Identification and instruments

Ari Hyytinen

Hanken and Helsinki GSE

March 2024

(This lecture: Part 1 and 2)

Table of contents

- **Part 1**: Econometric identification
  - Introduction
  - Basic concepts
- **Part 2**: Normalizations and special regressors in identification
  - Scale and location normalizations
  - Special regressors
- **Part 3**: Normalizations in discrete choice models
  - Scale of utility
  - Outside good
- **Part 4**: Instruments for and identification discrete choice demand models
  - Importance and intuition for non-parametric identification in BLP
  - Challenges
  - Logic of formal proof: Index – Inversion – Instruments

Literature for Part 1 and 2

- Identification: General

  - Lewbel, A. (2019) The Identification Zoo – Meanings of Identification in Econometrics, ***Journal of Economic Literature***, vol. 57, no. 4, 835-903

  - Matzkin, R.L. (2013) "Nonparametric Identification in Structural Economic Models," ***Annual Review of Economics***, Vol. 5, 457-486.

  - Matzkin, R.L. (2007) "Nonparametric Identification," Chapter 73 in ***Handbook of Econometrics***, Vol. 6b, (edited by J.J. Heckman and E.E. Leamer, Elsevier B.V.), 5307-5368.

Part 1

Brief introduction to
econometric identification

Introduction

- Econometric identification:
  - Model parameters (or features) of interest are uniquely determined from the observable population that data are drawn from
  - Lewbel (2019): > 20 types of identification concepts appear in the econometrics literature

Identification precedes estimation

- Identification $\rightarrow$ estimation $\rightarrow$ inference & testing
- The question of identification is separate from the question of estimation of a parameter using *finite* samples:
  - Identification, in general, is *not* about an estimator.
  - *Nor* is it about what happens in a given sample

Point identification of a parameter

- A couple of ways to think about (point) identification
  - Assume you'd have access to an "infinitely large sample" and ask, what you can learn about the parameter using such data?
    - Can the parameter be recovered uniquely using the infinite data?
  - Assume existence of certain population quantities, what you can learn about the parameter of interest using them?
    - There is no need to worry about sampling uncertainty (etc).

Identification vs. consistent estimator

- Connection between identification and consistent estimator
  - Point identification is a *necessary* condition for there to exist e.g. a *consistent estimator* for the target parameter
  - If the parameter is not point identified under certain assumptions, then consistent estimators for it do ***not*** exist under the same set of assumptions.

…

- Point identification is *not sufficient* for a consistent estimator to exist:
  - It only means that such an estimator *may* be available
- Yet, note the reverse possibility:
  - Finding a consistent estimator ($plim_{n\to\infty}\hat{\beta}= \beta_0$) may be a way to prove that a parameter is point identified

What we know and want to learn 1/2

- θ = what we want to estimate and learn about = "*theoretical estimand*"
  - unknown parameters / vectors / functions
  - corresponds to "*estimands*", i.e., the population values of estimators of the objects that we are interested in

Notation (θ, φ) from Lewbel (2019, JEL)
Terms "theoretical "vs "empirical" estimand adopted from Lundberg et al., American Sociological Review.

What we know and want to learn 2/2

- ϕ = what is knowable about the DGP from data = "*empirical estimand*"
  - ϕ is information that is assumed to be known, or that can be learnt from an unlimited amount of data one has.
  - Examples: conditional means, moments, distribution functions, quantiles, true regression coefficients, autocovariances.

The identification question

- **Given ϕ, what can we learn about θ?**

  - Can we logically deduce the unknown value of the parameter, θ, from what can be measured from the observed data, ϕ?

  - Are model parameters or features, as captured by θ, uniquely determined

  … from what we know $\Leftrightarrow$ from ϕ $\Leftrightarrow$ from the observable population that our data are drawn from?

…

- *Note*: Identification presumes that there are structural features, θ, that one wishes to uncover

  - θ are often abstract notions, not part of the data themselves

  - E.g., is consumer *demand function* identified?

    - Demand function is an abstract (theoretical) notion.

    - It allows asking counterfactual questions, like what the demand for a product *would* be *if* price was X% higher than it actually is, holding all other things constant?

Point vs partial identification

- Point identification:
  - If given what ϕ equals, we know the value θ equals $\rightarrow$ θ is point identified
  - Called also: “global identification”, “frequentist identification”
- Partial (set) identification:
  - If given what ϕ equals, we can say *something* about the value of θ but cannot determine its value exactly $\rightarrow$ θ is partially identified
    - E.g., we might be able to determine that θ falls in an interval.
  - Identification is not an “all-or-nothing” concept; see Tamer (2010).

Parametric identification

- Parametric identification:
  - θ is a finite set of constants and all values of ϕ correspond to values of a finite set of constants.
  - For example, let us consider point identification of the model parameters (β) in a linear regression
    - This is an example of a *continuous identifying mapping*.

Example

Linear regression, exogenous X -- in this example, θ is β.

Model

$$Y = \mathbf{X}\boldsymbol{\beta} + U; \quad E(U|\mathbf{X}) = 0,$$

Implication of the model

$$E(\mathbf{X}'Y) = [E(\mathbf{X}'\mathbf{X})]\boldsymbol{\beta}.$$

Inversion gives the equation (*)

$$\boldsymbol{\beta} = [E(\mathbf{X}'\mathbf{X})]^{-1}E(\mathbf{X}'Y).$$

We suppose that the following are knowable (ϕ):

$$E(\mathbf{X}'\mathbf{X}),\ E(\mathbf{X}'Y).$$

…

Equation (*)

$$\boldsymbol{\beta} = \left[E(\mathbf{X}'\mathbf{X})\right]^{-1} E(\mathbf{X}'Y).$$

Suppose E[X'X] is a non-singular matrix. Then β is uniquely determined and continuous function of E[X'X] and E[X'Y].

Equation (*) thus identifies β.

I.e., we have $\theta = f(\phi) \Leftrightarrow \beta = E[X'X]^{-1}E[X'Y]$.

$$\boldsymbol{\beta} = [E(\mathbf{X}'\mathbf{X})]^{-1}E(\mathbf{X}'Y).$$

...

- Identification $\rightarrow$ estimation: Given that β is identified, we can consider its *estimation*.
  - Identification does *not* ensure that there exists a consistent estimator.
- Data available: Suppose that we have a random sample from the probability distribution of Y and **X**.
- Analogy principle: Replace the unknown population expectations with sample averages.

$$\mathbf{m}_{XY} = n^{-1}\sum_{i=1}^{n}\mathbf{X}_i'Y_i \qquad \mathbf{m}_{XX} = n^{-1}\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{X}_i. \qquad \hat{\boldsymbol{\beta}}_{LS} = \mathbf{m}_{XX}^{-1}\mathbf{m}_{XY}.$$

…

- Consistency of the sample averages:

  - Consistency of the sample averages implies that when the sample size is sufficiently large, the averages converge to the corresponding population moments.

- Continuity of the identifying equation $\beta = E[X'X]^{-1}E[X'Y]$

  - Small changes in $E[X'X]$ and $E[X'Y]$ cause only small changes in β.

- Consistency of the sample averages + the continuity of the identifying equation $\rightarrow$ when the sample size is sufficiently large, $\hat{\boldsymbol{\beta}}_{LS}$ is arbitrarily close to $\boldsymbol{\beta} \rightarrow$ the LS estimator is consistent for β.

Another example

Linear IV regression, endogeous X and instruments Z

Model

$$Y = \mathbf{X}\boldsymbol{\beta} + U; \qquad E(U|\mathbf{Z}) = 0,$$

We suppose that the following are knowable (ϕ):

$$E(\mathbf{X}'\mathbf{Z}) \quad E(\mathbf{Z}'\mathbf{Z}) \quad E(\mathbf{Z}'Y)$$

Implications of the model

$$E(\mathbf{Z}'Y) = [E(\mathbf{Z}'\mathbf{X})]\boldsymbol{\beta},$$

$$E(\mathbf{X}'\mathbf{Z})\,[E(\mathbf{Z}'\mathbf{Z})]^{-1}E(\mathbf{Z}'Y) = E(\mathbf{X}'\mathbf{Z})\,[E(\mathbf{Z}'\mathbf{Z})]^{-1}\,[E(\mathbf{Z}'\mathbf{X})]\boldsymbol{\beta}.$$

Inversion gives the equation (**)

$$\boldsymbol{\beta} = \left\{E(\mathbf{X}'\mathbf{Z})\,[E(\mathbf{Z}'\mathbf{Z})]^{-1}E(\mathbf{Z}'\mathbf{X})\right\}^{-1}E(\mathbf{X}'\mathbf{Z})\,[E(\mathbf{Z}'\mathbf{Z})]^{-1}E(\mathbf{Z}'Y).$$

…

Equation (**)

$$\boldsymbol{\beta} = \left\{E(X'Z)\left[E(Z'Z)\right]^{-1}E(Z'X)\right\}^{-1}E(X'Z)\left[E(Z'Z)\right]^{-1}E(Z'Y).$$

If the inverse matrices on the RHS of this equation exist, β is uniquely determined. Equation (**) is a continuous identifying mapping.

Hence, β identified. It is a continuous function of the population moments.

Existence of a consistent IV estimator can be established using the same arguments as in the previous example.

Non-parametric identification

- Nonparametric identification:
  - θ consists of functions or infinite sets
  - Example: Think of
    - some function $m(X)$; or
    - joint density of ε and X, i.e., $f_{\varepsilon,X}$

Example

- Case when θ is a function rather than a vector:
  - DGP is iid (Y, X), and suppose that X is continuous, that U and X are independent (i.e., U _|_ X), and that Y = 1(X + U > 0).
  - We want to learn:
    - $\theta = F_U(u)$ which is the distribution function of U.

…

For any value x that X can take on:

$$E(Y \mid X = x)$$

$$= \Pr(X + U > 0 \mid X = x)$$

$$= \Pr(x + U > 0) = \Pr(U > -x)$$

$$= 1 - \Pr(U \leq -x) = 1 - F_U(-x).$$

- F_U can be recovered from $E(Y \mid X = x)$, i.e., function F_U is nonparametrically identified.
  - Note: $F_U(u)$ is only identified for values of u that are in the support of -x.

Another example

- Consider model

$$Y = g^*(X) + \varepsilon \qquad\qquad F^*_{\varepsilon,X}.$$

Assume the following are knowable: function $g^*(X)$ (e.g, has already been identified), marginal distribution F^*_X and conditional $F_{Y|X=x}(y)$.

Can we identify the joint cumulative distribuion function $F^*_{\varepsilon,X}$ i.e., true $\Pr(\varepsilon \leq e \text{ and } X \leq x)$?

…

$$\begin{aligned}\boxed{F_{Y|X=x}(y)} &= \Pr\left(Y \leq y|X = x\right) \\ &= \Pr\left(g^*(X) + \varepsilon \leq y|X = x\right) \\ &= \Pr\left(\varepsilon \leq y - g^*(x)\,|X = x\right) \\ &= F^*_{\varepsilon|X=x}\left(y - g^*(x)\right)\end{aligned}$$

This shows that the conditional $F^*_{\varepsilon \mid X = x}$ is identified.

Because the marginal distribution F^*_X is also identified, we can infer that $F^*_{\varepsilon,X}$ is identified.

Semiparametric identification

Semiparametric identification

$\approx$ identification that is neither completely parametric nor completely nonparametric

- θ may e.g. include a vector of constants *and* nuisance parameters that are functions

Example

- Consider random variables (Y, X, Z) and assume the observations are *iid.*

- Partially linear model: $Y = m(Z) + X'\beta + \varepsilon$ where m is an unknown function, β is a finite vector of parameters, and $E(\varepsilon \mid X, Z) = 0$.

- Unknown parameters θ: Vector of constants, β, and a function m(Z).

  - Identification of θ is semiparametric, because θ contains a parametric component (finite vector β) and a nonparametric component (function m(Z)).

…

- Semiparametric identification (”by construction”):

  - *Step 1*: Assume E[Y| Z] and E[X| Z] are knowable from data, i.e. included in ϕ.

  - *Step 2*: Because $Y - E[Y|Z] = (X-E(X|Z)'\beta + \varepsilon$, vector β is identified by construction:

    - $E[Y|Z] = m(Z) + E(X|Z)'\beta$
    - Regress $Y - E[Y|Z]$ on $(X-E(X|Z))$, works if var $(X|Z)$ is non-singular.

  - *Step 3*: Given β, function m(Z) identified non-parametrically, because $E(Y - X'\beta|Z) = E(m(Z) + \varepsilon|Z) = m(Z)$

What is knowable?

- More generally, what can be included in ϕ?
  - E.g., how do we know that ϕ includes the conditional expectation of Y given X, $E[Y \mid X]$?
  - Expectations of observed variables are knowable

    $\Leftrightarrow$

    Certain statistical properties of observable sample averages hold in the data (e.g., their unbiasedness, consistency).

…

- Circularity in the definition of identification:

  - We assume initially that ϕ is identified (known) to determine if θ is identified.

  - Assuming ϕ is knowable $\Leftrightarrow$ assuming ϕ is identified

    - This must be justified by assumptions about the underlying DGP (i.e., by the model)

…

- Examples of ϕ include the following:
  - Distribution of (Y, X) for IID observations
  - Reduced form linear regression coefficients
  - Conditional distribution of Y given D where D values determined by an experimental protocol
  - Means and autocovariances in stationary time series data
  - Transition probabilities if data assumed to follow a martingale process

Class discussion

- Suppose you are told that data (X_1, X_2,..., X_n) are *iid* random variables with a common cumulative distribution function.

- What is knowable to start with, i.e., what can be included in ϕ?

  - (Guess)

Model

- A model, M, corresponds to assumptions about and restrictions on the DGP:
  - Assumptions about the behaviour that generates the data
    - E.g., statistical (e.g. randomization) and behavioral assumptions (e.g., a set of equations describing behavior)
  - Assumptions about how the data are collected and measured.
    - E.g., assumptions about selection, measurement errors, and survey attrition.

…

- Model M imposes restrictions
  - on the possible values ϕ could take on
    - I.e., ϕ depends on the model (what is knowable depends on M)
  - on how ϕ and θ are related.
- θ is identified under the maintained model M if it is uniquely determined by the population distribution of observables.

…

- One cannot even ask whether θ is identified without a model (an abstraction)
  - Model M defines what this quantity, θ, is.
  - It is difficult to discuss most objects of interest in empirical economics without a model
  - Model may be based on economic theory or e.g. on hypothesized causal relations
    - (e.g., Rubin (1974) causal model)

…

- Trade-off: Stronger modelling assumptions may help to answer more intricate questions:

  - *Law of decreasing credibility*: The credibility of inference decreases with the strength of assumptions maintained.

  - Stronger assumptions may yield *clearer* conclusions, but are less credible.

    - E.g., contrast parametric vs non-parametric modelling choices, use of generic functional forms vs. linear relations

[Manski C. F., 2007, Identification for Prediction and Decision, Harvard Univ. Press]

…

- Clarifying note: Identification is not even about what we learn from an "infinitely large sample":

  - E.g. let M consists of utility functions that are maximized s.t. budget constraint and let ϕ = Demand functions and θ = Indifference curves.

  - Revealed preference theory $\rightarrow$ if ϕ known, point identification of θ.

  - A *separate* identification question: When can demand functions be identified from observed data?

…

- Example: As we will see later, the non-parametric identification of BLP model relies on a particular *non-parametric functional form* restriction

  - The non-parametric index restriction => 2×J instruments are sufficient to identify demand functions

  - Further parametric assumptions are stronger modelling assumptions that *reduce* the number of instruments that are needed for identification

    - (but may reduce the creditability of findings)

Model: Typical DGP assumptions

- ***DGP-assumption #1***: *iid* observations of data vector W, with n $\to \infty$.
  - Glivenko-Cantelli theorem: With such data, the distribution of W can be consistently estimated
  - Reasonable to assume that knowable ϕ includes the distribution function of W.

…

- ***DGP-assumption #2***: Randomized experiment:
  - Experimental protocol determines value of D (e.g., treatment indicator) for each observation.
  - Conditional on that value of D, data collected by randomly drawing an observation of Y, independent of other observations.
  - Reasonable to assume that knowable ϕ includes the conditional distribution function of Y given D.
    - Note: ϕ is only knowable for values of D that can be chosen by the experimenter.

- ***DGP-assumption #3***: Stationary time series data

  - Reasonable to assume that ϕ includes means, variances and autocovariances

  - Note: This does not automatically imply that ϕ includes higher moments

    - (they may be unstable over time)

Example

- What kind of a model (M) underlies RCTs and causal inference that they enable?

    - What is / could be M? What is / could be ϕ? What is / could be θ?

- E.g. consider an ideal RCT, with no non-compliance, no self-selection, no measurement error and let Y = outcome, and D = treatment

…

- Model, M, is implied by the ideal RCT + knowing (assuming) that data on realizations of (Y, D) are i.i.d. across individuals.

- Let θ be ATE: $\theta = E(Y(1) - Y(0))$, where $Y(t)$ is the outcome an individual would have if assigned $D = t$ (Rubin 1974).

- Given M, ϕ is the distribution of Y, D.

  - E.g., ϕ includes the conditional expectation of Y given D, $E[Y \mid D]$,

$\rightarrow$ $E[Y \mid D]$ is knowable from the data

…

- M is the set of all possible joint distributions of (Y,(1), Y(0), D).

  - Model of "an ideal RCT" implies that D determined randomly ("by a coin flip").

$\rightarrow$ This is a *restriction* on M: (Y(1), Y (0)) independent of D.

…

- Given M, θ is point identified:
  - $E(Y|D=1) - E(Y|D=0) = E(Y(1)|D=1) - E(Y(0)|D=0) = E(Y(1)) - E(Y(0)) = ATE = \theta$
  - Because, given M, there is no selection bias + no bias due to heterogenous treatment effects
  - Identification always entails a model: This applies also to causal inference

Proving point identification

- *Approach #1*: By construction, i.e., writing θ directly as a function of ϕ
  - All of the examples earlier
  - *A further example*: Directly prove consistency
    - Construct an estimator $\hat{\theta}$ and prove that, under the assumed DGP, the estimator is consistent.
    - The construction is $\operatorname{plim} \hat{\theta} = \theta$.

…

- *Approach #2*: Proving true θ is the unique solution to an optimization problem.
  - Example: Maximum likelihood (ML) with a concave population objective function $\rightarrow$ ML has a unique maximizing value
  - Identification follows if one can show that the unique maximiser in population is the true parameter value θ_0

…

- *Approach #3*: Showing the true θ (θ_0) is the unique fixed point in a contraction mapping based on M.

  - Example: A contraction mapping is used in Berry (1994) to prove that a necessary condition for identification (uniqueness in the error inversion step) holds in the BLP model.

…

- *Approach #4*: Applying characterizations of *observational equivalence* in some classes of models

  - see Matzkin (2008, 2013)

  - Recall: θ is point identified if each possible value of ϕ implies a unique value of θ in Θ (= the set of all possible values that the model says θ could be)

...

- Observational equivalence and identification:
  - Two possible values θ and θ^* are **observationally equivalent** if there exists a value of ϕ that could imply either θ or θ^*.
  - Identification using an observational equivalence argument:
    - θ is point identified **if** θ and θ^* being observationally equivalent **implies** that θ and θ^* are equal.
    - In other words, θ is point identified if there do *not* exist any pairs of possible values θ and θ^* (in Θ) that are different but observationally equivalent.

Summary

Common starting points for proving point identification

1. *Wright-Cowles identification*: ϕ is a set of reduced form population regression coefficients or population moments (e.g., E[X'X] and E[X'Y])

2. *Distribution based identification*: ϕ is the distribution function of an observable random vector Y

3. *Extremum based identification*: ϕ is the maximizer of some function (e.g. GMM or ML objective function)

Why point identification may fail?

- Non-mutually exclusive reasons:
  - Model incompleteness: E.g., variable relationships not fully specified, multiple equilibria.
  - Perfect collinearity / perfect dependence
  - Nonlinearity: Possible multiple solutions
  - Simultaneity
  - Endogeneity
  - Unobservability: E.g., counterfactuals

...

- See Horowitz (2019) for further discussion
- Discontinuities in non-parametric identification
    - (IV models)
- Ill-posed inverse problems in econometric identification

ANNUAL REVIEWS **Further**

Click here for quick links to Annual Reviews content online, including:

- Other articles in this volume
- Top cited articles
- Top downloaded articles
- Our comprehensive search

Ill-Posed Inverse Problems in Economics

Joel L. Horowitz

Department of Economics, Northwestern University, Evanston, Illinois 60208; email: joel-horowitz@northwestern.edu

Annu. Rev. Econ. 2014. 6:21–51

First published online as a Review in Advance on January 17, 2014

The *Annual Review of Economics* is online at economics.annualreviews.org

This article's doi: 10.1146/annurev-economics-080213-041213

JEL codes: C14, C26

Keywords

regularization, nonparametric estimation, density estimation, deconvolution, nonparametric instrumental variables, Fredholm equation

Abstract

A parameter of an econometric model is identified if there is a one-to-one or many-to-one mapping from the population distribution of the available data to the parameter. Often, this mapping is obtained by inverting a mapping from the parameter to the population distribution. If the inverse mapping is discontinuous, then estimation of the parameter usually presents an ill-posed inverse problem. Such problems arise in many settings in economics and other fields in which the parameter of interest is a function. This article explains how ill-posedness arises and why it causes problems for estimation. The need to modify or regularize the identifying mapping is explained, and methods for regularization and estimation are discussed. Methods for forming confidence intervals and testing hypotheses are summarized. It is shown that a hypothesis test can be more precise in a certain sense than an estimator. An empirical example illustrates estimation in an ill-posed setting in economics.

Class discussion

- In light of what we have covered, consider a standard linear regression model, under typical textbook assumptions
  - What is model M?
    - What restrictions does it imply?
  - What is knowable ϕ?
  - What is θ?
    - Discuss identification θ.

Part 2

This part is about normalizations and special regressors and how they are related to identification

Normalizations

Normalizations

- Identification requires often normalizations
  - If parameter restrictions (e.g. scaling of a parameter) can be made without loss of generality ("wlog") they are normalizations
  - In economics, "wlog" = if the parameter restriction does not affect economically meaningful parameters or summary measures

…

- A parameter restriction can be either a *free normalization* or imply a *behavioural restriction* (assumption)

  - The distinction depends on the model and research question

  - This is about how we use the model and interpret it.

...

- Note #1:

  - Even if a normalization is irrelevant for identification,

    it can affect numerical performance of estimators (e.g., convergence) and/or precision of estimates.

…

- Note #2:
  - Continuity, differentiability, monotonicity and other similar additional restrictions on functions in θ are behavioural restrictions (assumptions), not free normalizations.
    - How restrictive or consequential they are depends on the context;
    - See later our discussion on the conditions on invertability of mkt shares in BLP

Scale restriction / ”Up to scale”

- Identification “up to scale”:

  - Suppose all the elements θ in the identified set are proportional to true β, i.e. are of the form: $\beta^* = \beta / c$.

  - To identify β, a scale restriction is needed:

    - E.g., assume that the first element of β vector equals one (e.g. $\beta_1=1$).

    - Then $\beta^*_1 = 1/c$, which in turn allows recovering the remaining elements of β from β^*.

…

- A scale restriction is not always a free normalization
- In some models, the *level* of $X'\beta$ has *economic meaning*
  - **Example:** Y = 1 if willing to pay (WTP) more than C euros for a product or service.
  - Let $X'\beta + \varepsilon$ be WTP, where $\varepsilon \perp X$. Then $Y = 1(X'\beta + \varepsilon > C)$.
    - $E(Y \mid X) = g(X'\beta - C)$ where g is distribution of $-\varepsilon$ and where $(X'\beta - C)$ is the index being estimated.
  - Scaling is not a free normalization: $X'\beta + \varepsilon$ is WTP only if the coefficient of C is minus one.

Location restriction

- Location restrictions as a normalization
- Let $E(Y \mid X) = g(X + \alpha)$, g is an unknown function, α is an unknown scalar.
- Generally α is not identified, because $g(X + \alpha)$, is observationally equivalent to using $\alpha = 0$ and g^* such that $g^*(X) = g(X + \alpha)$.
  - Location normalizations may or may not be free, depending on the context.

Example: Linear index model

- Consider first a general linear index model:

  - $E\,(Y \mid X) = g(X'\beta)$,
  - g is strictly monotonically increasing
  - $E(XX')$ is non-singular

- What is known: ϕ is the joint distribution of Y and X

- What we want to know: θ is β

.

…

- Example: For Probit, g is assumed to be the cumulative *standard* normal distribution function ($g(X'\beta) = \Omega(x'\beta)$)
  - This implies that g is *known.*
- Result: In Probit, θ is identified.
  - $\Pr[Y = 1 \mid X] = E(Y \mid X) = g(X'\beta)$
  - Proof by construction:

$$\beta = [E(XX')]^{-1} E[X g^{-1}(E(Y \mid X))]$$

Normalizations in Probit

- To see the hidden normalizations in this ”identification” result, consider the following special case of the linear index model:

  - *Threshold crossing* binary choice model:

    - $Y = 1(\alpha + X'\beta + \varepsilon)$ where $\varepsilon \_|\_ X$.

  - Note: This is still a version of the general linear index model with $E(Y \mid X) = g(X'\beta)$ when g is the distribution function of $-(\alpha + \varepsilon)$.

...

- This threshold crossing binary choice model is equivalent to:
  - $Y = 1$ if $Y^* > 0$ and $Y = 0$ if $Y^* \leq 0$, with $Y^* = \alpha + X'\beta + \varepsilon$
  - $\Pr[Y = 1 \mid X=x] = \Pr[Y^* > 0 \mid X=x] = \Pr[\varepsilon > -X'\beta \mid X = x] = \Pr[\varepsilon < x'\beta]$
  - Which normalizations must be imposed into this model to obtain the Probit model?

…

- $\Pr[Y = 1 \mid X=x] = \Pr[\varepsilon < x'\beta] = \Omega(x'\beta)$

- Identification of the Probit coefficients based on both *location* and *scale normalization*

  - Probit assumes: $E(\varepsilon) = 0$ and $Var(\varepsilon) = 1$

...

- What do $E(\varepsilon) = 0$ and $Var(\varepsilon) = 1$ mean?
  - Scaling the variance of ε does not change the observed data, or choice.
  - Having mean zero for ε (i.e. threshold) is innocent as long as the model has a constant.
  - These determine the location and scale of $\alpha + X'\beta + \varepsilon$.
- Note, there are observationally equivalent normalizations: e.g., let ε have arbitrary mean and variance, but set: $\alpha = 0$, $\beta'\beta = 1$.
- We shall return to normalizations in discrete choice demand estimations

Special regressors

What is a special regressor?

- A special regressor
  - is an observed covariate
  - has properties that facilitate identification and estimation of econometric models
- E.g., non-parametric identification of BLP relies on existence of such a covariate.

Example

- Suppose an observed binary variable Y satisfies

$$Y = 1[V + W]$$

where V is the observed special regressor with a coefficient of one and where W is an unobserved latent variable.

The goal is identification and estimation of the distribution of W.

…

- Many models have this form Y = 1[V + W]
  - Compare this to: $Y = 1[\alpha + X'\beta + \varepsilon]$, which is the Probit model if ε is distributed the standard normal
  - A Probit does *not* require a special regressor, because it relies on parametric assumptions that allow uncovering β.
    - The distribution of ε has no free parameters and its shape is assumed.

…

- How does special regressor methods allow identification of the distribution of W?

  - They exploit the fact that if V is independent of W then variation in V changes Pr[Y = 1 | V] in a way that traces out the distribution of W.

  - Depending on the context, this can be done find out either the unconditional distribution of W or its distribution conditional on covariates.

Example

- Suppose we want to uncover the distribution of people's willingness to pay (WTP) W to preserve a piece of forest.

- Denote this distribution function as $F_W(w) = \Pr(W \leq w)$

- Data from a survey:

    - Random price P $\rightarrow$ is the respondent willing to pay $\geq$ P € to preserve the forest?

    - P is drawn independently of W.

…

- Here, $Y = 1[W - P]$, with P taking the role of a special regressor

- $E[Y = 1 \mid P = p] = \Pr[W > p] = 1 – \Pr[W < p] = 1 – F_W(p)$.

- Thus, $F_W(p) = 1 – E[Y = 1 \mid P = p]$.

  - This shows how variation in P allows identifying distribution of WTP.

…

- E.g., suppose 70% of the respondents said they would be not be willing to pay more than €50 to preserve the forest.

- In this example p = 50 and so 0.70 would be an unbiased estimate of $1 - E[D = 1 \mid p = 50] = F_W(50)$.

Hypotetical data

p	Share of willing to pay	Share of not willing to pay	1 - E[D = 1 | p]	F(W)	1-F(W)
40	1.000	0.000	0.000	0.000	1.000
41	0.991	0.009	0.009	0.009	0.991
42	0.950	0.050	0.050	0.050	0.950
43	0.900	0.100	0.100	0.100	0.900
44	0.800	0.200	0.200	0.200	0.800
45	0.700	0.300	0.300	0.300	0.700
46	0.600	0.400	0.400	0.400	0.600
47	0.500	0.500	0.500	0.500	0.500
48	0.450	0.550	0.550	0.550	0.450
49	0.400	0.600	0.600	0.600	0.400
50	0.300	0.700	0.700	0.700	0.300
51	0.290	0.710	0.710	0.710	0.290
52	0.250	0.750	0.750	0.750	0.250
53	0.200	0.800	0.800	0.800	0.200
54	0.190	0.810	0.810	0.810	0.190
55	0.160	0.840	0.840	0.840	0.160
56	0.110	0.890	0.890	0.890	0.110
57	0.900	0.100	0.100	0.100	0.900
58	0.500	0.500	0.500	0.500	0.500
59	0.025	0.975	0.975	0.975	0.025
60	0.000	1.000	1.000	1.000	0.000

Examples of usage

- The special regressor method can mean a variety of approaches. It has been used in e.g.
  - binary, ordered, and multinomial choice models
  - censored regression, selection and treatment models
  - truncated regression
  - binary and other nonlinear panel models with fixed effects
  - contingent valuation models
  - dynamic choice models
  - market equilibrium models of multinomial discrete choice (BLP)
  - models of games, including entry games and matching games

…

- Lewbel, 2014, An Overview of the Special Regressor Method
  - In: Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics, Edited by Ullah, Racine, and Su, Oxford University Press.