The Identification Zoo - Meanings of Identification in Econometrics: PART 1

Arthur Lewbel

Boston College

original 2015, heavily revised 2018

The Identification Zoo - Part 1 - sections 1, 2, and 3.

(These are notes to accompany the survey article of the same name in the Journal of Economic Literature).

Well over two dozen types of identification appear in the econometrics literature, including (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.

1. Introduction

Econometric identification really means just one thing:

Model parameters or features uniquely determined from the observable population that data are drawn from.

Goals:

1. Provide a new general framework for characterizing identification concepts

2. Define and summarize, with examples, the many different terms associated with identification.

3. Show how these terms relate to each other.

4. Discuss concepts closely related to identification, e.g., observational equivalence, normalizations, and the differences in identification between structural models and randomization based reduced form (causal) models.

Table of Contents:

Continued next slide

Table of Contents - continued:

Continued next slide

Table of Contents - continued:

Part 2 will have sections:

4. Coherence, Completeness and Reduced Forms

5. Causal Reduced Form vs Structural Model Identification

Part 3 will have sections:

6. Identification of Functions and Sets

7. Limited Forms of Identification

8. Identification Concepts that Affect Inference

9. Conclusions

Identifying Identification

Let θ be unknown parameters, vectors and/or functions.
θ is what we want to learn about, and hopefully, estimate.

Let ϕ be what is "knowable" about the data generating process (DGP) from data.

Example: θ is the vector of coefficients of traditional linear supply and demand curves. We can estimate linear reduced form regression coefficients. The probability limits of those regression coefficients are ϕ.

Example: with independent, identically distributed (IID) data, the distribution function of the data can be consistently estimated (the Glivenko–Cantelli theorem). So with IID data the distribution function is ϕ, and θ could include objects like structural model coefficients, elasticities, and error distributions.

Example: in an ideal randomized control trial (RCT) experiment, ϕ is the conditional distribution of the outcome given treatment, and θ could be the average of the treatment effect over some population.

The identification question: Given ϕ, which is what's knowable about the DGP, what can be learned about θ?

We say θ is identified, or more precisely, point identified, if given what ϕ equals, we would know the value that θ equals.

θ is partially identified if we can say something about it's value, but not know it exactly, given ϕ.

Identification logically precedes estimation, inference and testing.

Note: Previous definitions of identification all made specific (varying) assumptions about what ϕ was (Cowles, Sargan, Rubin, Newey-McFadden). This paper generalizes those by allowing ϕ to vary by context.

2. The Historical Roots of Identification

Before identification we need the notion of "ceteris paribus," that is, holding other things equal.

Formal application of this concept to economics attributed to Alfred Marshall (1890).

But earliest economic example is from William Petty (1662), "A Treatise of Taxes and Contributions:"

"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."

This may be the earliest example of identification: a claimed causal effect on prices.

Philip Wright (1915) defines the classic identification problem in economics, pointing 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.

Sewall Wright (1925) (Philip's son, a genetics statistician), invented causal path diagrams, and used them to construct an instrumental variables estimator, but likely for computational convenience instead of OLS, in a model of all exogenous regressors.

Earliest known solution to an identification problem in econometrics (linear regression using instrumental variables) is Philip Wright (1928), Appendix B, applying his son's methods.

Stock and Trebbi (2003) discuss whether Appendix B was actually written by Philip or Sewall. By stylometric analysis (statistical analysis of literary styles), they conclude that Philip Wright wrote Appendix B.

Aside: Sewall Wright's first application of causal path diagrams was to determine 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 looked at pig iron, the son studied actual pigs.

In addition to two different Wrights, two different Workings also worked on the subject

Holbrook Working (1925) and, more relevantly, Elmer J. Working (1927). Both wrote about statistical demand curves (Holbrook is the one for whom the Working-Leser Engel curve is named).

Jan Tinbergen (1930) proposed indirect least squares estimation, but like Sewall Wright, only for convenience not for solving identification.

Others on identification with simultaneity: 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.

Related important early work: Abraham Wald (1950), Henri Theil (1953), J. Denis Sargan (1958), Franklin Fisher (1966), and (using error restrictions) Karl G. Jöreskog (1970).

Milton Friedman (1953) critiques Cowles foundation work - warns against using different criteria to select models versus criteria to identify them.

A different problem: Causal Modeling - Identifying a treatment effect.

Identification based on randomization: Jerzy Neyman (1923), David R. Cox (1958), Donald B. Rubin (1978), many others.

In contrast to random selection, econometricians historically focused on cases where selection (who is treated or observed) and outcomes are correlated. Sources of correlation:

Simultaneity as in Trygve Haavelmo (1943). Pearl (2015) and Heckman and Pinto (2015) credit Haavelmo as the first rigorous treatment of causality in the context of structural econometric models.

Optimizing self selection as in Andrew D. Roy (1951).
Survivorship bias as in Abraham Wald (1943) - treatment assignment is random, but sample attrition is correlated with outcomes (WW II planes hit randomly, only ones hit in survivable spots return to be observed).
General models where selection and outcomes are correlated - James J. Heckman (1978).
Formal use of Causal Diagrams: Pearl (1988)

Another identification problem: identifying true linear regression coefficients when regressors are measured with error.

Robert J. Adcock (1877, 1878), and Charles H. Kummell (1879): measurement errors in "Deming regression", (popularized in stats lit by W. Edwards Deming 1943). Is regression that mins least squares errors measured perpendicular to the fitted line.

Corrado Gini (1921) gave an estimator for measurement errors in standard linear regression.
Ragnar A. K. Frisch (1934) was 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 partial or set identification.

3. Point Identification

In modern terminology, the standard notion of identification is called point identification (in other contexts, called global identification or frequentist identification).

Some early formal definitions of identification, structure and observational equivalence: Koopmans and Reiersøl (1950), Hurwicz (1950), Fisher (1966) and Rothenberg (1971). See Chesher (2008) for additional historical details on these classical identification concepts

In this survey I provide a general definition of point identification.

This new generalization maintains the intuition of existing classical definitions while encompassing a larger class of models than previous definitions

3.1 Introduction to Point Identification

Recall θ is unknown parameters, vectors and/or functions - what we want to learn about and hopefully, estimate.

Let ϕ be information that is assumed known, or that we could learn given an unlimited amount of whatever type of data we have.

Examples of ϕ: distribution functions, conditional means, quantiles, autocovariances, or true regression coefficients.

A model M imposes restrictions on the possible values ϕ could take on.

Simplest definition: Given the model M, parameter θ is *point identified* if θ is uniquely determined from ϕ.

Usually think of a model M as set of equations describing behavior.

More generally, a model corresponds to assumptions about and restrictions on the DGP.

This includes 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.

So, identification (even in purely experimental settings) *always* entails a model.

EXAMPLE: For scalars Y, X, and θ, model is that $Y = X\theta + e$ where $E\left(X^2\right) \neq 0$ and $E\left(eX\right) = 0$.

Assume ϕ, what we can learn from data, includes second moments of (Y, X).

Then θ is point identified: Have $\theta = E\left(XY\right) / E\left(X^2\right)$, which is a function of ϕ.

EXAMPLE: X is a treatment indicator. Model says X is determined by outcome of a coin flip.
Y is each individual's outcome. Observe realizations of (X, Y), independent across individuals.

Assume ϕ includes $E\left(Y \mid X\right)$. Let θ be the average treatment effect (ATE).

Given the model, θ is identified by the difference in means $\theta = E\left(Y \mid X = 1\right) - E\left(Y \mid X = 0\right)$.

Both of the examples assume expectations of observed variables are knowable, and so can be included in ϕ.

To justify the assumption, might appeal to statistical properties of (observable) sample averages:

Unbiasedness or (given a weak law of large numbers) consistency.

The definition of identification is somewhat circular:

Start by assuming something, ϕ, is identified to end by determining if something else, θ, is identified.

Assuming ϕ is knowable, or identified, must be justified by deeper assumptions regarding the underlying DGP (Data Generating Process).

Common DGP assumptions:

1. IID (Independently, Identically Distributed) observations of a vector W, with sample size $n \longrightarrow \infty$.
With such data can consistently estimate the distribution of W by the Glivenko–Cantelli theorem.
So reasonable to assume knowable ϕ is the distribution function of W.

2. Each observation of X is a value chosen by experiment.
Conditional on that value of X, randomly draw an observation of Y, (independent of other observations).
ϕ is the conditional distribution function of Y given X.
ϕ is only knowable for values of X that can be set by the experiment.

3. Stationary time series data: ϕ is variances and autocovariances
Not higher moments if they could be unstable over time.

ϕ depends on the model.

Example: In dynamic panel data models, the Arellano and Bond (1991) estimator is based on moments that are assumed knowable (can be estimated from data) and equal zero in the population.

Blundell and Bond (1998) provides additional moments (functional form information about the initial time period zero). Possible that θ is not identified with Arellano and Bond moments, but becomes identified if the model restricts ϕ by assuming Blundell and Bond moments also hold.

Example: experimental design, random assignment into treatment and control groups. Still need a model for identification of treatment effects. Typical model assumptions rule out measurement errors, sample attrition, censoring, social interactions, and general equilibrium effects.

Two types of DGP assumptions.

1. Assumptions regarding collection of data, e.g., selection, measurement errors, and survey attrition.

2. Assumptions regarding 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*.

Both types of assumptions determine the model M and what is knowable ϕ, and hence determine what identification is possible.

Identification logically precedes estimation. If θ is not point identified, then estimators for θ having some desirable properties (like consistency) will not exist.

However, identification does not by itself imply that estimators with any particular desired properties exist, only that they might.

Example: Suppose $\theta = E(X)$, and the DGP is such that θ is finite. With iid observations of X, we can show that $\theta = E(X)$ is identified.

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 in Section 8) are two situations where, despite being point identified, any estimator of θ will have some undesirable properties.

Big Data

In some ways, 'big data' is (or should be) about identification.

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 treated as if it were the population.

Identification deals precisely with what can be learned about the relationships among variables given the population.

3.2 Defining Point Identification

Recall ϕ is a set of constants and/or functions that we assume are known, or knowable, given the DGP.

Examples: ϕ could be:
i. the distribution of Y, X if IID observations.
ii. means and autocovariances in stationary data
iii. reduced form linear regression coefficients
iv. conditional distribution of Y given X where X values are set by experiment.
v. transition probabilities, if W follows a martingale process.

Previous definitions of point identification in the literature each started from a particular definition of ϕ. Examples:
– in Matzkin (2007, 2012), ϕ is a distribution function.
– In textbook linear supply and demand curves, ϕ is regression coefficients.

This survey generalizes and encompasses previous definitions by allowing ϕ to depend on context.

Recall parameters θ are a set of unknown constants and/or functions that characterize or summarize relevant features of a model.

θ can be anything we might want to estimate (θ will generally be estimands, i.e., population values of estimators of objects that we want to learn about).

Examples θ could include regression coefficients, the sign of an elasticity, an average treatment effect, or an error distribution.

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.

Rough definitions of observational equivalence and of point identification:

Two possible values $\overline{\theta}$ and $\widetilde{\theta}$ are *observationally equivalent* if there exists a value of ϕ that could imply either $\overline{\theta}$ and $\widetilde{\theta}$.

θ is *point identified* if $\overline{\theta}$ and $\widetilde{\theta}$ being observationally equivalent implies $\overline{\theta}$ and $\widetilde{\theta}$ are equal.

That is, θ is *point identified* if each possible value of ϕ implies a unique value of θ.

The remainder of this subsection (which can be skipped if one's primary interest is in later sections) defines point identification a little more precisely.

A more mathematically rigorous definition is provided in the Appendix of the Identification Zoo survey.

Definitions:

A *model* M is a set of functions or sets that satisfy some given restrictions.

M can include restrictions on regression functions, distribution functions of errors or other unobservables, utility functions, payoff matrices, or information sets.

A *model value* $m \in M$ is an element of M. So m is a particular value of the functions, matrices, and sets that comprise the model.

Example: If $Y_i = g(X_i) + e_i$, then M could be 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 M could be restricted: e.g., require linearity $g(X_i) = a + bX_i$. Other possible restrictions: $var(e_i)$ finite, $E(e_i \mid X) = 0$.

Each model value $m \in M$ generally implies a unique data generating process (DGP). Exceptions are incoherent models - see section 4.

Assume each model value $m \in M$ implies a particular value of ϕ and of θ.

Violations of this assumption can lead to incoherence or incompleteness - see section 4.

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 m that yield both the given values of ϕ and of θ.

Let Θ denote the set of all possible values that the model says θ could be.

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.

θ and $\widetilde{\theta}$ observationally equivalent means there exists a ϕ and model values m and $\widetilde{m}$ such that m implies the values ϕ and θ, and $\widetilde{m}$ implies the values ϕ and $\widetilde{\theta}$..

Definition of 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 $\widetilde{\theta}$ in Θ that are different but observationally equivalent.

Let θ_0 be the unknown true value of θ.

The particular value θ_0 is point identified if θ_0 not observationally equivalent to any other θ in Θ.

But we don't know which of the possible values of $\theta \in \Theta$ is θ_0.

So 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.

Showing identification in theory

We have defined point identification of parameters θ.

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 the parameters θ).

Identification of the model implies identification of any model parameters θ.

We define the model M, so we could in theory
1. enumerate every $m \in M$,
2. list every ϕ and θ that is implied by each m, and thereby determine every $s(\phi, \theta)$
3. check every value of every pair of structures $s(\phi, \theta)$ 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.

Misspecification

Let ϕ_0 be the value of ϕ that corresponds to the true DGP.

A model M is defined to be *misspecified* if there does not exist any model value $m \in M$ that yields ϕ_0.

Misspecification means that what we can observe about the true DGP, which is ϕ_0, cannot satisfy the restrictions of the model M.

If our model M is not misspecified, then there exists a model value m_0 which implies ϕ_0.

What is the true model value? What is meant by truth of a model, since models only approximate the real world?

We avoid that question, by just saying that, whatever the "true" model value m_0 is, it has the property of not conflicting with what we can potentially observe or know, which is the true ϕ_0.

Additional Definitions

Ensuring point identification can require ruling out some potential values of θ.

Local and generic identification are examples (are discussed in more detail later)

Local identification of θ means that there exists a neighborhood of θ such that, for all values $\widetilde{\theta}$ in this neighborhood (other than the value θ) θ is not observationally equivalent to $\widetilde{\theta}$.

Generic identification means that set of values of θ in Θ that cannot be point identified is a very small subset (formally, a measure zero subset) of Θ.

θ_0 is said to be *set identified* (or *partially identified*) if there exist some values of $\theta \in \Theta$ that are not observationally equivalent to θ_0.

The only time a parameter θ is not set identified is when all $\theta \in \Theta$ are observationally equivalent.

The *identified set* is the set of all values of $\theta \in \Theta$ that are observationally equivalent to θ_0.

Point identification of θ_0 is when the identified set contains only one element, which is θ_0.

Parametric identification is where θ is a finite set of constants, and all values of ϕ correspond to values of a finite set of constants.

Nonparametric identification is where θ consists of functions or infinite sets.

Other cases are called *semiparametric identification*, e.g., θ includes both a vector of constants and some functions.

3.3 Examples and Classes of Point Identification

Example 1: a median.
M is set of possible distributions of continuous W with strictly monotonically increasing distribution functions $F(w)$.
DGP is IID draws of W. Each ϕ is an F function.
Each model value m happens to correspond to a unique value of ϕ.

Let θ be the median of W.
The structure $s(F, \theta)$ has one element if $F(\theta) = 1/2$, is empty otherwise.
No pair $\theta \neq \widetilde{\theta}$ are observationally equivalent, because $F(\theta) = 1/2$ and $F\left(\widetilde{\theta}\right) = 1/2$ implies $\theta = \widetilde{\theta}$.
θ is identified because it's the unique solution to $F(\theta) = 1/2$. Knowing F, we can determine θ.

Example 2: Linear regression.

DGP is observations of Y, X where Y is a scalar, X is a $K-$vector.
Observations of Y, X might not be IID.
ϕ is first and second moments of X and Y. Assumed finite, constant across observations.
M is the set of joint distributions of e, X that satisfy $Y = X'\theta + e$, $E(Xe) = 0$ for an error term e.
$s(\phi, \theta)$ is nonempty when moments comprising ϕ satisfy $E[X(Y - X'\theta)] = 0$ for the given θ.

If restrict M by assuming $E(XX')$ is nonsingular, then θ identified by $\theta = E(XX')^{-1} E(XY)$.
Otherwise $\widetilde{\theta} = E(XX')^{-} E(XY)$ for different pseudoinverses $E(XX')^{-}$ are observationally equivalent.

Identification is parametric: θ and ϕ are finite vectors.
Would be semiparametric if, e.g., assumed ϕ was distribution of Y, X under IID data, and parameter set included the distribution function of e.

Example 3: treatment.

DGP: Assign treatment $T = 0$ or $T = 1$, generate an outcome Y.
Y, T independent across individuals. ϕ is distribution of Y, T.

Rubin (1974) causal notation: Random $Y(t)$ is the outcome an individual would have if assigned $T = t$.
θ is the average treatment effect (ATE), defined by
$\theta = E(Y(1) - Y(0))$.
M is the set of all possible joint distributions of $Y(1)$, $Y(0)$, and T.
A restriction on M: Rosenbaum and Rubin's (1983) assumption that $(Y(1), Y(0))$ is independent of T.
Rubin (1990) calls this unconfoundedness, is equivalent to random assignment of treatment.

θ is identified because unconfoundedness implies that
$\theta = E(Y \mid T = 1) - E(Y \mid T = 0)$.

Heckman, Ichimura, and Todd (1998): a weaker sufficient condition for identification of θ is the mean unconfoundedness assumption that $E\left(Y\left(t\right) \mid T\right) = E\left(Y\left(t\right)\right)$.

Without some form of unconfoundedness, θ might not equal $E\left(Y \mid T = 1\right) - E\left(Y \mid T = 0\right)$.

More relevantly for identification, without unconfoundedness, different joint distributions of $Y\left(1\right)$, $Y\left(0\right)$, and T (i.e., different model values m) might yield the same joint distribution ϕ, but have different values for θ.

Those different θ values would then be observationally equivalent to each other, and so we would not have point identification.

Above can all be generalized to allow for covariates. The key for identification is not a closed form expression like $E\left(Y \mid T = 1\right) - E\left(Y \mid T = 0\right)$ for θ. The key is a unique value of θ for each possible ϕ.

Example 4: linear supply and demand.

In each time period, demand $Y = bX + cZ + U$ and supply $Y = aX + \varepsilon$.
Y quantity, X price, Z income, U, ε mean zero errors, independent of Z.
Each model value m consists of a particular joint distribution of Z, U, and ε in every time period.
These distributions could change over time.

ϕ could be 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.
Solving for the reduced form coefficients: $\phi_1 = ac/(a-b)$ and $\phi_2 = c/(a-b)$.

Demand $Y = bX + cZ + U$ and Supply $Y = aX + \varepsilon$.
Let θ be a, the supply price coefficient.

A given structure $s(\phi, \theta)$ contains all model values m that satisfy $\theta = a$, $\phi_1 = ac/(a-b)$, and $\phi_2 = c/(a-b)$.

If $c \neq 0$, then $\phi_1/\phi_2 = a$, so $s(\phi, \theta)$ is empty if $c \neq 0$ and $\phi_1/\phi_2 \neq \theta$. Otherwise, $s(\phi, \theta)$ contains many elements m, because there are many different possible distributions of Z, U, and ε that can go with each such ϕ and θ.

θ is not identified unless we add the restriction $c \neq 0$ which implies that $\phi_2 \neq 0$. Otherwise any θ and $\widetilde{\theta}$ will be observationally equivalent when $\phi = (0, 0)$.

Example 5: latent error distribution.

DGP is IID (Y, X). ϕ is the joint distribution of Y, X.
M is the set of joint distributions of X, U satisfying: X is continuous, $U \perp X$, and $Y = I(X + U > 0)$.
$\theta = F_U(u)$, the distribution function of U.

For any value x that X can take on, we have $E(Y \mid X = x) =$
$\Pr(X + U > 0 \mid X = x) = \Pr(x + U > 0) = 1 - \Pr(U \leq -x)$
$= 1 - F_U(-x)$.

So function F_U is nonparametrically identified; it can be recovered from $E(Y \mid X = x)$.
But $F_U(u)$ is only identified for values of u that are in the support of $-X$.

This is the logic behind the identification of Lewbel's (2000) special regressor estimator (see section 6.4 later).

Many identification arguments begin with one of three cases:
1. ϕ is a set of reduced form regression coefficients
2. ϕ is a data distribution, or
3. ϕ is the maximizer of some function.

These starting points are common enough to deserve names, so I will call these classes
1. Wright-Cowles identification,
2. Distribution Based identification, and
3. Extremum Based identification.

Wright-Cowles Identification

Associated with Philip and Sewall Wright and with Cowles foundation,
Concerning linear systems like supply and demand equations.

Y a vector of endogenous variables
X a vector of exogenous variables (regressors and instruments).

ϕ is the matrix of population reduced form linear regression coefficients,
i.e., the coefficients obtained from a linear projection of Y onto X.

M is structural linear equations.
Restrictions defining M 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.

θ is a set of structural model coefficients we wish to identify.
Examples: θ could be the coefficients of one equation, e.g., the demand equation.
θ could be all the coefficients in the structural model
θ could just be a single price elasticity.
θ could be some function of coefficients, like an elasticity.

For each possible ϕ, θ, a structure $s(\phi, \theta)$ is all model values m having structural coefficients equal θ and reduced form coefficients equal ϕ.

Identification of θ requires that there can't be any $\widetilde{\theta} \neq \theta$ that has the same matrix of reduced form coefficients ϕ that θ could have.

Note there could be multiple values of ϕ consistent with any given θ.

A convenient feature of Wright-Cowles identification:
It can be applied to time series, panel, or other DGP's with dependence across observations.
Only require that reduced form linear regression coefficients have some well defined limiting value ϕ.

Identification of linear models sometimes combines restrictions on structural coefficients with restrictions on $cov(errors|X)$.
Then need to expand definition of ϕ to include both reduced form coefficients and $cov(errors|X)$.
Assumes $cov(errors|X)$ is knowable.

When ϕ includes these error covariances, Identification is then sometimes possible without exclusion based instruments. Examples: LISREL model of Jöreskog (1970), heteroskedasticity based identification of Lewbel (2012, 2018).

Distribution Based Identification

Equivalent to the definition of identification given by Matzkin (2007, 2012). Also see Hsiao (1983).

Assumes ϕ is the distribution function of an observable random vector Y (or the conditional distribution function of Y given a vector X).

Definition derived from Koopmans and Reiersøl (1950), Hurwicz (1950), Fisher (1966), and Rothenberg (1971).

In these earlier references, model implied ϕ was in a known parametric family, so ϕ could be estimated by maximum likelihood.

Suitable for IID data, where ϕ would be nonparametrically knowable by the Glivenko-Cantelli theorem.

Could also apply to non-IID DGP's, if the distribution is sufficiently parameterized.

θ could be:
parameters of a parameterized distribution function
features of ϕ like moments or quantiles, possibly conditional.
constants or functions describing a behavioral or treatment model generating data drawn from the distribution ϕ.

θ is point identified if it's uniquely determined from knowing the distribution function ϕ.

Note the difference:

Distribution based identification assumes an entire distribution function is knowable.

Wright-Cowles just assumes features of the first and second moments of data are knowable.

Extremum Based Identification:

Following Sargan (1959, 1983), Amemiya (1985), and Newey and McFadden (1994).

Extremum estimators maximize an objective function, such as GMM, MLE, or least squares.

In Extremum based identification, each model value m is associated with the value of a function G.

ϕ is set of values of vectors or functions ζ that maximize $G(\zeta)$.

θ is 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 $G(\zeta)$.

Connection to extremum estimation:

Consider example of an extremum estimator that maximizes an average with IID data:
Assume $\widehat{\phi}$ equals the set of all ζ that maximize $\sum_{i=1}^{n} g\left(W_i, \zeta\right) / n$
where IID W_i are observations of an observable data vector,
and g is a known function.

e.g. if $-g\left(W_i, \zeta\right)$ is a squared error term this would be a least squares estimator.

if g is the probability density of W_i, this would be a maximum likelihood estimator.

Would then define G by $G\left(\zeta\right) = E\left(g\left(W_i, \zeta\right)\right)$.

More generally G would be the probability limit of the extremum objective function.

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.

Wright-Cowles identification can be 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 complicated DGP's, since it only requires that maximizing values of a given objective function G be knowable.

In Extremum based identification nothing about the DGP is assumed to be known other than the maximizing values of the objective function G.

An advantage: that's the identification one needs to establish asymptotic properties of any given extremum based estimator.

A drawback: Doesn't say anything about whether θ could have been identified from other features of the underlying DGP that might be knowable in practice.

Example: DGP is IID observations of a bounded scalar W and $\theta_0 = E(W)$. Applying Distribution based identification, have that θ_0 is identified.

But consider Extremum based identification with $G(\zeta) = -E\left[(W - |\zeta|)^2\right]$.
Is maxed by $\phi = \{\theta_0, -\theta_0\}$. So θ_0 and $-\theta_0$ are observationally equivalent, and θ is not identified, using this G.

Here G failed to account for other info that would be knowable given IID data.

Failure of extremum based identification can be due either to more fundamental nonidentification in the DGP, or due to the particular choice of objective function.

This problem typically does not apply to Distribution based parametric identification. Since (given regularity) conditions, the likelihood function contains all of the info about parameters that is available in the population.

However, this issue can arise in Wright-Cowles identification.
By defining ϕ just in terms first and second moments, Wright-Cowles ignores potential additional info in the DGP.

Example: Lewbel (1997b) uses some information in third moments to obtain identification in models containing mismeasured covariates without instruments (other examples in section 3.7).

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 Proving Point Identification

In the earlier examples, identification was proved "by construction:" Writing θ directly as a function of ϕ:

example 1: $\theta = F^{-1}(1/2)$

example 2: $\theta = E(XX')^{-1} E(XY)$

example 3: $\theta = E(Y \mid X = 1) - E(Y \mid X = 0)$

example 5: $\theta = F_{U|Z}(u \mid z) = 1 - E(Y \mid X = -u, Z = z)$.

Typical (especially in the statistics literature) is to directly prove consistency. Construct an estimator $\widehat{\theta}$ and prove that, under the assumed DGP, $\widehat{\theta}$ is consistent.

This is a special case of identification by construction, where the construction is $\theta = plim\ \widehat{\theta}$.

In example 2 above $\widehat{\theta}$ would be the standard ordinary least squares regression coefficient. Others are similar.

Caution: Some proofs of consistency either implicitly or explicitly assume identification. E.g. Theorem 2.1 of Newey and McFadden (1994) proves the consistency of extremum estimators. But it includes extremum based identification as one of its assumptions.

An example of proving identification by proving consistency:

DGP is IID observations of a vector W.
$F(w)$ is the distribution function of W, evaluated at the value w.
The empirical distribution function is $\widehat{F}(w) = \sum_{i=1}^{n} I(W_i \leq w)/n$
$\widehat{F}$ estimates the probability that $W \leq w$ by counting observations of W_i that are less than w.

The Glivenko–Cantelli theorem: If W_i are IID, $\widehat{F}(w)$ is a uniformly consistent estimator of $F(w)$.

This gives identification of the function $F(w)$ by construction, taking the probability limit of $\widehat{F}(w)$

This justifies the starting assumption with IID data that what is knowable, ϕ, is the distribution function of W.

1. 'By construction' is the commonest way to prove identification.

Other methods are:

2. Proving true θ is the unique solution to an optimization problem.
Example: maximum likelihood with a concave population objective function (e.g., probit, see Haberman 1974).

3. Applying characterizations of observational equivalence in some classes of models. See Roehrig (1988) and Matzkin (2008).

4. Showing the true θ is the unique fixed point in a contraction mapping based on M.
Example: The BLP model (Berry, Levinsohn, and Pakes 1995) doesn't quite do this, but a contraction mapping is used to prove that a necessary condition for identification, uniqueness in the error inversion step, holds.
Example: Pastorello, Patilea, and Renault (2003) use a fixed point Extremum based identification assumption for their proposed latent backfitting estimator.

For many examples of applying these methods to prove identification, see Matzkin (2005, 2007, 2012).

In some cases, it is possible to empirically test for identification.

These are generally tests of extremum based identification, based on the behavior of associated extremum based estimators.

Examples: Cragg and Donald (1993), Wright (2003), Inoue and Rossi (2011), Bravo, Escanciano, and Otsu (2012), and Arellano, Hansen, and Sentana (2012).

Point identification can be defined without reference to any data at all!

Example: Model M consists of (regular) utility functions maximized under a linear budget constraint.

$\phi =$ Demand functions, $\theta =$ Indifference curves.

Revealed preference theory of Samuelson (1938,1948), Houthakker (1950), and Mas-Colell (1978) shows point identification of θ.

Note again the sense in whch definitions of identification can be a bit circular or recursive: Start by assuming something is identified (demand functions) to show that something else is identified (indifference curves), given the revealed preference assumptions.

A separate question would then be when or whether demand functions can be identified from observed data.

3.5 Common Reasons for Failure of Point Identification

Typical (somewhat overlapping) reasons identification fails or is difficult to attain:

1. model incompleteness,
2. perfect collinearity or dependence,
3. nonlinearity,
4. simultaneity,
5. endogeneity,
6. unobservability.

1. Incompleteness: variable relationships not fully specified (more about completeness and coherence later).

Example: games having multiple equilibria, without or unknown equilibrium selection rule.

2. Perfect collinearity.
Let $Y_i = a + bX_i + cZ_i + e_i$. If X_i is linear in Z_i, can't identify a, b, c.

Perfect dependence:
Can't identify the function $g(X, Z) = E(Y \mid X, Z)$ if $X = h(Z)$.

3. Nonlinearity can cause multiple solutions:
Example: $Y = (X - \theta)^2 + e$ with $E(e) = 0$.
Then true θ_0 satisfies $E\left(Y - (X - \theta)^2\right) = 0$
True θ_0 is one of two roots of $E\left(Y - X^2\right) + 2E(X)\theta - \theta^2 = 0$.
Identification needs more info, e.g., maybe knowing sign of θ.
Without more data, θ is locally identified.

4. Simultaneity: X and Y being determined jointly or simultaneously

Classical Cowles foundation analysis of identification.

supply curve: $Y = aX + \varepsilon$
demand curve $Y = bX + cZ + U$

Y is log quantity, X is log price, Z is log income, errors $E\left[(\varepsilon, U) \mid Z\right] = 0$.

For simplicity, assume all variables mean zero, so no constant terms.

supply: $Y = aX + \varepsilon$ demand: $Y = bX + cZ + U$

Moments we have in ϕ are related by:

$$E(ZY) = E(ZX)\, a$$
$$E(ZY) = E(ZX)\, b + E(ZZ)\, c$$

a is identified by $a = E(ZY)/E(ZX)$ if $E(ZX) \neq 0$.
But for b and c all we have is $E(ZX)(a-b) = E(ZZ)\, c$

Equate supply and demand to get $E(ZX) = E(ZZ)\, c/(a-b)$.

a is identified by $a = E(ZY)/E(ZX)$ if $E(ZX) \neq 0$.
b and c are not identified since 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 $\widetilde{U} = \lambda U + (1-\lambda)\, \varepsilon$.

Graphical interpretation:
variation in instrument Z moves the demand curve
intersection of the two curves at varying Z values trace out the supply slope.
have no information about demand slope
Essentially, only see one point on the demand curve.

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

5. Endogeneity is the general problem of regressors being correlated with errors.

Simultaneity is one source of endogeneity.

Endogeneity can arise in other ways as well:.

Measurement errors

Sample selection

Correlated heterogeneity

example: Production function error is an unobserved factor of production such as entrepreneurship, may correlate with other factors of production.

example: Wage equation error is an individual's ability or drive, correlates with other factors that determine wages, like education.

6. Unobservability.

Many concepts we would like to estimate are unobservable.
Counterfactuals: what an untreated individual's outcome would have been had they been treated.
Other Examples: random utility parameters, dynamic model state variables, production efficiency frontiers.

Other concepts in theory observable, but difficult to measure.
Examples: individual's bargaining power within a household.
individuals information set in a game or a dynamic optimization

Must combine assumptions and observable data to identify functions of unobservable (or unobserved) variables.

Examples:

Point identify compensating and equivalent variation to bound (set identify) unobserved true consumer surplus.

Assume unconfoundedness to overcome unobservability of counterfactuals in identification of ATE (Average Treatment Effects).

3.6 Control Variables

"I controlled for that."

Common response to a potential identification question (particularly in simple regressions and in Difference-in-Difference analyses). What does it mean?

Let θ be a parameter measuring the effect of one variable X on another variable Y.

The effect identified by the model may not equal the desired θ because of other so-called "confounding" connections between X and Y.

Adding a control is including another variable Z in the model to fix the problem.

The idea: Z explains the confounding relationship between X and Y.

By putting Z in the model we statistically hold Z fixed, and thereby "control" for the alternative, unintended connection between X and Y.

Fixing Z is assumed to maintain the ceteris paribus condition.

Example: X physical exercise, Y is weight gain. A control Z would be participant's age.

Example: "Parallel trends" assumption in difference-in-difference models. Assumes time and group dummies Z control for all confounding relationships between X and Y other than the desired causal treatment effect.

Including a covariate Z does NOT always fix the identification problem. Can make the problem worse! Two reasons why:

1. Functional form

Unless we have a structural model of how Z affects Y, should include Z in the model in a highly flexible (ideally nonparametric) way. Otherwise, model might be misspecified.

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 causal like LATE or difference-in-difference. Exception is "saturated" model.

Second, bigger reason why Including a covariate Z does NOT always fix the identification problem. Can make the problem worse:

2. Endogeneity

If Z is endogenous, fixing the omitted variable problem introduces endogeneity.

In the causal diagram literature have "confounders" and "colliders." See, e.g., Pearl (2000, 2009).

When Z is confounder, including it controls the problem.

When Z is a collider, including it can ruin identification of the causal effect of X on Y.

Example: A wage equation

Y is wage, X is a gender dummy. Looking to identify, e.g., wage discrimination.

Confounding problem: women may choose different occupations from men, and occupation affects wages.

Should we include occupation Z in the model as a control?

Z is endogenous: wages that are offered to men and to women affect their occupation choice.

Z is a collider. Unless we have a proper instrument for Z, the coefficients on both X and Z will still be biased (inconsistent) in general.

Same problem can affect Difference-in-Difference models:

The dummies and other covariates in these models are supposed to be 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 well known Berkson (1946) and Simpson (1951) paradoxes.

Bottom line: either implicit or explicit considerations of underlying structure is needed to convincingly argue that covariates intended to serve as controls will actually function as they are intended.

3.7 Identification by Functional Form

Definition: Identification based on assuming some functions in the model have specific parametric or semiparametric forms.

Example: Heckman (1978) selection model: $Y = (b'X + U)\, D$ and $D = I\,(a'X + \varepsilon \geq 0)$
Observe Y, D, and a vector X. Unobserved errors U and ε are independent of X.

b is identified if have excusions (some elements of vector a known to equal zero). Alternatively, can be identified by known functional form of U and ε jointly normal.

Actually, even without exclusions, error normality is much stronger than needed. Just nonlinearity in D is almost sufficient. See Dong (2012) and Escanciano, Jacho-Chávez, and Lewbel (2016).

Nonlinearity can help identification by functional form. Example:
supply curve: $Y = dX^2 + aX + \varepsilon$
demand curve $Y = bX + cZ + U$
errors $E[(\varepsilon, U) \mid Z] = 0$.

We still have no exogenous regressors in the supply curve to use as instruments for the demand curve.

We only have the single exogenous Z in the demand curve and two coefficients (that of X and X^2) to identify in the supply curve.

Despite the apparent shortage of instruments, both equations are identified! How?

supply: $Y = dX^2 + aX + \varepsilon$, demand: $Y = bX + cZ + U$

Why are both identified now? Nonlinearity.

To see identification here equate supply and demand to get
$dX^2 + (a - b) X + cZ + \varepsilon - U = 0$,
solving for X yields the reduced form equation:

$$X = \left(b - a \pm \left((b - a)^2 - 4 (cZ + \varepsilon - U) d \right)^{1/2} \right) / 2d.$$

X is linear in $(Z + \gamma)^{1/2}$ for some γ
X^2 is linear in Z and $(Z + \gamma)^{1/2}$.
Assume $Z^{1/2}$ is correlated with $(Z + \gamma)^{1/2}$
Then $Z^{1/2}$ and Z are usable instruments for both equations.

One can also see the identification graphically. Nonlinear supply means as Z shifts the demand curve, one sees intersections of supply at different points along the demand curve, tracing out the slope of the demand curve.

Formal proof of identification depends on showing that the equations
$E\left(Y - dX^2 - aX \mid Z = z\right) = 0$
$E\left(Y - bX - cZ \mid Z = z\right) = 0.$
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.
Example: identification would fail if Z only took the values zero and one.

Idea of Identifying linear coefficients by exploiting nonlinearity elsewhere greatly generalizes.

Example: model $Y = h\left(Z'b, g\left(Z\right)\right) + \varepsilon$ and $X = g\left(Z\right) + U$
Functions h and g are unknown
joint distribution of ε and U are unknown, independent of Z.

Models like these can arise with endogenous regressors or with sample selection.

Escanciano, Jacho-Chávez, and Lewbel (2016) show that the coefficients b and the functions h and g can generally be identified in this model.

Key requirement is linear $Z'b$ and nonlinear g.

Identification by functional form, another example:

$$Y = a + bX + cZ + U$$

Assume X is endogenous (correlated with U) and have no exclusion assumption (no outside instruments). Lewbel (2012, 2018) exploits heteroskedasticity instead of nonlinearity in the X equation to identify the model.

Regress $X = \gamma + \delta Z + e$, then let $R = \left(Z - \overline{Z}\right)\widehat{e}$. Under some heteroskedasticity assumptions, R is a valid instrument for X.

Example assumptions that work: X is mismeasured, so U contains both model and measurement error. True model error in the Y equation is homoskedastic, e is heteroskedastic. This also works for some kinds of factor models.

Measurement error Identification by functional form:

Let $Y = a + bX + U$ where X has classical measurement error. Assume true X independent of model and measurement errors.

Reiersøl (1950) showed a and b are identified as long as either the true X or the true errors are NOT normal. Normality is the worst possible functional form for identification with measurement error!

Lewbel (1997b) shows, if measurement error is symmetric and true X is asymmetric, then $\left(X - \overline{X}\right)^2$ is a valid instrument for X. Empirical example is regression of patent counts on R&D expenditures.

Schennach and Hu (2013) show Reiersøl extends to identify $Y = g(X) + U$ with mismeasured X for any function g except for a few specific functional forms of g and of the error distributions. In these models independence of the true error has strong identifying power.

Identification by functional form or by constructed instruments depends on relatively strong modeling assumptions.

When available, better to use 'true' outside instruments (based on theory based exclusions).

Causal inference proponents go further, accepting only randomization as a valid source of exogenous variation.

Good use of constructed instruments or identification by functional form? For testing and robustness.

In practice, one often can't be sure if assumptions needed for outside instrument validity (exclusions) are satisfied.

Even with randomization, assumptions for validity, like no measurement errors correlated with treatment, can be violated (see section 5).

For testing: can combine functional form based moments (e.g. constructed instruments) with moments based on "true" instruments to get overidentification.

Use the results for instrument validity testing, robustness checks, and efficiency.

Example: $Y = a + bX + cZ + U$

Z is exogenous, X is endogenous, outside variable W a proposed instrument for X, and $R = \left(Z - \overline{Z}\right)\widehat{e}$ is Lewbel's (2012) heteroskedasticity based constructed instrument.

2SLS using both R and W as instruments for X is overidentified - test jointly for validity of the instruments by Sargan (1958) and Hansen (1982) J-test.

If rejected, then model misspecified or an instrument is invalid. Else have increased confidence in both W and R. Both might then be used in estimation to maximize efficiency.

Or just check robustness of estimated effects to various possible true and constructed instruments.

More confidence that estimated effects are reliable if different sources of identification (randomization, exclusions, constructed moments) all agree.

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

Models are often contain sets of equalities involving θ, e.g., moment conditions like $E\left[g\left(W,\theta\right)\right]=0$,
where W is data and g are known functions.

Many estimators are based on moment equalities: OLS, 2SLS, GMM, and first order conditions from extremum estimators like MLE score functions.

Suppose identification of a vector θ is based on equalities.
θ is *exactly* identified if removing any one equality loses identification,
θ is *overidentified* if θ can still be identified after removing one or more equalities.
θ is *underidentified* if don't have enough equalities to identify θ.

For θ a $J-$vector, usually need J equations to exactly identify θ. Number of equations equal number of unknowns is the *order condition* for identification.

In linear models, identification also requires a *rank condition* on the matrix that θ multiplies.

General rank conditions for nonlinear models exist, based on the rank of relevant Jacobian matrices. See Fisher (1959, 1966), Rothenberg (1971), Sargan (1983), Bekker and Wansbeek (2001), and section 8.1 later on local identification.

Often satisfying the order condition may suffice for generic identification (see section 8.2 later).

When parameters are overidentified, can usually test validity of the moments used for identification.

Intuitively, 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 exploit all the moments simultaneously. E.g. Sargan (1958) and Hansen (1982) J-test.

Arellano, Hansen, and Sentana (2012) discuss testing for underidentification.

Terminology defined in this subsection is generally from the Cowles foundation era, e.g., the term 'order condition' dates back at least to Koopmans (1949).

Above terminology is for θ a vector, not a function. Chen and Santos (2015) extend to define a notion of semiparametric local overidentification.

End of Part 1

Part 2 will have sections:

4. Coherence, Completeness and Reduced Forms
5. Causal Reduced Form vs Structural Model Identification

Part 3 will have sections:

6. Identification of Functions and Sets
7. Limited Forms of Identification
8. Identification Concepts that Affect Inference
9. Conclusions