Hurwicz (1950).

In principle, establishing nonparametric or semiparametric identification entails the same steps as parametric identification, establishing that different values of θ are not observationally equivalent.[12] In practice establishing identification in these models can be much more difficult. For example, we cannot simply count the number of known and unknown functions to establish an order condition, the way one can count known and unknown parameters in parametric models.

To illustrate some differences, consider the nonparametric instrumental variables model. Here the goal

[12] When θ contains functions, as in nonparametric identification, a technical issue can arise in the choice of metric or norm used to compare values θ and $\widetilde{\theta}$. Depending on context, we might define any two values as being equal as long as any difference between them has measure zero.

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 θ to identify is the function $m(X)$, what is knowable, ϕ, 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 $\int_{supp(Y,X|Z=z)} (Y - m(X))\, dF(Y, X \mid z) = 0$ 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 θ 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 ϕ in terms of θ, in the same way that the sums (like expectations of discrete random variables) may be solved for vectors θ 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 examples where identification takes the form of a unique solution to an integral equation are random coefficients as in Beran, Feuerverger, and Hall (1996) and many of the structural models in Matzkin (2007, 2012). Additional examples of semiparametric identification are given in Section 6.4 below.

6.2 Set Identification

Recall that we let θ denote the parameters we wish to identify, Θ is the set of all possible values of θ, and θ_0 is the unknown true value of θ. *Partial Identification* broadly refers to the analysis of situations where ϕ provides some information about parameters θ, but not enough information to actually point identify θ.. As defined in Section 3.2, the true parameter value θ_0 is said to be *set identified* if there exist some values of $\theta \in \Theta$ that are not observationally equivalent to θ_0. So the only time a parameter θ is not set identified is when all $\theta \in \Theta$ are observationally equivalent. For set identified parameters, the *identified set* is defined to be the set of all values of $\theta \in \Theta$ that are observationally equivalent to θ_0. Point identification of θ_0 is therefore the special case of set identification in which the identified set contains contains only one element, which is θ_0. As with point identification, since we don't know ahead of time which value of θ is θ_0, to prove that set identification holds we would need to show that for any $\widetilde{\theta} \in \Theta$, there exist some values of $\theta \in \Theta$ that are not observationally equivalent to $\widetilde{\theta}$. For more details formalizing the definition of set identification, see Chesher and Rosen (2017).

To illustrate, let W be the fraction of a consumer's total budget that is spent on food. The model M is that we observe W and budget shares are nonnegative and sum to one. Suppose that what is knowable, ϕ, is the distribution of W in the population (e.g., our data might be an IID sample of consumers). Then we can identify $\beta = E(W)$, the expected budget share for food in the population. But suppose the θ that we want to know is the expected budget share of clothing, not food. What we can know about θ from this model is that θ must lie between zero and $1 - \beta$. In this simple example we can say θ is set identified, and the identified set is the interval $[0, 1 - \beta]$. This set also provides bounds on the possible values of θ. Sometimes a parameter can be set identified without also providing bounds. *Sharp bounds* are defined to be the tightest bounds possible on θ, given M and ϕ.

An important tool used for studying set identification is the theory of random sets. See, e.g., Beresteanu, Molchanov, and Molinari (2011), Bontemps, Magnac, and Maurin (2012), and Chesher and Rosen (2017).

The earliest example of a set identification analysis appears to be Frisch (1934). Other early examples are Reiersøl (1941) on correlated errors, Marschak and Andrews (1944) on production functions, Fréchet

(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 Θ.

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 $\theta = E(Y^*)$ where the scalar random variable Y^* can take on any value, but what is knowable, ϕ is the distribution of Y, defined by $Y = I(-b \leq Y^* \leq b) Y^*$ 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 θ 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, θ 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 θ 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 θ 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 and inference of set identified objects are complicated, generally requiring multiple tuning parameters and penalty functions. See, e.g., Chernozhukov, Hong, and Tamer (2007). Development of more tractable

[13]This example of non-robust identification is also an illustration of nongeneric identification, the flip side of generic identification as discussed in a later section. Different values of b correspond to different DGP's for Y. If we think b could have taken on any value in some interval, then the set of DGP's for which the model is not identified is a positive measure subset of the set of possible DGP's, and so is not generically identified.

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

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 $g\left(X'\beta\right)$ 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 $\widetilde{\beta} = \beta/c$, $\widetilde{g}(z) = g(cz)$, and $\widetilde{\theta} = \{\widetilde{g},\ \widetilde{\beta}\}$. Then for any X we have $E(Y \mid X) = g\left(X'\beta\right) = \widetilde{g}\left(X'\widetilde{\beta}\right)$, which shows that θ and $\widetilde{\theta}$ 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 β. Suppose that all of the elements of the identified set for β are proportional to the true β, so all have the form of $\widetilde{\beta} = \beta/c$. Then we would say that β is *identified up to scale*. An early example in the semiparametric literature is Powell, Stock, and Stoker (1989), who argue that their weighted average derivative estimates are meaningful as linear index coefficients up to scale.

To have any chance of point identifying β, we need to impose a restriction on the model that rules out all values of $c \neq 1$. For example, we might assume that the first element of β equals one, or that the length of the vector β equals one. Given a β that satisfies one of these scale restrictions, any observationally equivalent alternative of the form $\widetilde{\beta} = \beta/c$ for $c \neq 1$ will violate this restriction, and so will be ruled out. Suppose β is identified up to scale without a scale restriction. If we impose one of these scale restrictions, then β will be point identified.

In the linear index model, there's a sense in which restricting the scale of the vector β doesn't entail any loss of generality, because any rescaling of β can simply be absorbed into the definition of g. For example, suppose we imposed the restriction that the length of the vector β equals one. Then if we find any β and g that makes $E(Y \mid X) = g\left(X'\beta\right)$, we can just let $c = \sqrt{\beta'\beta}$, let $\widetilde{\beta} = \beta/c$, and let $\widetilde{g}(z) = g(cz)$, thereby creating a new $\widetilde{\beta}$ and $\widetilde{g}$ that still satisfies the model restriction $E(Y \mid X) = \widetilde{g}\left(X'\widetilde{\beta}\right)$ but in addition satisfies the scale restriction that the length of the vector $\widetilde{\beta}$ equals one. In this situation, we would call the scale restriction a *normalization.*

Loosely speaking, calling a restriction on the model a normalization implies that the restriction does not in any way limit or restrict behavior. Sometimes normalizations are specifically called "free" normalizations to emphasize this point, or are explicitly said to be imposed without loss of generality. Formally, whether a restriction can really be considered a free normalization or is instead an actual behavioral restriction, depends in part on how we will use and interpret the model.

A restriction is really a free normalization, imposed without loss of generality, if economically meaningful parameters or summary measures are unaffected by the restriction. For example, suppose in the linear index model instead of defining the parameters of interest θ to be β and g, we instead define

$\theta = \{g\left(X'\beta\right), \beta g'\left(X'\beta\right)\}$ where g' denotes the derivative of the function g. This θ equals the conditional mean function $E\left(Y \mid X\right)$ and its derivatives (i.e., marginal effects) $\partial E\left(Y \mid X\right)/\partial X$. This says that what we really want to estimate (what is economically meaningful) are not the coefficients themselves, but rather the predicted value (mean) of Y given X, and the marginal effects of X on this predicted value. Now let $\widetilde{\beta} = \beta/c$, and let $\widetilde{g}\left(z\right) = g\left(cz\right)$, and so $\widetilde{\theta} = \{\widetilde{g}\left(X'\widetilde{\beta}\right), \widetilde{\beta}\widetilde{g}'\left(X'\widetilde{\beta}\right)\}$ for any positive c. Then $\{\widetilde{g}, \widetilde{\beta}\}$ is observationally equivalent to $\{g, \beta\}$ but $\theta = \widetilde{\theta}$. In words, if β is identified up to scale, then the economically meaningful parameters θ are point identified. So in this case we would say that a scale normalization, like restricting the length of β to equal one, really is a free normalization, since it has no effect on θ.

In contrast to this choice of θ, an arbitrary scale restriction on β is not a free normalization if the level of $X'\beta$ itself has some economic meaning, or is itself a parameter of interest. For example, suppose our model is a willingness to pay study, where an individual reports $Y = 1$ if they would be willing to pay more than V dollars for a product or service, and reports $Y = 0$ otherwise. Suppose an individual's willingness to pay function is given by $X'\beta + e$ where e is an unobserved mean zero error that is independent of X. This means that $X'\beta$ is the expected number of dollars an individual would be willing to pay for the product. Then $Y = I\left(X'\beta + e \geq V\right)$ and therefore $E\left(Y \mid X, V\right) = g\left(X'\beta - V\right)$ where g is the distribution function of $-e$. Note that in this application we may think of $X'\beta - V$ rather than just $X'\beta$ as the index being estimated, and the proper scaling to impose is to set the coefficient of V equal to minus one. Any other scaling would yield $X'\widetilde{\beta} \neq X'\beta$ and would therefore not correctly identify the mean willingness to pay. In this application, scale is not a free normalization, but fortunately economic theory tells us the right scaling, and so β is identified. Nonparametric and semiparametric identification and estimation of general willingness to pay models like this are studied in more detail in Lewbel, Linton, and McFadden (2012).

Scale restrictions, whether they are free normalizations or not, are commonly needed to point identify semiparametric models. Also common are location normalizations. Suppose $E\left(Y \mid X\right) = g\left(X + \alpha\right)$, where g is an unknown function and α is an unknown scalar. Then generally α is not identified, because

given any $\widetilde{\alpha}$ we can define an observationally equivalent $\widetilde{g}$ such that, $\widetilde{g}(X+\widetilde{\alpha}) = g(X+\alpha)$. 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, $Y = I\left(\alpha + X'\beta + e \geq 0\right)$ 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 $E(Y \mid X) = g\left(X'\beta\right)$ where g is the distribution function of $-(\alpha + e)$. 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 $\alpha + X'\beta + e$. We could instead have imposed the location and scale restrictions that $\alpha = 0$, $\beta'\beta = 1$, 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 α and β instead of on error distribution parameters like $E(e)$ and $var(e)$. This is done both because it often simplifies identification and estimation, and because estimates of moments like $var(e)$ will sometimes converge at slower rates than estimates of α or β. More generally, though irrelevant for identification, the choice of normalization can affect the precision with which parameters are estimated.

When comparing parametric to semiparametric estimators, or comparing two different semiparametric estimators, it is important that the results from both be cast in the same location and scale normalization to make the estimates comparable. Or alternatively, one may just compare summary measures that do not depend on the choice of normalization, like estimated conditional means or average derivatives.

Another common use of normalizations for identification are in nonseparable error models. These are models where error terms appear inside general functions, as opposed to showing up additively. Suppose, e.g., that $Y = g(X, e)$ where both the function g and the distribution of an unobserved continuously distributed scalar error e are unknown. This is observationally equivalent to the model $Y = G(X, h(e))$

for an unknown strictly monotonic, invertible function h, where the function $G(X, v)$ is defined by $G(X, v) = g\left(X, h^{-1}(v)\right)$. Here identification of g may require assuming that the entire distribution function of e is known. Common examples are to assume that e is uniform on $[0, 1]$ or is a standard normal. This is a free normalization when one only wishes to use the model to summarize the conditional distribution of Y given X. Otherwise, if we wish to interpret g as a structural model, this normalization generally corresponds to a strong behavioral restriction. See Lewbel (2007c) for details. Matzkin (2007, 2012) provides multiple examples of nonseparable model identification results using this type of normalization.

Returning to the example of threshold crossing models, when derived from utility maximization such models embody an additional normalization to location and scale. Suppose $\alpha_y + X'\beta_y + e_y$ is the utility one receives from making the choice $Y = y$, for y equal to zero or one. Utility maximization then means choosing $Y = I\left(\alpha + X'\beta + e \geq 0\right)$, where $\alpha = \alpha_1 - \alpha_0$, $\beta = \beta_1 - \beta_0$, and $e = e_1 - e_0$. This means we can interpret $\alpha + X'\beta + e$ as being the utility obtained from choosing $Y = 1$ if we assume the restrictions or normalizations that α_0, β_0, and e_0 all equal zero. With these restrictions, choice zero is sometimes called the outside option. In a static discrete choice model, these restrictions are generally free normalizations, as long as one is careful to interpret $\alpha + X'\beta + e$ as the utility of choice $Y = 1$ relative to the utility of choice $Y = 0$. However, in dynamic discrete choice models, restricting the outside option to have zero utility implies that the outside option has the same utility in every time period. This is then no longer a free normalization, but instead corresponds to a real restriction on preferences and hence on behavior. See, e.g., Rust (1994) and Magnac and Thesmar (2002).

Another normalization associated with utility is the arbitrary choice of cardinalization of ordinally identified utility levels. In standard utility maximization, one chooses a quantity vector x to maximize a utility function $U(x)$ under some constraints (e.g., a budget constraint). The revealed preference theory of Samuelson (1938, 1948), Houthakker (1950), and Mas-Colell (1978) provides conditions under which, given demand functions, indifference curves (i.e., level sets) associated with $U(x)$ are identified, though we can't identify the actual utility or happiness level $U(x)$ associated with each indifference curve. This is

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 $Y = I\left(\alpha + X'\beta + e \geq 0\right)$ is observationally equivalent to $Y = I\left(g\left(\alpha + X'\beta + e\right) \geq g\left(0\right)\right)$ where g is any strictly monotonically increasing function. Without more information, we could therefore never tell if one's actual utility level were $\alpha + X'\beta + e$ or $g\left(\alpha + X'\beta + e\right)$ 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 θ 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 the normalization is free, can affect the precision and even rate of convergence of estimates, despite being irrelevant for identifcation. Finally, in addition to normalizations, semiparametric and nonparametric identification also often requires additional behavioral restrictions on the model. Common restrictions include continuity, differentiability, or monotonicity of functions in θ.

6.4 Examples: Some Special Regressor Models

Return now to example 5 from Section 3.3, regarding identification of a latent error distribution. Here we consider the same model, except now we allow for the presence of additional covariates Z. The DGP is IID observations of scalar random variables Y, X, Z, so ϕ is the joint distribution of Y, X, Z. The model consists of the restrictions that the conditional distribution of X given Z is continuous, and that $Y = I\left(X + U > 0\right)$ where the unobserved $U \perp X \mid Z$. Consider identification of the function $F_{U|Z}\left(u \mid z\right)$, the conditional distribution function of U given $Z = z$. In this model $E\left(Y \mid X = x, Z = z\right) = \Pr\left(X + U > 0 \mid X = x, Z\right.$ $\Pr\left(x + U > 0 \mid Z = z\right) = 1 - \Pr\left(U \leq -x \mid Z = z\right) = 1 - F_{U|Z}\left(-x \mid z\right)$. This shows that the function $F_{U|Z}\left(u \mid z\right)$ is identified, since it can be recovered from the function $E\left(Y \mid X = x, Z = z\right)$, which itself

is identified given ϕ. 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 $\theta = E(U)$. 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 $E(U) = \int_{-\infty}^{\infty} u dF_U(u)$. 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 $-b \leq u \leq -a$. 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 $a = -\infty$ or $b = \infty$, then we could place bounds on $E(U)$, giving us set identification, and if both $a = -\infty$ and $b = \infty$, 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 X equals one is a scale normalization (assuming the effect of X on Y is positive). We could have instead given X an unknown coefficient and normalized the variance of e to be one, as is usual in the probit model. Suppose that the model also includes the large support assumption that the support of U given Z is contained in the support of $-X$ given Z. Then the unknown function g is identified because $F_{U|Z}(u \mid z)$ is identified for all u and $g(z) = E(U \mid Z = z) = \int_{supp(u|z)} u dF_{U|Z}(u \mid z)$.

As above, the large support restriction on the special regressor X is only needed because we wanted to identify a mean, specifically, $g(Z) = E(U \mid Z)$. Large support would not be needed to identify other features of the distribution of U. For example, if we still let $U = g(Z) + e$ but now define $g(Z)$ to be the conditional median rather than the conditional mean of U given Z, then the support of $-X$ would only need to include a neighborhood of the median of U given Z to identify $g(Z)$.

Another parameter we could identify in this model would be an elasticity function like $\partial E(\ln g(Z)) / \partial \ln Z$ (which is identified given g). Other economically interesting parameters we could identify are the distribution of e conditional on Z, and the model's "average structural function" as defined in Blundell and Powell (2004). For a general model of Y as a function of covariates Z and an error e, the average structural function is defined as what the function $E(Y \mid X)$ would have been if the conditional distribution of the error, $F_{e|Z}$, were replaced with its marginal distribution F_e. In the special regressor model, $F_{e|Z}$ is identified by $F_{e|Z}(e \mid Z) = F_{U|Z}(g(z) + e \mid z)$. Then, given $F_{e|Z}$, we can calculate the unconditional distribution F_e, and using F_e the average structural function is then identified, since in this application it is by definition given by $\int_{supp(e)} I(X + g(Z) + e > 0) dF_e(e)$. For related results and associated estimators, see Chen, Khan, and Tang (2016) and Lee and Li (2018). Note that in all these examples the identification is by

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 $E(U \mid Z) = Z'E(e)$ we can identify $E(e)$, From $E\left(U^2 \mid Z\right) = Z'E\left(ee'\right)Z$ we can typically identify $E\left(ee'\right)$, and similarly all the moments of e can (with some regularity) be identified, by looking at the corresponding conditional moments of U given Z.

Now, instead of a linear model, consider the binary choice random coefficients model $Y = I\left(Xe_x + Z'e_z > 0\right)$, where e_x is a scalar random coefficient e_z is a vector of random coefficients. Assume that $e_x > 0$ and let $e = e_z/e_x$ (this is a scale normalization as in Section 4.3). This model can then be equivalently written as $Y = I\left(X + Z'e > 0\right)$ where e is now the vector of random coefficients. In this model $U = Z'e$ is no longer observed, but now $F_{U|Z}$ is identified by the special regressor X, instead of being identified by observing U directly. So again the distribution of e is identified. Ichimura and Thompson (1998) and Gautier and Kitamura (2013) also show identification of the binary choice random coefficients model, but under a different scale normalization. This special regressor based identification is extended to the multinomial choice setting by Fox and Gandhi (2016).

Special regressors are used to identify games with discrete strategies in Lewbel and Tang (2014), and semiparametric generalizations of the BLP (Berry, Levinsohn, and Pakes 1995) model as in Berry and Haile (2014). In Berry and Haile the special regressors are never referred to by that name but are instead just called $x_{jt}^{(1)}$. Many of the results in Matzkin (2015) have a similar unstated connection to special regressors, e.g., her equations (2.2) and (2.4) make her so-called "exclusive regressors" X_g become special

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 θ_0, recall that point identification of θ_0 means that there is no other $\theta \in \Theta$ (the set of all possible values of θ, according to the model) that is observationally equivalent to θ_0. Since the true value of θ_0 is unknown, proving point identification requires that no distinct pairs of values θ and $\widetilde{\theta}$ in Θ 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 θ_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

identification. Intuitively, local identification of θ_0 means that θ_0 is identified among alternative values of θ, if we restrict attention only to alternative values that are very close to the true value θ_0.

Formally, *local identification* of θ_0 means that there exists a neighborhood of θ_0 such that no $\theta \in \Theta$ exists in this neighborhood that is both unequal to θ_0 and observationally equivalent to θ_0. Since we don't know what θ_0 equals, proving local identification requires showing that, for any two values θ and $\widetilde{\theta}$ in Θ that are observationally equivalent to each other but not equal to each other, there must exist a neighborhood of $\widetilde{\theta}$ that does not contain θ.

Note that this idea of local identification differs from (and predates), the term local as used in LATE (local average treatment effect). In LATE, local means that what is being identified is the mean parameter value for a particular subpopulation. In contrast, the word local in local identification refers to identification of a population (not a subpopulation) parameter, but only relative to values the population parameter might take on in a neighborhood of the true value.

To illustrate the difference between local and global identification, suppose $m(x)$ is a known continuous function. (equivalently, assume $m(x)$ is point identified in some model with some given ϕ). Suppose also that all we know about the parameter θ we wish to identify is that the true θ_0 is a real valued scalar (so Θ is the real line) that satisfies $m(\theta_0) = 0$. Consider three possible cases:

Case 1: Suppose we know $m(x)$ is strictly monotonic. Then θ (if it exists) is globally identified, because strict monotonicity ensures that only one value of θ can satisfy the equation $m(\theta) = 0$.

Case 2: Suppose m is known to be a J'th order polynomial for some integer J. Then θ may not be globally identified, because there could exist up to J different values of θ that satisfy $m(\theta) = 0$. However, in this case θ is locally identified, because for any value of θ that satisfies $m(\theta) = 0$, there always exists a neighborhood of θ that is small enough to not contain any of the other roots of $m(x) = 0$. More generally, if a real valued parameter vector θ is set identified, and the set has a finite number of elements (in this example, no more than J elements), then that parameter must be locally identified.

Case 3: Suppose all we know about m is that it is continuous. Then θ might not be locally identified, because $m(x)$ could equal zero for all values of x in some interval, and any given θ in that interval will be

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 θ is set identified, and the set has a finite number of elements, then θ 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 θ 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 θ 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 generalized to other parametric models by Rothenberg (1971) and Sargan (1983). Letting θ be a J vector of structural parameters and ϕ be a set of reduced form parameters, these authors consider models that take the form $r(\phi(\theta), \theta) = 0$ for some known vector valued function r. Letting $R(\theta) = r(\phi(\theta), \theta)$, the equation $R(\theta) = 0$ characterizes all of the restrictions that are imposed by the model.

A sufficient condition for local identification of θ in this class of models is that $R(\theta)$ be differentiable and that the rank of $\partial R(\theta)/\partial\theta$ equals J. Sargan (1983) calls any violation of this condition *first order (lack of) identification*, and observes that there exist nonlinear models in which θ is locally identified despite first order lack of identification. Implications of this first order underidentification for inference are discussed by Lee and Chesher (1986), Dovonon and Renault (2009), and Arellano, Hansen and Sentana (2012).

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 θ 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 $\widetilde{\Theta}$ be a subset of Θ, defined as follows: Consider every $\theta \in \Theta$. If θ is observationally equivalent to any other $\widetilde{\theta} \in \Theta$, then include θ in $\widetilde{\Theta}$. This construction means that if θ_0 takes on a value that is in $\widetilde{\Theta}$ then θ_0 is not point identified, otherwise θ_0 is point identified. Proving point identification for any value θ_0 might take on requires that $\widetilde{\Theta}$ be empty. Following McManus (1992), the parameter θ is defined to be *generically identified* if $\widetilde{\Theta}$ is a measure zero subset of Θ.

To interpret what generic identification means, imagine that nature chooses a value θ_0 by randomly picking an element of Θ. Assume all elements of Θ are equally likely to be picked, so nature is drawing from a uniform distribution over the elements of Θ. Generic identification means that there is a zero probability that nature chooses a value for θ_0 that is not point identified.

In models that are systems linear equations (as in Wright-Cowles identification), generic identification is closely related to the order condition for identification. Consider a system of linear regression equations in which the order condition for identification is satisfied. Identification then only requires that a rank condition holds, which can be expressed by saying that a certain, say J by J, matrix of coefficients is

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 $E\left(XX'\right)$ 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.

8 Identification Concepts That Affect Inference

For the most part, identification is treated as a precursor to estimation. We first assume or establish identification, then consider inference and properties of specific estimators given identification of the estimands. However, there are situations where the nature of the identification affects inference. These include concepts like identification at infinity, weak identification, and ill-posed inverse problems. All are forms of point identified parameters, but each describes situations where the nature of the identification affects inference. These are considered identification (rather than estimation) issues, because they are not properties of specific estimators, but rather refer to features of the underlying population that affect the behavior of any estimator one might propose.

In this section we summarize these identification concepts that affect inference. However, it should be noted that some previously discussed concepts are also related to inference, e.g., extremum based identification applies only to objects that are defined in terms of the associated extremum estimators, and inference differs greatly for parameters that are only set identified from those that are point identified.

8.1 Weak vs. Strong Identification

Informally, weak identification arises in situations that are, in a particular way, close to being not point identified. Both weakly identified and strongly identified parameters are point identified. One might not want to call these concepts identification at all, because weak vs. strong identification does not refer to the question of whether the parameters θ are uniquely determined by the knowable information ϕ. The difference between weak and strong identification instead refers to issues associated with inference. They are referred to as forms of identification because they are features of the underlying model and associated DGP, and hence affect any possible estimator we might propose.

Perhaps the earliest recognition of a weak identification problem is Sargan (1983), who wrote, "we can conjecture that if the model is almost unidentifiable then in finite samples it behaves in a way which is difficult to distinguish from the behavior of an exactly unidentifiable model." Other early work that at least hints at the problem include the Handbook of Econometrics chapters by Phillips (1983) and Rothenberg

(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 $\widehat{X}$, and some or all of the model coefficients may be weakly identified if the matrix $E\left(\widehat{X}X'\right)$ 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 $E\left(XX'\right)$ instead of $E\left(\widehat{X}X'\right)$ 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 $Y = \theta X + U$ and $X = \beta Z + V$ 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 $\beta \neq 0$, the parameter θ is identified by $\theta = E(ZY)/E(ZX)$. A corresponding estimator would replace these expectations with sample averages, yielding the standard linear instrumental variables estimator. However, since $E(ZX) = \beta E\left(Z^2\right)$, if β is close to zero then $E(ZX)$ will be close to zero, making θ weakly identified. But how close is close? Small errors in the estimation of $E(ZX)$ will yield large errors in the estimate of θ. The bigger the sample size, the more accurately $E(ZX)$ can be estimated, and hence the closer β can be to zero without causing trouble.

To capture this idea asymptotically, econometricians pretend that the true value of β 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 $X = \beta_n Z + U_x$, where $\beta_n = bn^{-1/2}$ for some constant b. The larger n gets, the smaller the coefficient β_n becomes. This gives us a model where θ 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 β_n is not identified, so tests and confidence regions for θ have been developed that are robust to weak instruments, that is, they do not depend on consistent estimation of β_n. See, e.g., Andrews, Moreira, and Stock (2006) for an overview of such methods.

In the econometrics literature, saying that a parameter θ in a model "is" weakly identified, means that the asymptotics being employed for inference on θ are derived assuming relevant parameters drift towards zero (or more generally towards any value that would result in nonidentification). For example, in the above model an econometrician would say θ is strongly identified when the asymptotics he or she uses are based on $\beta_n = \beta$, and would say θ is weakly identified when the asymptotics used for inference are based on (or at least allow for) $\beta_n = bn^{-1/2}$.

We rarely believe that parameters actually drift towards zero as n grows. Rather, when we assume parameters are weakly identified we are expressing the belief that the drifting parameter model provides a better asymptotic approximation to the true distribution than standard asymptotics do.

Other related terms in the weak instruments literature including *nearly-weak* identification, *semi-strong* identification and *nonstandard-weak* identification. See, e.g., Andrews and Guggenberger (2015), Cheng

(2015), and references therein. These refer to models where parameters, or their impact on θ, 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 $Y = Y^* D$ where D is a binary variable equal to zero or one, Y^* is a latent unobserved variable that is independent of Z, and $\lim_{z\to\infty} E\left(D \mid Z = z\right) = 1$. The goal is identification and estimation of $\theta = E\left(Y^*\right)$. 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, θ 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 θ is identified by $\theta = \lim_{z\to\infty} E\left(Y \mid Z = z\right)$. 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 θ by $\widehat{\theta} = \sum_{i=1}^{n} w\left(n, Z_i\right) Y_i / \left(\sum_{i=1}^{n} w\left(n, Z_i\right)\right)$ where $w\left(n, Z_i\right)$ are weights that grow to infinity with Z in some way. An example would be letting $\widehat{\theta}$ equal the average value of Y just for people that have $Z > c_n$ where $c_n \to \infty$ as $n \to \infty$ (see Heckman 1990 and Andrews and Schafgans 1998), or letting $w\left(n, Z_i\right)$ be the inverse of the density of Z as in Lewbel (2007b).

As first shown by Chamberlain (1986), the general problem with identification at infinity is that esti-

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 $\theta = E\left(Y^*\right)$ 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

parameter values are assumed to vary (typically drift toward zero) with the sample size, in order to obtain good approximations to the true precision with which they can be estimated in moderate sized samples. In contrast, estimates of irregularly identified parameters converge slowly for more or less the same reasons that nonparametric regressions converge slowly, that is, they are based on a vanishingly small subset of the entire data set. The rates of convergence of irregularly or thin set identified parameters can vary widely, from extremely slow (such as $\log n$) up to $n^{1/2}$, depending on details regarding the shape of the density function in the neighborhood of the identifying point. See Khan and Tamer (2010) for examples.

Sometimes proofs of identification rely on thin set identification arguments, but it is possible that such parameters are still strongly identified, because the structure contains more information about the parameter than is being used in the identification proof. For example, Heckman and Honore (1989) showed that without parameterizing error distributions, the nonparametric Roy model or competing risks model can be identified by observations where covariates drive individual risks to zero. Thus their proof is only of thin set identification. However, Lee and Lewbel (2013) later showed that this model is actually strongly identified, using information over the entire DGP, not just data where individual risks go to zero.

Another class of examples of thin set identification is what Hill and Renault (2011) call *eventual identification*. They use this term to describe models where asymptotic trimming is used to obtain limit normal inference based on means of thick tailed distributions.

8.3 Ill-Posed Identification

Suppose that parameters θ are point identified. Problems of ill-posedness arise when the connection from ϕ to θ is not sufficiently smooth. For example, suppose that our DGP is n IID observations W_i of a vector W, so ϕ corresponds to the distribution function of W. Let $F(w)$ denote this distribution function evaluated at the point $W = w$. Recall the Glivenko–Cantelli theorem that a uniformly consistent estimator of F is the empirical distribution function, defined as $\widehat{F}(w) = \sum_{i=1}^{n} I(W_i \leq w)/n$. Suppose we want to estimate a parameter vector or function θ, where $\theta = g(F)$ for some known g. Then θ is point identified, by construction. Moreover, if g is continuous, then $\widehat{\theta} = g\left(\widehat{F}\right)$ would be a consistent estimator for θ.

However, if g is discontinuous, then $\widehat{\theta}$ 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, θ, and ϕ.

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 $\widehat{F}(w)$ 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 $\widehat{f}(w) = \left(\widehat{F}(w+b) - \widehat{F}(w-b)\right)/(2b)$, which for a small value of the bandwidth b is a smooth (regularized) but biased approximation of $dF(w)/dw$. Shrinking b to zero as the sample size grows is what asymptotically shrinks the bias in this estimator to zero. Other examples of regularization are the kernel and bandwidth choices in nonparametric kernel regressions, and the selection of basis functions and the number of terms to include in nonparametric sieve estimators. One might interpret the curse of dimensionality, where rates of convergence slow as the dimension of nonparametric estimators rises, as a property of the regularization required to overcome ill-posedness.

Depending on the application, the degree of ill-posedness can range from mild to moderate to severe. In the case of one dimensional nonparametric density or regression estimation, ill-posedness results in just a moderate slowing in the rate of convergence, typically $n^{2/5}$ instead of $n^{1/2}$. However, in other contexts, such as nonparametric instrumental variables (see, e.g., Newey and Powell 2003), ill-posedness can be severe, causing very slow convergence rates like $\ln(n)$.

[14] More precisely, continuity of the mapping g depends on the chosen norms. Ill-posedness arises when the norm selected for proving consistency results in discontinuity of g.

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 θ 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 θ, and obtain a posterior distribution, regardless of whether θ 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 θ to be *essentially identified* (with respect to a given prior distribution) if θ is not point identified, but would be point identified if the model included the additional assumption that θ is drawn from the given prior distribution. The idea is that unidentified parameters can become identified by adding restrictions to a model. Imposing the restriction that θ be a draw from a given distribution function could be an example of such a restriction. For parameters that are not point identified, Gustafson shows that the behavior of Bayes estimators can depend heavily on whether they are essentially identified or not for the given prior.

A parameter vector θ is defined to be *Bayes identified* if its posterior distribution differs from its prior

distribution. That is, θ is said to be Bayes identified as long as data provides any information at all that updates the prior. This definition of Bayesian identification goes back at least to Florens, Mouchart, and Rolin (1990). A formal definition and analysis of Bayesian identification is provided by Florens and Simoni (2011).

Typically, a parameter that is point identified will also be Bayes identified. Parameters that are set rather than point identified are also generally Bayes identified, since data that provides information about the identified set will typically also provide information that updates ones prior. An example is given by Moon and Schorfheide (2012), who show that if a parameter θ is set identified, where the set is an interval, then the support of the posterior distribution of θ will generally lie inside the identified interval. The intuition is that, even if data tells us nothing about where θ could lie inside the identified interval, the prior does provide information inside this interval, and the posterior reflects that information. See also Gustafson (2015) for more on Bayes estimation in partially identified models.

9 Conclusions

Identification is a rapidly growing area of research within econometrics, as the ever expanding zooful of different terms for identification indicates. Many factors are driving this growth. The rise of big data, with increasingly larger data sets and faster computation, has enabled the specification and estimation of correspondingly larger and more complicated models, requiring increasingly more elaborate identification analyses. Examples of modern models with complicated identification issues include games and auctions, social interactions and network models, forward looking dynamic models, models with nonseparable errors (typically assigning behavioral meaning to unobservables), and general equilibrium models.

At the other extreme, the so-called credibility revolution has lead to a widespread search for sources of randomization in either constructed or natural experiments, with the goal of gaining identification of interpretable if not structural parameters. Like structural models, causal models are also becoming increasingly sophisticated, and a similar search for novel methods of identification exists in the reduced form literature. Regression kink design and the construction of synthetic controls are recent examples.

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 ϕ is a data distribution function.

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

include objects like regression functions, distribution functions of errors or other unobservables, utility functions, payoff matrices, or information sets. Define a *model value* $m \in M$ to be an element of M. So m would correspond to a particular value of the functions, matrices, and sets that comprise the model. For example, if we had the regression model $Y_i = g(X_i) + e_i$, then M could consist of the set of possible regression functions g and the set of possible joint distributions of the regressor X_i and the error term e_i for all i in the population. The elements of the set M could be restricted with a requirement like saying the regression function must be linear, so $g(X_i) = a + bX_i$ for some constant a and b. Other possible restrictions could be that each e_i has finite variance and is mean zero conditional on regressors.

Each model value $m \in M$ implies a data generating process (DGP) (an exception is for incoherent models, which can have values that do not correspond to any DGP). Define ϕ to be a set of constants and/or functions that we assume are known, or knowable, given the DGP. For example, if the DGP consists of independent, identically distributed (IID) observations of the vector $W_i = (Y_i, X_i)$, then ϕ could be the distribution function of W_i, because given data with a large enough number of such observations, this distribution function could, with arbitrarily high probability, be estimated to any desired degree of accuracy. When observations are not independent, as in time series data or with social interactions, then ϕ could consist just of features of the DGP that we assume are knowable, such as means and autocovariances in stationary data, or reduced form linear regression coefficients. If some elements of W (call them X) are determined by experimental design and others, called Y, are observed outcomes given X, then under suitable experimental design ϕ could be the conditional distribution function of Y given X, evaluated at the values of X that could be set by the experimenter.

Since any (coherent) model value m implies a DGP, each m must also imply a value ϕ. Let $\phi = \Pi(m)$ be the function (or more generally, the mapping) that defines the particular observable ϕ that corresponds to any given model value $m \in M$. Let ϕ_0 be the value of ϕ that corresponds to the true DGP, that is, the DGP that actually generates what we can observe or know. The model M is defined to be *misspecified* if $\phi_0 \neq \Pi(m)$ for any $m \in M$. Misspecification implies that what we can observe about the true DGP, which is ϕ_0, cannot satisfy the restrictions of the model M. If the model is not misspecified, then the true

model value m_0 must satisfy $\phi_0 = \Pi(m_0)$. If more than one value of $m \in M$ satisfies $\phi_0 = \Pi(m)$, 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 ϕ_0.

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

Assume that there is a unique value of θ associated with each model value m (violation of this assumption relates to the coherence and completeness conditions; see Section 4 for details). Let $\theta = \Delta(m)$ be the function or mapping that defines the particular parameter value θ that corresponds to the given model value m. The true parameter value θ_0 satisfies $\theta_0 = \Delta(m_0)$.

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

Similar to Θ, define $\Phi = \{\phi \mid \phi = \Pi(m) \text{ where } m \in M\}$. So Φ is the set of all ϕ that are possible given the model M. Any ϕ that is not in Φ is ruled out by the model. While the set Θ embodies or describes all the restrictions on the parameters θ that are implied by the model, the set Φ embodies all of the observable restrictions that are implied by the model (assuming that what we can observe is ϕ). The functions Π and Δ are $\Pi \;\; M \to \Phi$ and $\Delta \;\; M \to \Theta$.

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

Two sets of parameter values θ and $\widetilde{\theta}$ are defined to be *observationally equivalent* in the model M if

there exists a $\phi \in \Phi$ such that $s(\phi, \theta) \neq \varnothing$ and $s\left(\phi, \widetilde{\theta}\right) \neq \varnothing$. Equivalently, θ and $\widetilde{\theta}$ are observationally equivalent if there exist model values m and $\widetilde{m}$ in M such that $\theta = \Delta(m)$, $\widetilde{\theta} = \Delta(\widetilde{m})$, and $\Pi(m) = \Pi(\widetilde{m})$. Roughly, θ and $\widetilde{\theta}$ observationally equivalent means there exists a value ϕ such that, if ϕ is true, then either the value θ or $\widetilde{\theta}$ 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 $\widetilde{\theta} \in \Theta$, having θ and $\widetilde{\theta}$ be *observationally equivalent* implies $\theta = \widetilde{\theta}$. 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 ϕ_0, and $\phi_0 = \Pi(m_0)$. We therefore can't distinguish between m_0 and any other m for which $\Pi(m) = \Pi(m_0)$, and so we can't distinguish between $\theta = \Delta(m)$ and $\theta_0 = \Delta(m_0)$ if any such m exists. And, since we don't know before hand which of the possible values of ϕ will be the ϕ_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 $\widetilde{\theta}$ 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 $\Delta(m)$ is observationally equivalent to some $\Delta(\widetilde{m})$. 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 $\widetilde{m}$ in M are observationally equivalent. This means that for any $m \in M$ and $\widetilde{m} \in M$, if $\Pi(m) = \Pi(\widetilde{m})$ then $m = \widetilde{m}$. Having the model be point identified is sufficient, but stronger than necessary, to also have any $\theta \in \Theta$ be point identified.

11 References

Adcock, R. J. (1877), "Note on the Method of Least Squares," The Analyst (Annals of Mathematics), 4(6), 183-184

Adcock, R. J. (1878), "A Problem in Least Squares," The Analyst (Annals of Mathematics), 5(2),

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.

Angelucci, M. and T. De Maro (2016), "Programme Evaluation and Spillover Effects," J. of Development Effectiveness, 8, 22-43.

Andrews, D. W. K., and P. Guggenberger (2015), "Identification- and Singularity-Robust Inference for Moment Condition Models," Unpublished Manuscript, Cowles Foundation, Yale University

Andrews, D. W. K., M. J. Moreira, and J. H. Stock (2006), "Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression," Econometrica, 74(3) 715-752.

Andrews, D. W., & Schafgans, M. M. (1998), "Semiparametric estimation of the intercept of a sample selection model." The Review of Economic Studies, 65(3), 497-517.

Andrews, I., M. Gentzkow, and J. M. Shapiro (2017), "Measuring the Sensitivity of Parameter Estimates to Estimation Moments," Quarterly Journal of Economics, 132(4), 1553-1592.

Andrews, I., M. Gentzkow, and J. M. Shapiro (2018), "On the Informativeness of Descriptive Statistics for Structural Estimation," Harvard University Working Paper.

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

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996), "Identification of Causal Effects Using Instrumental Variables," Journal of the American Statistical Association, 91, 444-455.

Angrist, J. D., and J-S. Pischke. (2008), Mostly harmless econometrics: An empiricist’s companion.

Princeton university press.

Aradillas-Lopez, A. (2010), "Semiparametric estimation of a simultaneous game with incomplete information," Journal of Econometrics, 157(2), 409-431.

Arellano, M. (2003), “Panel data econometrics,” Oxford University Press.

Arellano, M. and S. Bond, (1991), "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations". Review of Economic Studies 58, 277–297.

Arellano, M., L. P. Hansen, and E. Sentana (2012), "Underidentification?" Journal of Econometrics, 170:2, 256–280.

Balke, A. and J. Pearl, (1997), "Bounds on Treatment Effects from Studies with Imperfect Compliance," Journal of the American Statistical Association 92, 1171-1176.

Banerjee, A. V. and E. Duflo (2009), "The Experimental Approach to Development Economics," Annual Review of Economics, Annual Reviews, 1(1), 151-178.

Baum, C. and A. Lewbel (2019), “Advice on Using Heteroscedasticity Based Identification,” The Stata Journal, Forthcoming.

Behrman, J. R., and P. E. Todd (1999)," Randomness in the Experimental Samples of PROGRESA," International Food Policy Research Institute (IFPRI) report 2, Washington, DC: IFPRI.

Bekker, P.A. and T.J. Wansbeek, (2001), Identification in Parametric Models, in: B.H. Baltagi (ed.), Companion in Theoretical Econometrics, Chapter 7, 144-161, Blackwell, Oxford.

Beran, Feuerverger, and Hall (1996), "On Nonparametric Estimation of Intercept and Slope Distributions in Random Coeffient Regression," Annals of Statistics, 24, 2569-2692.

Beran, R. and P. W. Millar (1994), “Minimum Distance Estimation in Random Coefficient Regression Models,” Annals of Statistics, 22, 1976–1992.

Beresteanu, A., I. Molchanov, and F. Molinari (2011), "Sharp Identification Regions in Models With Convex Moment Predictions," Econometrica, 79(6), 1785-1821.

Berkson, J. (1946), "Limitations of the Application of Fourfold Table Analysis to Hospital Data". Biometrics Bulletin. 2(3), 47–53.

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

Berry, S., J. Levinsohn, and A. Pakes (1995), “Automobile Prices in Market Equilibrium,” Econometrica, 60, 889–917.

Blume, L., W. Brock, S. Durlauf, and Y. Ioannides (2011), "Identification of Social Interactions," in Handbook of Social Economics, J. Benhabib, A. Bisin, and M. Jackson, eds., Amsterdam: North Holland.

Bound, J., Jaeger, D. A., & Baker, R. M. (1995), "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak," Journal of the American statistical association, 90(430), 443-450.

Blundell, R., Smith, R. J. (1994), “Coherency and estimation in simultaneous models with censored or qualitative dependent variables,” Journal of Econometrics, 64(1), 355-373.

Blundell, R., Bond, S. (1998), “Initial conditions and moment restrictions in dynamic panel data models,” Journal of econometrics, 87(1), 115-143.

Blundell, R. W., Powell, J. L. (2004), “Endogeneity in semiparametric binary response models”˙The Review of Economic Studies, 71(3), 655-679.

Bontemps, C., T. Magnac, and E. Maurin, (2012), "Set Identified Linear Models," Econometrica, 80(3) 1129-1155.

Bresnahan, T. F., Reiss, P. C. (1991), “Empirical models of discrete games,” Journal of Econometrics, 48(1), 57-81.

Brock, W. A. and S. N. Durlauf (2001), “Interactions-based Models,” Handbook of Econometrics 5, 3297-3380 (J. J. Heckman and E. Leamer, Eds.) Amsterdam: North-Holland.

Brown, B. W. (1983), “The identification problem in systems nonlinear in the variables,” Econometrica, 51, 175-196.

Box, G. E. P. (1979), "Robustness in the strategy of scientific model building", in Launer, R. L.; Wilkinson, G. N., Robustness in Statistics, Academic Press, pp. 201–236.

Bravo, F., J. C. Escanciano, and T. Otsu (2012), "A Simple Test for Identification in GMM under

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

Chesher, A. (2007),“Identification of nonadditive structural functions,” In Blundell, R., Newey, W.K., Persson, T. eds. Advances in Economics and Econometrics; Theory and Applications, Ninth World Congress, Vol. III. Cambridge University Press. New York.

Chesher, A. and A. M. Rosen (2017), "Generalized Instrumental Variable Models," Econometrica 85(3), 959-989.

Chiappori, P. A., & Ekeland, I. (2009), "The Microeconomics of Efficient Group Behavior: Identification," Econometrica, 77(3), 763-799.

Conlon, C. T. and J. H. Mortimer (2016), Efficiency and Foreclosure Effects of Vertical Rebates: Empirical Evidence. NBER Working Paper No. 19709.

Cosslett, S. R. (1983), "Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model," Econometrica, 51(3), 765-782.

Cox, D. R., (1958), Planning of Experiments. New York: Wiley.

Cragg, J. and S. G. Donald (1993), "Testing identifiability and specification in instrumental variable models," Econometric Theory, 9, 222–240.

Dagenais, M. G. (1997), “A simultaneous probit model,” Unpublished manuscript, Universites d’Aix-Marseille II et III.

Deming, W. E. (1943), Statistical adjustment of data. Wiley, NY

Dong, Y. (2010), "Endogenous Regressor Binary Choice Models Without Instruments, With an Application to Migration," Economics Letters, 107(1), 33-35.

Dong, Y, and A. Lewbel (2015), "Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models,” Review of Economics and Statistics, 97(5) 1081-1092.

Dovonon, P., and E. Renault (2009), “GMM Overidentification Test With First Order Underidentification,” University of North Carolina Working Paper.

Durbin, J. (1954), "Errors in variables", Review of the International Statistical Institute, 22, (1), 23-32.

Engel, E (1857), Die Productions- und Consumtionsverhaltnisse des Königreichs Sachsen" in Zeitschrift des Statistischen Bureaus des Königlich-Sächsischen, Ministerium des Innern, No. 8 u. 9, 1-54. Reprinted

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.

Fox, J. T., and Gandhi, A. (2016), "Nonparametric identification and estimation of random coefficients in multinomial choice models." The Rand Journal of Economics, 47(1), 118-139.

Fréchet, M. (1951), "Sur les tableaux de corrélation dont les marges sont données," Annales de l’Université de Lyon. Section A: Sciences mathématiques et astronomie 9: 53–77.

Freedman, D. A. (2005), Statistical Models: Theory and Practice. Cambridge University Press.

Friedman, M. (1953) "The Methodology of Positive Economics" in, Essays In Positive Economics, Chicago: Universiy of Chicago Press, 3-43.

Frisch, R. (1934), "Statistical confluence analysis by means of complete regression systems," Vol. 5, Universitetets Økonomiske Instituut.

Frölich, M. and M. Huber (2017), "Direct and indirect treatment effects – causal chains and mediation analysis with instrumental variables" Journal of the Royal Statistical Society: Series B 79(5), 1645-1666.

Gourieroux, C., Laffont, J. J., Monfort, A. (1980), “Disequilibrium econometrics in simultaneous equations systems,” Econometrica, 48, 75-96.

Gautier, E. and Kitamura, Y. (2013), "Nonparametric Estimation in Random Coefficients Binary

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 Heterogeneity," The Review of Economics and Statistics, 88(3), 389-432.

Heckman, J. J. and E. J. Vytlacil (2007), "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation," Handbook of Econometrics, in: J.J. Heckman & E.E. Leamer (ed.), Handbook of Econometrics, edition 1, volume 6, chapter 70 Elsevier.

Hill, J. B., and E. Renault (2011), "Generalized Method of Moments with Tail Trimming," Unpublished manuscript.

Hirano, K., G. W. Imbens and G. Ridder (2003) "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," Econometrica, 71(4), 1161-1189.

Hoover, K. D. (2008), "Causality in Economics and Econometrics," The New Palgrave Dictionary of

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,"

Econometric Theory, 13(01), 32-51.

Lewbel, A. (1997b), “Constructing instruments for regressions with measurement error when no additional data are available, with an application to patents and R&D," Econometrica, 65, 1201–1213.

Lewbel, A. (2000), “Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables,” Journal of Econometrics, 97(1), 145-177.

Lewbel, A. (2007a), "Coherence and Completeness of Structural Models Containing a Dummy Endogenous Variable," International Economic Review, 48, 1379-1392.

Lewbel, A. (2007b), “Endogenous selection or treatment model estimation,” Journal of Econometrics, 141(2), 777-806.

Lewbel, A. (2007c),"Modeling Heterogeneity," in Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress (Econometric Society Monographs), Vol. III, chapter 5, Richard Blundell, Whitney K. Newey, and Torsten Persson, editors, Cambridge: Cambridge University Press, .

Lewbel, A. (2012), “Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models,” Journal of Business and Economic Statistics, 30, 67-80.

Lewbel, A. (2014), “An Overview of the Special Regressor Method,” in the Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics, Co-edited by Aman Ullah, Jeffrey Racine, and Liangjun Su, 38-62, Oxford University Press.

Lewbel, A. (2018), “Identification and Estimation Using Heteroscedasticity Without Instruments: The Binary Endogenous Regressor Case,” Economics Letters, 165, 10-12.

Lewbel, A., Y. Dong, and T. T. Yang, (2012), “Viewpoint: Comparing Features of Convenient Estimators for Binary Choice Models With Endogenous Regressors,” Canadian Journal of Economics, 45, 809-829.

Lewbel, A., O. Linton, and D. McFadden, (2011), “Estimating features of a distribution from binomial data,” Journal of Econometrics, 162(2), 170-188.

Lewbel, A. and X. Tang, (2015), “Identification and Estimation of Games with Incomplete Information Using Excluded Regressors,” Journal of Econometrics, 189, 229-244

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 76 945–978

Matzkin R.L. (2012),“Nonparametric identification,” online update of Handbook of Econometrics, Vol 6B, ed J.J. Heckman, E.E. Leamer, pp 5307–5368 Amsterdam: Elsevier

Matzkin R.L. (2015), “Estimation of Nonparametric ModelsWith Simultaneity,” Econometrica, 83, 1–66.

McManus, D. A. (1992), “How common is identification in parametric models?” Journal of Econometrics, 53(1), 5-23.

Mill, J. S. (1851) A System of Logic, Ratiocinative and Deductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation, 3rd. ed., vol. I. London: John W. Parker.

Moon, H. R. and F. Schorfheide, (2012) "Bayesian and Frequentist Inference in Partially Identified Models, Econometrica, 80(2) 755-782.

Newey, W. K., McFadden, D. (1994), “Large sample estimation and hypothesis testing,” Handbook of econometrics, 4, 2111-2245.

Newey, W. K., Powell, J. L. (2003), “Instrumental variable estimation of nonparametric models,” Econometrica, 71(5), 1565-1578.

Neyman, J. (1937), “Outline of a theory of statistical estimation based on the classical theory of probability,” Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences, 236(767), 333-380.

Neyman, J., K. Iwaszkiewicz, and St. Kolodziejczyk, (1935), "Statistical Problems in Agricultural

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.

Powell, J. L. (1994), “Estimation of semiparametric models,” Handbook of econometrics, 4, 2443-2521.

Powell, J. L., J. Stock, and T. M. Stoker (1989), "Semiparametric Estimation of Index Coefficients," Econometrica, 57, 1403-1430.

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

Reiersøl, O. (1945), Confluence Analysis by Means of Instrumental Sets of Variables. Uppsala: Almquist & Wiksell.

Reiersøl, O. (1950), “Identifiability of a linear relation between variables which are subject to error,” Econometrica, 18, 375-389.

Rigobon, R., (2003), "Identification Through Heteroskedasticity, The Review of Economics and Statistics," 85(4), 777-792.

Robinson, P. M. (1988), “Root-N-consistent semiparametric regression,” Econometrica, 56, 931-954.

Roehrig, C. S. (1988), “Conditions for identification in nonparametric and parametric models,” Econometrica, 56, 433-447.

Rosenbaum, P. R., Rubin, D. B. (1983), “The central role of the propensity score in observational studies for causal effects,” Biometrika, 70(1), 41-55.

Rosenbaum, P. R.,(2007), “Interference Between Units in Randomized Experiments,” Journal of the American Statistical Association, 102, 191-200.

Rosenzweig, M. and Udry, C. (2016), "External Validity in a Stochastic World," NBER Working Paper No. 22449.

Rothenberg, T. J. (1971), “Identification in parametric models,” Econometrica, 39, 577-591.

Rothenberg, T. J. (1984), “Approximating the distributions of econometric estimators and test statistics,” Handbook of econometrics, 2, pp 881-935

Roy, A. (1951), "Some thoughts on the distribution of earnings," Oxford Economic Papers, 3(2), 135–146.

Rubin, D. B. (1974), “Estimating causal effects of treatments in randomized and nonrandomized studies,” Journal of educational Psychology, 66(5), 688.

Rubin, D. B. (1980), “Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment," Journal of the American Statistical Association, 75(371), 591-593.

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. (1923), “On the Application of Probability Theory to Agricultural Experiments," Essay on Principles, Section 9. published in english in (1990) Statistical Science 5(4), 465–472, translated by Dorota M. Dabrowska and Terence P. Speed.

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.