Notes to accompany lectures 1-5 of ECON0108

2022-2023

Andrew Chesher

Department of Economics, University College London

CeMMAP: The Centre for Microdata Methods and Practice

www.cemmap.ac.uk

1. Introduction[1]

In econometric analysis since the 1940's, economic data are regarded as realisations of random variables which have well defined probability distributions.[2] The statistical methods of econometrics which are the main focus in many econometrics textbooks are designed to extract information about features of these probability distributions.

Different economic data generating processes, or *structures*, can give rise to identical probability distributions for observable random variables. Such structures are described as *observationally equivalent*. If there can be observationally equivalent structures then even complete knowledge of the exact probability distribution of observed variables may not give unambiguous information about some features of the economic data generating process. This is the problem addressed in the study of *identification*.

This ambiguity may not arise if there are restrictions that rule out as inadmissible certain observationally equivalent structures. This is the job done by an econometric model. An econometric model is just a set of restrictions which place limitations on the properties of admissible structures.

Suppose that for each probability distribution of observable outcomes a model admits just one structure. Then the model identifies the structure. Models with this powerful property are often highly restrictive. In practice one is usually interested in one or a few features of a structure which can perhaps be identified by less restrictive models.[3] If a model restricts admissible structures such that within any set of observationally equivalent structures a feature is constant then the model identifies that feature. If a model restricts admissible structures such that within any set of observationally equivalent structures a feature takes values in a set then the model set identifies, or partially identifies that feature.It is good to have a clear idea of which features of structures are interesting and which are not when constructing a model.

In practice econometric models contain many restrictions which are not essential for identification of some structural features. There may be for example restrictions which facilitate estimation (e.g. parametric restrictions or semiparametric index restrictions) or inference (e.g. restrictions on existence of moments). The modern study of identification aims to understand which of the restrictions of a model have essential identifying power for particular structural features. This leads to the study of nonparametric identification.

It is important to realise that identifiability of structural features depends on what can be observed. Faced with a problem in which there is only identification of some feature of interest using a highly restrictive model one might consider expending effort changing data collection procedures so that more is observed.

---

[1] These (incomplete) notes have evolved while preparing short courses given at Oxford (2004), Copenhagen (2004), Brown (2004), Yale (2008), Uppsala (2008), UCLA (2011), UCL (2011), Renmin University of China (2012), Boston Unversity (2015), Jinan University (2018), Shanghai Jiaotong University (2018) and the core PhD course in econometrics at University College London. I thank participants for helpful comments.

[2] This is the approach set out in Haalvemo (1942).

[3] Note how we talk about a model "identifying a structure". It is common to hear people talk, incorrectly, of a model being (or not being) identified. It is models that do the identifying and structures that are (or are not) identified by a model.

2. Econometric models

We consider economic processes that deliver values of endogenous outcomes given values of exogenous variables whose values are unchanged by the process. We mainly focus on static models which do not place restrictions on the dynamics of the process delivering outcomes. Mcuh of what follows extends todynamic models very straighforwardly.

Economics providesinformation about some of the properties of data generating economic processes. The knowledge that economics gives us about data generating processes is embodied in econometric models. Most econometric models also embody some restrictions that do not flow from economics.

As an example consider an econometric model which might be used in the study of the *returns to schooling*.

Suppose we are interested in the determination of a labour market outcome, say the log wage, W, and a measure of investment in schooling (S), say years of schooling, given a value of another specified characteristic (X) of an individual. Here is an example of the structural equations of a model for the process generating wage and schooling data given a value of X.[4]

$$W = \alpha_0 + \alpha_1 S + \alpha_2 X + \varepsilon_1 + \lambda\varepsilon_2 \tag{2.1}$$

$$S = \beta_0 + \beta_1 X + \varepsilon_2 \tag{2.2}$$

The term ε_2 is unobserved and allows individuals with identical values of X to have different values of S, something likely to be seen in practice. We could think of ε_2 as a measure of "ability". This term also appears in the log wage equation, expressing the idea that higher ability people tend to receive higher[5] wages, other things being equal. The term ε_1 is also unobserved and allows people with identical values of S, X and ε_2 to receive different wages, again, something likely to occur in practice.

In econometric models unobservable, latent, terms like ε_1 and ε_2 are specified as *random variables*, varying across individuals (in this example) with *probability distributions*.[6] Typically an econometric model will place restrictions on these probability distributions as well as on the functions which transform realisations of latent variables into values of observable outcomes. In this example a model could require ε_1 and ε_2 to have expected value zero and to be uncorrelated with X.

This model's structural functions embody very strong restrictions, being linear in variables and parameters and triangular in form in that S appears in the structural equation for W but W does not appear in the structural equation for S.

The terms α_0, α_1, α_2, λ, β_0 and β_1 are unknown *parameters* of this model. A *particular* data generating process that conforms to this model will have equations as set out above with *particular* numerical values of the parameters and *particular* distributions for the unobservable ε_1 and ε_2. We will call such a fully specified data generating process a *structure*.

Each structure implies a particular probability distribution for W and S given X and the statistical tools of econometrics can inform us about this distribution. Of course it may be that

[4] Card (1995, 2001) presents an economic model of wage determination and schooling choice leading to a structural form like this one and those considered later. Becker and Chiswick (1966), Chiswick (1974), Chiswick and Mincer (1972), Mincer (1974) are seminal works on this problem.

[5] If $\lambda > 0$.

[6] An approach set down in Haavelmo (1944).

the structure that generates the data we see does not conform to the restrictions embodied in the model so far described. In that case the model is misspecified.

In the example above the structural equations are linear and involve a finite number of parameters. Typically economic argument does not lead to a linear specification so one might consider a less restrictive specification. For example in the wage-schooling example one might entertain a model in which structural equations have the nonadditive form:

$$W = h_1(S, X, \varepsilon_1, \varepsilon_2) \tag{2.3}$$
$$S = h_2(X, \varepsilon_2) \tag{2.4}$$

for some functions h_1 and h_2, perhaps with some restrictions imposed on these functions, for example smoothness or monotonicity restrictions. The linear model set out earlier is a special case of this model.

Or one might propose an index restriction

$$\begin{aligned} W &= h_1(\theta(S, X), \varepsilon_1, \varepsilon_2) \\ S &= h_2(X, \varepsilon_2) \end{aligned}$$

where $\theta(\cdot, \cdot)$ is a scalar valued function, with all the functions, h_1, h_2 and θ, left nonparametrically specified.

In the context of wage-schooling models which incorporate these various types of structural equations we may be interested in the value of the *returns to schooling*. In the linear model this is the value of the parameter α_1. In the nonadditive model it is $\nabla_s h_1(s, x, e_1, e_2)$ potentially depending on the values, s, x, e_1, and e_2, of respectively S, X, ε_1, and ε_2.

We now consider whether, if data were generated by a structure admitted by one of these models, it could be informative about the value of the returns to schooling.

This question arises because *distinct* structures may imply the *same* probability distribution for the observable random variables. Such structures are termed *observationally equivalent*. Data, in whatever quantity, can only be informative about the probability distribution of the observable random variables and functionals of that probability distribution (for example quantile functions, mean regression functions and so forth).

If, across observationally equivalent structures, a structural feature (e.g. α_1 in the linear case, or the function $\nabla_s h_1(s, x, e_1, e_2)$ in the nonparametrically specified case) takes *different* values then *no amount* of data can be informative about the value of that structural feature. We talk then of the structural feature being *not identified*. If an econometric model is sufficiently restrictive then among any set of observationally equivalent structures a structural feature takes a common value and the structural feature is point identified.[7]

To see how observationally equivalent structures can arise, return to the linear wage-schooling model, and consider what happens when we substitute for ε_2 in the log wage equation using $\varepsilon_2 = S - \beta_0 - \beta_1 X$ which is implied by the schooling equation. After collecting terms we obtain the following.

$$\begin{aligned} W &= (\alpha_0 - \lambda\beta_0) + (\alpha_1 + \lambda)\, S + (\alpha_2 - \lambda\beta_1)\, X + \varepsilon_1 \\ S &= \beta_0 + \beta_1 X + \varepsilon_2 \end{aligned}$$

[7] There may be restrictions sufficient to identify the value of a structural feature up to an interval, but insufficient to achieve point identification. We encounter this situation when we study identification in problems with discrete variation. See Chesher (2003), Manski (2003).

Write the wage equation as

$$W = \gamma_0 + \gamma_1 S + \gamma_2 X + \varepsilon_1$$

$$\begin{aligned} \gamma_0 &= \alpha_0 - \lambda\beta_0 \\ \gamma_1 &= \alpha_1 + \lambda \\ \gamma_2 &= \alpha_2 - \lambda\beta_1 \end{aligned}$$

and note that for any given values of β_0 and β_1 there are many values of α_0, α_1, α_2 and λ which yield identical values of γ_0, γ_1, and γ_2.

Consider that, if we knew a few values of W, S, X and *also* of ε_1 and ε_2 then we could *deduce* the values of β_0, β_1, γ_0, γ_1, and γ_2. But we could never tell what values of α_0, α_1, α_2 and λ had generated the data. Clearly *no* statistical analysis of data in the context of this model can be informative about the values of the parameters of the structural wage equation. A more restrictive model is required if these parameters are to be identified.

If economic theory required that $\alpha_2 = 0$ and $\beta_1 \neq 0$ then there would be only one set of values of α_0, α_1 and λ which could produce a given set of values of W and S given a particular set of values of X, ε_1 and ε_2.[8]

Considering the schooling equation, note that only one value of β_0 and β_1 could generate a particular set of values of S given a particular set of values of X and ε_2.[9] The values of β_0 and β_1 are identified, *if* we know the values taken by ε_2.

When, as is always the case, we do *not* observe the values of ε_2 the linear model for schooling is not sufficiently restrictive as it stands to allow identification of the schooling equation's parameters. For example, if large values of ε_2 tend to be associated with large values of X, then data on X and S alone cannot distinguish the impact of X and ε_2 on S. The parameter β_1 measures the impact of X on S, ε_2 held constant and in the absence of further restrictions the marginal impacts of X and ε_2 on S cannot be determined when values of ε_2 are not observed.

Clearly to achieve identification of the parameters of the schooling equation there must be some restriction on the *co-variation* of latent variable ε_2 and X.

Such restrictions can take many forms. For example we might propose that in admissible economic processes the conditional expectation or median of ε_2 given $X = x$ to be invariant with respect to changes in x. Conditional moment and quantile restrictions of this sort are often used to motivate econometric *estimators*. Note that here they arise as restrictions on economic processes which achieve identification of features of a process. Their use in motivating estimators arises precisely because of their identifying power.

We might impose a more severe restriction and suppose that in admissible economic processes the entire conditional distribution of ε_2 given $X = x$ is invariant with respect to x. This strong independence restriction is frequently considered in models with nonparametrically specified, nonseparable equations like (2.3) and (2.4).

Some of the restrictions that constitute an econometric model may flow from economic arguments - for example that W does not appear in (2.1) or (2.3), a restriction really driven by the timing of decisions. Others, for example statistical independence of X and ε_2, may not flow from economic arguments, although there are cases in which an independence restriction can be

[8] Unless special sets of values of X arise.

[9] Unless the X data take special sets of values, for example each of the 100 values of X is identical.

justified by economic argument, for example in the study of the operation of efficient financial markets.

An important point is that identification of values of features of economic processes *always* rests on some set of non-trivial restrictions and in any econometric model with identifying power there is always a subset of restrictions which no amount of data can refute. This observation leads to the following conclusions.

1. All interpretations of econometric analysis are contingent on a set of restrictions holding, a set of restrictions whose appropriateness cannot be determined from data alone.

2. When considering models with which to identify interesting structural features we should seek to understand what *minimally* restrictive conditions are required to achieve identification.

Conclusion 2 leads me to consider in these lectures, less and less restrictive models in sequence, with the aim of discovering what fundamental restrictions lead to identification of particular structural features.

While minimally restrictive models are of great interest because they reveal where fundamental identifying power resides, they are rarely useful on their own in econometric practice. When we come to the point of extracting information from data we have to employ more restrictive models than are required to achieve identification of the structural features of interest. For example we may employ parametric models because there is insufficient information in data to allow accurate estimation using nonparametric models. Such models may be good approximations to the economic process that generates data, but they may not be. If a structural feature can be identified in the absence of some of the restrictions embodied in the econometric model used to extract information from data then there may be scope for detecting failure of restrictions not essential for identification. From this observation flows the econometric study of misspecification.

Section 3 sets out in formal terms many of the concepts introduced in this section. I then consider the analysis of identification when equations are restricted to be linear and then relax the linearity restriction, considering nonparametrically specified models with additive latent variates and then nonseparable nonparametric models. Finally I consider additive and non-additive models in which variables exhibit discrete variation and an important class of partially identifying instrumental variable models.

3. Structures, models and identification

This Section[10] makes precise the definitions of various concepts and states and proves a Lemma which is helpful in determining whether a model identifies a structural characteristic.

Following Hurwicz (1950)[11], a *structure* is defined as:

[10] Some of this material appears in Chesher (2007b) which gives a brief review of the study of identification. Section 14 of these notes gives references to some of the significant contributions.

[11] The Cowles Commission Monograph number 10 in which Hurwicz (1950) and some other fine papers appear is available to download at the Cowles Foundation website.

1. a system of equations delivering a unique value of a vector outcome, $Y = \{Y_m\}_{m=1}^{M}$ given a value of a vector covariate, $X = \{X_k\}_{k=1}^{K}$ and a value of a vector of unobservable random variables, $\varepsilon = \{\varepsilon_r\}_{r=1}^{R}$, and,

2. a conditional distribution function, $F_{\varepsilon|X}$ for the unobservables given the covariates,

3. such that, the conditional distribution function of outcomes given covariates, $F_{Y|X}$ is well defined for all values x of X.

In this definition Hurwicz considers only *complete* structures. Complete structures are structures that satisfy condition 1, delivering a *unique* value of Y given any value of X and ε. Plenty of econometric models admit structures that are incomplete, for example models of equilibrium behaviour in which there may be multiple equilibria and models with a single structural equation and more than one endogenous outcome. At this early stage of our development we consider complete models, defined as models which admit only complete structures.

The definition of a particular structure requires a full numerical specification of a system of equations and a conditional distribution $F_{\varepsilon|X}$.

A *structural characteristic*[12] is a functional $\theta(S)$ of a structure, S, for example: the value of a parameter if there is a parametric specification, the value of a partial derivative of a structural function at a given point.

Data are generated by some structure, we know not which, and we wish to discover the value of a characteristic, or feature, of the data generating structure. Many structures with different values of a structural characteristic may generate identical conditional distribution functions, $F_{Y|X}$.[13] Structures which generate the same conditional distribution function for Y given X for all values of X are said to be *observationally equivalent*.

Data generated by a structure are informative about $F_{Y|X}$, but cannot alone distinguish one observationally equivalent structure from another. If the value of a structural characteristic varies within observationally equivalent structures then the value cannot be identified.[14] So, in order to identify the value of a structural characteristic the class of admissible structures must be restricted so that there is no variation in the value of the characteristic within observationally equivalent structures.

The term "*model*" is used to describe a set of restrictions defining admissible structures. A model is a proper subset of the class of all structures, for example all structures in which the equations are restricted to be linear and $F_{\varepsilon|X}$ is multivariate normal independent of X. Complete models admit only complete structures. We will consider incomplete models later.

A model *identifies* a characteristic, $\theta(S)$ in a structure S_0 if that characteristic is the same in all structures which are admitted by the model and observationally equivalent to S_0 (Koopmans and Reiersøl (1950)). A characteristic $\theta(S)$ is *uniformly identified* by a model if it is identifiable for every structure S admitted by the model.

Here are formal definitions.

[12] The term "structural characteristic" seems to be due to Koopmans and Reiersøl (1950). Hurwicz (1950) uses the term "criterion".

[13] $F_{Y|X}$ refers to the collection of conditional distrbutions $\{F_{Y|X=x} : x \in \mathcal{R}_X\}$ where $\mathcal{R}_X$ is the support of X.

[14] If that variation is limited to a set of values then the value of the structural characteristic can be partially, or "set' identified.

- *A model M identifies a structure $S_0 \in M$ if for all $S \in M$ with $S \neq S_0$, $F^S_{Y|X} \neq F^{S_0}_{Y|X}$. A model is uniformly identifying for the structures it admits if it identifies each structure that it admits.*
- *A model M identifies a a feature θ of a structure S_0, $\theta(S_0)$, if $\theta(S) = \theta(S_0)$ for all $S \in \{S : (S \in M) \wedge (F^S_{Y|X} = F^{S_0}_{Y|X})\}$. A model M uniformly identifies θ if it identifies $\theta(S)$ for all $S \in M$.*

It is helpful to have a simple means of determining whether a model uniformly identifies a structural characteristic. This is provided by the following Lemma.

Lemma 1. *Consider a model, let S^a be the set of admissible structures such that $\theta(S) = a$ and let A be the set of all values of $\theta(S)$ generated by admissible structures. Let $F^S_{Y|X}$ denote the conditional distribution function generated by a structure S. Suppose there exists a functional of the conditional distribution function of Y given X, $G(F_{Y|X})$, such that for each $a \in A$, $G(F^S_{Y|X}) = a$ for all $S \in S^a$. Then $\theta(S)$ is uniformly identified by the model.*

Proof of Lemma 1. *Consider any value of $a_0 \in A$ and any structure S_0 with $\theta(S_0) = a_0$ and let S^*_0 be the set of structures observationally equivalent to S_0. Consider any $S' \in S^*_0$ and let $\theta(S') = a'$. If a functional G with the stated property exists then $G(F^{S'}_{Y|X}) = a'$ and $G(F^{S_0}_{Y|X}) = a_0$. Since S' and S_0 are observationally equivalent $F^{S'}_{Y|X} = F^{S_0}_{Y|X}$ and therefore $a' = a_0$. Therefore, if a functional G with the stated property exists then, for any $a_0 \in A$, all structures observationally equivalent to any structure S_0 with $\theta(S_0) = a_0$ have the same value, a_0, of the structural characteristic, and so $\theta(S)$ is uniformly identified by the model.*

If, for some model and some structural feature, a functional with the stated property can be found then uniform identification of the structural feature by the model is assured and there is a clear route to estimation of the value of the structural feature *via* the analog principle using $\hat{\theta}(S) = \mathcal{G}(\hat{F}^S_{Y|X})$.

Properties of such estimators will depend on restrictions additional to those required to achieve identification. Accurate estimation may not be feasible unless $\mathcal{G}$ is sufficiently smooth. If small changes in $F_{Y|X}$ can lead to large changes in $\mathcal{G}(F_{Y|X})$ then convergence of estimators will be slow.

A model is *overidentifying* for a characteristic $\theta(S)$ of a structure S if there exist *distinct* functionals, $\mathcal{G}^* \neq \mathcal{G}^+$ say, with the property set out in Lemma 1. When this situation arises then, for any value of $\theta(S_0)$, say a^0, $\mathcal{G}^*(F^{S_0}_{Y|X}) = a_0 = \mathcal{G}^+(F^{S_0}_{Y|X})$. Put informally, when a model is overidentifying there are at least two distinct ways to deduce the value of $\theta(S)$ from knowledge of the distribution $F^S_{Y|X}$.

In this situation the model implies a restriction on $F^S_{Y|X}$ which may not be satisfied exactly by estimates of $F^S_{Y|X}$ unless the restriction is imposed when estimation is done, that is it can be that $\mathcal{G}^*(\hat{F}^S_{Y|X}) \neq \mathcal{G}^+(\hat{F}^S_{Y|X})$ even though $\mathcal{G}^*(F^S_{Y|X}) = \mathcal{G}^+(F^S_{Y|X})$. Efficient estimation could be achieved by imposing the restriction $\mathcal{G}^*(F^S_{Y|X}) = \mathcal{G}^+(F^S_{Y|X})$ at the point of estimation ensuring that $\mathcal{G}^*(\hat{F}^S_{Y|X}) = \mathcal{G}^+(\hat{F}^S_{Y|X})$. Alternatively, conflicting estimates $\mathcal{G}^*(\hat{F}^S_{Y|X}) \neq \mathcal{G}^+(\hat{F}^S_{Y|X})$ could be efficiently resolved using a minimum distance procedure.

If the restrictions embodied in a model are not satisfied by a structure S then it may be that $\mathcal{G}^*(F^S_{Y|X}) \neq \mathcal{G}^+(F^S_{Y|X})$. So, when $\mathcal{G}^*(\hat{F}^S_{Y|X})$ is farther from $\mathcal{G}^+(\hat{F}^S_{Y|X})$ than sampling variation

in the estimates leads one to expect, the possibility that the data generating structure does not satisfy the restrictions of the model being employed must be considered. This leads to so called tests of "overidentifying restrictions".

4. Set identification

The type of identification discussed above is *point identification*. When a structural feature is point identified by a model the value of the feature can be deduced from knowledge of the joint distribution of outcomes given covariates. Important cases arise in which a model does not point identify a structural feature but embodies restrictions sufficient to allow the joint distribution of outcomes given covariates to convey some information about the value of the feature, for example the information that it lies in a set of values.

There is the following definition of set identification. A model *set identifies* a characteristic, $\theta(S)$ in a structure S_0 if that characteristic has a value that lies in a set A_0 in all structures which are admitted by the model and observationally equivalent to S_0. Of course a model is not useful (for understanding θ) unless the set of values A_0 is a proper subset of the set of all potential values of the feature. Here is a formal definition.

- *A model M set identifies $\theta(S_0)$ to within A_0 if $\theta(S) \in A_0$ for all $S \in \{S : (S \in M) \wedge (F^S_{Y|X} = F^{S_0}_{Y|X})\}$.*

Point identification of $\theta(S_0)$ arises as a special case when the set A_0 is a singleton.

Here is a simple example of a context in which set identification arises.[15] Let Y be a random variable and let X be a binary covariate. Suppose the value of Y is only observable when $X = 1$ and that the value of X is always observable.[16] Let $g(\cdot)$ be a function. Consider the structural feature $E[g(Y)]$ which is required to exist in what follows.

The Law of Total Probability implies that

$$E[g(Y)] = E[g(Y)|X = 1]P[X = 1] + E[g(Y)|X = 0]P[X = 0]$$

in which $E[g(Y)|X = 1]$, $P[X = 1]$ and $P[X = 0]$ are point identified. Suppose a model restricts $E[g(Y)|X = 0]$ to lie in a set Γ. Then the model set identifies $E[g(Y)]$ as follows.

$$E[g(Y)] \in \{E[g(Y)|X = 1]P[X = 1] + \gamma P[X = 0] : \gamma \in \Gamma\}$$

Now let $\Gamma = [\gamma_L, \gamma_U]$, a closed interval. Then there is interval identification as follows.

$$E[g(Y)] \in [E[g(Y)|X = 1]P[X = 1] + \gamma_L P[X = 0], [E[g(Y)|X = 1]P[X = 1] + \gamma_U P[X = 0]] \tag{4.1}$$

Note that defining $g(\cdot)$ as the indicator function:

$$g(Y) \equiv 1[Y \leq y] \equiv \begin{cases} 1 & Y \leq y \\ 0 & Y > y \end{cases}$$

[15] This example and others are discussed in Manski (2003).

[16] This is remiscent of the "sample selection model" often arising in applied microeconometrics. In that context there is usually interest in the structural feature $E[g(Y)|Z = z]$ where Z is a list of covariates which also condition the probability $P[X = x|Z]$.

the result (4.1) can be used to develop bounds on distribution functions, because then $E[g(Y)] = P[Y \leq y]$ (which is assured to exist), and on inversion, bounds on quantile functions. The argument above could be carried out conditional on other observable covariates.

Hurwicz (1950) provides a definition of set identification. Manski (2003) gives many examples of problems in which set identification arises and many references. Some examples arising in structural econometrics appear later - see Chesher (2005, 2010, 2013), Chesher, Rosen and Smolinski (2013) and Chesher and Rosen (2017).

5. Misspecification

If no structure admitted by a model delivers the conditional probability distributions $F_{Y?X}$ delivered by a process then the model is misspecified for that process. In this circumstance identified sets of structural features will be empty sets. A misspecified model may become well specified if one or more of its restrictions are dropped or weakened.

6. Linear Models

First consider a simple, and very restrictive, linear model for a single outcome, Y, in which there are k covariates[17] arrayed in a k-element column vector $X = [X^1, \ldots, X^k]'$ and a single latent unobservable random variable, U. The model restricts the relationship between Y, X and U to be linear, as follows.[18]

$$Y = X'\beta + U \tag{6.1}$$

We consider alternative restrictions on the conditional distribution of U given X and ask whether the resulting models identify the value of β. It may seem obvious that they do - but it is worth working carefully through a simple and familiar case in preparation for more difficult problems.

6.1. Conditional expectation restrictions

First consider the model consisting of (6.1) and the following restriction on the *conditional* expected value of U given $X = x$

$$E[U|X = x] = 0, \quad x \in A \tag{6.2}$$

where $\mathcal{A} \subseteq \Re^k$. Clearly, for all $x \in \mathcal{A}$,

$$E[Y|X = x] = x'\beta$$

[17] By "covariates" I mean variables upon whose values we may condition.
[18] Note that Y and X are random variables, not data.

and with k values of $x \in \mathcal{A}$, $x_1, \ldots, x_k$, and

$$X_k = \begin{bmatrix} x_1' \\ \vdots \\ x_k' \end{bmatrix}$$

then

$$\bar{Y}_k = X_k \beta$$

where

$$\bar{Y}_k = \begin{bmatrix} E[Y|X = x_1] \\ \vdots \\ E[Y|X = x_k] \end{bmatrix}$$

and, if $\tilde{x}$ has full rank,

$$\beta = X_k^{-1} \bar{Y}_k.$$

Consider all structures satisfying (6.1) and (6.2) in which $\beta = b$, a particular numerical value. Note that, in the language of the Lemma of the previous section, b is a functional of a structure, that is for a structure S, $b = \theta(S)$.

We have shown that in structures admitted by the model with $\beta = b$ the functional of the conditional distribution of Y given X, $X_k^{-1}\bar{Y}_k$ is equal to b. Therefore if a full rank X_k exists then the model (6.1) and (6.2) identifies the value of β. Whether or not the set $\mathcal{A}$ contains values of X showing sufficient variation to allow identification depends on the richness of the support of X and on the satisfaction of the conditional expectation condition (6.2). When we come to consider identification using nonparametric models the issue of the support of X becomes important.

If there is more than one matrix X_k with rank k then (the value of) β is over identified. Suppose there are $n > k$ values of X for which (6.2) is satisfied and now let

$$X_n = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix}$$

which is an $n \times k$ matrix and let

$$\bar{Y}_n = \begin{bmatrix} E[Y|X = x_1] \\ \vdots \\ E[Y|X = x_n] \end{bmatrix}.$$

Then

$$\bar{Y}_n = X_n \beta$$

and for any $n \times k$ matrix $h(X_n)$ whose elements may be functions of elements of X_n

$$h(X_n)'\bar{Y}_n = h(X_n)'X_n\beta$$

and so, if the $k \times k$ matrix $h(X_n)'X_n$ has rank k,

$$\beta = \left(h(X_n)'X_n\right)^{-1} h(X_n)'\bar{Y}_n. \tag{6.3}$$

Choosing $h(X_n) = X_n$ gives

$$\beta = \left(X_n'X_n\right)^{-1} X_n'\bar{Y}_n \tag{6.4}$$

and replacing $\bar{Y}_n$ by $y = [y_1, \ldots, y_n]'$, a vector of realisations of $Y = [Y_1, \ldots, Y_n]'$, the value y_i being obtained when $X = x_i$, leads to the Ordinary Least Squares estimator of (the value of) β.

Choosing $h(X_n) = \Omega_n^{-1}X_n$ leads to:

$$\beta = \left(X_n'\Omega_n^{-1}X_n\right)^{-1} X_n'\Omega_n^{-1}\bar{Y}_n$$

and replacing $\bar{Y}_n$ by $y = [y_1, \ldots, y_n]'$ leads to a Generalised Least Squares estimator and then we might interpret Ω_n as the conditional covariance matrix of Y given $X = x_1, \ldots, X = x_n$.

When β is overidentified data can be informative about the correctness of the model (6.1) and (6.2) for the data generating structure. For example we could investigate the correctness of the linear specification, but *only* while maintaining the covariation restriction (6.2).

Suppose that X_n has rank less than k; this is the "perfect collinearity" situation described in elementary econometrics textbooks. Then β cannot be point identified, but set identification may be possible. For example if two elements of x, say x^1 and x^2 (with coefficients β_1 and β_2) are such that $x^2 = \alpha x^1$ then $\beta_1 + \alpha\beta_2$ (and other elements of β) may be point identified but β_1 and β_2 can only be set identified with

$$(\beta_1, \beta_2) \in \{(b_1, b_2) : b_1 + \alpha b_2 = c\}$$

where c is the point identified value of $\beta_1 + \alpha\beta_2$.

6.2. Unconditional expectation restrictions

The model set out in the previous Section has the implication

$$E[Y|X = x] = x'\beta$$

for $x \in \mathcal{A}$, some specified set.

Now suppose this conditional expectation restriction holds for all $x \in \mathcal{R}_X$ which is the support of X, that is $\mathcal{A} = \mathcal{R}_X$. For any k-element vector valued function $h(X)$

$$E[h(X)Y|X = x] = h(x)x'\beta$$

for all $x \in \mathcal{R}_X$, and taking expectations with respect to X (supposing that the required expectations exist)

$$E[h(X)Y] = E[h(X)X']\beta$$

and if $E[h(X)X']$ has full rank (k) there is the following identifying correspondence.

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

Note that this correspondence comes from a "zero correlation" restriction $E[h(X)U] = 0$ which is implied[19] by $E[U|X = x] = 0$ for all $x \in \mathcal{R}_X$ but is weaker, for example allowing the possibility that $E[U|X = x] = s(x)$ for some function $s(x)$ which is not necessarily zero as long as $E[h(X)s(X)] = 0$.

The Ordinary Least Squares (OLS) estimator can be motivated as an analogue estimator developed from an identifying correspondence built on an unconditional moment restriction. Consider the model

$$Y = X'\beta + U \quad E[XU] = 0, \quad \text{rank}\,(E[XX']) = k$$

which implies

$$\beta = E[XX']^{-1}E[XY]$$

and the OLS estimator arises on replacing the expected values here by sample averages.

$$\hat{\beta} = (n^{-1}X_n'X_n)^{-1}(n^{-1}X_n'Y_n)$$

Sampling properties of the estimator depend on additional (non-identifying) restrictions.

6.3. Identification via extremum problems

Consider b_{opt} defined as

$$b_{opt}(\tilde{y}, \tilde{x}, \Omega) \equiv \arg\min_b (\tilde{y} - \tilde{x}b)'\Omega^{-1}(\tilde{y} - \tilde{x}b)$$

where Ω is a $n \times n$ positive definite matrix. Suppose that X_n has rank k. Then the $k \times k$ matrix $X_n'\Omega^{-1}X_n$ also has rank k so its inverse exists. Define

$$\bar{b} \equiv \left(X_n{}'\Omega_n^{-1}X_n\right)^{-1} X_n{}'\Omega_n^{-1}\bar{Y}_n.$$

It is now shown that $b_{opt}(\bar{Y}_n, X_n, \Omega_n) = \bar{b}$.

Define

$$S(b) \equiv (\bar{Y}_n - X_n b)'\Omega^{-1}(\bar{Y}_n - X_n b)$$

and rewrite as follows.

$$S(b) = ((\bar{Y}_n - X_n\bar{b}) + X_n(\bar{b} - b))'\Omega^{-1}((\bar{Y}_n - X_n\bar{b}) + X_n(\bar{b} - b))$$

Expanding there is the following expression.

$$\begin{aligned} S(b) &= \left(\bar{Y}_n - X_n\bar{b}\right)'\Omega^{-1}\left(\bar{Y}_n - X_n\bar{b}\right) \\ &\quad +2(\bar{b} - b)'X_n'\Omega^{-1}\left(\bar{Y}_n - X_n\bar{b}\right) \\ &\quad +(\bar{b} - b)'X_n'\Omega^{-1}X_n(\bar{b} - b) \end{aligned}$$

The first term does not involve b and so may be neglected in determining b_{opt}. The second term is zero because

$$\begin{aligned} X_n'\Omega^{-1}\left(\bar{Y}_n - X_n\bar{b}\right) &= X_n'\Omega^{-1}\left(I - X_n\left(X_n'\Omega^{-1}X_n\right)^{-1}X_n'\Omega^{-1}\right)\bar{Y}_n \\ &= X_n'\Omega^{-1}\bar{Y}_n - X_n'\Omega^{-1}\bar{Y}_n \end{aligned}$$

[19] Supposing that the required expectations exist.

so that term is also irrelevant to the determination of b_{opt}. Thus there is

$$b_{opt}(\bar{Y}_n, X_n, \Omega_n) \equiv \arg\min_b (\bar{b} - b)' X_n' \Omega^{-1} X_n (\bar{b} - b)$$

and clearly the solution is $b_{opt} = \bar{b}$. This is the expression (6.3) above with $h(X_n) \equiv \Omega^{-1} X_n$.

We showed that gave an expression for the parameter β in terms of features of the conditional distribution of outcomes given covariates. We have now seen how the parameter β can be identified as the (unique) solution to an extremum problem involving such features of distributions.

6.4. Maximum likelihood

The maximum likelihood estimator can be regarded as an analogue estimator based on an identifying correspondence involving an extremum condition.

Consider a parametric econometric model which requires that the probability distribution of a random variable Y is entirely determined by the value of a vector of parameters θ. Specifically, consider the case in which the model states that a discrete random variable Y with support $\mathcal{R}_Y$ has some particular proper probability mass function $p(y;\theta)$ for some value $\theta_0 \in \Theta$ where

$$P[Y = y] = p(y;\theta), \quad y \in \mathcal{R}_Y$$

and $p(y;\theta) \neq 0$ is twice differentiable with respect to θ for all $y \in \mathcal{R}_Y$ and$\mathcal{R}_Y$ does not depend on θ.

For every value of θ probabilities add to one across $\mathcal{R}_Y$ so there is, for any value θ

$$\sum_{y \in \mathcal{R}_Y} p(y;\theta) = 1$$

and on differentiating

$$\sum_{y \in \mathcal{R}_Y} \nabla_\theta p(y;\theta) = 0$$

where "∇_θ" takes first partial derivative with respect to θ.

It follows that θ_0 satisfies

$$\sum_{y \in \mathcal{R}_Y} p(y;\theta_0) \nabla_\theta \log p(y;\theta_0) = 0 \tag{6.5}$$

that is, the expected value of $\nabla_\theta \log p(Y;\theta_0)$ is zero when Y has a distribution given by $p(y;\theta)$ with $\theta = \theta_0$.

$$E[\nabla_\theta \log p(Y;\theta_0) | Y \sim p(y;\theta_0)] = 0$$

If θ_0 is the *only* value of θ such that this expected value is zero, that is

$$\{\theta_0\} = \{\theta : E[\nabla_\theta \log p(Y;\theta) | Y \sim p(y;\theta_0)] = 0\}$$

then this is an identifying correspondence. If the set is not a singleton then the model is partially identifying.

When $Y \sim p(y;\theta_0)$ the function

$$L(Y,\theta) \equiv \log p(Y;\theta)$$

has expected value

$$E[L(Y,\theta)|Y \sim p(y;\theta_0)] = \sum_{y\in\mathcal{R}_Y} p(y;\theta_0)\log p(y;\theta)$$

and

$$\begin{aligned}\nabla_\theta E[L(Y,\theta)|Y \sim p(y;\theta_0)] &= \sum_{y\in\mathcal{R}_Y} p(y;\theta_0)\nabla_\theta \log p(y;\theta) \\ &\quad \sum_{y\in\mathcal{R}_Y} p(y;\theta_0)\left(\frac{1}{p(y;\theta)}\nabla_\theta p(y;\theta)\right).\end{aligned}$$

If the second derivative matrix

$$\begin{aligned}\nabla_{\theta\theta'} E[L(Y,\theta)|Y \sim p(y;\theta_0)] &= \sum_{y\in\mathcal{R}_Y} p(y;\theta_0)\nabla_{\theta\theta'} \log p(y;\theta) \\ &= \sum_{y\in\mathcal{R}_Y} p(y;\theta_0)\left(\frac{1}{p(y;\theta)}\nabla_{\theta\theta'} p(y;\theta) - \left(\frac{1}{p(y;\theta)}\nabla_\theta p(y;\theta)\right)\left(\frac{1}{p(y;\theta)}\nabla_\theta p(y;\theta)\right)'\right)\end{aligned}$$

is negative definite at $\theta = \theta_0$ then $\theta = \theta_0$ is the unique maximiser of the expected value of the log likelihood function and there is the following identifying correspondence.

$$\{\theta_0\} = \{\theta : \arg\max_{\theta^*} E[L(Y,\theta^*)|Y \sim p(y;\theta_0)] = \theta\}$$

The maximum likelihood estimator then arises as an analogue estimator, maximising the sample average of $L(Y,\theta)$, that is the sample *log likelihood function.*

$$\hat\theta_{ML} = \arg\max_\theta \left(\frac{1}{n}\sum_{i=1}^n L(Y_i,\theta)\right) = \frac{1}{n}\sum_{i=1}^n \log p(Y_i;\theta)$$

On the negative definiteness of the Hessian here, since

$$\sum_{y\in\mathcal{R}_Y} p(y;\theta_0)\left(\frac{1}{p(y;\theta)}\nabla_{\theta\theta'} p(y;\theta)\right)\Bigg|_{\theta=\theta_0} = \sum_{y\in\mathcal{R}_Y} \nabla_{\theta\theta'} p(y;\theta)|_{\theta=\theta_0} = 0$$

there is

$$\nabla_{\theta\theta'} E[L(Y,\theta)|Y \sim p(y;\theta_0)]|_{\theta=\theta_0} = -\sum_{y\in\mathcal{R}_Y} p(y;\theta_0)\left(\nabla_\theta \log p(y;\theta)|_{\theta=\theta_0}\right)\left(\nabla_\theta \log p(y;\theta)|_{\theta=\theta_0}\right)'$$

which is negative definite provided $\nabla_\theta \log p(Y;\theta)|_{\theta=\theta_0}$ has a full rank covariance matrix.

6.5. Instrumental variables

Consider the model with

$$Y = X'\beta + U \qquad E[ZU] = 0$$

where Z is a list of r variables, $Z = [Z^1, \ldots, Z^r]$. If $E[ZX']$ exists there is

$$E[ZY] = E[ZX']\beta$$

which identifies β if $E[ZX']$ has rank k. If $r = k$ there is

$$\beta = E[ZX']^{-1}E[ZY]$$

which motivates the simple Instrumental Variable (IV) estimator.

For $r \geq k$ and if $E[ZZ']$ has full rank then

$$E[ZZ']^{-1}E[ZY] = E[ZZ']^{-1}E[ZX']\beta$$

and

$$E[XZ']E[ZZ']^{-1}E[ZY] = E[XZ']E[ZZ']^{-1}E[ZX']\beta$$

leading to

$$\beta = \left(E[XZ']E[ZZ']^{-1}E[ZX']\right)^{-1} E[XZ']E[ZZ']^{-1}E[ZY]$$

which delivers the Generalised Instrumental Variable Estimator (GIVE) as an analogue estimator on replacing expected values of moment arrays (e.g. $E[XZ']$) by sample averages of squares and cross-products of realisations of random variables (e.g. $n^{-1}X_n'Z_n$).

6.6. Conditional quantile and other restrictions

Identification of β could be obtained by using different restrictions on the covariation of observed explanatory and latent variables. First consider restrictions on conditional quantiles.

We must first define the quantiles of a random variable. For $\tau \in (0, 1)$ the τ-quantile of a scalar random variable, A, with distribution function F_A is defined as follows[20],

$$Q_A(\tau) = \inf\{q \in \mathbb{R} | F_A(q) \geq \tau\}$$

and note that such quantiles are equivariant with respect to monotone transformations, that is, if h is a non-decreasing function on $\mathcal{R}_Y$ then

$$Q_{h(A)}(\tau) = h(Q_A(\tau)).$$

This τ-quantile is well defined whenever A has a proper distribution function, including cases in which A is a discrete random variable and the equivariance property applies in such cases.

For a continuous random variable, A, $Q_A(\tau)$ is the unique solution to

$$\tau = F_A(Q_A(\tau)).$$

[20] The distribution function is defined as: $F_A(a) = P[A \leq a]$.

The conditional τ-quantile of A given a vector of covariates $B = b$ is analogously defined as

$$Q_{A|B}(\tau|b) = \inf\{q \in \mathcal{R}_Y | F_{A|B}(q|b) \geq \tau\}$$

where $F_{A|B}$ is the conditional distribution function[21] of A given $B = b$, and the equivariance property

$$Q_{h(A,B)|B}(\tau|b) = h(Q_{A|B}(\tau|b), b)$$

applies for all b for which $h(a, b)$ is a nondecreasing function of a.

For a continuous random variable A, the conditional τ-quantile of A given $B = b$ is the unique solution to

$$\tau = F_{A|B}(Q_{A|B}(\tau, b)|b).$$

Now suppose that, instead of (6.2) there is the conditional *median* restriction

$$Q_{U|X}(0.5|x) = 0, \quad x \in \mathcal{A} \tag{6.6}$$

then, by the equivariance property of quantiles which implies

$$Q_{Y|X}(0.5|x) = x'\beta + Q_{U|X}(0.5|x)$$

we have

$$Q_{Y|X}(0.5|x) = x'\beta$$

and if there are k values of vector $x \in \mathcal{A}$, $x_1, \ldots, x_k$ such that

$$X_k = \begin{bmatrix} x_1' \\ \vdots \\ x_k' \end{bmatrix}$$

has rank k then

$$Q_Y(X_k) = X_k\beta$$

where

$$Q_Y(X_k) \equiv \begin{bmatrix} Q_{Y|X}(0.5|x_1) \\ \vdots \\ Q_{Y|X}(0.5|x_k) \end{bmatrix}$$

and

$$\beta = X_k^{-1} Q_Y(\tilde{x}).$$

Arguing as before, if a full rank X_k exists then the model (6.1) and (6.6) identifies the value of β.

Other covariation restrictions can produce models that identify β. For example suppose ε given $X = x$ is continuously distributed with a unimodal density function and let $M_{\varepsilon|X}(x)$ be the location of the mode of the conditional density function of U given $X = x$. Since the location of the mode of a density is equivariant under affine transformation,

$$M_{Y|X}(x) = x'\beta + M_{U|X}(x)$$

[21] That is $F_{A|B}(a|b) \equiv P[A \leq a|B = b]$.

the restriction

$$M_{U|X}(x) = 0, \quad x \in A$$

in place of (6.2) will, with sufficient values of x in $\mathcal{A}$ satisfying the conditional mode restriction, lead to identification of the value of β, following the argument given above.

7. Linear simultaneous equations models

Now let Y be a m-vector of outcomes, $Y = [Y^1, \ldots, Y^m]'$ with X defined as before and consider *complete* models which restrict Y to be a unique solution to the simultaneous equations

$$Y' = Y'\Gamma + X'B + U'$$

where $U' = [U^1, \ldots, U^m]$ is a m-vector of unobserved latent random variables.

The requirement that the solution for Y be unique implies that $I_m - \Gamma$ is nonsingular and Y can be expressed as

$$Y' = X'\Pi + V'$$

where $\Pi = B(I_m - \Gamma)^{-1}$ is $k \times m$ and $V' = U'(I_m - \Gamma)^{-1}$.

Consider the restriction $E[XU'] = 0$ (this is a $k \times m$ matrix of zeros) which implies that $E[XV'] = 0$ and so

$$E[XY'] = E[XX']\Pi \tag{7.1}$$

and if $E[XX']$ has rank k then $E[XX']^{-1}$ exists and there is the identifying correspondence

$$\Pi = E[XX']^{-1}E[XY']$$

and the model identifies the matrix of reduced form coefficients, Π.

Since Π is identifiable, elements of Γ and B are identifiable if they can be deduced from Π. The model implies no restriction involving Γ and B other than (7.1) so this is the only route to identification of the parameters of the structural equations.

Values of elements of Γ and B cannot be deduced from knowledge of the values of the elements of Π without further restrictions since Π has $k \times m$ elements but Γ and B have in total $m^2 + km$ elements. In Γ, m elements can be fixed by normalisation, for example the m leading diagonal elements can be set equal to zero leaving $m(m-1) + km$ free elements in Γ and B. This suggests that if there are $m(m-1)$ restrictions on the elements of Γ and B then the remaining elements may be identifiable.

We have

$$\Pi(I_m - \Gamma) = B$$

and considering the coefficients in one (the ith) of the simultaneous equations we have

$$\Pi\delta_i = \beta_i$$

where δ_i is the ith column of $I_m - \Gamma$ and β_i is the ith column of B. Consider just *exclusion* restrictions which require certain elements of Γ and B to be zero and arrange δ_i and β_i so that restricted elements appear first and rearrange the elements of Π accordingly. Then

$$\begin{bmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{21} & \Pi_{22} \end{bmatrix} \begin{bmatrix} \delta_i^1 \\ e_i \end{bmatrix} = \begin{bmatrix} \beta_i^1 \\ 0 \end{bmatrix}$$

where e_i is a vector with a 1 followed by zeros and δ_i^1 and β_i^1 contain the unrestricted elements of δ_i and β_i, m_i^1 and k_i^1 element vectors respectively. We have

$$\Pi_{21}\delta_i^1 = -\Pi_{22}e_i$$

which can be solved for δ_i^1 if Π_{21} has rank m_i^1 and the equation

$$\Pi_{11}\delta_i^1 + \Pi_{12}e_i = \beta_i^1$$

will then yield a value for β_i^1.

The *rank condition* on Π_{21} can only be satisfied if the number of rows of Π_{21} is at least equal to m_i^1. The number of rows in Π_{21} is equal to the number of restricted elements in β_i, so we can conclude that, if there are at least as many covariates excluded from the ith equation as there are outcome variables with unrestricted coefficients in the ith equation, then the unrestricted elements in the ith equation are identified. Because this necessary condition relates to the row order of a matrix it called an *order* condition.

As an example, in the restricted two equation model

$$\begin{aligned} Y^1 &= \gamma_{21}Y^2 + \beta_{11}X^1 + \beta_{21}X^2 + U^1 \\ Y^2 &= \gamma_{12}Y^1 + \beta_{12}X^1 + U^2 \end{aligned}$$

and with the restriction on the conditional expectation of U given X, the parameters γ_{12} and β_{12} are identified as long as β_{21} is not zero (for if it were the rank condition would not hold) but the remaining parameters are not. Clearly the order condition is satisfied because there is one covariate excluded from the second equation which is equal to the number of outcome variables in that equation with unrestricted coefficients. The condition $\beta_{21} \neq 0$ arises from the rank condition. To see this note that

$$I_2 - \Gamma = \begin{bmatrix} 1 & -\gamma_{12} \\ -\gamma_{21} & 1 \end{bmatrix}$$

and so

$$(I_2 - \Gamma)^{-1} = \frac{1}{1-\gamma_{12}\gamma_{21}} \begin{bmatrix} 1 & \gamma_{12} \\ \gamma_{21} & 1 \end{bmatrix}$$

giving, with

$$B = \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & 0 \end{bmatrix}$$

the equation

$$\Pi = \frac{1}{1-\gamma_{12}\gamma_{21}} \begin{bmatrix} \beta_{11}+\beta_{12}\gamma_{21} & \beta_{11}\gamma_{12}+\beta_{12} \\ \beta_{21} & \beta_{21}\gamma_{12} \end{bmatrix}.$$

This is already arranged appropriately for consideration of the restrictions imposed on the second equation, that is:

$$\frac{1}{1-\gamma_{12}\gamma_{21}} \begin{bmatrix} \beta_{11}+\beta_{12}\gamma_{21} & \beta_{11}\gamma_{12}+\beta_{12} \\ \beta_{21} & \beta_{21}\gamma_{12} \end{bmatrix} \begin{bmatrix} -\gamma_{12} \\ 1 \end{bmatrix} = \begin{bmatrix} \beta_{12} \\ 0 \end{bmatrix}. \tag{7.2}$$

Now consider what happens if we know the elements of the matrix Π thus

$$\Pi = \begin{bmatrix} \pi^*_{11} & \pi^*_{12} \\ \pi^*_{21} & \pi^*_{22} \end{bmatrix}$$

then we have

$$-\gamma_{12}\pi^*_{21} + \pi^*_{22} = 0$$

which yields

$$\gamma_{12} = \frac{\pi^*_{22}}{\pi^*_{21}}$$

as long as $\pi^*_{21} \neq 0$, but, looking back at Π written in terms of the parameters of interest (7.2) we see that π^*_{21} will be zero in structures having $\beta_{21} = 0$. In such structures the values of γ_{12} and β_{12} are not identified[22].

Applying this to the wage-schooling example introduced earlier[23] we see that in the linear model (2.1) and (2.2) with the conditional expectation restriction (6.2) the values of the wage equation's parameters are identified if $\alpha_2 = 0$ and $\beta_1 \neq 0$, that is if the covariate X does not feature in the wage equation but does affect the choice of schooling. Covariates with this property are sometimes called *instrumental variables*.

These results and others on identification in linear simultaneous equations models were given a full and formal exposition in Koopmans, Rubin and Leipnik (1950).

8. Nonlinear additively separable models

Economics rarely provides detailed information about the functional forms of economic relationships and we do not want the crucial issue of identification to hang on a slender thread. So it is interesting to ask what conditions are required for identification of features of nonparametrically specified structural functions.

We start by considering models like (2.3) and (2.4) but with the restriction that the latent variables appear *additively*. That takes us one small step from the linear simultaneous equations system studied in the previous Section and prepares us for the study of identification in nonseparable systems studied in the Section 10.

The model considered now has the following structural equations.

$$\begin{aligned} W &= h_1(S, X) + \varepsilon_1 \\ S &= h_2(X) + \varepsilon_2 \end{aligned}$$

Here W and S are observable scalar outcomes and, for the sake of simplifying the exposition, X is a *scalar* observable covariate. The variates ε_1 and ε_2 are not observable. Note that ε_2 is excluded from the first structural equation in this specification, but ε_1 and ε_2 will be allowed to be jointly dependent.

[22] Try carrying out this analysis for the first equations's coefficients adding the further restriction $\beta_{11} = 0$.

[23] Think of U_1 as $\varepsilon_1 + \lambda\varepsilon_2$, and of U_2 as ε_2. The elements of Y are W and S and there is a single covariate so $k = 1$.

Now add to the model the following covariation restrictions.

$$\begin{aligned} E_{\varepsilon_1|\varepsilon_2 X}[\varepsilon_1|e_2, x] &= E_{\varepsilon_1|\varepsilon_2}[\varepsilon_1|e_2] \\ E_{\varepsilon_2|X}[\varepsilon_2|x] &= E_{\varepsilon_2}[\varepsilon_2] \end{aligned}$$

For the moment suppose these restrictions hold for all values of e_2 and x. Of course these expectations are required to exist.

Take the expectation of W conditional on $\varepsilon_2 = e_2$ and $X = x$, and of S conditional on $X = x$, giving, with the covariation restrictions imposed:

$$\begin{aligned} E_{W|\varepsilon_2 X}[W|e_2, x] &= h_1(h_2(x) + e_2, x) + E_{\varepsilon_1|\varepsilon_2}[\varepsilon_1|e_2] \\ E_{S|X}[S|x] &= h_2(x) + E_{\varepsilon_2}[\varepsilon_2] \end{aligned}$$

and then, noting that conditioning on $\varepsilon_2 = e_2$ and $X = x$ is identical to conditioning on $S = s$, defined as

$$s \equiv h_2(x) + e_2$$

and $X = x$, rewrite so that on the left hand sides of the equations we have expectations conditional on $S = s$ and $X = x$ as follows.

$$E_{W|SX}[W|s, x] = h_1(s, x) + E_{\varepsilon_1|\varepsilon_2}[\varepsilon_1|e_2 = s - h_2(x)] \tag{8.1}$$

$$E_{S|X}[S|x] = h_2(x) + E_{\varepsilon_2}[\varepsilon_2] \tag{8.2}$$

Data are informative about the expectations (8.1) and (8.2) which are well defined functionals of the conditional distribution function of (W, S) given X. The identifiability of features of the structural functions hangs on whether knowledge of those features can be deduced from knowledge of $E_{W|SX}[W|s, x]$ and $E_{S|X}[S|x]$ at the values of x that are available.

We proceed focussing on identification of the *partial derivatives* of the structural function h_1. Were the structural equations to be linear in variables these would be the parameters whose identification was studied in the previous Section.

Define[24]

$$\gamma(s, x) \equiv \nabla_{e_2} E_{\varepsilon_1|\varepsilon_2}[\varepsilon_1|e_2]|_{e_2 = s - h_2(x)}.$$

Taking partial derivatives with respect to s and x in (8.1) and (8.2) gives

$$\nabla_s E_{W|SX}[W|s, x] = \nabla_s h_1(s, x) + \gamma(s, x) \tag{8.3}$$

$$\nabla_x E_{W|SX}[W|s, x] = \nabla_x h_1(s, x) - \nabla_x h_2(x)\gamma(s, x) \tag{8.4}$$

$$\nabla_x E_{S|X}[S|x] = \nabla_x h_2(x) \tag{8.5}$$

Of course the structural functions must be restricted so that the partial derivatives that appear here exist.

If x exhibits continuous variation then data are informative about the values of the derivatives on the left hand sides of these equations.[25] Under what additional restrictions can the values of

[24] ∇_{e_2} denotes partial derivative with respect to e_2.

[25] Note that continuous variation is necessary for identification of structural partial *derivatives* in the absence of further (e.g. parametric) restrictions on the structural functions. However with only discrete variation it may be possible to identify structural partial *differences*, a point taken up in later.

the structural functions' derivatives on the right hand sides of these equations be deduced from knowledge of those left hand side partial derivatives?

A moments thought shows that an additional restriction is required because these three linear equations (in partial derivatives) involve four unknowns, $\nabla_s h_1$, $\nabla_x h_1$, $\nabla_x h_2$ and γ. One restriction which resolves this difficulty is $\gamma(s,x) = 0$. In this case, S is "locally exogenous" because ε_1 and ε_2 do not covary (in a sense) at $\varepsilon_2 = s - h(x)$. In that case the required partial derivatives are identified directly by the partial derivatives of the conditional expectations. When $\gamma(s,x) \neq 0$ we need an alternative restriction to achieve identification.

Suppose there is the restriction: $\nabla_x h_1(s,x) = 0$ which limits the impact of variation in X on h_1 at (x,s). With this restriction the equations simplify to

$$\begin{aligned} \nabla_s E_{W|SX}[W|s,x] &= \nabla_s h_1(s,x) + \gamma(s,x) \\ \nabla_x E_{W|SX}[W|s,x] &= -\nabla_x h_2(x)\gamma(s,x) \\ \nabla_x E_{S|X}[S|x] &= \nabla_x h_2(x) \end{aligned}$$

from which, as long as $\nabla_x h_2(x) \neq 0$ (that is, X *does* affect h_2 at $X = x$),

$$\gamma(s,x) = -\frac{\nabla_x E_{W|SX}[W|s,x]}{\nabla_x E_{S|X}[S|x]}$$

and

$$\nabla_s h_1(s,x) = \nabla_s E_{W|SX}[W|s,x] + \frac{\nabla_x E_{W|SX}[W|s,x]}{\nabla_x E_{S|X}[S|x]}.$$

This demonstrates the identifiability of the partial derivative $\nabla_s h_1(s,x)$ under the restrictions imposed. Of course $\nabla_x h_1(s,x)$ is trivially identified since the model now embodies the restriction $\nabla_x h_1(s,x) = 0$. Consider the following points.

1. The argument above could be local to one or more particular values of X. Then one would be considering identification of values of structural partial derivatives at particular values of arguments of the structural functions.

2. If local identification of that sort is of interest then one could achieve identification under weaker covariation restrictions of the form, e.g., $\nabla_x E_{\varepsilon_1|\varepsilon_2 X}[\varepsilon_1|e_2,x] = 0$ at the value of x of interest, with $e_2 = s - h_2(x)$. Similarly we only require $\nabla_x h_1(s,x) = 0$ at values of s and x of interest.

3. The expression for $\nabla_s h_1(s,x)$ is analogous to the expression obtained in a similarly restricted linear simultaneous equations model. In that model, taking this formula and replacing derivatives of conditional expectations by OLS estimates of coefficients in linear regressions of W on S and X and of S on X produces the Indirect Least Squares estimator. You might care to check that this is indeed the case.

4. Similar formulae apply if more than one covariate X appears in the structural function for S and is excluded from the structural function h_1. Then there is overidentification (subject to $\nabla_{x_i} h_2(x) \neq 0$ for each X_i). A minimum distance procedure could be used to efficiently combine the competing estimates that could be produced in this case.

5. The argument above applies when h_1 and/or h_2 are restricted to parametric families. In that case partial derivatives above are known up to a finite number of parameters.

6. Equation (8.1) suggests analogue estimation in which W is "regressed" on $h_1(s, x)$ and $g(s - \hat{h}_2(x))$ where $\hat{h}_2$ is an estimate of h_2. This can be done in parametric and nonparametric contexts.[26] The function g might be parametrically or nonparametrically specified. This is sometimes called the "control function" approach to estimation in the presence of endogeneity. Contrast this with the "fitted value" approach to estimation in the presence of endogeneity in which W is "regressed" on $h_1(\hat{h}_2(x), x)$. The "fitted value" approach has an analogue estimation interpretation when h_1 is a linear function of s, but otherwise does not generally produce a consistent estimator.

9. Marginal covariation restrictions and completeness conditions

In the model used in the previous Section an endogenous variable appeared embedded in a nonlinear function. The model employed an *iterated* covariation condition, as follows.

$$E_{\varepsilon_1|X}[\varepsilon_1|x] = E_{\varepsilon_1}[\varepsilon_1] = c_1 \tag{9.1}$$
$$E_{\varepsilon_2|X}[\varepsilon_2|x] = E_{\varepsilon_2}[\varepsilon_2] = c_2 \tag{9.2}$$

The iterated condition leads to pointwise identification of the structural function and points to relatively simple analogue estimation procedures. The marginal condition can be more awkward to employ in practice. Neither set of conditions implies the other but if (9.1) and (9.2) are replaced by "ε_2 and X are independently distributed" the iterated condition implies the marginal condition.

To study the use of the marginal condition in a simple context suppose that in the example of the previous section X does not appear in the structural function h_1.

$$\begin{aligned} W &= h_1(S) + \varepsilon_1 \\ S &= h_2(X) + \varepsilon_2 \end{aligned}$$

Under the condition (9.1) there is

$$\forall x: \quad E[W|x] = \int h_1(s) f_{S|X}(s|x) dx + c_1 \tag{9.3}$$

in which the conditional density of S given X, $f_{S|X}$, is identified and the conditional expectation, $E[W|x]$, is identified. The second equation $S = h_2(X) + \varepsilon_2$ is irrelevant.

The function h_1 is identified (up to a location shift) if there are conditions on h_1 and on the distribution of S given X and the distribution of X sufficient to ensure a unique solution to the integral equation (9.3) with c_1 normalised to say zero. This requires that the support of X be at least as rich as the support of S, and in particular that X be continuously distributed if S is continuous.

[26] That is h_1 and/or h_2 and/or g can be parametrically or nonparametrically specified.

When S has a discrete distribution with M points of support $s_1, \ldots, s_M$ and X has K points of support $x_1, \ldots, x_K$ and with c_1 normalised to zero there is:

$$\begin{bmatrix} E[W|x_1] \\ \vdots \\ E[W|x_K] \end{bmatrix} = \begin{bmatrix} P[S=s_1|X=x_1] & \cdots & P[S=s_M|X=x_1] \\ \vdots & \ddots & \vdots \\ P[S=s_1|X=x_K] & \cdots & P[S=s_M|X=x_K] \end{bmatrix} \begin{bmatrix} h_1(s_1) \\ \vdots \\ h_1(s_M) \end{bmatrix}$$

which can be solved uniquely for the unknowns: $h_1(s_1), \ldots, h_1(s_M)$ when the right-hand side matrix has rank M. Clearly $K \geq M$ is required. When $K > M$ there may be overidentification.

Conditions under which there is a unique solution to (9.3) are known as *completeness conditions*. Here is one such condition. Suppose the probability distribution of S and X is such that the following condition holds - here the integral is definite over the support of S given $X = x$.

$$\forall x: \quad \int r(s) f_{S|X}(s|x) dx = 0 \implies \forall s: \quad r(s) = 0 \tag{9.4}$$

If the distribution of S and X is such that the condition (9.4) holds then for any c_1 there is a unique function $h_1(\cdot)$ satisfying (9.3).

10. Nonadditive models

Now consider identification of structural partial derivatives in the following model in which unobservables may not be additively separable.

$$W = h_1(S, X, \varepsilon_1) \tag{10.1}$$
$$S = h_2(X, \varepsilon_2) \tag{10.2}$$

The structural functions h_1 and h_2 are restricted to be strictly monotonically varying in respectively ε_1 and ε_2. Note that this restriction was satisfied in the model of the previous Section. The functions are normalised to be increasing in these variables.

Restrictions on conditional *expectations* are not helpful now that the unobservable variables appear embedded in what may be nonlinear functions. As earlier we focus on conditional covariation restrictions and now consider the identifying power of quantile restrictions.

Recall the equivariance under monotone transformation property of quantiles. Let x_i denote an element of x. With the monotonicity conditions imposed in this Section we have, considering the S equation:

$$Q_{S|X}(\tau_s|x) = h_2(x, Q_{\varepsilon_2|X}(\tau_s|x))$$

and with the weak covariation restriction: $\nabla_{x_i} Q_{\varepsilon_2|X}(\tau_s|x) = 0$ there is

$$\nabla_{x_i} Q_{S|X}(\tau_s|x) = \nabla_{x_i} h_2(x, Q_{\varepsilon_2|X}(\tau_s|x))$$

which serves to identify the x_i-partial derivative of h_2 at $X = x$, $\varepsilon_2 = Q_{\varepsilon_2|X}(\tau_s|x)$, subject of course to there being continuous variation in x_i. See Matzkin (2003). The covariation restriction could be local to a particular value of x in which case there is local identification of the partial derivative at (x, e_2).

With this idea to hand we can proceed much as in the previous Section. There is the following.

$$\begin{aligned} Q_{W|\varepsilon_2 X}(\tau_w|e_2,x) &= h_1(h_2(x,e_2),x,Q_{\varepsilon_1|\varepsilon_2 X}(\tau_w|e_2,x)) && (10.3)\\ Q_{S|X}(\tau_s|x) &= h_2(x,Q_{\varepsilon_2|X}(\tau_s|x)) && (10.4)\end{aligned}$$

Set

$$\begin{aligned} e_2 &\equiv Q_{\varepsilon_2|X}(\tau_s|x)\\ s &\equiv Q_{S|X}(\tau_s|x)\end{aligned}$$

and note that $s = h_2(x.e_2)$, and therefore $e_2 = h_2^{-1}(x,s)$ where h_2^{-1} is the inverse function satisfying[27]

$$a = h_2(x, h_2^{-1}(x,a)).$$

In (10.3) and (10.4), on switching conditioning from (ε_2, X) to (S, X), there is the following.

$$\begin{aligned} Q_{W|SX}(\tau_w|s,x) &= h_1(s,x,Q_{\varepsilon_1|\varepsilon_2 X}(\tau_w|h_2^{-1}(x,s),x)) && (10.5)\\ Q_{S|X}(\tau_s|x) &= h_2(x,Q_{\varepsilon_2|X}(\tau_s|x)) && (10.6)\end{aligned}$$

Manipulating these two results it can be shown that the following correspondence identifies the partial derivative of the structural function.

$$\nabla_s h_1(s,x,e_1) = \nabla_s Q_{W|SX}(\tau_w|s,x) + \frac{\nabla_{x_i} Q_{W|SX}(\tau_w|s,x)}{\nabla_{x_i} Q_{S|X}(\tau_s|x)} \qquad (10.7)$$

Chesher (2003) gives the following derivation of (10.7) involving manipulation of inverse functions. The derivation given in Section 10.2 is easier to follow.

10.1. A derivation using inverse functions

Define:

$$e_1 \equiv Q_{\varepsilon_1|\varepsilon_2 X}(\tau_w|e_2,x)$$

and consider partial x_i-derivatives of the identifiable conditional quantile functions $Q_{W|SX}(\tau_w|s,x)$ and $Q_{S|X}(\tau_s|x)$. Impose the covariation restrictions

$$\begin{aligned} \nabla_{x_i} Q_{\varepsilon_1|\varepsilon_2 X}(\tau_w|\bar{e}_2,x))|_{\bar{e}_2=e_2} &= 0\\ \nabla_{x_i} Q_{\varepsilon_2|X}(\tau_s|x) &= 0\end{aligned}$$

and the "local exclusion" restriction

$$\nabla_{x_i} h_1(s,x,\bar{e}_1)|_{\bar{e}_1=e_1} = 0.$$

Define

$$\gamma(e_2,x) \equiv \nabla_{\varepsilon_2} Q_{\varepsilon_1|\varepsilon_2 X}(\tau_w|e_2,x).$$

[27] Note that in the additively separable model $h_2^{-1}(x,s)$ has the form: $s - h_2(x)$.

There are the following partial derivatives.

$$\begin{aligned}
\nabla_s Q_{W|SX}(\tau_w|s,x) &= \nabla_s h_1(s,x,e_1) + \gamma(e_2,x)\nabla_s h_2^{-1}(x,s) \\
\nabla_{x_i} Q_{W|SX}(\tau_w|s,x) &= \gamma(e_2,x)\nabla_{x_i} h_2^{-1}(x,s) \\
\nabla_{x_i} Q_{S|X}(\tau_s|x) &= \nabla_{x_i} h_2(x,e_2)
\end{aligned}$$

The ratio of the last two of these derivatives is

$$\begin{aligned}
\frac{\nabla_{x_i} Q_{W|SX}(\tau_w|s,x)}{\nabla_{x_i} Q_{S|X}(\tau_s|x)} &= \gamma(e_2,x)\left(\frac{\nabla_{x_i} h_2^{-1}(x,s)}{\nabla_{x_i} h_2(x,e_2)}\right) &(10.8)\\
&= -\gamma(e_2,x)\nabla_s h_2^{-1}(x,s) &(10.9)
\end{aligned}$$

and therefore there is the formula (10.7) which serves to identify the structural partial derivative of h_1 with respect to S at $S = s = Q_{S|X}(\tau_s|x)$ and $\varepsilon_1 = e_1$.

The result in (10.9) arises because

$$s = h_2(x, h_2^{-1}(x,s))$$

and so there is

$$\begin{aligned}
1 &= \nabla_{\varepsilon_2} h_2(x,e_2)\nabla_s h_2^{-1}(x,s) \\
0 &= \nabla_{x_i} h_2(x,e_2) + \nabla_{\varepsilon_2} h_2(x,e_2)\nabla_{x_i} h_2^{-1}(x,s)
\end{aligned}$$

giving

$$\begin{aligned}
\nabla_s h_2^{-1}(x,s) &= \frac{1}{\nabla_{\varepsilon_2} h_2(x,e_2)} \\
\nabla_{x_i} h_2^{-1}(x,s) &= -\frac{\nabla_{x_i} h_2(x,e_2)}{\nabla_{\varepsilon_2} h_2(x,e_2)}
\end{aligned}$$

and so:

$$\frac{\nabla_{x_i} h_2^{-1}(x,s)}{\nabla_{x_i} h_2(x,e_2)} = -\nabla_s h_2^{-1}(x,s).$$

These manipulations are similar to those in the previous Section but use conditional quantile functions rather than conditional expectation functions. The latter would not have been helpful in the nonseparable model of this Section but the conditional quantile function approach could, of course,have been used in the previous Section.

10.2. An alternative derivation

Starting from (10.5) and (10.6) there is, setting $s = h_2(x, Q_{\varepsilon_2|X}(\tau_s|x))$:

$$Q_{W|SX}(\tau_w|Q_{S|X}(\tau_s|x),x) = h_1(s,x,Q_{\varepsilon_1|\varepsilon_2 X}(\tau_w|Q_{\varepsilon_2|X}(\tau_s|x),x)).$$

The derivative of the left hand side with respect to x_i is as follows.

$$\text{L: } \nabla_{\tilde{s}} Q_{W|SX}(\tau_w|\tilde{s},x)|_{\tilde{s}=Q_{S|X}(\tau_s|x)}\nabla_{x_i} Q_{S|X}(\tau_s|x) + \nabla_{x_i} Q_{W|SX}(\tau_w|\tilde{s},x)|_{\tilde{s}=Q_{S|X}(\tau_s|x)}$$

If h_1 does *not* vary through its x argument with x_i and if $Q_{\varepsilon_1|\varepsilon_2 X}$ does not vary with x_i then the x_i-derivative of the right hand side is as follows.

$$\text{R: } \nabla_{\tilde{s}} h_1(\tilde{s}, x, Q_{\varepsilon_1|\varepsilon_2 X}(\tau_w|Q_{\varepsilon_2|X}(\tau_s|x), x))|_{\tilde{s}=h_2(x,Q_{\varepsilon_2|X}(\tau_s|x))} \nabla_{x_i} h_2(x, Q_{\varepsilon_2|X}(\tau_s|x))$$

Since

$$\nabla_{x_i} Q_{S|X}(\tau_s|x) = \nabla_{x_i} h_2(x, Q_{\varepsilon_2|X}(\tau_s|x))$$

there is, on dividing the derivative L by $\nabla_{x_i} Q_{S|X}(\tau_s|x)$ and the derivative R by $\nabla_{x_i} h_2(x, Q_{\varepsilon_2|X}(\tau_s|x))$, the result (10.7).

10.3. Estimation and discussion

Estimation of identifiable structural derivatives in a nonseparable model can be conducted by applying the formula (10.7) to estimates of derivatives of conditional quantile functions of W given S and X and of S given X. This is a quantile regression based analogue of the "Indirect Least Squares" estimator. Minimum distance methods can come to the rescue when there is overidentification. For more detail see Chesher (2003) and for a study of global identification of nonseparable structural functions and a different approach to estimation, see Imbens and Newey (2009). An alternative approach to estimation under the identification conditions set out above is considered in Ma and Koenker (2004) and in Lee (2004).

11. Discrete variation in covariates

Nonparametric identification of partial derivatives requires there to be continuous variation in covariates. But frequently in econometrics we encounter discrete covariates, for example the binary "quarter of birth" instrumental variables of Angrist and Krueger (1991).

As always we must be clear about the structural feature in which we are interested. Here we focus on structural partial *differences*. Consider the structural equations:

$$W = h_1(S, X, \varepsilon_1) \tag{11.1}$$
$$S = h_2(X, \varepsilon_2) \tag{11.2}$$

in which ε_1 and ε_2 are continuously distributed unobservable variates, h_2 is a strictly monotonic (normalised increasing) in ε_2 and h_1 is monotonic (normalised nondecreasing) in ε_1. One implication of these conditions is that S given X is continuously distributed.

In the context of this model examples of structural partial differences are

$$h_2(x', e_2) - h_2(x'', e_2)$$

and

$$h_1(s', x, e_1, e_2) - h_1(s'', x, e_1)$$

where the arguments of these functions are particular numerical values. We will focus on the second of these.

Note that, if X has continuous variation and h_2 is differentiable with respect to its first argument then

$$\lim_{\substack{x' \to x \\ x'' \to x}} \frac{h_2(x', e_2) - h_2(x'', e_2)}{x' - x''} = \nabla_X h_2(X, \varepsilon_2)|_{X=x, \varepsilon_2 = e_2}$$

which is the X-partial derivative of the h_2 function evaluated at x and e_2. So consideration of identification of partial differences can lead to results on identification of partial derivatives *via* limiting arguments if structural equations are restricted to be sufficiently smooth and variables show continuous variation.

As in the previous Section consider a quantile based covariation restriction, as follows.

Covariation condition. For τ_1 and τ_2 both in $(0, 1)$, there is a quantile invariance in the distributions of ε_1 given ε_2 and X and of ε_2 given X, namely:

$$\begin{aligned} Q_{\varepsilon_2|X}(\tau_2, x') &= Q_{\varepsilon_2|X}(\tau_2, x'') \equiv e_2, \text{ say,} \\ Q_{\varepsilon_1|\varepsilon_2 X}(\tau_2, e_2, x') &= Q_{\varepsilon_1|\varepsilon_2 X}(\tau_2, e_2, x'') \equiv e_1, \text{ say.} \end{aligned}$$

Now consider the import of these conditions. First note that, because of the equivariance property of quantiles:

$$\begin{aligned} Q_{S|X}(\tau_2, x') &= h_2(x', e_2) \\ Q_{S|X}(\tau_2, x'') &= h_2(x'', e_2) \end{aligned}$$

and so the partial difference of the h_2 function:

$$h_2(x', e_2) - h_2(x'', e_2) = Q_{S|X}(\tau_2, x') - Q_{S|X}(\tau_2, x'')$$

which is identified because on the right hand side we have a well defined functional of the conditional distribution of S given X.

Now consider the wage equation and substitute for S giving

$$W = h_1(h_2(X, \varepsilon_2), X, \varepsilon_1)$$

then fix $X = x'$, $\varepsilon_2 = e_2$ giving

$$W(x', \varepsilon_1, e_2) = h_1(h_2(x', e_2), x', \varepsilon_1)$$

and note that, by the equivariance property of quantiles

$$Q_{W|\varepsilon_2 X}(\tau_1, e_2, x') = h_1(h_2(x', e_2), x', e_1).$$

But, and this is now a familiar argument, conditioning on $\varepsilon_2 = e_2$ and $X = x'$ is identical to conditioning on $S = h_2(x', e_2) = s'$, say, and $X = x'$ and so

$$Q_{W|SX}(\tau_1, s', x') = h_1(s', x', e_1).$$

A similar argument applies at $X = x''$ giving

$$Q_{W|SX}(\tau_1, s'', x'') = h_1(s'', x'', e_1)$$

where $s'' = h_2(x'', e_2)$.

Define the following structural difference.

$$\Delta = h_1(s', x', e_1) - h_1(s'', x'', e_1)$$

Note this is *not* a *partial* difference because the s *and* x arguments vary. This point is returned to shortly.

We can conclude that

$$\begin{aligned}\Delta &= Q_{W|SX}(\tau_1, s', x') - Q_{W|SX}(\tau_1, s'', x'') \\ &= Q_{W|SX}(\tau_1, Q_{S|X}(\tau_2, x'), x') - Q_{W|SX}(\tau_1, Q_{S|X}(\tau_2, x''), x'')\end{aligned}$$

and then that Δ is identifiable because in the second line, on the right hand side we have a well defined functional of the conditional distribution of W and S given X. This second line follows because, again by the equivariance property of quantiles

$$\begin{aligned}s' &= h_2(x', e_2) = Q_{S|X}(\tau_2, x') \\ s'' &= h_2(x'', e_2) = Q_{S|X}(\tau_2, x'').\end{aligned}$$

As it stands

$$\Delta = h_1(s', x', e_1) - h_1(s'', x'', e_1)$$

is *not* a partial difference of the h_1 function because two of the arguments change value as we go from x' to x''. The model must be further restricted if Δ is to be a partial difference. Clearly we require that the h_1 function be insensitive to variations in its X argument and the following *order condition* completes the restrictions defining a model that identifies a partial difference of the h_1 function.

Order condition: The function h_1 satisfies one or both of the equations:

$$h_1(s', x', e_1) = h_1(s', x'', e_1) \tag{11.3}$$

$$h_1(s'', x', e_1) = h_1(s'', x'', e_1). \tag{11.4}$$

This final condition ensures that the identifiable Δ is:

1. the partial difference
$$h_1(s', x'', e_1) - h_1(s'', x'', e_1)$$
if (11.3) holds,
2. the partial difference
$$h_1(s', x', e_1) - h_1(s'', x', e_1)$$
if (11.4) holds, and,
3. equal to both of these partial differences which are identical if both (11.3) and (11.4) hold.

Finally, note that if $s' = s''$ then the partial difference, Δ, is trivially zero. It can only be nonzero if the following *rank condition* holds.

Rank condition: The function h_2 satisfies

$$h_2(x', e_2) \neq h_2(x'', e_2).$$

11.1. Discussion

The argument above shows that features of nonseparable structures can be identified under rather weak conditions. The development follows Chesher (2007b) where there are some additional results.

The following points are worth considering.

1. There is one strong restriction that has not been commented on till now. The model we have considered has exactly as many latent unobservable random variables as there are observable stochastic outcomes. If there were measurement error obscuring the value of S in addition to variation induced by ε_2 then the identification result would not follow.

2. The result will not go through if S has a discrete distribution except in special cases which are not econometrically attractive. This is taken up at the start of the next Section.

3. The quantile invariance conditions might hold for some quantiles (e.g. conditional medians) but not for others (e.g. if there were heteroskedastic variation in the latent variables).

4. The quantile invariance conditions could be imposed in situations when conditional expectations do not exist, for example in applications in financial econometrics.

5. The identification results suggest estimation via conditional quantile function estimation[28].

12. Conditional independence restrictions

Conditional independence restrictions are widely used in modern econometric models. In this Section I consider a few examples. In practical applications one should carefully consider whether the restrictions are plausible and whether the objects identified are of practical interest.

12.1. Various average functions

Altonji and Matzkin (2005) consider models for scalar Y as follows.

$$Y = m(X, U) \qquad U \perp\!\!\!\perp X|Z$$

We read this "U and X are independently distributed ($\perp\!\!\!\perp$) conditional on (|) Z". In terms of probability density functions (if the various variables are continuous) we have

$$f_{UXZ} = f_{U|Z} f_{X|Z} f_Z.$$

Here U is unobservable and it may be a vector, that is have dimension exceeding 1. X is continuously distributed and may also be a vector but in this exposition I will consider the scalar case. The function m is restricted to be differentiable and various expected values are required to exist.

[28] See Koenker and Bassett (1978) and Koenker (2005).

Consider identification of the Local Average Response function (LAR), $\beta(x)$ defined as follows.

$$\begin{aligned}\beta(x) &\equiv E_{U|X}[\nabla_x m(x,U)|X=x] \\ &= \int \nabla_x m(x,u) f_{U|X}(u|x) du\end{aligned}$$

Here $f_{U|X}(u|x)$ is the conditional probability density function of U given $X = x$.

Note this does not arise on differentiating

$$E[Y|X=x] = \int m(x,u) f_{U|X}(u|x) du$$

because

$$\nabla_x E[Y|X=x] = \int \nabla_x m(x,u) f_{U|X}(u|x) du + \int m(x,u) \nabla_x f_{U|X}(u|x) du.$$

But consider

$$E[Y|X=x, Z=z] = \int m(x,u) f_{U|XZ}(u|x,z) du$$

which can be written as:

$$E[Y|X=x, Z=z] = \int m(x,u) f_{U|Z}(u|z) du$$

because $U \perp\!\!\!\perp X|Z$. It follows that

$$\nabla_x E[Y|X=x, Z=z] = \int \nabla_x m(x,u) f_{U|Z}(u|z) du$$

and taking expectations over Z given $X = x$:

$$\begin{aligned}E_{Z|X}[\nabla_x E[Y|X=x,Z]] &= \int \left(\int \nabla_x m(x,u) f_{U|Z}(u|z) du \right) f_{Z|X}(z|x) dz \\ &= \int \nabla_x m(x,u) \left(\int f_{U|XZ}(u|x,z) f_{Z|X}(z|x) dz \right) du \\ &= \int \nabla_x m(x,u) f_{U|X}(u|x) du \\ &= \beta(x)\end{aligned}$$

Here is a simple example. With

$$E[Y|X=x, Z=z] = \alpha + \beta_1 x + \beta_2 z + \beta_3 xz$$

$$E[Z|X=x] = g(x)$$

there is

$$\nabla_x E[Y|X=x, Z=z] = \beta_1 + \beta_3 z$$

and so $\beta(x) = \beta_1 + \beta_3 g(x)$.

Here is another average local response:

$$\gamma(x) \equiv E_U[\nabla_x m(x,U)] = \int \nabla_x m(x,u) f_U(u) du.$$

Since

$$\nabla_x E[Y|X=x,Z=z] = \int \nabla_x m(x,u) f_{U|Z}(u|z) du$$

there is

$$\begin{aligned} E_Z[\nabla_x E[Y|X=x,Z=z]] &= \int \left(\int \nabla_x m(x,u) f_{U|Z}(u|z) du \right) f_Z(z) dz \\ &= \int \nabla_x m(x,u) \left(\int f_{U|Z}(u|z) f_Z(z) dz \right) du \\ &= \int \nabla_x m(x,u) f_U(u) du \\ &= \gamma(x). \end{aligned}$$

The Average Structural Function (ASF)

$$\mu(x) \equiv E_U(m(x,U)] = \int m(x,u) f_U(u) du$$

has received some attention - see for example Blundell and Powell (2003). Since $U \perp\!\!\!\perp X|Z$

$$E[Y|X=x,Z=z] = \int m(x,u) f_{U|Z}(u|z) du$$

and taking expectation with respect to Z there is the following.

$$\begin{aligned} E_Z E[Y|X=x,Z=z] &= \int \left(\int m(x,u) f_{U|Z}(u|z) du \right) f_Z(z) dz \\ &= \int m(x,u) \left(\int f_{U|Z}(u|z) f_Z(z) dz \right) du \\ &= \int m(x,u) f_U(u) du \\ &= \mu(x) \end{aligned}$$

Some questions:

1. What are $\gamma(x)$ and $\mu(x)$ in the simple example.
2. Why might $\beta(x)$, $\gamma(x)$ and $\mu(x)$ be interesting structural features?
3. Could we have done all the analysis above with U discrete?

12.2. Treatment effect models

An individual is treated ($D = 1$) or not treated ($D = 0$). The outcomes for the individual are Y_1 if she is treated and Y_0 if she is not treated. We observe realisations of D and of Y defined as follows.

$$Y = DY_1 + (1 - D)Y_0.$$

Note that realisations of Y_0 and Y_1 are never available for any single individual. Either an individual is treated and we see a realisation of Y_1 or she is not treated and we observe a realisation of Y_0. Accordingly no aspect of the dependence between Y_1 and Y_0 (e.g. a conditional mean such as $E[Y_1|Y_0 = y_0]$) can be identified.

There may be interest in the Average Treatment Effect (ATE)

$$\mu \equiv E[Y_1 - Y_0]$$

or in the Average Effect of Treatment on the Treated (ATT):

$$\mu_1 \equiv E[Y_1 - Y_0|D = 1].$$

If $(Y_1, Y_0) \perp\!\!\!\perp D$ (which arises if each individual is equally likely to be selected for treatment - i.e. if treatment allocation is randomised) then

$$E[Y|D = 1] = E[Y_1|D = 1] = E[Y_1]$$

$$E[Y|D = 0] = E[Y_0|D = 0] = E[Y_0]$$

so

$$E[Y|D = 1] - E[Y|D = 0] = \mu$$

but otherwise not.

We may expect a treatment indicator to be correlated with potential outcomes in the absence of full randomisation. For example in medical contexts a doctor may assign to treatment those that she thinks will benefit most (e.g. those that have a high value of Y_1 or of $Y_1 - Y_0$). When individuals self select into treatment we can expect those who perceive the benefits of treatment to be higher to be more likely to select the treatment option.

It is common to find the following conditional independence restriction imposed

$$(Y_1, Y_0) \perp\!\!\!\perp D|Z$$

and, in the absence of parametric restrictions, there is the support condition that for each[29] z

$$P[D = 1|Z = z] \in (0, 1).$$

Here Z is a list of observable characteristics of individuals and descriptors of the environment in which decisions are made. The restriction required that all the dependence between the potential outcomes and the selection indicator arise entirely through dependence of each on the observed variables Z.

[29] That is for no z is $P[D = 1|Z = z] = 0$ or $P[D = 1|Z = z] = 1$.

Under this condition

$$E[Y|D=1,Z=z] = E[Y_1|D=1,Z=z] = E[Y_1|Z=z]$$

$$E[Y|D=0,Z=z] = E[Y_0|D=0,Z=z] = E[Y_0|Z=z]$$

so the ATE is identified as follows.

$$E_Z[E[Y|D=1,Z=z] - E[Y|D=0,Z=z]] = \mu$$

Identification of the ATT proceeds in a similar fashion.

See Lectures 1 and 2 of the Imbens-Wooldridge CeMMAP 2009 lectures: "New Developments in Econometrics" for further details.[30]

12.3. Control functions

Consider the model, with V *scalar*, g *strictly* increasing in V and

$$\begin{aligned} Y &= h(X,U) \\ X &= g(Z,V) \end{aligned}$$

with $(U,V) \perp\!\!\!\perp Z$.

We now allow the possibility that U is a vector and/or discrete and that h is not monotonic in U.

In this model, $V = g^{-1}(Z,X)$ and g^{-1} is identified, and there is the following independence condition.

$$U \perp\!\!\!\perp X|V$$

This arises because $(U,V) \perp\!\!\!\perp Z$ implies $U \perp\!\!\!\perp Z|V$ and thus that, given V, U is independent of any function of Z and V, in particular $X = g(Z,V)$. Note that this argument does not rely on V being scalar or on g being strictly monotone.

Since there is conditional independence given V we can proceed as in Section 12.1 replacing conditioning on Z by conditioning on V. For example the average structural function:

$$\mu(x) \equiv E_U[h(x,U)]$$

is identified by:

$$E_V[E_{Y|XV}[Y|X=x,V]]$$

It is important to understand that this is not useful in practice if V cannot be identified. This rules out cases with discrete X and cases in which V is not a scalar. Set identification is possible with discrete X if additional monotonicity restriction is imposed, see Chesher (2005). Perhaps there is progress to be made along similar lines with vector V.

[30] Access at http://www.cemmap.ac.uk/resources/resources25.php.

13. Incomplete models

Incomplete models are models that admit incomplete structures. An incomplete structure does not always deliver a *unique* value for endogenous outcomes given values of observed and unobserved exogenous variables.

Throughout this Section let Y denote observed endogenous variables, let Z denote observed exogenous variables with support $\mathcal{R}_Z$, let U denote unobserved exogenous variables.

A leading example of an incomplete model is a model specifying a single structural equation with more than one endogenous variable, for example

$$Y_1 = \alpha Y_2 + \beta Z_1 + U$$

with the dependence of U and $Z = (Z_1, Z_2)$ restricted. This can be thought of as one equation in a system of equations that uniquely determine Y_1 and Y_2 (and maybe other endogenous variables).

Another important example is a model of strategic interaction admitting structures with multiple equilibria with no selection mechanism specified. The simultaneous firm entry model studied in Tamer (2003) is a widely considered case. Here Y_1 and Y_2 are binary indicating the presence (1) or not (0) of firms 1 and 2 in a market with:

$$\begin{aligned} Y_1 &= 1[\delta_1 Y_2 + \beta_1 Z_1 + U_1 \geq 0] \\ Y_2 &= 1[\delta_2 Y_1 + \beta_2 Z_2 + U_2 \geq 0] \end{aligned}$$

$U = (U_1, U_2)$ and Z independently distributed and the linear indexes interpreted as expected profits involving observed and unobserved cost shifters. In many applications one would expect δ_1 and δ_2 to be negative (why?) and in that case there are admissible structures deliver a unique solution for Y_1 and Y_2 *except* when (U_1, U_2) fall in the rectangle

$$[-\beta_1 Z_1, -\beta_1 Z_1 - \delta_1] \times [-\beta_2 Z_2, -\beta_2 Z_2 - \delta_2]$$

in which case $Y = (0, 1)$ and $Y = (1, 0)$ are both solutions. Check that this is indeed the case.

In the absence of a selection mechanism which determines which solution will be observed, this is an incomplete model. A selection mechanism could be simply that one of the two solutions is chosen with a constant probability p. More complicated selection mechanisms allow the selection probability to depend on the values of observed and unobserved exogenous variables. But a plausible selection mechanism may be difficult to specify, requiring knowledge of the process not available to the researcher.

Incomplete models are attractive because inferences obtained using them are not susceptible to misspecification of the selection mechanisms and other model elements one would have to employ to produce a complete model. Incomplete models have been used since the very earliest days of econometrics.

13.1. Sets and structures

An incomplete structure delivers a *set* of values of endogenous variables which may not be singleton. In some models, complete or incomplete, at a particular value of observed exogenous variables, a particular value of the endogenous variables can be delivered by all members of a

set of values of unobserved exogenous variables. This happens for example when endogenous outcomes are discrete or high dimensional. To see this consider the values of U that deliver a particular outcome. say $Y = (0, 0)$ in the simultaneous firm entry model.

To accommodate these possibilities it is convenient to characterize the structural relationships embodied in a structure using a function h with the property that the values of (Y, Z, U) that can arise are the following set of values.

$$\mathcal{L}(y, z, u; h) \equiv \{(y, z, u) : h(y, z, u) = 0\}$$

In the linear model at the start of this Section

$$h(y, z, u) = y_1 - \alpha y_2 - \beta z_1 - u$$

and

$$h(y, z, u) = (y_1 - \alpha y_2 - \beta z_1 - u)^2$$

would do as well. Write down a function h suitable for the simultaneous firm entry model.

In this notation a structure comprises a pair $(h, \mathcal{G}_{U|Z})$ where

$$\mathcal{G}_{U|Z} \equiv \{G_{U|Z}(\cdot|z) : z \in \mathcal{R}_Z\}$$

is a collection of conditional distributions of U given $Z = z$. This is the notation and set up in Chesher and Rosen (2017) which gives a very general treatment of the identifying power of incomplete (and complete) models.

Models place restrictions on h, for example the parametric linear restriction of the linear model above and the threshold crossing form with linear indexes of the simultaneous firm entry model. Models place restrictions on the conditional distributions $\mathcal{G}_{U|Z}$ - for example requiring U and Z to be mean independent or stochastically independent. In the latter case $\mathcal{G}_{U|Z} = \{G_U\}$ a singleton set.

Define $\mathcal{Y}(u, z)$ as the set of values of Y obtained when $Z = z$ and $U = u$ according to structural function h, as follows.

$$\mathcal{Y}(u, z; h) \equiv \{y : h(y, z, u) = 0\}$$

The set $\mathcal{Y}(U, z)$ obtained when $U \sim G_{U|Z}(\cdot|z)$ is a *random set.*[31] A random set is characterized by the collection of random variables whose realizations lie in the random set with probability 1. Such random variables are called the *selections* of the random set.[32]

13.2. Observational equivalence

Consider some specific collection of distributions of Y given Z,

$$\mathcal{F}_{Y|Z} \equiv \{F_{Y|Z}(\cdot|z) : z \in \mathcal{R}_Z\}$$

[31] Molchanov (2005) is a wide ranging review of random set theory. Very little of that material is necessary to understand the work discussed here.

[32] Think of a simple example of a random set, namely in interval on the real line $\mathcal{T} = [T_1, T_2]$. Examples of selections of this random set are the random variables $A = \lambda T_1 + (1 - \lambda)T_2$ where $\lambda \in [0, 1]$ and λ could itself be a random variable.

delivered by an economic process. We assume the observation process is such that $\mathcal{F}_{Y|Z}$ is identified.

We define two structures $(h', \mathcal{G}'_{U|Z})$ and $(h'', \mathcal{G}''_{U|Z})$ as *observationally equivalent* with respect to $\mathcal{F}_{Y|Z}$ if for all $z \in \mathcal{R}_Z$, $F_{Y|Z}(\cdot|z) \in \mathcal{F}_{Y|Z}$ is the distribution of a selection of $\mathcal{Y}(u, z; h')$ when $U \sim G'_{U|Z=z} \in \mathcal{G}'_{U|Z}$ and also the distribution of a selection of $\mathcal{Y}(u, z; h'')$ when $U \sim G''_{U|Z=z} \in \mathcal{G}''_{U|Z}$. When the sets $\mathcal{Y}(U, Z; h)$ are singleton with probability 1 this accords with the definition given earlier for complete models.

A structure $(h, \mathcal{G}_{U|Z})$ is in the identified set delivered by a model $\mathcal{M}$ and a collection of distributions $\mathcal{F}_{Y|Z}$ if and only if the structure is admitted by $\mathcal{M}$ and for all $z \in \mathcal{R}_Z$, $F_{Y|Z}(\cdot|z) \in \mathcal{F}_{Y|Z}$ is the distribution of a selection of $\mathcal{Y}(U, z; h)$ when $U \sim G_{U|Z}(\cdot|z) \in \mathcal{G}_{U|Z}$.

This is a rather cumbersome definition for the many econometric models that place restrictions on the dependence of unobserved U and Z. Define $\mathcal{U}(y, z)$ - the set of values of U that can give rise to $Y = y$ when $Z = z$ according to structural function h.

$$\mathcal{U}(y, z; h) \equiv \{u : h(y, z, u) = 0\}$$

We call this a U level set of the function h.

Using a simple duality argument Chesher and Rosen (2017) show that a structure $(h, \mathcal{G}_{U|Z})$ is in the identified set delivered by a model $\mathcal{M}$ and a collection of distributions $\mathcal{F}_{Y|Z}$ if and only if the structure is admitted by $\mathcal{M}$ and, for all $z \in \mathcal{R}_Z$, $G_{U|Z}(\cdot|z) \in \mathcal{G}_{Y|Z}$ is the distribution of a selection of the random set $\mathcal{U}(Y, z; h)$ when $Y \sim F_{Y|Z}(\cdot|z) \in \mathcal{F}_{Y|Z}$.

There are many ways in which to characterize the selectionability property of a probability distribution $G_{U|Z}(\cdot|z)$ relative to a random set $\mathcal{U}(Y, z; h)$.[33] Particularly useful is the characterization given by Artstein's inequality (Artstein (1983)). Chesher and Rosen (2017) use this to obtain the following characterization of an identified set.

Proposition 13.1. *A structure $(h, \mathcal{G}_{U|Z})$ is in the identified set delivered by a model $\mathcal{M}$ and a collection of distributions $\mathcal{F}_{Y|Z}$ if and only if the structure is admitted by $\mathcal{M}$ and, for all $z \in \mathcal{R}_Z$ and all sets $\mathcal{S} \in \mathsf{Q}(z, h)$, a collection of core determining sets,*

$$G_{U|Z}(\mathcal{S}|z) \geq \mathbb{P}[\mathcal{U}(Y, z; h) \subseteq \mathcal{S}|Z = z] \tag{13.1}$$

where $G_{U|Z}(\mathcal{S}|z)$ is the probability mass placed on the set S by the probability law $G_{U|Z}(\cdot|z)$ and the probability law used in calculating the probability on the right-hand side is $F_{Y|Z}(\cdot|z) \in \mathcal{F}_{Y|Z}$. The collection of core determining sets $\mathsf{Q}(z, h)$ comprises certain unions of the sets $\mathcal{U}(y, z; h)$ with $y \in \mathcal{R}_{Y|Z=z}$.

It can be shown that the weak inequalities in (13.1) are *equalities* in complete models and in models in which with probability 1 realizations of the sets $\mathcal{U}(Y, Z; h)$ are singleton.

Another characterization using the Aumann expectation of a random set is useful when there are restrictions on the conditional moments of U given Z and these lead to many of the classical identification results.[34] For example, suppose there is the restriction $E[U|Z = z] = 0$ for all

[33] A probability distribution of a point valued random variable is selectionable with respect to a random set if it is the distribution of a selection of the random set.

[34] The Aumann expectation of a random set is the closure of the set of expected values of all of the random set's selections that have finite expected values.

		$\mathbb{P}[\mathcal{U}(Y,z;h) \subseteq \mathcal{S}\|Z=z]$	
$\mathcal{S}$	$G_U(\mathcal{S})$	$g(0) \leq g(1)$	$g(0) \geq g(1)$
$[0, g(0)]$	$g(0)$	$p_{00}(z)$	$p_{00}(z) + p_{01}(z)$
$[0, g(1)]$	$g(1)$	$p_{00}(z) + p_{01}(z)$	$p_{01}(z)$
$[g(0), 1]$	$1 - g(0)$	$p_{10}(z) + p_{11}(z)$	$p_{10}(z)$
$[g(1), 1]$	$1 - g(1)$	$p_{11}(z)$	$p_{10}(z) + p_{11}(z)$

Table 13.1: Values of probabilities in the inequalities defining the identified set of threshold functions in the binary outcome IV model

$z \in \mathcal{R}_Z$. The identified set of structural functions comprises all functions h such that, for all $z \in \mathcal{R}_Z$, zero is an element of the Aumann expectation of $\mathcal{U}(Y, Z; h)$. Details are given in Chesher and Rosen (2017). All the characterizations discussed are characterizations of *sharp* identified sets, that is: *all* and *only* admissible structures that can deliver the conditional distributions of Y given Z under consideration lie in the sets. Identified sets for structural features are obtained by projection.

13.3. Application: a binary outcome IV model

There is the following model for a binary outcome Y_1 with endogenous explanatory variable Y_2

$$Y_1 = s(Y_2, U) \equiv \begin{cases} 1 & , \quad g(Y_2) \leq U \\ 0 & , \quad g(Y_2) \geq U \end{cases} \qquad \text{and} \quad U \perp\!\!\!\perp Z$$

with $U \sim Unif(0, 1)$. In this case there is

$$h(Y, Z, U) = Y_1 - s(Y_2, U)$$

and there are the following U level sets.

$$\begin{array}{ll} \mathcal{U}((0,0), z; h) = [0, g(0)] & \mathcal{U}((0,1), z; h) = [0, g(1)] \\ \mathcal{U}((1,0), z; h) = [g(0), 1] & \mathcal{U}((1,1), z; h) = [g(1), 1] \end{array}$$

The sets employed in the characterization of the identified set for the function g are precisely these level sets.[35] The value of $G_{U|Z}(\mathcal{S}|z) = G_U(\mathcal{S})$ (because of the independence restriction) is shown in Table 13.1. The values are just the lengths of the intervals since U is uniformly distributed.

Let $p_{ij}(z)$ denote $P[Y_1 = i \wedge Y_2 = j|Z = z]$. The value of the probability $\mathbb{P}[\mathcal{U}(Y, z; h) \subseteq \mathcal{S}|Z = z]$ involves these probabilities and depends on the relative magnitude of $g(0)$ and $g(1)$. The values are given in Table 13.1.

Using the inequalities (13.1) and the values of probabilities given in Table 13.1 we obtain the result that the identified set for the threshold function $g(\cdot)$ comprises functions $g(\cdot)$ such that

[35] Unions of sets with lower bounds 0 and unions of sets with upper bounds 1 deliver one or other of the simple level sets. Other unions either deliver the whole unit interval or a disconnected pair of intervals, neither of which produce informative inequalities.

either:

$$(g(0) \leq g(1)) \wedge \inf_{z \in \mathcal{R}_Z} (p_{01}(z) + p_{00}(z)) \geq g(0) \geq \sup_{z \in \mathcal{R}_Z} p_{00}(z) \wedge \inf_{z \in \mathcal{R}_Z} (1 - p_{11}(z)) \geq g(1) \geq \sup_{z \in \mathcal{R}_Z} (p_{00}(z) + p_{01}(z))$$

or:

$$(g(0) \geq g(1)) \wedge \inf_{z \in \mathcal{R}_Z} (1 - p_{10}(z)) \geq g(0) \geq \sup_{z \in \mathcal{R}_Z} (p_{00}(z) + p_{01}(z)) \wedge \inf_{z \in \mathcal{R}_Z} (p_{01}(z) + p_{00}(z)) \geq g(1) \geq \sup_{z \in \mathcal{R}_Z} p_{01}(z)$$

the "sup" and "inf" operations arising because the instrumental variable Z is *excluded* from the threshold function g.

Identified sets are graphed for some cases in Chesher (2013). When instrumental variables have a weak effect on Y_2 the sets can be disconnected. The continuous X, binary Y case is studied in Chesher (2010) which does not use random set theory methods. Chesher and Rosen (2013) gives an application to estimation of the average treatment effect of changing family size on married female's employment using the data employed in Angrist and Evans (1998). Disconnected sets arise here when using the weak "same sex" instrument but not when using the strong "twins" instrument.

Work out how this analysis can be extended to the IV probit model in which[36]

$$Y_1 = \begin{cases} 1 & , \quad \beta_0 + \beta_1 Z_1 + \alpha Y_2 \leq U \\ 0 & , \quad \beta_0 + \beta_1 Z_1 + \alpha Y_2 \geq U \end{cases} \quad \text{and} \quad U \perp\!\!\!\perp (Z_1, Z_2) \quad \text{and} \quad U \sim N(0,1).$$

14. Some significant milestones and recent work

The study of identification attracted the attention of the pioneers of econometrics. Significant contributions can be found in Working (1925), Working (1927), Tinbergen (1930), Frisch (1934, 1938), Haavelmo (1944), Hurwicz (1950), Koopmans, Rubin and Leipnik (1950), Koopmans and Reiersøl (1950), Wald (1950), Fisher (1959, 1961, 1966), Wegge (1965) and Rothenberg (1971). Apart from Hurwicz (1950) and Koopmans and Reiersøl (1950) these studies focussed almost exclusively on the identifying power of parametric models.

Roehrig (1988), extending the work of Brown (1983), considered nonparametric global identification of smooth structural functions under the restriction that latent variates are distributed independently of covariates. Newey and Powell (1988), Newey, Powell and Vella (1999), Pinkse (2000), and Darolles, Florens and Renault (2000) study nonparametric models with additive latent variables which satisfy mean independence conditions.

[36] You can find the answer in Chesher and Rosen (2018).

Brown and Matzkin (1996) study the nonparametric global identification of primitive functions, for example production or utility functions, associated with nonseparable simultaneous equations systems when latent variables and covariates are restricted to be independently distributed. Altonji and Matzkin, (2003) study global identification in nonseparable panel data models with endogeneity under conditional exchangeability assumptions. Imbens and Newey (2003) give results on global identification of structural functions in triangular systems, like those for the most part considered here, when latent variates and covariates are restricted to be independently distributed. They relax Roehrig's (1988) smoothness restriction and consider identification and estimation of various average structural functions. Blundell and Powell (2003) and Chesher (2004a) provide surveys of research on identification as it stood in the early 2000's.

Identification is considered from a conditional quantile perspective in the absence of endogeneity in Matzkin (2003). The quantile based analysis in earlier sections extends this work to problems in which there is endogeneity. Matzkin (2008) gives new results on identification in complete nontriangular systems correcting an error in Brown (1983) and Roehrig (1988). Chernozhukov and Hansen (2005) and Chernozhukov, Imbens and Newey (2007) study identification of nonseparable functions using marginal independence restrictions. Chesher (2010, 2013) studies (set) identification under marginal independence restrictions when the outcome variable is discrete. Chesher (2010) is the first paper to consider IV models with discrete outcomes.

Manski (2003) summarises work on set identification and gives many useful references. Much the work on set identification stems from the research of Manski and his associates, starting in the late 1980's, although there was consideration of the topic much earlier, e.g. Hurwicz (1950). Very little of Manski's work concerns structural econometric models.

The models considered in these notes require the number of latent variates to be equal to the number of observable outcomes. But many models employed in microeconometric practice violate this restriction. It is interesting to ask what minimal additional restrictions secure identification of interesting structural features when this restriction on the number of latent variates is relaxed? Chesher (2009) considers the power of index restrictions in this context. Schennach (2004, 2007) and Altonji and Matzkin (2005) have some interesting results in measurement error and panel data contexts. Chesher, Rosen and Smolinski (2013) study the identifying power of instrumental variable restrictions in a multiple discrete choice model in which the latent variable is high-dimensional. Chesher and Rosen (2017) characterize identified sets delivered by a wide class of incomplete and complete models admitting discrete or continuous outcomes and unobservables, and scalar or nonscalar heterogeneity. That paper introduces Generalized Instrumental Variable (GIV) models which allow relaxation of the commonly imposed restriction (see e.g. Newey and Powell (1988, 2003), Chernozhukov and Hansen (2005), Matzkin (2008)) requiring unobservable variables to be a single-valued function of observable variabes. A more relaxed exposition of GIV models and their applications is in Chesher and Rosen (2018, 2020)[37]. An application to models for ordinal outcomes, specifically measures of social wellbeing as captured in the Moving to Opportunity experiment, is in Chesher and Rosen (2019). An appliication to a dynamic model of market structure is in Berry and Compiani 2021).

[37] You can down load the 2018 CeMMAP working paper for free. It is very similar to the final published Handbook of Econometrics chapter.

References

Altonji, Joseph and Rosa L. Matzkin (2005): "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *Econometrica*, 73, 1053-1102.

Angrist, Joshua D., and William N. Evans (1998): "Children and their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *Amercan Economic Review*, 88, 450-487.

Angrist, Joshua D., and Alan B. Krueger (1991): "Does compulsory school attendance affect schooling and earnings?" *Quarterly Journal of Economics*, 106, 979-1014.

Amemiya, Takeshi (1982): "Two stage least absolute deviations estimators," *Econometrica*, 50, 689-711.

Artstein, Zvi (1983): "Distributions of random sets and random selections," *Israel Journal of Mathematics*, 46, 313-324.

Becker, Gary S., and Barry R. Chiswick (1966): "Education and the distribution of earnings," *American Economic Review*, 56, 358-369

Berry, Steven T., and Giovani Compiani (2021): "Empirical Models of Industry Dynamics with Endogenous Market Structure," *Annual Review of Economics*,13, 309-334.

Blundell, Richard W., and James L Powell (2003): "Endogeneity in Nonparametric and Semiparametric Regression Models," in *Advances in Economics and Econonometrics: Theory and Applications, Eighth World Congress, Vol. II*, M. Dewatripont, L.P. Hansen and S.J. Turnovsky, eds. Cambridge: Cambridge University Press.

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

Brown, D.J., and R.L. Matzkin (1996): "Estimation of nonparametric functions in simultaneous equations models, with an application to consumer demand," mimeo, Department of Economics, Northwestern University.

Card, David (2001): "Estimating the returns to schooling: Progress on some persistent econometric problems," *Econometrica*, 69, 1127-1160.

Card, David (1995): "Earnings, ability and schooling revisited," in *Research in Labour Economics, Volume 14*, ed., S.Polachek. Greenwich, Conn: JAI Press.

Chernozhukov, Victor and Christian Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245-261.

Chernozhukov, Victor, Guido W. Imbens and Whitney K. Newey (2007): "Instrumental Variable Identification and Estimation of Nonseparable Models via Quantile Conditions," *Journal of Econometrics*, 139, 4-14.

Chesher, Andrew (2002b): "Semiparametric identification in duration models," Centre for Microdata Methods and Practice Working Paper 20/02.

Chesher, Andrew, (2003): "Identification in nonseparable models," *Econometrica*, 71, 1405-1441.

Chesher, Andrew, (2004a): "Identification of sensitivity to variation in endogenous variables," The A.W.H. Phillips Lecture, presented at the Australasian Meetings of the EconometricSociety Melbourne, July 7th 2004, Centre for Microdata Methods and Practice Working Paper CWP10/04.

Chesher, Andrew, (2004b): "Identification in additive error models with discrete endogenous variables," Centre for Microdata Methods and Practice Working Paper CWP11/04.

Chesher, Andrew, (2005): "Nonparametric identification under discrete variation," *Econometrica*, 73, 1525-1550.
Chesher, Andrew, (2007a): "Identification of nonadditive structural functions," in: *Advances in Economics and Econonometrics: Theory and Applications, Ninth World Congress, Vol. III*, R Blundell, W.K. Newey and T. Persson, eds. Cambridge: Cambridge University Press.
Chesher, Andrew, (2007b): "Instrumental Values," *Journal of Econometrics*, 139, 15-34.
Chesher, Andrew, (2009): "Excess heterogeneity, endogeneity and index restructions," *Journal of Econometrics*, 152, 35-47.
Chesher, Andrew, (2010): "Instrumental variable models for discrete outcomes," *Econometrica*, 78, 575-601.
Chesher, Andrew, (2013): "Semiparametric structural models of binary response: shape restrictions and partial identification," *Econometric Theory*, 29, 231-266.
Chesher, Andrew and Adam Rosen (2012): "Simultaneous equations models for discrete outcomes: coherence, completeness, and identification," CeMMAP Working Paper CWP21/12.
Chesher, Andrew and Adam Rosen (2013): "What do instrumental variable models deliver with discrete dependent variables?", *American Economic Review: Papers and Proceedings*, 103, 557-562, and online Annex.
Chesher, Andrew and Adam Rosen (2017): "Generalized Instrumental Variable Models," *Econometrica*, 85, 959-989.
Chesher, Andrew and Adam Rosen (2018): "Generalized Instrumental Variable Models, Methods and Applications," CeMMAP Working Paper CWP43/18, chapter prepared for the *Handbook of Econometrics Volume 7A*.
Chesher, Andrew and Adam Rosen (2020): "Generalized Instrumental Variable Models, Methods and Applications," *Handbook of Econometrics Volume 7A*, edited by Steven N. Durlauf, Lars Peter Hansen, James J. Heckman, Rosa L. Matzkin, Chapter 1. pages 1-110, Elsevier, Amsterdam.
Chesher, Andrew and Adam Rosen (2019): "Estimating Endogenous Effects on Ordinal Outcomes," CeMMAP Working Paper CWP 66/19. Instrumental Variable Models, Methods and Applications," CeMMAP Working Paper CWP43/18.
Chesher, Andrew, Adam Rosen and Konrad Smolinski (2013): "An Instrumental Variable Model of Multiple Discrete Choice," *Quantitative Economics*, 4, 157-196.
Chiswick, Barry R. (1974): *Income inequality: regional analyses within a human capital framework*. New York: Columbia University Press.
Chiswick, Barry R., and Jacob Mincer (1972): "Time series changes in personal income inequality," *Journal of Political Economy*, 80, S34-S66.
Das, Mitali, (2005): "Instrumental Variables Estimators for Nonparametric Models with Discrete Endogenous Regressors," *Journal of Econometrics*, 124, 335-361.
Fisher, Franklin M. (1959): "Generalization of the rank and order conditions for identifiability," *Econometrica*, 27, 431-447.
Fisher, Franklin M. (1961): "Identifiability criteria in nonlinear systems," *Econometrica*, 29, 574-590.
Fisher, Franklin M. (1966): *The identification problem in econometrics*, New York: McGraw Hill.
Haavelmo, T.M. (1944): "The probability approach in econometrics," *Econometrica*, 12, Supplement, July 1944, 118 pp.

Hurwicz, Leonid, (1950): "Generalization of the concept of identification," in *Statistical inference in dynamic economic models*. Cowles Commission Monograph 10, New York, John Wiley.[38]
Imbens, Guido and Whitney K. Newey (2009): "Identification and estimation of triangular simultaneous equations models without additivity," *Econometrica*, 77, 1481-1542.
Lee, Sokbae (2007): "Endogeneity in Quantile Regression Models: a Control Function Approach," *Journal of Econometrics*, 141, 1131-1158.
Koenker, R.W., (2005): *Quantile Regression*, Cambridge: Cambridge University Press.
Koenker, Roger and Gilbert Bassett Jr. (1978): "Regression quantiles," *Econometrica*, 46, 33-50.
Koopmans, Tjalling C., and Olaf Reiersøl (1950): "The identification of structural characteristics," *Annals of Mathematical Statistics*, 21, 165-181.
Koopmans, Tjalling C., H. Rubin and R.B. Leipnik (1950): "Measuring the equation systems of dynamic economics," in *Statistical inference in dynamic economic models*. Cowles Commission Monograph 10, New York, John Wiley.[39]
Koopmans, Tjalling C., and O. Reiersøl (1950): "The identification of structural characterstics," *Annals of Mathematical Statistics*, 21, 165-181.
Ma, Lingjie and Roger W. Koenker (2006): "Quantile Regression Methods for Recursive Structural Equation Models," *Journal of Econometrics*, 134, 471-506.
Manski, Charles F., (1988): *Analog estimation methods in econometrics*, New York: Chapman and Hall.
Manski, Charles F., (2003): *Partial identification of probability distributions*, Heidelberg: Springer Verlag.
Matzkin, Rosa L., (2003): "Nonparametric estimation of nonadditive random functions," *Econometrica*, 71, 1339-1375.
Matzkin, Rosa L., (2008): "Identification in nonparametric simultaneous equations models," *Econometrica*, 76, 945-978.
Mincer, Jacob (1974): *Schooling experience and earnings*. New York: Columbia University Press.
Molchanov, Ilya (2005): *Theory of random sets*. London: Springer-Verlag.
Newey, Whitney K., and James L. Powell (1988): "Instrumental Variables Estimation for Nonparemtric Models," mimeo, Department of Economics, MIT. Available as CeMMAP Working Paper CWP07/17.
Newey, Whitney K., and James L. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71(5), 1565-1578.
Newey, Whitney K., James.L. Powell, and Frank. Vella (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica* 67, 565-603.
Pinkse, Joris (2000): "Nonparametric two-step regression functions when regressors and errors are dependent," *Canadian Journal of Statistics*, 28, 289-300.
Roehrig, Charles S. (1988): "Conditions for identification in nonparametric and parametric models", *Econometrica* 56, 433-447.
Rothenberg, Thomas J. (1971): "Identification in parametric models," *Econometrica*, 39, 577-591.

[38] Available online at: http://cowles.econ.yale.edu/P/cm/m10/m10-04.pdf .
[39] Available online at: http://cowles.econ.yale.edu/P/cm/m10/m10-02.pdf.

Schennach, Susanne M., (2004): "Estimation of Nonlinear Models with Measurement Error," *Econometrica*, 72, 33-75.
Schennach, Susanne M., (2007): "Instrumental Variable Estimation of Nonlinear Errors-in-Variables Models," *Econometrica*, 75, 201-239.
Tamer, Elie (2003): "Incomplete Simultaneous Discrete Response Model with Multiple Equilibria," *The Review of Economic Studies*, 70, 147-165.
Tinbergen, Jan (1930), Bestimmung und Deutung von Angebotskurven: Ein Beispiel, *Zeitschrift für Nationalökonomie* 1, 669-679.
Wald, Abraham (1950): "Note on identification of economic relations," in *Statistical inference in dynamic economic models*. Cowles Commission Monograph 10, New York, John Wiley.
Wegge, Leon (1965): "Identifiability Criteria for a System of Equations as a Whole," *The Australian Journal of Statistics*, 7, 67-77.
Working, Elmer.J. (1927): "What do statistical 'demand curves' show?" *Quarterly Journal of Economics*, 41,212-235.
Working, Holbrook (1925): "The statistical determination of demand curves," *Quarterly Journal of Economics*, 39, 503-543.