The Identification Zoo - Meanings of Identification in Econometrics: PART 3

Arthur Lewbel

Boston College

2019

The Identification Zoo - Part 3 - sections 6, 7, 8, and 9.

(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:

1. Introduction

2. Historical Roots of Identification

3. Point Identification

3.1 Introduction to Point Identification

3.2 Defining Point Identification

3.3 Examples and Classes of Point Identification

3.4 Proving Point Identification

3.5 Common Reasons for Failure of Point Identification

3.6 Control Variables

3.7 Identification by Functional Form

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

4. Coherence, Completeness and Reduced Forms

Continued next slide

Table of Contents - continued:

5. Causal Reduced Form vs Structural Model Identification

- 5.1 Causal or Structural Modeling? Do Both
- 5.2 Causal vs Structural Identification: An Example
- 5.3 Causal vs Structural Simultaneous Systems
- 5.4 Causal vs Structural Conclusions

6. Identification of Functions and Sets

- 6.1 Nonparametric and Semiparametric Identification
- 6.2 Set Identification
- 6.3 Normalizations in Identification
- 6.4 Examples: Some Special Regressor Models

Continued next slide

Table of Contents - continued:

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

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

9. Conclusions

Part 1 had sections:

1. Introduction

2. Historical Roots of Identification

3. Point Identification

Part 2 had sections:

4. Coherence, Completeness and Reduced Forms

5. Causal Reduced Form vs Structural Model Identification

6. Identification of Functions and Sets

This section discusses:
1. Nonparametric and semiparametric identification
2. Set identification
3. Normalizations in identification (frequently used, rarely discussed)
4. Examples of these (special regressor).

Sections 1. and 2. will be brief, since many good surveys of them already exist, e.g.

Powell (1994) focuses on semiparametrics.
Chesher (2007) nonadditive models with nonseparable errors
Matzkin (2007, 2012) economic and functional restrictions to identify vectors and functions.
Tamer (2010) for set identification.

6.1 Nonparametric and Semiparametric Identification

Parametric identification: θ is a finite set of constants, and all possible values of ϕ also correspond to different values of a finite set of constants.

Nonparametric identification: θ is infinite dimensional vectors or functions.

Example: Assume IID W_i, so knowable ϕ is the distribution function $F(W)$. Let θ be the density function $f(W) = \partial F(W)/\partial W$. Is nonparametrically identified by construction.

Example: Let ϕ be the joint distribution $F(Y, X)$ (e.g., if DGP is IID Y_i, X_i). Model $Y = m(X) + e$ with $E(e \mid X) = 0$. Then $m(X)$ is nonparametrically identified by construction - it can be recovered from $F(Y, X)$.

Have $m(X) = E(Y \mid X) = \int_{supp(Y|X)} Yf(Y \mid X)\, dY$ and the conditional density $f(Y \mid X)$ is identified from the joint distribution function $F(Y, X)$.

Semiparametric Identification: θ includes both vectors and functions, or θ vectors and ϕ functions.

Semiparametric models are (loosely) models containing vectors of parameters of interest, but also contain unknown functions, which may be nuisance functions.

Example (Härdle and Stoker 1989) Average derivative vector $\theta = E\left[\partial m\left(X\right)/\partial X\right]$ where $m\left(X\right) = E\left(Y \mid X\right)$, with iid Y, X.

Example : Given iid Y, X, Z, a partially linear model (see, e.g., Robinson 1988) is $Y = m\left(X\right) + Z'\beta + e$ with $E\left(e \mid X, Z\right) = 0$ and $var\left(X \mid Z\right)$ nonsingular. Following Robinson (1988), identify β from linear regression $\left(Y - E\left(Y \mid X\right)\right) = \left(Z - E\left(Z \mid X\right)\right)'\beta$. Then given β, m is identified by $m\left(X\right) = E\left(Y - Z'\beta \mid X\right)$.

Other early examples of semiparametric identification are Manski (1975, 1985) and Cosslett (1983).

Consider: ordinary linear regression $Y = X'\beta + e$ with $E(e \mid X) = 0$.

If θ is β and ϕ is first and second moments of Y, X like in Wright-Cowles, then is parametric identification.

If θ is β and distribution of e, and ϕ is distribution of Y, X, then is semiparametric identification.

As this shows, distinctions between parametric and semiparametric identification can be somewhat arbitrary. See Powell (1994) for further discussion of this point.

These types of distinctions can be traced back at least to Hurwicz (1950).

In theory, same steps for semi or nonparametric identification as for parametric:

Establish that different values of θ are not observationally equivalent.

In practice, can be much harder. e.g., parametric order condition involves counting number of equations and vs number of unknowns. Doesn't work when the unknowns themselves are also functions.

Consider: nonparametric instrumental variables:

Model is $Y = m(X) + e$, $E(e \mid Z) = 0$ for instruments Z, some regularity.
θ is $m(X)$ and knowable ϕ is distribution $F(Y, X \mid Z)$.

$E(e \mid Z) = 0$ for all $z \in supp(Z)$ implies
$\int_{supp(Y,X|Z=z)} (Y - m(X))\, dF(Y, X \mid z) = 0.$
θ is identified if this integral equation can be uniquely solved for $m(X)$.

Newey and Powell (2003) show that identification here is equivalent to an example of statistical completeness

Have identification if integral equation $\int_{supp(Y,X|Z=z)} (Y - m(X))\, dF(Y, X \mid z) = 0$ can be uniquely solved for $m(X)$.

This identification is an example of statistical completeness (see Newey and Powell 2003).

Here identification corresponds to uniqueness of the solution of an integral equation.

If Y, X, and Z were all discrete would reduce to parametric identification. Then this integral equation would reduce to a matrix equation. Identification would only require nonsingularity of a moment matrix, as in linear instrumental variables regression.

Method of identifying vector θ by inverting matrices often can be extended to identify function θ (solving integral equations) by "operator 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.

Concepts like completeness and injectivity are crucial to identification, as infinite dimensional analogues to the invertibility of matrices.

Examples:
Schennach (2007), identifying nonparametric regressions with mismeasured regressor,
random coefficients as in Beran, Feuerverger, and Hall (1996),
many of the structural models in Matzkin (2005, 2007, 2012),
Additional examples of semiparametric identification are given in Section 6.4 below.

Often model restrictions are moments, which take the form of integrals. Semiparametric identification then corresponds to integral equations having a unique solution.

Examples are random coefficients models as in Beran, Feuerverger, and Hall (1996), measurement error problems as in Schennach (2007), and many structural models in Matzkin (2005, 2007, 2012).

Additional examples of semiparametric identification are given in section 6.4 below.

6.2 Set Identification.

Partial Identification generally refers to where ϕ provides info about θ, but not enough info to identify θ.

Set Identification: The true θ_0 is *set identified* if there exist some values of $\theta \in \Theta$ that are not observationally equivalent to θ_0.

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

Point identification is when the identified set only contains one element.

We don't know which value of θ is θ_0. To prove set identification, need to show that for any $\widetilde{\theta} \in \Theta$, there exist some values of $\theta \in \Theta$ that are not observationally equivalent to $\widetilde{\theta}$.

See also Chesher and Rosen (2017).

Example of Set Identification:

Let W be the fraction of a consumer's total budget that is spent on food. Model M is observe W. Budget shares are nonnegative and sum to 1. What is knowable, ϕ, is the distribution of W in the population.

Then can identify $\beta = E\left(W\right)$, the expected budget share for food.

We want θ, the expected budget share for clothing.

θ is not point identified from this model. The identified set for θ is the interval $\left[0, 1-\beta\right]$.

This set also provides *bounds* on θ (inequalities that θ satisfies). Set identification does not always yield bounds.

Sharp bounds are tightest bounds possible on θ given M and ϕ.

Early examples of set identification:

Frisch (1934) on measurement errors, 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.

Bounds on behavior from economic theory: revealed preference inequalities of Afriat (1967) and Varian (1982, 1983).

Systematic study of Set identification: Manski (1990, 1995, 2003).

Interest in set identification grew when methods for inference on estimators of set identified parameters began to be developed, as in Manski and Tamer (2002).

Much modern literature on set identification focuses on:

1. Derivation of sharp bounds.
2. Verifying that one has obtained the smallest possible identified set.
3. Finding cases where the identified set is very small relative to Θ.
4. Improving methods for estimation and inference.

An important tool for studying set identification is the theory of random sets, see Beresteanu, Molchanov, and Molinari (2011), Bontemps, Magnac, and Maurin (2012), and Chesher and Rosen (2014).

Typical reasons for set rather than point identification:

Model incompleteness (often taking the form of inequality constraints).

Limited observation of data, like regressors censored, discretized, mismeasured, or observations missing not at random. See, e.g., Manski and Tamer (2002) and references therein.

Economic theory that only provides inequalities rather than equalities on behavior, e.g., Pakes, Porter, Ho, and Ishii (2015).

Presence of unobserved variables (e.g. counterfactuals, random utility parameters, unobserved state variables). Schennach (2014) provides a general technique for deriving observable moments that characterize the identified set in models defined by moments over both observables and unobservables.

Same as reasons for nonidentification.

Some (e.g., Chesher and Rosen 2017) argue that economic theory rarely provides enough restrictions to point identify parameters of interest.
Claim much econometric literature devoted to poorly motivated tricks to obtain point identification.
They essentially argue that set identification should be treated as the usual situation.

A less extreme view: make needed assumptions to obtain point identification.
Then, examine what happens to the identified set when the strongest or least defensible point identifying assumptions are dropped.

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.

N*on-robust identification (*Khan and Tamer 2010): Defined as when a point identified θ loses even set identification when an identifying assumption is relaxed.

Example: $\theta = E\left(Y^*\right)$, random scalar Y^* can be any real number.
DGP is iid observations of $Y = I\left(-b \leq Y^* \leq b\right) Y^*$ for some constant b.
Observed Y is a censored version of the true Y^*. ϕ is the distribution of Y.
The no censoring assumption that $b = \infty$, is non-robust because
a. with the assumption θ is point identified.
b. without the assumption θ could be any real number.

Intuition: 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.

Set identification is an area of active research in econometric theory
But it is only applied rarely in empirical work. Why?

Estimation and inference on sets is complicated.
generally requires multiple tuning parameters and penalty functions.
See, e.g., Chernozhukov, Hong, and Tamer (2007).

Freedman (2005), Angrist and Pischke (2008), etc., argue that modern econometrics is too complicated; too difficult to discern or assess the plausibility of underlying identifying assumptions and to implement estimators.

But irony? Removing complicated identifying assumptions leads to set identification, which requires even more complicated econometrics for identification, estimation and inference.

6.3 Normalizations in Identification

Often identification, particularly non- and semi-parametric identification, requires *normalizations*.

Example: Linear index model $E\left(Y \mid X\right) = g\left(X'\beta\right)$, g is strictly monotonically increasing, $E\left(XX'\right)$ is nonsingular, ϕ is the joint distribution of Y and X.

Linear regression is a linear index model where $g\left(\cdot\right) = \cdot$.
Probit is a linear index model where g is the cumulative standard normal distribution function.
Logit, and many censored and truncated regression models are also linear index models.

Case 1: g is known and $\theta = \beta$. Then θ is identified.
Proof is by construction: $\beta = \left[E\left(XX'\right)\right]^{-1} E\left[Xg^{-1}\left(E\left(Y \mid X\right)\right)\right]$.

Case 2: g is unknown and $\theta = \{g, \beta\}$.

For any positive constant c we can define $\widetilde{\theta} = \{\widetilde{g}, \widetilde{\beta}\}$ by $\widetilde{\beta} = \beta/c$ and $\widetilde{g}(z) = g(cz)$.

Then for any X we have $\widetilde{g}\left(X'\widetilde{\beta}\right) = g(X'\beta)$ so θ is observationally equivalent to $\widetilde{\theta}$.

This shows θ is not identified. At best it is set identified, where the set includes $\widetilde{\theta}$ for any $c > 0$.

Suppose all the elements $\widetilde{\theta}$ in the identified set have $\widetilde{\beta}$ that is proportional to β.

So all have the form of $\widetilde{\beta} = \beta / c$. Then say β is *identified up to scale*.

To actually identify β, we need a scale restriction like assuming the first element of β equals one, or that the length of the vector β equals one.

Restricting scale of β might be without loss of generality (wlog) here because the scaling of β is absorbed into the definition of g.

Loosely, restrictions are called normalizations when they are wlog, i.e., if economically meaningful parameters or summary measures are unaffected by the restriction.

Sometimes the term "free" normalization is used to emphasize a wlog restriction.

When is a restriction a free normalization vs a behavioral restriction? It depends on how we use and interpret the model.

Example: instead of $\theta = \{\beta, g\}$, let $\theta = \{g(X'\beta), \beta g'(X'\beta)\}$ where g' denotes the derivative of the function g.

This θ is $E(Y \mid X)$ and its derivatives (marginal effects) $\frac{\partial E(Y|X)}{\partial X}$.

As before $\{\widetilde{g}, \widetilde{\beta}\}$ differs from, but is observationally equivalent to, $\{g, \beta\}$. But now $\theta = \widetilde{\theta}$.

β is only identified up to scale, but the economically meaningful parameters θ are point identified.

Here the scale normalization is a free normalization, wlog.

A scale restriction is *not* a free normalization *if* the level of $X'\beta$ is assigned some economic meaning, like dollars.

Example: willingness to pay (WTP). $Y = 1$ if willing to pay more than V dollars for a product or service.
Let $X'\beta + e$ be WTP, where $e \perp X$.

Then $Y = I\left(X'\beta + e \geq V\right)$.
so $E\left(Y \mid X\right) = g\left(X'\beta - V\right)$ where g is distribution of $-e$.

Here $X'\beta - V$ is the index being estimated.
Scaling is not free. $X'\beta + e$ is WTP only if the coefficient of V equals minus one.

For general semiparametric identification and estimation of willingness to pay models see Lewbel, Linton, and McFadden (2012).

Scale restrictions or normalizations are common in semiparametric models (An early example is Powell, Stock, and Stoker 1989).

Also common are *location* restrictions or normalizations.

Let $E(Y \mid X) = g(X + \alpha)$, g is an unknown function, α is an unknown scalar.

Generally α is not identified, because observationally equivalent to $\alpha = 0$ using $\widetilde{g}$ such that $\widetilde{g}(X) = g(X + \alpha)$.

As with scaling, location normalizations may or may not be free, depending on context.

Example: threshold crossing binary choice model
$Y = I(\alpha + X'\beta + e \geq 0)$ where $e \perp X$.
Is a special case of a linear index model, has $E(Y \mid X) = g(X'\beta)$ where g is the distribution function of $-(\alpha + e)$.

Identification here requires both location and scale normalization.
Parametric probit assumes $E(e) = 0$ and $var(e) = 1$.
This uniquely determines the location and scale of $\alpha + X'\beta + e$.

We could instead (observationally equivalently) impose $\alpha = 0$, $\beta'\beta = 1$, and let e have arbitrary mean and variance.

In semiparametric models, usually impose normalization on α and β, not $E(e)$ and $var(e)$.

Why? often simplifies identification and estimation. Latter sometimes converge slower.

Nonseparable error models.

$Y = g(X, e)$, the function g and the distribution of an unobserved continuously distributed scalar error e are unknown.

Observationally equivalent to $Y = G(X, h(e))$ for any unknown strictly monotonic, invertible function h, where $G(X, v) = g\left(X, h^{-1}(v)\right)$.

Generally, identification of g requires assuming that the entire distribution function of e is known.
Typical is assuming e is uniform on $[0, 1]$ or is a standard normal.

Is this a free normalization? Yes, IF only care about conditional distribution of Y given X.

Otherwise, if g as a structural model, this normalization is a strong behavioral restriction.

See Lewbel (2007c) for details. Matzkin (2007, 2012) provides multiple examples.

Another normalization:

Let utility from choice $Y = y$ be $\alpha_y + X'\beta_y + e_y$ for $y = 0, 1$.

Utility maximization means choose $Y = I\left(\alpha + X'\beta + e \geq 0\right)$, where $\alpha = \alpha_1 - \alpha_0$, $\beta = \beta_1 - \beta_0$, and $e = e_1 - e_0$.

Interpret $\alpha + X'\beta + e$ as utility from $Y = 1$ if assume the normalization that utility of the "outside option" $\alpha_0 + X'\beta_0 + e_0 = 0$.

In static discrete choice model, is generally a free normalization.

But in dynamic discrete choice models, identification typically assumes the outside option has the same utility in every time period. This is not free, imposes real restrictions on preferences and hence on behavior. See, e.g., Rust (1994) and Magnac and Thesmar (2002).

Another normalization: choice of cardinalization of ordinally identified utility levels.

Choose a quantity vector x to maximize utility $U(x)$ under some constraints (e.g., a budget constraint).

Revealed preference, Samuelson (1938, 1948), Houthakker (1950), Mas-Colell (1978): Given demand functions, indifference curves (i.e., level sets) associated with $U(x)$ are identified.

Actual utility or happiness level $U(x)$ associated with each indifference curve is not identified. Utility is identified up to an arbitrary monotonic transformation, which we may call a normalization.

Similarly, $Y = I(\alpha + X'\beta + e \geq 0)$ is observationally equivalent to $Y = I(g(\alpha + X'\beta + e) \geq g(0))$ for strictly monotonically increasing g. Without more info, can't tell if one's actual utility level is $\alpha + X'\beta + e$ or $g(\alpha + X'\beta + e)$ for any such g.
As before, whether choice of g corresponds to a free normalization or to a behavioral restriction depends on context.

Final normalization notes:

When comparing parametric and semiparametric estimation:
Either recast θ in the same normalization, or compare summary measures like marginal effects that are independent of normalization.

Also, choice of normalizations can affect precision of estimates, even if irrelevant for identifcation.

Common additional restrictions on functions in θ include continuity, differentiability, and/or monotonicity. These are behavioral restrictions, not normalizations.

6.4 Examples: Some Special Regressor Models

Return to example 5 from section 3.3, now with covariates:
Model is $Y = I(X + U > 0)$ for an unobserved U, where $U \perp X \mid Z$, and $X \mid Z$ is continuous.
ϕ is the joint distribution of Y, X, Z (from, e.g., IID DGP).
$\theta = F_{U|Z}(u \mid z)$, the conditional distribution function of U given $Z = z$.

By construction:
$E(Y \mid X = x, Z = z) = \Pr(X + U > 0 \mid X = x, Z = z)$
$= \Pr(x + U > 0 \mid Z = z) = 1 - \Pr(U \leq -x \mid Z = z) =$
$1 - F_{U|Z}(-x \mid z)$

So $F_{U|Z}(u \mid z) = 1 - E(Y \mid X = -u, Z = z)$ is identified for values of u that $-X$ can equal.

Intuition of this special regressor (see Lewbel 1997, 2000, 2014):
Distribution of latent error U can be identified if model contains $U + X$, since variation in X moves Y the same way that variation in U does.

Example: Set Identification of the Latent Mean.

For simplicity let Z be empty, want to identify $E(U)$.

Suppose for now supp(U) is the whole real line.

Given $Y = I(X + U > 0)$, can identify $F_U(-x)$.

If supp(X) is whole real line, then $F_U(u)$ is identified everywhere, and $E(U) = \int_{-\infty}^{\infty} u dF_U(u)$.

But if supp(X) bounded to $a \leq X \leq b$, then $F_U(u)$ only identified for $-b \leq u \leq -a$.

In this case, $E(U)$ is NOT even set identified. It's non-robust identification.

No bounds on $E(U)$, because F_U could have mass arbitrarily far below $-b$ or above $-a$.

Other features of F_U are point or set identified, like quantiles.

Set or point identification restored with assumptions about tails of U, e.g. bounded support, or Magnac and Maurin's (2007) tail symmetry.

If either $a = -\infty$ or $b = \infty$, then we get bounds on $E(U)$: if $a = -\infty$, $E(U) \leq b$.

Example: General Binary Choice With Special regressor.

Now let $U = g(Z) + e$ 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 probit.
Instead, $g(Z)$ is an unknown function, latent e has unknown distribution.
Note coefficient of X equals one is a scale normalization.
Assume "large support:" $-X$ can equal any value U can be.

Then $g(Z)$ is identified, because then $F_{U|Z}(u \mid z)$ is identified for all u and $g(z) = \int_{supp(u)} u dF_{U|Z}(u \mid z)$.
Needed large support because a mean $g(Z) = E(U \mid Z)$ depends on the entire distribution.

If instead had defined $g(Z) = med(U \mid Z)$ would only have needed support of $-X$ to include a neighborhood of $med(U \mid Z)$ to identify $g(Z)$.

Examples of other parameters we could identify in this model:

An elasticity function like $\partial E\left(\ln g(Z)\right)/\partial \ln Z$. Is identified given g.

The distribution of e conditional on Z. Is identified by $F_{e|Z}\left(e \mid Z\right) = F_{U|Z}\left(g(z) + e \mid z\right)$.

The average structural function (ASF) as in Blundell and Powell (2004), defined as what the function $E\left(Y \mid X\right)$ would have been if the conditional distribution of the error, $F_{e|Z}$, were replaced with its marginal distribution F_e.

Here, given $F_{e|Z}$ we can calculate the unconditional F_e,
and then ASF $= \int_{supp(e)} I\left(X + g(Z) + e > 0\right) dF_e\left(e\right)$.
see Chen, Khan, and Tang (2016) and Lee and Li (2018).

Example: Binary Choice With Random Coefficients.

preliminary result: linear model random coefficients:
Suppose observe iid continuous U_i, Z_i, so $F_{U|Z}$ is identified.
Suppose $U = Z'e$ where e is a vector of random coefficients.
Distribution F_e unknown. Is it nonparametrically identified?
Yes, with some regularity assumptions. See Beran and Millar (1994) and Beran, Feuerverger, and Hall (1996).
Intuition: $E(U \mid Z) = Z'E(e)$ identifies $E(e)$,
$E\left(U^2 \mid Z\right) = Z'E(ee')Z$ identifies $E(ee')$, etc.
(real proof works through the characteristic function).

Example continued: Binary Choice With Random Coefficients.

Now, instead of linear, consider $Y = I\left(Xe_x + Z'e_z > 0\right)$,
where e_x and the vector e_z are random coefficients.
Assume that $e_x > 0$ and let $e = e_z / e_x$ (this is a scale normalization).
Model is same as $Y = I\left(X + Z'e > 0\right)$
e is now the vector of random coefficients.

Here U defined by $U = Z'e$ is NOT observed,
But now $F_{U|Z}$ is identified by special regressor X,
So as in the linear model, F_e is identified.

Ichimura and Thompson (1998) and Gautier and Kitamura (2013) directly proved identification of random coefficients in $Y = I\left(Xe_x + Z'e_z > 0\right)$. Used a different scale normalization from above.
Special regressor random coefficients extended to multinomial choice setting by Fox and Gandhi (2016).

Special regressor identification in other discrete choice models:
Games with discrete strategies: Lewbel and Tang (2014)
Semiparametric generalizations of BLP (Berry, Levinsohn, and Pakes 1995): Fox and Gandhi (2012), Berry and Haile (2014).

Identification theorems for binary choice models that predate special regressors, but that can be reinterpreted as special: Cosslett (1983), Manski (1985), Horowitz (1992), and Lewbel (1997a).

In Berry and Haile (2014) the special regressors are hidden, they're called $x_{jt}^{(1)}$.
In Matzkin (2015), equations (2.2) and (2.4) make what she calls 'exclusive regressors' X_g special.
Gautier and Kitamura (2013) above also did not make the connection to special regressor identification.

Special regressors surveys

Lewbel (2014) - survey provides derivations and intuition.
Lewbel, Dong, and Yang (2012) - provides comparisons between special regressors, control functions, maximum likelihood, and linear probability models.
Dong and Lewbel (2015) - provides simple ways to implement special regressor estimation.

7. Limited Forms of Identification

It's often hard to prove point identification.

Possible responses:
1. Set identification (estimation and inference is hard).
2. Ignore the problem and assume identification (trouble if wrong).

A third way: Prove some more limited form of identification holds.

Then the leap of faith in assuming point identification is not as large.

Limited forms of identification include local identification and generic identification. These are each necessary conditions for point identification.

7.1 Local Identification

Point identification of θ_0 requires that no other $\theta \in \Theta$ be observationally equivalent to θ_0.

Since we don't know what θ_0 is, proving point identification requires showing that no two different values θ and $\widetilde{\theta}$ in Θ be observationally equivalent. *Global identification.*

In contrast, *Local identification* of θ_0 means that there exists a neighborhood of θ_0 such that, for all values θ in this neighborhood (other than the value θ_0) θ_0 is not observationally equivalent to θ.

Proving local identification requires showing that this holds replacing θ_0 with any $\widetilde{\theta} \in \Theta$.

Note: Local identification differs from (and predates), the term local as used in LATE (local average treatment effect). In LATE, local means the mean parameter value for a particular subpopulation.

Examples: Knowable, ϕ, is a continuous function $m(x)$ for real, scalar x. Suppose $m(\theta) = 0$. Consider three possible cases:

Case a: m is strictly monotonic. Then θ is globally identified. Strict monotonicity ensures only one value of θ can satisfy $m(\theta) = 0$.

Case b: m is a J'th order polynomial for some integer J. Then θ typically not globally identified. Up to J different values of θ that satisfy $m(\theta) = 0$. But θ is locally identified. A neighborhood of the true θ can be small enough to exclude all other roots of $m(x) = 0$.

Case c: m is just continuous. Then θ might not even be locally identified, because $m(x)$ could equal zero for all values of x in some interval.

Suppose 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, but we can't rule out the possibility of a finite number of local optima. Then one might still show local 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 true parameter values to know that the correct θ should lie in a particular region.

Example: In Lewbel (2012), a parameter is a coefficient in a simultaneous system of equations, and the set has two elements, one positive and one negative.

In this case have only local identification, unless economic model is sufficient to determine the sign of the parameter a priori (e.g., it is the price coefficient in a supply equation).

The notion of local identification is described by Fisher (1966) for linear models.

Generalized to other nonlinear parametric models by Rothenberg (1971) and Sargan (1983), as follows:

Let θ be a J - vector of structural parameters, and ϕ be set of reduced form parameters

Assume the model implies $r\left(\phi\left(\theta\right),\theta\right)=0$ for some known vector valued function r. Let $R\left(\theta\right)=r\left(\phi\left(\theta\right),\theta\right)$.

A sufficient condition for local identification of θ is that $R\left(\theta\right)$ be differentiable and rank$\left(\partial R\left(\theta\right)/\partial\theta\right)=J$. Sargan (1983) calls violation of this condition *first order (lack of) identification*.

For parametric models that can (if identified) be estimated by maximum likelihood, this rank condition is equivalent to information matrix at θ_0 being 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) extend to local identification in models defined by conditional moment restrictions.

Chen and Santos (2015) provide a concept of local overidentification for a class of semiparametric models.

7.2 Generic Identification

Like local identification, generic identification is a necessary condition for point identification that is sometimes easier to prove.

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 $\widetilde{\Theta}$.

If θ_0 takes on a value in $\widetilde{\Theta}$ then θ_0 is not point identified.

Global identification 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.

Interpreting generic identification:

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.

Example: Let ϕ be second moments of X, Y. Model is $Y = X'\beta + e$ with $E(e \mid X) = 0$.

β is generically identified if the probability is zero that nature chooses a distribution function for X with the property that $E(XX')$ is singular.

Generic identification is closely related to the order condition.

Consider a linear regression system where the order condition holds.
If a certain coefficient matrix B is nonsingular, then the rank condition holds, giving identification.

In the set of all possible values for B, if the subset that is singular has measure zero, then the order condition suffices for generic identification of the model.

Another example: suppose iid of Y, X are observed with
$X = X^* + U$,
$Y = X^* \beta + e$,
unobserved model error e, unobserved measurement error U, and unobserved true covariate X^* are mutually independent with mean zero.

Despite not having instruments, β is identified when Y, X have any joint distribution except a normal (Reiersøl 1950).

So β is generically identified if the set of possible Y, X distributions is sufficiently large, e.g., if e could have been drawn from any continuous distribution.

Given the same independence of U, e, and X^*, Schennach and Hu (2013) show that 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 could say 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.

In many social interactions models, showing point identification is intractible, but one can establish generic identification. See, e.g., the survey by Blume, Brock, Durlauf, and Ioannides (2011).

8. Identification Concepts that Affect Inference

Identification preceeds estimation. We first show identification, then consider properties of estimators.

However, sometimes the nature of the identification affects inference.

The way θ is identified may impact properties of any possible estimator $\widehat{\theta}$.

These identification concepts that affect inference include weak identification, Identification at infinity, and ill-posedness.

Note many previously discussed identification concepts also have implications for estimation, including:
Extremum based identification,
set vs point identification.

8.1 Weak vs. Strong Identification

Roughly, weak identification is when θ is, in a particular way, close to being not point identified.

Both weakly identified and strongly identified parameters are point identified (θ is uniquely determined given ϕ).

These should not, strictly speaking, be called identification.

These instead refer to specific difficulties relating to asymptotic inference.

They're called forms of identification because they arise from features of the underlying model and associated DGP, and hence effect any possible estimator we might propose.

Historical notes on weak identification:

First recognition of weak identification may be this: "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." - Sargan (1983)

Handbook of Econometrics chapters by Phillips (1983) and Rothenberg (1984) hint at the issue.

Bound, Jaeger, and Baker (1995) specifically raise 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).

Usual source of weak identification: low correlations among variables used to attain identification, like instrument Z and covariate X.

Parameters are not identified if the correlation is zero. Identification is weak (or the instruments Z are weak) when this correlation is close to zero.

First stage of 2SLS is regress X on Z to get fitted values $\widehat{X}$.

Some coefficients may be weakly identified if $E\left(\widehat{X}X'\right)$ is ill conditioned (nearly 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.

Key feature of weak identification is NOT imprecisely estimates with large standard errors (though they do have that feature).

Rather, standard asymptotics poorly approximate the true precision of parameter estimates when identification is weak (and higher order asymptotics don't help, since they too depend on precise parameter estimates).

Recall all asymptotics are merely approximations to true small sample distributions.

Compare: nonparametric regressions are imprecise with large standard errors (due to slow convergence rates by the curse of dimensionality).

But nonparametric regressions are not weakly identified, because standard asymptotic theory (e.g., a CLT with slow rates) adequately approximates their true estimation precision.

Suppose that θ is identified, and can be estimated at rate root-n using an extremum estimator (maximizing some objective function, e.g., least squares or GMM or maximum likelihood).

If some parameters are weakly identified, then any objective function used to identify and estimate them will be close to flat in some directions.

Flatness yields imprecision, but more relevantly, flatness also means that standard errors and t-stats (either analytic or bootstrapped) will be poorly estimated. They depend on the inverse of a matrix of derivatives of the objective function. Flatness makes this matrix close to singular.

Weak identification resembles multicollinearity, which in linear regression means $E(XX')$ is ill-conditioned.

Like multicollinearity, it is *not* that parameters either "are" or "are not" weakly identified. Weakness of identification depends on the sample size. For big enough n, standard asymptotic approximations must become valid.

This is why weakness of identification is often judged by rules of thumb rather than formal tests.

Staiger and Stock (1997) suggest the rule of thumb for linear two stage least squares models that should be concerned about potential weak instruments if the F-statistic on the excluded regressors in the first stage is less than 10.

Two separate things:
1. Parameters that are weakly identified at reasonable sample sizes.
2. The models econometricians use to deal with weak identification.

In real data, weak identification disappears when n gets sufficiently large. This makes it difficult to provide a general asymptotic theory to deal with the problem.

Econometrician's trick is an alternative asymptotic theory known as drifting parameter models.

Consider $Y = \theta X + U$ and $X = \beta Z + V$, data are iid observations of scalars Y, X, Z. Errors $E(UZ) = E(VZ) = 0$, for simplicity all variables are mean zero.

If $\beta \neq 0$, then θ is identified by $\theta = E(ZY)/E(ZX)$, corresponding to standard IV. Since $E(ZX) = \beta E\left(Z^2\right)$, if β is close to zero then both $E(ZY)$ and $E(ZX)$ are close to zero, making θ weakly identified. But how close is close?

The bigger the sample size, the more accurately $E(ZX)$, and $E(ZY)$ can be estimated, and hence the closer β can be to zero without causing trouble.

To capture this idea asymptotically, pretend true β is not a constant, but instead is $\beta_n = bn^{-1/2}$. The larger n gets, the smaller β_n becomes.

This gives us a model that has the weak identification problem at all sample sizes, and so can be analyzed asymptotically.

If it were true that $\beta_n = bn^{-1/2}$, then b and hence β_n would generally not be identified, so tests and confidence regions for θ have been developed that are robust to weak instruments, that is, they do not depend on estimating β_n. See, e,g., Andrews, Moreira, and Stock (2006) for an overview of such methods.

In the literature, when we say θ "is" weakly identified, we mean we are providing asymptotic theory for inference that allows for identification to be based on a parameter that drifts towards zero.

So in the above model would say θ is strongly identified when the asymptotics we use 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) $\beta_n = bn^{-1/2}$.

Usually, we don't believe that parameters are actually drifting towards zero as *n* grows. Rather, when we assume weak identification, we're expressing the belief that the drifting parameter model provides better asymptotic approximations to the true distribution than standard asymptotics.

Other related terms include *semi-strong* identification and *nonstandard-weak* identification. See, e.g., Andrews and Guggenberger (2014). These and other variants have parameters drift to zero at other rates, or have models with a mix of drifting, nondrifting, and purely unidentified parameters.

8.2 Identification at infinity or zero; Irregular and Thin set identification

Based on Chamberlain (1986) and Heckman (1990) Identification *at infinity* is when identification is based on the distribution of data at points where one or more variables goes to infinity.

Suppose iid scalar random variables Y, D, Z.
Assume $Y = Y^* D$, Z independent of Y^*,
Y^* is a latent unobserved variable
D is binary, $\lim_{z\rightarrow\infty} E\left(D \mid Z = z\right) = 1$.
The goal is identification and estimation of $\theta = E\left(Y^*\right)$.

This is a selection model, Y is selected (observed) only when $D = 1$. So D could be a treatment indicator, Y^* is the outcome if one is treated, θ is the average outcome if everyone 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.

$Y = Y^* D$, $Z \perp Y^*$, Y^* is latent, D is binary,
$\lim_{z \to \infty} E(D \mid Z = z) = 1$, $\theta = E(Y^*)$.

D and Y are correlated, so θ is identified only by $\lim_{z \to \infty} E(Y \mid Z = z)$.

Y^* and D may be correlated, $E(Y)$ confounds the two. But everyone who has Z infinite is treated, so looking at the mean of just those people eliminates the problem.

In real data we would estimate θ as the average value of Y just for people that have $Z > c$ for some chosen c, and then let $c \to \infty$ as the sample size grows (Heckman 1990 and Andrews and Schafgans 1998)

Or could estimate θ as a weighted average of Y with weights that get arbitrarily large as Z gets arbitrarily large (e.g., Lewbel 2007b).

First shown by Chamberlain (1986): Estimators based on identification at infinity usually converge slower than root-n.

Same estimation problem arises whenever identification is based on a Z taking on a value or range of values that has probability zero.

Kahn and Tamer (2010) call this *thin set* identification.

Example: Manski's (1985) and Horowitz's (1992) binary choice model, estimated by maximum score methods, is thin set identified. It gets identification only from info at one point (the median) of a continuously distributed variable.

Irregular identification is when thin set identification leads to slower than root-n rates of estimation. See Kahn and Tamer (2010) and Graham and Powell (2012).

Not all thin set identified or identification at infinity parameters are irregular. For example, estimates of $\theta = E\left(Y^{*}\right)$ in the selection problem can converge at rate root-n if Z has a strictly positive probability of equaling infinity.

More subtly, the 'impossibility' theorems of both Chamberlin 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.

Thick tails (infinite variance) can overcome the impossibility. This is one way that the Lewbel (2000) special regressor estimator can converge at rate root n.

Irregular identification is *not* the same as weak identification.

Irregularly identified parameters converge slowly for essentially the same reasons that nonparametric regressions converge slowly; at each point they're based on a vanishingly small subset of the entire data set.

Also like nonparametric regression, given suitable regularity conditions, standard methods can usually be used to derive asymptotic theory for irregularly identified parameters.

However, the rates of convergence of thin set identified parameters can vary widely, from extremely slow up to root n, depending on details regarding the shape of the density function in the neighborhood of the identifying point.

Sometimes it's easier to prove a parameter is identified by looking at a thin set, but the parameters may still be strongly identified, because the model and ϕ contain more information about the parameter than is being used in the proof.

Example: Heckman and Honore (1989) showed that without parameterizing error distributions, a competing risks model is identified by data where covariates can drive individual risks to zero. Lee and Lewbel (2013) later showed that this model is actually strongly identified, using information over the entire DGP, not just where individual risks go to zero.

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

Notes on irregular identification:

Identification at infinity of the selection model given above is discussed by Chamberlain (1986), Heckman (1990), Andrews and Schafgans (1998), Lewbel (2007), and Khan and Tamer (2010).

Khan and Tamer point out that the average treatment effect model of Hahn (1998) and Hirano, Imbens and Ridder (2003) is generally irregularly identified, and so will not attain the parametric root n rates derived by those authors unless the instrument has thick tails

Similarly, Lewbel's (2000) special regressor estimator for binary choice requires the special regressor to have thick tails (or other restrictions) to avoid being irregular.

7.3 Ill-Posed Identification

Let θ be point identified from knowable information ϕ in a model M. Ill-posedness arises when the connection from ϕ to θ is not sufficiently smooth.

Ill-posedness is an identification concept like weak identification or identification at infinity: it's a feature of θ, ϕ, and M, and hence a property of the population, that affects inference for any estimator one might propose..

The concept of well-posedness (the opposite of ill-posedness) is due to Hadamard (1923). See Horowitz (2014) for a survey.

Example: iid observations of W, so ϕ is the distribution function of W, denote it $F(w)$.

Glivenko–Cantelli theorem: $\widehat{F}(w) = \sum_{i=1}^{n} I(W_i \leq w)/n$ is uniformly consistent, asymptotically $n^{1/2}$ normal estimator of F.

Suppose $\theta = g(F)$ for some known g. Then θ is (point) identified.

If g is continuous, then $\widehat{\theta} = g\left(\widehat{\phi}\right)$ is consistent.

If g is not continuous, then $\widehat{\theta}$ is generally not consistent, and we have ill-posedness.

When ill-posed, consistent $\widehat{\theta}$ requires "regularization," smoothing out the discontinuity in g.
Regularization introduces bias, and consistency needs a way to asymptotically shrink this bias to zero.
Result: ill-posedness leads to slower convergence rates.

Example:
Estimation of the density function $f(w) = dF(w)/dw$.
There is not a continuous g such that $f = g(F)$.
Can't take derivative of $\widehat{F}(w) = \sum_{i=1}^{n} I(W_i \leq w)/n$

Rosenblatt-Parzen Kernel density estimation is regularization. Example:

$\widehat{f}(w) = \left(\widehat{F}(w+b) - \widehat{F}(w-b)\right)/(2b)$. Need $b \to 0$ as $n \to \infty$.
$\widehat{F}$ converges $n^{1/2}$, kernel $\widehat{f}$ is rate $n^{2/5}$.

Nonparametric regression (kernel or sieve based) is similarly ill-posed. Regularization is kernel and bandwidth choice, or in the selection of basis functions and the number of terms to include in nonparametric sieve estimators.

In some problems, ill-posedness is severe, causing very slow convergence rates, like $\ln(n)$.

Examples of problems that are often severely ill-posed are:

Nonparametric instrumental variables, e.g., Newey and Powell (2003),

Probability density estimation in models containing mismeasured variables.

Random coefficient models where the distribution of the random coefficients is nonparametrically estimated.

7.4 Bayesian and Essential Identification

Point identification is sometimes called frequentist identification or sampling identification, to contrast with Bayesian identification.

Bayesian: parameter θ is random, has a prior and a posterior.

Lindley (1971): point identification is irrelevant for Bayesians: has no effect on whether one can specify a prior and obtain a poserior. But Poirier (1998) discusses implications of failure of point identification for updating priors.

But there are still notions of identification that are relevant for Bayes estimation.

Gustafson (2005): Defines θ to be *essentially identified* (with respect to a given prior) if

1. θ is not point identified, but

2. θ would be point identified if the model included the additional assumption that θ is drawn from the given prior distribution.

Intuition: Unidentified parameters can become identified by adding restrictions to a model.
Imposing that θ be a draw from a given distribution function could be an example of such a restriction.

Gustafson shows that the behavior of Bayes estimators can depend heavily on whether they are essentially identified or not.

Florens, Mouchart, and Rolin (1990), Florens and Simoni (2011): θ is *Bayes identified* if its posterior differs from its prior distribution.

So θ is Bayes identified if data provides any info that updates the prior.

Point identification implies Bayes identification, because then the population updates the prior to a degenerate distribution.

Set identified parameters are also typically Bayes identified. Enough info to confine θ to a set should be enough to update a prior.

Example: Moon and Schorfheide (2012): When θ is set identified, the support of the posterior will typically lie *inside* the identified set.

Intuition: The data tells us nothing about where θ could lie inside the identified set, but the prior provides information inside the set, and the posterior reflects that information.

See Gustafson (2015) book for more on Bayes estimation in partially identified models.

9. Conclusions

Why is identification a rapidly growing area within econometrics (as evidenced by the growing terminology)?

The rise of big data and computation creating ever more complicated models needing identification.

Examples of increasingly complex models: games and auction models, social interactions and network models, forward looking dynamic models, and models with nonseparable errors assigning behavioral meaning to unobservables.

At the other extreme, the so-called credibility revolution: search for sources of randomization in constructed or natural experiments.

Like structural models, causal models are becoming more sophisticated. 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 for proving identification.

Identification theorems tend to be highly model specific.

Some general techniques for proving identification:
Control function methods as in Blundell and Powell (2004),
Special regressors as in Lewbel (2000),
Contraction mappings, fixed point theorems as in Berry, Levinsohn, and Pakes (1995),
Completeness as in Newey and Powell (2003),
Observational equivalence characterizations as in Matzkin (2005).
Characterizations of moments over unobservables as in Schennach (2014).

Development of more such general techniques and principles would be a valuable area for future research.

Identification, big data, and machine learning

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's treated as if it were the population.

Identification deals precisely with what can be learned given the population, i.e., given big data.

A valuable area for future research: connections between methods used to establish identification and techniques used to analyze big data.

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.

Given the increasing recognition of the importance of identification in econometrics, the Identification zoo is likely to keep expanding.