

Identification Analysis and Evidence Science

Andrew Chesher

Centre for Microdata Methods and Practice, IFS and UCL

Evidence Seminar: October 5th 2006

ABSTRACT

Economics has a set of tools, the tools of identification analysis, which are designed to determine conditions under which data (evidence) can be brought to bear unambiguously on particular propositions. Identification analysis was developed in the 1930's and 1940's in earliest days of econometrics and it is still a very active subject.

In this talk I will set out the elements of identification analysis as practiced in economics and ask: do the tools and constructions employed in identification analysis offer a framework within which to build a science of evidence?

Here are some rough notes to accompany the talk.

1 Econometrics and the evidence problem in economics

Econometrics is the name given to the part of economics that deals with the measurement of economic phenomena and the use of economic data in understanding human behaviour and the consequences of economic interactions. So it deals with the business of processing and interpreting economic evidence. Econometrics is also concerned with the design of measurement instruments although this has not had so much prominence historically.

Some econometric analysis uses experimental data but there are severe limits (in part due to costs) on the scale and complexity of experiments. Consequently most economic data come from observation of individuals or groups of individuals and their responses to their economic and social environment and to changes in it.

Many aspects of individuals and their environment go unmeasured or are imperfectly measured. Individuals and organisations are instrumental in determining their own situation and, possibly imperfectly, optimise their responses. When they do this their relative valuations of current and future, certain and risky, outcomes are influential as are their views of the likelihood of future outcomes and their knowledge of their current situation. All of these elements vary across individuals and perhaps over time.

With many aspect of individuals preferences and environment unmeasured there is much scope for ambiguity in the interpretation of the limited record provided in economic data.

A very early concern in econometrics was the problem of determining the conditions under which economic data are capable of particular interpretations. This is the subject studied under the heading “identification”. It is distinct from, and logically prior to, the problem of inference.

2 Examples of identification problems

Here are some examples problems where identification issues arise.

2.1 Demand supply and market equilibrium

Desired demand for (e.g.) flowers (e.g. by households at some location) is a decreasing function of the price at which flowers are sold and demand side factors (e.g. income, timing of religious festivals).

Desired supply of flowers (by sellers at some location) is an increasing function of the price at which flowers are sold and supply side factors (e.g. weather conditions, transport costs).

Buyers and sellers come together in a market and the price at which transactions occur is the price at which desired demand and supply are equal. This process continues day after day and data record for each day the price at which flowers are bought and sold, the amount bought and sold and demand and supply side factors. Under what circumstances can the data reveal information about either demand functions or supply functions or both?

2.2 Returns to schooling

People choose investments in schooling which depend on ability and family factors. They later earn wages that depend on their schooling outcomes and ability. We see data on schooling outcomes, wages and family factors for a sample of people, but there is no record of ability. Under what conditions can this data reveal information about the “pure” (*ceteris paribus*) effect of schooling on wages, ability held constant?

2.3 Smoking and mortality

People have different life expectations and smoking causes diseases which rarely affect the young. Perhaps people with the view that their life expectation is relatively short are over-represented amongst smokers. Can data on smoking behaviour and mortality alone reveal the effect of smoking on mortality? What additional measurements could resolve this problem and under what conditions?

2.4 Assessing the benefit of training programmes

Aimed at the long term unemployed, training programmes are designed to speed return to work and promote longer job tenure. Those who view programmes as well suited to their needs may be more inclined to join and remain in training programmes. Under what conditions can data on participation in programmes and subsequent labour market histories be informative about the efficacy of training programmes?

2.5 Identification and evidence

Thinking of data, or functions of data, as “evidence” it seems at first sight as if identification analysis as practiced in econometrics may have relevance for evidence science. But is that the case? I will return to this question. First I will try to set out the bare bones of identification analysis. Even if there is nothing here for evidence science it may be interesting to see how this problem is approached in economics.

3 Identification analysis

3.1 Measurement and random variation

We imagine a process generating data on a list of outcomes Y under circumstances characterised by a list of unobservable variables U and observable variables X . From time to time and/or from person to person X and U vary and as a result Y varies. Values of Y and X (but not U) are recorded.

The various variables are taken to be *measurable* and are regarded as *random variables* with well defined *probability distributions*.¹ Data are regarded as realisations of these random variables, drawn from these probability distributions. Variables could be binary indicating possession (or not) of some characteristic or occurrence or otherwise of some event, so measurability is not perhaps as severe a restriction as it might seem at first sight.

Data are informative about the joint distribution F_{YX} but in some cases one works entirely with the conditional distribution $F_{Y|X}$ allowing that X may not be a random variable at all. For data to be informative about probability distributions there has to be a sense in which information accrues as data are amassed.

3.2 Processes

Given a value of X and U , a value of Y is generated as a unique solution to an equation

$$g(Y, X, U) = 0.$$

Formally g is a vector valued function. In the wage (Y_1), schooling (Y_2) example g could have the form

$$\begin{aligned} Y_1 &= \theta Y_2 + U_1 + \lambda U_2 \\ Y_2 &= \beta X + U_2 \end{aligned}$$

where U_2 denotes ability.

Uniqueness of the solution for Y is important. The probability distribution of X and U together with a function g that solves uniquely for Y induce a *probability distribution* for Y and X .

A particular process is called a *structure*. A structure S consists of a particular function g^S and a particular probability distribution of U and

¹We may make an exception for X .

$X, F_{U|X}^S: S = \{g^S, F_{U|X}^S\}$. Each structure S implies exactly one probability distribution $F_{Y|X}^S$. If one is working conditionally on X then there will be $S = \{g^S, F_{U|X}^S\}$ which implies exactly one *conditional* probability distribution $F_{Y|X}^S$.

3.3 Observational equivalence

When appraising evidence we are interested in the particular structure S^* generating the data (evidence) before us. This data is informative about the probability distribution $F_{Y|X}^{S^*}$ associated with S^* . In identification analysis one asks: under what conditions is knowledge of the probability distribution implied by the structure informative about the structure or features of it.

The question arises because, although for each S there is exactly one $F_{Y|X}^S$, there may be many structures S which imply a particular $F_{Y|X}$, that is there may be $S_1 \neq S_2$ such that $F_{Y|X}^{S_1} = F_{Y|X}^{S_2}$. Structures S_1 and S_2 such that $S_1 \neq S_2$ and $F_{Y|X}^{S_1} = F_{Y|X}^{S_2}$ are said to be *observationally equivalent* (OE).

If two structures are observationally equivalent then no amount of data can be informative about the identity of the structure which generated the data. If there are propositions which are true in one structure and false in another and both structures are observationally equivalent then the proposition can never be confirmed or denied on the basis of evidence contained in data.

3.4 Models

Models are restrictions that define which structures are admissible, that is could have generated data.

For example the restrictions: g is linear in its arguments and U and X are independently distributed random variables is a model. With $Y = (Y_1, Y_2)$ and $X = (X_1, X_2)$ and $U = (U_1, U_2)$ and with Greek symbols denoting constants this model could be expressed as follows.

$$\begin{aligned} \text{(A). } Y_1 &= \alpha_{12}Y_2 + \beta_{11}X_1 + \beta_{12}X_2 + U_1 \\ \text{(B). } Y_2 &= \alpha_{21}Y_1 + \beta_{21}X_1 + \beta_{22}X_2 + U_2 \end{aligned}$$

(C). (U_1, U_2) and (X_1, X_2) are independently distributed

Models define sets of structures.

Consider a model M and a structure $S_0 \in M$. If there is no $S_1 \in M$ (with $S_1 \neq S_0$) such that $F_{YX}^{S_1} = F_{YX}^{S_0}$ then M identifies S_0 . If M identifies all structures $S \in M$ then M is a *uniformly identifying model*. There is a one-to-one correspondence between the structures admitted by a uniformly identifying model and the probability distributions they imply.

If we add the restrictions $\{\beta_{12} = \beta_{21} = 0, \beta_{11} \neq 0, \beta_{22} \neq 0\}$ to the model set out above then the model is uniformly identifying if X_1 and X_2 vary sufficiently. The restrictions $\{\alpha_{12} = \alpha_{21} = 0\}$ also render the model uniformly identifying.

Current econometric research on identification studies models in which there are not parametric restrictions.

3.5 Restrictiveness of models and just identifying models

If $M_1 \subset M_0$ then M_0 admits structures which are not admitted by M_1 and M_1 is more restrictive than M_0 . If M_0 is a uniformly identifying model and $M_1 \subset M_0$ then M_1 is also a uniformly identifying model.

Suppose M_0 is a uniformly identifying model and that no model M with $M_0 \subset M$ is uniformly identifying. Then M_0 is said to be a *just identifying model*. A just identifying model loses its (uniform) identifying power if any of its restrictions are relaxed.

3.6 Falsifiability of models

If every structure *not admitted* by a model is observationally equivalent to some structure *admitted* by the model then the model cannot be falsified. A model M is falsifiable if there exists a structure not admitted by the model which is observationally distinguishable from every structure admitted by the model, that is if there exists $S^* \notin M$ such that for every $S \in M$, $F_{YX}^{S^*} \neq F_{YX}^S$. All non-falsifiable models are just identifying. There exist just identifying models that are falsifiable.

3.7 Features of structures

Often there is interest only in some feature of a model, for example a coefficient in a linear equation system, or the sign of such a coefficient (e.g. α_{12}) above. A model may identify a structural feature when it is not uniformly

identifying. This happens when, even though a model admits observationally equivalent structures, a feature is *invariant* within any set of observationally equivalent structures.

Formally let $\theta(S)$ be a feature of structure S , consider a model M and a structure $S_0 \in M$. A model M identifies $\theta(S_0)$ if $\theta(S_0) = \theta(S)$ for every S in M which is observationally equivalent to S_0 . If this is true for every $S_0 \in M$ then the model uniformly identifies θ .

If a model M_0 identifies a structural feature and $M_1 \subset M_0$ then the more restrictive model M_1 also identifies the structural feature. Suppose that M_1 identifies a structural feature and that there is no M_0 with $M_1 \subset M_0$ such that M_0 identifies the structural feature. Then M_1 is a *just identifying* model for that structural feature.

There may be many just identifying models for a structural feature. One can sometimes obtain a just identifying model M_1 from another just identifying model M_0 by relaxing one restriction of M_0 while tightening another restriction. If M_0 and M_1 are both just identifying models for a structural feature then $M_0 \not\subseteq M_1$ and $M_1 \not\subseteq M_0$.

3.8 Set identification

The discussion so far has been about *point* identification. Consider a structural feature and a model and suppose that within any set of observationally equivalent structures admitted by the model the structural feature lies in a *set*. In this case the model *set identifies* the structural feature.

Set identification and inference on set identified structural features is an active research field in econometrics. There is an example in Chesher (2005a).

3.9 How can the identifying power of a model be determined?

3.9.1 Identification of structures

The power of a model for identifying structures is shown by demonstrating that no two structures admitted by the model generate the same distribution of Y and X (or of Y given X if the analysis is conditional on X). This can be easy to do when a model is parametrically specified. An example is provided by the model below if augmented by the restriction that all variables are jointly normally distributed.

3.9.2 Identification of features of structures

It can be proved² that if, for a model M , there exists a functional of the distribution function of Y and X , $\mathcal{G}(F_{YX})$, such that for all $S \in M$, when $\theta(S) = a$, $\mathcal{G}(F_{YX}) = a$, then M identifies θ .

Here is an example of this in action.

There is the model:

$$(A). Y = \alpha X + U$$

(B). For $z \in \mathcal{Z}$, $E[U|Z = z] = c$, $E[X|Z = z]$ exists, is finite and varies with z in which note there is no requirement that X and U be uncorrelated or independent. For $(z_1, z_2) \in \mathcal{Z}$ there is:

$$\begin{aligned} E[Y|Z = z_1] &= \alpha E[X|Z = z_1] + c \\ E[Y|Z = z_2] &= \alpha E[X|Z = z_2] + c \end{aligned}$$

and so if $E[X|Z = z_1] \neq E[X|Z = z_2]$:

$$\alpha = \frac{E[Y|Z = z_1] - E[Y|Z = z_2]}{E[X|Z = z_1] - E[X|Z = z_2]}.$$

Consider some value a . Every structure with $\alpha = a$ implies a distribution for Y , X and Z such that

$$\frac{E[Y|Z = z_1] - E[Y|Z = z_2]}{E[X|Z = z_1] - E[X|Z = z_2]} = a$$

the left hand side expression here being the required functional \mathcal{G} .

4 Identification and Evidence Science

A few incomplete remarks.

1. Considerations of identification arise before considerations of inference (in the statistical sense). When studying identification one asks whether information about propositions could ever be gleaned from any amount of data no matter how large.

²Chesher (2006).

2. Identification is achieved by imposing restrictions. Some restrictions may be based on evidence (got from earlier observation of a process) and others are necessarily beliefs and cannot be falsified.
3. Identification can be achieved by changing the nature of measurement - e.g. measuring more accurately, or measuring different aspects of a process.
4. In a “structure” one has a complete specification of the way in which observable outcomes are generated. Typically one is interested in some feature of a structure.
5. If observationally equivalent structures are permitted then there is the possibility of alternative explanations of phenomena. Features of structures can be identified when structures themselves cannot. One requires that across observationally equivalent structures the value of the feature of interest does not vary (much).
6. To use this apparatus one must be prepared (able) to think in terms of processes that generate outcomes that are (can be modelled as) realisations of random variables.

REFERENCES

Below (1-4) are some of the papers on identification produced during the Evidence project and then a few of the classical references Reference (2) has a more extensive bibliography.

- (1). CHESHER, ANDREW (2006): “Instrumental Values,” forthcoming in *Journal of Econometrics*, revision of Centre for Microdata Methods and Practice Working Paper CWP 17/02.
- (2). CHESHER, ANDREW (2003): “Identification in Nonseparable Models,” *Econometrica*, 71, 1405-1441.
- (3). CHESHER, ANDREW (2005a): “Nonparametric identification under discrete variation,” *Econometrica*, 73, 1525-1550.
- (4). CHESHER, ANDREW (2005b): “Identification with Excess Heterogeneity,” Centre for Microdata Methods and Practice Working Paper CWP 19/05, presented at the 2006 European Meeting of the Econometric Society, Vienna.

- (5). FISHER, FRANKLIN M. (1959): "Generalization of the rank and order conditions for identifiability," *Econometrica*, 27, 431-447.
- (6). FISHER, FRANKLIN M. (1961): "Identifiability criteria in nonlinear systems," *Econometrica*, 29, 574-590.
- (7). FISHER, FRANKLIN M. (1966): *The identification problem in econometrics*, New York: McGraw Hill.
- (8). HAAVELMO, TRYGVE, M. (1944): "The probability approach in econometrics," *Econometrica*, 12, Supplement, July 1944, 118 pp.
- (9). HURWICZ, LEONID (1950): "Generalization of the concept of identification," in *Statistical inference in dynamic economic models*. Cowles Commission Monograph 10, New York, John Wiley.
- (10). KOOPMANS, TJALLING C., AND OLAV REIERSØL (1950): "The identification of structural characteristics," *Annals of Mathematical Statistics*, 21, 165-181.
- (11). KOOPMANS, TJALLING C., HERMAN RUBIN AND ROY 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.
- (12). ROTHENBERG, THOMAS J. (1971): "Identification in parametric models," *Econometrica*, 39, 577-591.