Journal of Econometrics and Statistics
Vol. 2, Issue 1, 2022
DOI: https://doi.org/10.46791/jes.2022.v02i01.04

Causality in Econometric Modeling
From Theory to Structural Causal Modeling

Renzo ORSI[a], Michel MOUCHART[b], and Guillaume WUNSCH[c]
[a] *Department of Economics, University of Bologna, Italy*
[b] *Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA) and CORE, UCLouvain, Belgium*
[c] *Demography, UCLouvain, Belgium*

Article History
Received : 29 January 2022; Revised : 28 February 2022; Accepted : 19 March 2022; Published : 30 June 2022

To cite this paper
Renzo ORSI, Michel MOUCHART & Guillaume WUNSCH (2021). Causality in Econometric Modeling from theory to Structural Causal Modeling. *Journal of Econometrics and Statistics.* 2(1), 61-90.

Abstract

This paper examines different approaches for assessing causality as typically followed in econometrics and proposes a constructive perspective for improving statistical models elaborated in view of causal analysis. Without attempting to be exhaustive, this paper examines some of these approaches. Traditional structural modeling is first discussed. A distinction is then drawn between model-based and design-based approaches. Some more recent developments are examined next, namely history-friendly simulation and information-theory based approaches. Finally, in a constructive perspective, structural causal modeling (SCM) is presented, based on the concepts of mechanism and sub-mechanisms, and of recursive decomposition of the joint distribution of variables. This modeling strategy endeavors at representing the structure of the underlying data generating process. It operationalizes the concept of causation through the ordering and role-function of the variables in each of the intelligible sub-mechanisms.

Keywords: structural modeling, exogeneity, causality, model-based and design-based approaches, recursive decomposition, history-friendly simulation, transfer entropy.
JEL Classification: C01, C03, C15, C18, C51, C54

Aknowledgements. The authors are deeply grateful to J.H.Drèze, V.Ginsburgh and F. Russo for their quite interesting comments and suggestions on a first version of this paper. Comments and suggestions from an anonymous reviewer on a former version of this paper are gratefully acknowledged and have led the authors to clarify the rationale behind the order of the presentation.

1 Introduction

Economic Theory and Econometric Modeling

This paper presents a critical view of causal assessment as typically done in econometrics and proposes a constructive approach for improving statistical models elaborated for causal analysis.

Following some prominent textbooks in econometrics, economic theory and econometric modeling are closely related. According to J. Johnston (1972), for example, econometrics translates the propositions of economic theorists into a mathematical form and then sees if the data confirm these *a priori* propositions. In their popular textbook, James Stock and Mark Watson (2003, p.3) define econometrics, at a broad level, as "the science and art of using economic theory and statistical techniques to analyze economic data". Marno Verbeek (2004, p.2) puts it succinctly by saying that econometrics is the interaction between economic theory, observed data, and statistical methods; in his words: "econometricians formulate a statistical model, usually based on economic theory, confront it with the data, and try to come up with a specification that meets the required goals". More recently, in his well-known textbook, Jeffrey Wooldridge (2013, p.1) defines econometrics as follows: "Econometrics is based upon the development of statistical methods for estimating economic relationships, testing economic theories, and evaluating and implementing government and business policy". He also stresses the fact that econometrics deals mainly with observational, *i.e.* non-experimental, economic data.

Widening the economic approach to non-economic fields, Gary Becker (1976, p.5) asserts that the economic approach can provide a unified framework for understanding all human behavior. For him, "The combined assumptions of maximizing behavior, market equilibrium, and stable preferences, used relentlessly and unflinchingly, form the heart of the economic approach". He has applied this approach to various aspects of human activity, such as crime and punishment, or marriage, fertility, and the family.

From this small sample of references, one could thus tentatively put forward that econometrics is, in the first place, the science of testing economic theories, possibly translated beforehand- following Wooldrige - into mathematically formulated economic models. One notices that this approach to econometrics takes economic theory as pre-established with respect to the data to be analyzed and is used as a device for interpreting the data. Said differently, economic theory in an econometric model acts as an "interpretative tale", *i.e.* a "story" providing an interpretation to empirical findings that is not built in a systematic way to analyze a particular data set. This would give to economic theory some "universal " validity. For example, an econometric model for analyzing the final demand of a specific good might be based on an economic theory of inter-temporal utility maximization. Such a theory is typically developed without reference to a particular population of interest and is conceived to be congruent with general economic theory without being embedded in a particular period or space, *i.e.* in a specific context.

Structural Models

There is a significant literature on causality in econometrics and, more generally, in economics. Some basic principles go back to David Hume and John Stuart Mill, but it is generally considered that a significant step in the causal approach in econometrics dates from the development of structural models in the 1930s by Jan Tinbergen, and especially from the work of the Cowles Commission (in particular the works of Koopmans and Klein) which came out in 1950 (LeRoy, 2006; see Hoover, 2008, for an overview). Other relevant ideas can be traced back to Trygve Haavelmo in the late 1940s, with among others his probability approach to econometrics, and to Hermann Wold who published in 1954 an important paper on causality and econometrics. For Wold (1954), the supreme tool is the controlled experiment, but most econometric works deal with non-experimental observations. In the latter case, Wold adheres to a recursive system approach, pioneered by Tinbergen, forming causal chains. Conditional probability distributions can be interpreted as causal relations and, taken together, the conditional distributions constitute a joint probability distribution recursively decomposed.

Most economists would agree with Wold that the gold standard for drawing inferences is the randomized controlled experiment, though the external validity of controlled experiments can be questioned (see *e.g.* Athey and Imbens, 2017). Actually, a large share of empirical work in econometrics relies on observational data where, *inter alia*, the possibility of confounding, or loss of exogeneity, has to be taken into account. In the case of non-experimental data, many econometricians now refer to the potential outcomes or Neyman-Rubin causal model, where the effect of a treatment is compared to the effect of its counterfactual. This approach has been criticized by James Heckman (2008), among others, who considers that the potential outcomes model is a black box device, postulating counterfactuals without modeling the factors determining the outcome and without any discussion of the theory that could explain the outcome. One can add that the same criticism can be leveled at the randomized controlled experiment too. For Heckman, and others, scientific models need to "go into the black box" (Heckman's terms) and explore the mechanisms producing the effects.

The Role of Time

One of the issues discussed in causal modeling is the role of historical time. For John Hicks (1980), among others, it is undeniable that the cause-effect relation has some reference to time. Experimental science, in its nature, is out of historical time, as time is irrelevant for the significance of an experiment. This is not the case in economics: it is the past that provides the economist with facts, which are used to make generalizations. In this sense economics, according to Hicks, would not be far off from history as a science, though history is focused on the past and economics on the present and the future. Economics is concerned with human actions and decisions. No decision made now can affect what has happened in the past. From this, Hicks draws an interesting conclusion. With respect to the past, one can be fully determinist: there are no events in the past that one may not attempt to explain.

Causal Models

Causal assessment in econometric models is necessarily based on associations among variables. However, econometric models based solely on associations cannot lead to causal statements; see *e.g.* Moneta and Russo (2014). Such models measure statistical covariation among variables and may be useful *e.g.* for forecasting. In order to yield causal explanation, Russell Davidson (2015, p.1200), among others, stresses the fact that a model should embody theory *and* that theory provides an explanation; following Davidson: "a theory must explain by proposing a mechanism, or in other words a *causal* chain". Moneta and Russo (2014) consider that a causal model, in econometrics, is actually an "augmented" statistical model, incorporating causal information. In order to make causal claims, it is necessary to go beyond the usual characteristics of a statistical model. One has to take into account the fact that a causal model should spell out the structure of the underlying data generating process.

Policymakers and causal evidence

In recent years, empirical econometrics has shifted its paradigm towards an increasing focus on causality, as policymakers have increasingly demanded evidence-based policies. However, policymakers do not show a preference between experimental and observational studies. In particular, policymakers are preoccupied with whether research results are relevant to their specific local context, without discerning the nature of the studies carried out, as research results are communicated to policymakers as headlines of impact estimates with very little illustration of the research design. From our point of view this result is surprising since from a causal inference perspective, experimental studies are limited in generalizability compared to observational studies. However, experiments may be the preferred research design for policy makers if the objective is to promote policies that have been shown to improve outcomes in the past. Nowadays policymakers face pressure to use research evidence to inform their decisions. We believe that policymakers are more likely to incorporate research evidence into their decision-making processes if they are enabled to follow and understand how this evidence was obtained in the research presented to them. In this paper we will review the different approaches to conducting causal inference analyses and try to highlight the limitations and merits of each, in the belief that the evidence provided by empirical research will be used by policy makers in the real world.

Contents

A goal of all sciences is to explain the events and processes they observe. When possible, scientists set up controlled experiments for this purpose, in which all variables are held constant except for the variable deemed to produce the effect. In sciences dealing with human beings, the randomized controlled trial is a well-known example of such an experimental approach where an intervention being tested (*e.g.*, a drug) is given to a target group (the experimental group) and the outcome is compared to that of a comparison group (or control group) that does not receive the intervention. Controlled experiments are also conducted in economics, but this strategy is impossible in many situations for practical or for ethical reasons. Economists then have to rely on non-experimental (or

observational) methods to infer causal knowledge from the facts.

Without intending to be exhaustive, this paper presents some approaches aiming at deriving knowledge in economics in non-experimental conditions, by way of various econometric methodologies. Section 2 discusses traditional structural estimation. A distinction is then made, in Section 3, between model-based and design-based approaches. The following sections present two more recent developments, namely history-friendly simulation (Section 4) and information theory-based approaches (Section 5). In Section 6, the paper develops a structural causal modeling methodology based on the concepts of mechanism and sub-mechanisms, and on the recursive decomposition of the global mechanism. Discussion and conclusions are presented in Section 7.

In particular, the paper supports the view that a thorough scientific understanding of the cause-effect relations in economics requires an elucidation of the sub-mechanisms leading from the causes to the effects. From a policy viewpoint however, policymakers can base their decisions on weaker claims, such as those made by the non-structural models discussed in this paper. In this case the paths from cause to effect remain a black box, but this does not necessarily hinder the impact of the policies if it can be shown that a variation in the cause produces a variation in the effect, controlling for possible confounders. Indeed, even in a randomized controlled trial the route from intervention to outcome actually remains unrevealed.

2 Traditional structural modeling and causality

2.1 A long tradition

Structural modeling has a long tradition in econometrics and is a result of the attempts to bridge theory and empirical findings in economics, as a development of the innovative work done by the Cowles Commission, with a particular reference to Marschak (1953) and Koopmans (1953). In structural models, economic theory is used in order to explain how a set of variables, say y, is related to a set of supposedly exogenous variables named x. The basic issue for the Cowles Commission was the development of systems of simultaneous equations reflecting an economic theory explaining the joint production of a set of variables.

An analysis of the relationship between the theoretical model and the statistical model is crucial in framing the problem of causality. More specifically, making assumptions about the causal structure of an economic model, in the absence of a statistical model, may lead to questionable results. In trying to build a bridge linking the two models, one starts from the general consideration that any economic phenomenon for which one wants to conduct an empirical analysis should be conceived as stochastic. The empirical model supposed to represent this economic phenomenon should be based on two different sources of closely related information: on the one hand, an information coming from the economic theory of reference (the substantive source) and on the other hand a purely empirical source of information (the data). Quoting Haavelmo (1944,

p.*iii*): "The method of econometric research aims, essentially, at a conjunction of economic theory and actual measurements, using the theory and technique of statistical inference as a bridge pier". Integrating these two different sources of information, in the phase of the specification of the empirical model, is a problem that has proved of fundamental importance over time.

The probabilistic approach, initially suggested by Haavelmo (1944) and later developed by several others, among them Spanos (2006), considers these two sources of information as complementary but separately treated: in the form of a statistical model using the data for measurement purposes and of a theoretical (or conceptual) model based on economic theory and called a structural model. This is mainly due to the fact that the statistical model is initially considered as a particular configuration of a stochastic process, namely the data generating process (DGP) underlying the data (y, x) that is derived, under specific conditions, from the structural model under study.

Considering a phenomenon as stochastic typically involves the action of a high number of latent, or not measurable, factors. For this reason, economic theory is often safeguarded by a series of *ceteris paribus* clauses. These clauses may be justified in a partial equilibrium framework but may be questionable in a general equilibrium framework. Moreover, these clauses, as long as they are not verifiable, may act as wishful thinking for the model-builder, with the possibility of underlying endogeneity for some of the variables involved in the *ceteris paribus* clauses, thus jeopardizing the validity of the model.

The idea behind this approach, further developed in Subsection 2.2, is that any economic theory, when trying to predict the behavior of one or more variables y, defines the main aspects to be modeled by selecting the principal factors of influence, say those described by the set of variables x. The model-builder is aware of the fact that there may be numerous other factors, say ζ, not necessarily observable, but potentially relevant, that could affect the behavior of the variables y under study, but that for convenience can be ignored, at least initially.

2.2 Theoretical model and statistical model

Theoretical model

Following Spanos (2006), among others, in the classical econometric approach a theoretical model can be written in the form:

$$y_t = f^*(x_t, \zeta_t) \tag{1}$$

where t stands for an identifier of the units of observation. The term $f^*(.)$ indicates the postulated behavioral relationship for y_t, which depends both on the primary factors of influence, represented by the variables in x_t, and on the factors not observable or judged not to be of prime importance, captured by ζ_t.

In order to obtain a specification useful for empirical analysis, the usual econometric practice decomposes (1) into a term depending only on x_t and a

"residual term " depending on both x_t and ζ_t:

$$y_t = f(x_t) + \epsilon(x_t, \zeta_t) \tag{2}$$

where $f(x_t)$ no longer indicates a complete behavioral relationship, but rather a postulated structural component of y_t. The term $\epsilon(x_t, \zeta_t)$ represents the structural error that is considered to be a function of both x_t and ζ_t. According to equation (2), the structural error term

$$\epsilon(x_t, \zeta_t) = y_t - f(x_t) \tag{3}$$

represents the unmodeled influence. In order to derive from this structural model a manageable statistical model, it is convenient to assume that this error term behaves as a white noise uncorrelated with the systematic component $f(x_t)$.

This may be achieved by assuming the following properties:

$$\begin{aligned} &(i)E[\epsilon(x_t, \zeta_t) \mid x_t] = 0, \quad (ii)E[\epsilon(x_t, \zeta_t)^2 \mid x_t] = \sigma^2 \\ &(iii)E[\epsilon(x_t, \zeta_t)\epsilon(x_k, \zeta_k) \mid x_t, x_k] = 0 \ \ \forall\, t \neq k \quad (iv)E[\epsilon(x_t, \zeta_t) f(x_t)] = 0 \end{aligned} \tag{4}$$

Thus, this classical structural modeling approach provides an idealized description of the economic phenomena of interest in the form reported in (2). The specification of the structural model, as considered above, involves several choices. Among them, more explicitly:

1. The definition of the subject of interest from an economic point of view, and the variables (y_t, x_t) which best represent it.

2. The fundamental aspects of the phenomenon that have to be modeled, such as, for example, the characteristics of the conditional distribution, the conditional expectation and/or the conditional variance *i.e.* assumptions in (4) imply that $E[y_t \mid x_t] = f(x_t)$ and $V[y_t \mid x_t] = \sigma^2$.

3. The extent to which the inferences derived from the structural model are relevant for the phenomenon of interest, in particular for policy purposes.

It is worth noting that model (2) aims at approximating the actual DGP. The validity of this approximation crucially depends upon the characteristics of the residual term, but the properties (i) - (iv), in (4) above, cannot be directly verified since the error term is not observable. To render them testable one needs to embed the structural model into a statistical model, a crucial step that often goes unnoticed. The characteristics of this embedding depend upon the nature of the data, namely whether the data (y_t, x_t) are collected as the outcome of an experiment or are observational in nature.

Statistical Model

When elaborating a statistical model from equation (2), economists would agree that the three variables (y_t, x_t, ζ_t) have not been generated in an isolated world

but are rather part of a complex world characterized by a large set of other variables, say W_t. The role of the statistical model is to characterize the distribution of $(y_t, x_t, \zeta_t \mid W_t)$. The essential purpose of the *ceteris paribus* clauses is to assume that the validity of the model is limited to a fixed value of W_t. Here, W_t represents standing conditions and variables specific to the context that can be left out of the model specifying the DGP of y_t; see Wunsch, Mouchart and Russo (2018).

From a statistical point of view, writing $\epsilon(x_t, \zeta_t)$ or $\epsilon(x_t, W_t)$ may be viewed as meaning that ϵ is a *deterministic* function of x_t and ζ_t or W_t, but not that ϵ is a random variable possibly associated with ζ_t or W_t. This point raises the issue of a deterministic as opposed to a stochastic view of the real world, an issue not to be discussed in this paper.

When dealing with experimental data, experimental designs may ensure that the *ceteris paribus* clauses are satisfied and that the joint distribution of (y_t, x_t, ζ_t) is implicitly conditional to a fixed value of W_t. In such a case, ϵ_t is no longer a (deterministic) function of x_t and ζ_t but rather a random variable, along with y_t and x_t. In equation (2) the value of y_t is determined by (x_t, ϵ_y), but now, in the statistical model, $f(x_t)$ becomes an expectation of y_t conditional on x_t. A simple way of obtaining this reinterpretation of $f(x_t)$ is to assume:

$$(\epsilon_t \mid x_t) \sim IID(0, \sigma^2) \tag{5}$$

The main consequence when analyzing experimental data, *i.e.* data obtained through experimental designs, is that the structural model (2) is transformed into the statistical model below:

$$y_t = f(x_t) + \epsilon_t \tag{6}$$

In this model the error term is different, in nature, from the error term of the structural model (2) since it is no longer a deterministic function of (x_t, ζ_t) . In summary, when the statistical model has been obtained as a result of an experimental design, it represents an idealized probabilistic description of a stochastic process $\{(y_t, x_t); t \in T\}$ generating the data (y_t, x_t), this description being in the form of a consistent set of probabilistic assumptions. This ensures that the data represent a typical realization of the data generating process.

When data (y_t, x_t) are instead observed as a result of an ongoing DGP in a non-experimental situation, the error term can not typically be assumed to be free of (x_t, ζ_t) in (3).

Economic theory and statistical inference

The connection between economic theory and statistical inference allows facing new classes of problems, models and methods of inference. Many of these are typical of econometrics and this is mainly due to the non-experimental nature of most economic data. In addition to specifying relevant variables y and x, a fundamental reason to make use of economic theory is to spell out how institutional and economic conditions influence the relationships between y and x. An emblematic case is that of the simultaneous equations model and the related

identification problem. For example, one can regress market quantity on price, but this does not necessarily mean that the parameters of a market demand function have been properly estimated. In this case, the conditions of supply should be made explicit as well as the source of the error in the estimating equation.
More specifically, consider a two-equation model representing the supply and demand of a specific good and let this model be estimated from the observed equilibrium of price and quantity. Such a model does not explain the mechanism leading to the equilibrium of the market but only represents the result of this mechanism. This model leads to a statistical model characterizing the joint distribution of two endogenous variables, namely price and quantity, conditionally on the exogenous variables, appearing in the supply and demand equations, but does not permit identifying a causality connection between price and quantity.

2.3 Empirical econometrics

Structural empirical econometrics is an empirical analysis based on a statistical model derived from a structural economic model. A typical example is a structural model in the spirit of the Cowles Commission. Structural empirical econometrics is carried out with the aim of understanding facts and mechanisms that regulate the functioning and the behavior of a complex system, such as an economic system. From this point of view, economics is not different from other sciences. In astrophysics, for example, mechanisms are inferred from sometimes very indirect observations. Similarly to astrophysics, economics is a science with little capacity to carry out controlled experiments. A similar argument can be held for the social sciences in general where controlled experiments are often impossible for practical or ethical reasons. In economics, it is not always easy to build coherent economic theories that can explain the data well. Any economic fact can potentially be explained by several alternative economic theories that are proposed to account for the same phenomenon.

Moreover, in economics, observational data are periodically revised, some economic variables of interest are unobserved and their measured indicators can differ significantly according to the measurement hypotheses adopted. Besides, as mentionned above, different researchers can come up with different economic theories for the same set of data. This in itself is not necessarily bad, since studying how models differ from each other can help to better understand the functioning of the system. What we want to emphasize is that, in economics, a purely data-driven approach can rarely lead to fully satisfactory results in an explanatory perspective.

On the theoretical side, economic theory continually provides us with invaluable insights into the behavior of agents and individual choices, the functioning of markets, the functioning of macroeconomic systems and much more.

Empirical econometrics is an empirical analysis guided by economics but in the absence of a complete structural model. A typical example may be found in financial econometrics where a basic objective is to reproduce past observations from a simulation based on an empirical descriptive model.

Descriptive econometrics aims at building a coherent and significant description of economic data, such as a distribution of income with a precise concept of income and of the underlying population. Beyond the immediate use that one can make of it (*e.g.* for taxation purposes), the description can also be seen as a step towards the specification of a more complete and theoretically grounded model for policy analysis (for instance, in order to reduce economic inequalities). When one can identify a policy parameter in an empirical model, this means that some economic structure has been injected into the descriptive model. It may happen that structural models contain parts that are directly specified in reduced form if we are not interested in modeling the sub-mechanisms generating the exogenous variables.

Neverthelss, there is a significant distinction between the two different ways of doing empirical analysis. Non-structural econometric analysis was boosted by the fact that the complications resulting from the specification of a structural model are not always deemed useful. Little effort has been made, in the past, to demonstrate the alleged superiority of structural empirical econometrics. Recent literature on experimental econometrics is mainly concerned with the identification of causal effects inferred from natural- or quasi-experimental data.

2.4 Difficulties

Structural modeling in economics is a complicated task. Among the problems must be counted the fact that economic theories may be quite complex in their formulation and may be difficult to translate into estimable relationships. Structural models rarely admit a convenient closed form solution that can be estimated by means of regressions or other standard statistical and econometric methods. Specific tools including numerical analysis and computer programming are required, along with a thorough understanding and knowledge of the data and context. Another element that affects structural modeling is the possible lack of data on the constructs or quantities y and x for a given economic theory. This can considerably complicate inference and in any case puts crucial limits on what can be obtained from the available data. And economic theory does not always give the researcher everything (s)he needs to develop a model, with the consequence that an important part of the specification is left to the researcher's intuition and creativity.

Nevertheless, for an econometrician, the appeal to economic theory is not only useful but necessary if one wants to assess causalilty from the estimated model or use it for developing alternative scenarios; see Section 4. Structural econometric models make use of economic and statistical assumptions to identify economic quantities from the conditional density $f(y \mid x)$. This approach endeavors at making clear which economic assumptions are necessary, but not always sufficient, to allow causal inference.

In a nutshell, strong assumptions are needed if one wants to go beyond a simple summary of the data, for infering causal relations, testing theories, estimating fundamental parameters or making predictions about the effects of

policy changes on the behavior of economic agents.

2.5 No model is "true"

The absence of relevant data can considerably complicate estimation and restrict what the researcher can do with what is available, but nevertheless this does not always make a correct empirical analysis impossible. Pure measurement models consider inference as a process of testing hypotheses and uncovering the "true" parameters. Such models are certainly not the best method for a proper empirical analysis. Models are, by definition, simplifications and approximations of economic reality, and as such they are all necessarily imperfect as literal descriptions of this economic reality. Among these wrong models, some provide better approximations of economic reality than others, and hopefully over time these models will become more realistic and the approximations will improve.

Thus, looking for a good approximation in econometric modeling would be an appropriate strategy for the analysis of structural estimation. Even though structural models are likely to be misspecified and are based on questionable assumptions of functional form, some of these models can nevertheless produce useful results. Their results are often better than those of purely descriptive models that avoid the formulation and estimation of a behavioral economic model. The collective belief among many economists that all structural models are falsified (in Popper's sense) and therefore that none of them are any good, contributes to discrediting structural modeling in favor of the simplistic view of descriptive econometrics. On the one hand, one may accept some of the fundamental limits of economic theory, and among them an extreme difficulty of modeling the general environment for the functioning of an economy. On the other hand, the identification problem is one of the most important limits that must be faced when doing structural estimation, though it should not be concluded that the existence of this limit leads to the futility of the structural approach.

3 Model-based and Design-based approaches

Two alternative approaches

In econometric practice one often refers to model-based and design-based as two alternative approaches for elaborating a specific statistical model; see Koch and Gillings (2004) and Sterba (2009).

Model-based inference relies on a structural model specification with the aim of carrying out an empirical analysis of economic phenomena and making inferences about structural parameters. The theoretical model, though uncertain, is proposed under the guidance of economic theory and progressively specified. Inference on the unknown parameters is based on observational data.

In the *design-based* approach, the theoretical model is not specified while the relevant statistical model is specified and is based on a sample of observational data, drawn from the population, which are treated as experimental data, or

quasi- experimental data. From a methodological point of view, this is basically a descriptive approach, rather than a structural one. The estimators are obtained according to the sample design and available auxiliary information.

Models and causality

Quasi-experimental research can be seen as opposed to the research carried out on structural models. The experimentation is based on observational data, and researchers look for what is called a "natural experiment", *i.e.* a situation where individuals are assigned to treatment or control groups as if it were a real experiment. Estimates of causal effects are obtained by observing the results in outcomes between groups, where comparisons are made by using the regression discontinuity when the selection rules are known, or by using the difference-in-differences method by exploiting the variation in the timing of policy changes across individuals.

In the current applied econometric literature, one notes that a transition from models to methods is in progress. Quasi-experimental methods, such as "difference-in-difference", "regression discontinuity" and other related methods, have had the effect of overshadowing the role of economic theory in the specification of a model. In particular, this transition occurred in applied microeconomics such as to allow Angrist and Pischke (2010) to say that a "credibility revolution" is underway, capable of providing credible answers to the fundamental questions of policymakers about the evaluation of government programs.

More explicitly, an explanation of the difference between model-based and design-based approaches in the context of causality can be found by referring to the prevailing literature on the subject. There are two opposing views: one view, supported among others by Heckman (2008), states that causality should be model-based, in the sense that causality only exists within the framework of a specific theory that says "X causes Y". An opposing view, among supporters of which one can list Holland (1986) and Rubin (1974), states that causality is designed-based, in the sense that a claim of causality requires that it must be possible to design a manipulation which identifies whether "X causes Y". More explicitly, statistical units are randomly selected for different levels of treatments X and the consequent levels of the outcome Y are observed.

Those from the design approach emphasize the importance of causal identification, along with a careful specification of the causal parameters, the importance of which may be central when examining the impact of policies. On the other hand, those from the model approach emphasize the basic role of theory and the subsequent external validity.

It is important to stress some essential differences between the two approaches. For what concerns model specification, design-based approaches make use of simplified models with no formal derivation of the DGP for the observed data. Structural model-based approaches, such as presented in Section 2, derive the DGP from the assumed model. The parameter of interest for a design-based approach only concerns the estimation of parameters in a simplified model such as $E(y \mid X) = X\beta$. The parameter of interest for a model-based approach considers a formal model in the form of $g(y, X, \theta) = 0$ focusing on estimating θ.

Such a model is not always operational for a direct estimation of θ and may be supplemented by a linear approximation $L(y \mid X) = X\beta(\theta)$. This may raise the question of the relative relevance between β and θ.

As far as identification is concerned, design-based modelers believe that testing of assumptions is sufficient to meet the requirements for causal identification. Model-based users, in order to ensure causal identification, need to make explicit the functional form along with stochastic assumptions and any restrictions suggested by theory.

As far as reliance on the model is concerned, design-based studies attempt to explore basic predictions or, alternatively, try to evaluate a policy program. Model-based analysis usually aims at estimating unknown parameters and performing some kind of policy analysis or out-of-sample predictions.

Choosing one or the other

The choice between these two approaches depends upon the amount and type of information available, the objectives for which the model is proposed and, last but not least, on the experience of the investigator. It should be emphasized, however, that each of these approaches is based on different assumptions, and consequently can lead to different results regarding causality and inference in general. The plausibility of the results mainly depends upon the validity of the hypotheses specified, as well as on the properties of the statistical model used.

Concerning the design-based approach, a real risk associated with this class of models is what is called overfitting. This happens when the model explains the idiosyncrasies in the data instead of capturing the underlying relationship. In other words, one can run into situations where the model attempts to extract more information from the data than the one that is inherent to the data themselves. This happens because the final model is obtained after repeated regressions aimed at identifying the control variables to be included into the model. Since the model is repeatedly re-estimated, and since the data unavoidably have random errors, one ends up fitting the model to the noise in the sample, *i.e.* to the idiosyncrasies specific to the sample.

This problem arises when one cannot rely on randomized experiments and the data available are observational data that are not the outcome of a natural experiment. Much of the data in social sciences are of this nature. In the presence of overfitting, a statistical model mainly describes random errors instead of the underlying relationship. It will therefore be unsuitable for replicating the results out of sample, since its validity is based on features that only exist in the sample and not in the population that the sample is thought to represent. In this case, not only causality is missed, but even correlations are doubtful. External validity is thus put into question, one of the main goals of econometric modeling being to provide general results, *i.e.* results which are replicable outside the sample.

Endogeneity

Considering the model-based approach, a problem frequently encountered in this type of approach is that of endogeneity. In particular, when significant determi-

nants of the dependent variable Y, which are correlated with the X's, are not included in the model, an endogeneity problem is typically encountered. The main consequences are that the model parameters are not properly identified, while the direct estimates of the coefficients are biased and inconsistent. Therefore no causal relationship can be claimed in the presence of endogeneity, and correlations found cannot be correctly interpreted. To solve the problem, one needs to rethink the model in the perspective of simultaneous equations models, include data on missing variables and re-estimate the model. This challenge will be developed in Section 6.4.

In cases in which the researcher can control the data generating process, as happens by running well-designed controlled experiments such as double-blind randomized trials, most endogeneity problems disappear. In randomized experiments, the causality between right and left hand-side variables is assumed, since individuals, or sample elements, are randomly assigned to different values of the Xs. As widely argued, the difference between an estimated model based on observational data and a model based on quasi-experimental data is due to the fact that since the quasi-experiment has controlled for a large number of factors, the values of the Xs are deemed to be exogeneously (possibly, randomly) assigned. But at the same time this is also a problem, since this is not recognized as a true randomization.

Without attempting to be exhaustive, the following sections present two more recent examples dealing with causality in economics, respectively history-friendly simulation and information-theory based approaches.

4 History-friendly simulation and causality

Computer simulations

The notable increase in computing power and the simultaneous reduction in the cost of computers has made it possible to make wide use of computer-intensive methods such as computer simulations, presently applied in many branches of science. Computer simulations are employed for a variety of purposes, but in this section we examine an interesting field in the economic literature, known under the heading of "*evolutionary economics*" or "*history-friendly simulation*" (see for example Malerba and Orsenigo, 2002). This approach uses simulation as a tool for reproducing stylized facts. The purpose is to examine the relationships between the factors that are supposed to be the driving forces of a social process, the structure of which one wants to know, as it usually remains largely unknown or insufficiently understood.

History-friendly simulation seeks to offer a possibility for explaining observed historical developments with the aid of a simulation model, as long as one interprets the historical occurrence of an economic phenomenon as a realization of a stochastic process. Here, simulations provide a powerful tool allowing the derivation of the implications of a theory that are impossible to observe in practice. In a virtual world, such as the one offered by simulation, a history can be rerun in view of obtaining repeated realizations of a stochastic process, thus

acquiring repeated outcomes enabling the assessment of the degree of variability of these outcomes. These types of studies are known as *path-dependency* and have been proposed, among others, by Arthur (1994).

Repeating a simulation experiment with the same initial conditions and different parameter values generates a number of observations of the process that would otherwise not be observed. Virtual observations are thus created and these can be analyzed using conventional statistical tools. Given that the results obtained from the simulations can be manipulated by choosing the model and the parameters that reproduce the desired results, it is obvious that clear rules concerning how and under what conditions the simulations should be conducted are crucial for interpreting simulated histories. In particular, simulated histories have been used to study possible causal relationships between variables.

An example

In a study of the impact of chemists on the development of the dye industry in Germany, Brenner and Murmann (2003) propose a method that enables the researcher to make counterfactual analyses and study the effects of possible causal relations using simulation results. They also develop a sort of guideline for how such simulations should be carried out. Actually, simulation models make possible to run counterfactual analyses. Most explanations provided to understand social development contain, at least implicitly, the concept that if a certain action had -or had not - been taken, or if some factor had -or had not- been adopted, the outcome of the process would have been different.

The methodology proposed by Brenner and Murmann (2003) is developed in successive steps. The first step consists in specifying a model suitable for representing a good approximation of the process studied, where the parameters are estimated on the basis of available empirical evidence. Since the available information does not yield the exact values of all parameters, the second step consists in performing groups of simulations based on parameter values randomly selected within a predetermined range. This exercise is conducted by fixing the values of a subset of parameters and by varying, by means of simulation, only those that are explicitly the object of study. Therefore, a subset of parameter settings will be the same for all runs within a group of simulations, while other parameters vary within the group. The primary interest of conducting these experiments is studying possible effects of causes within a DGP. If the simulations are run on the basis of a structural model such as the one developed in Section 6 of this paper, one could say that there is a causal impact when a change of a specific state of the process leads to a particular characteristic of a later state of the process. Indeed, if the underlying model is deemed to be structural, one is actually studying the effects of a cause by varying only the parameter of interest and leaving all other parameters constant.

It is evident that to study causal effects following the simulation exercise described above, one needs to specify the relevant variables and the parameter settings. As pointed out by Brenner and Murmann (2003), simulation exercises are not fruitful unless they focus on a restricted set of variables that play a central role in the social process examined. In a nutshell, after having identified the

subset of explanatory variables and their interrelations, a simulation experiment is carried out in which these variables are varied in a systematic way. For each parameter value of an explanatory variable, several simulation runs are carried out. In doing so, one can obtain a sufficiently high number of simulations to ensure a correct statistical analysis.

On the one hand, unlike traditional simulation methods, the one described above offers the possibility of deriving knowledge concerning effects of causes that cannot be obtained on the basis of the observed real data. On the other hand, nothing ensures that these simulations produce robust knowledge concerning causal effects. The more the simulation experiment is well designed, *i.e.* based on a structural model considered to be causal, and the trials are numerous, the more confident one can be in the fact that the simulations produce a robust and useful information for investigating the consequences of causal relations.

5 Information-theory based approaches for causality analysis

Transfer entropy

Science philosophers such as Salmon (1984) have suggested that the impact of a cause on an outcome can be considered as the propagation of causal influence from the cause to the outcome. For Collier (1999), causation would thus be a transfer of information. Based on these ideas, information theory provides a wide variety of approaches for measuring causal influence among, for example, multivariate time series.

The transfer entropy approach (see Schreiber, 2000), based on transition probabilities containing the information on causality between two variables, was proposed for distinguishing driving from responding elements in a causal relationship. Barnett *et al.*(2009) have shown that, for normally distributed variables with linear relationships, the objectives of Granger causality (Granger, 1969) and transfer entropy are equivalent. However, it should be noted that these two approaches are based on different elements: Granger causality is based on autoregressive (AR) processes for which there may exist a problem of model identification, while the transfer entropy method is an information-theoretic approach that does not need assumptions on the structure of the process. It is based on the concept of Shannon-entropy and is suitable for linear and nonlinear relations. Its key assumption is that the sampled data should follow a well-defined probability distribution. Both approaches are however essentially descriptive, as they are not based on a structural modeling of the data generating process.

Transfer entropy measures the amount of information transferred from a variable X to another variable Y. This transfer information is deemed to represent the total causal influence from X on Y. The transfer entropy from X to Y can be interpreted as the gain obtained when using past information on both X and Y to predict the future of Y compared to only using the past information

of Y.

Direct or indirect influence

Since it may be difficult to distinguish whether this influence is direct or indirect through some intermediate variables, in order to establish whether the connection is of a direct or indirect type, the *direct transfer entropy* concept has been proposed in the literature by Duan *et al.*(2013) and others. From an operational point of view, the *difference direct transfer entropy* for discrete valued random variables has been introduced as an extension of the transfer entropy. Also, standardized measures for difference transfer entropy and difference direct transfer entropy have been introduced to measure respectively the connectivity strength of causality and direct causality. The detection of a direct information flow can be formulated as a problem of hypothesis testing. Considering the direct causality from X to Y, and admitting the possibility of an indirect causality from X to Y through a third variable Z, testing may be carried out by using the simulation approach for constructing resampling data or surrogate data.

The following section deals with a general framework for assessing causality in economics, *i.e.* structural causal modeling, based on the recursive decomposition of the data generating process seen as the mechanism and sub-mechanisms producing the outcomes.

6 Structural causal modeling: mechanisms, recursive decomposition, and causality

This section develops a general approach, *i.e.* structural causal modeling (SCM), that does not contradict the methods in the previous sections, but presents a more global view of the concept of causality. This general framework provides a sound basis for causal analysis if one wishes to go beyond description or prediction, with the purpose of explaining and understanding correlations among variables in a mechanistic perspective.

6.1 Mechanism and sub-mechanisms

Causation as generating process

That correlation, or statistical association, does not imply causality is an accepted position, though a large part of empirical work searching for cause-effect relations is actually based on statistical associations. Blossfeld (2009) has called this view the *causation as robust dependence* approach. In this perspective, as discussed by Cox (1992), a variable X is a plausible cause of another variable Y if the dependence between the two cannot be eliminated by introducing in the analysis additional variables. In this case, it is impossible to be sure that all relevant variables have been controlled for. Moreover, as Blossfeld (2009) stresses, because covariates are often correlated, parameter estimates depend

upon the specific set of variables included in the statistical model. In order to go beyond this *causation via association* approach, as Cox (1992) calls it, one needs what Blossfeld (2009) has coined a *causation as generative process* approach. As developed in Mouchart, Wunsch and Russo (2016 a, b), one should characterize the properties of the underlying data generating process, *i.e.* the mechanism behind the data. More generally, in a perspective of explanation and policy intervention, one needs understanding the plausible mechanism and sub-mechanisms generating the data in a particular context and during a specific period of time.

In the preceeding sections, the concepts of structural model and of causality do not necessarily correspond to the perspective developed in the previous paragraph. For example, Granger causality does not correspond to a classic definition of causality (see Little (2011) from a causal realist perspective) but to a concept of "self-predictivity" (see Florens, Mouchart and Rolin 1990, p.255) in the sense that prediction does not require structural modeling and that Granger causality is rather a concept of sufficiency to predict. To give another example, structural models in economics do not often fully spell out the mechanisms generating the data in a particular context and during a specific period of time. Economic theory, in this case, is considered as universally valid. In this section we attempt to waive these restrictions.

A causal framework

The following paragraphs propose, in the context of statistical models, a modelling strategy to operationalize the concept of causation as a generating process. The starting point is a set of variables X along with a statistical model in the form of a set of probability distributions

$$\mathcal{M} = \{P_X^\theta \quad \theta \in \Theta\} \tag{7}$$

where θ is a parameter characterising a probability distribution and $\mathcal{M}$ represents a set of plausible hypotheses concerning the data generating process (DGP). Representing the DGP by probability distributions characterized by a parameter, implies that what is "explained" by the statistical model is embodied in the parameter whereas what is not explained is embodied in the stochastic component of the probability distributions. For more details, see Mouchart and Orsi (2016).

In order to be "structural", the model should specify a plausible structure of the underlying DGP, relatively to a well-specified population of reference. Mouchart, Russo and Wunsch (2010) identify three main features of a structural causal model (SCM):

(i) A recursive decomposition of the joint distribution interpretable as an ordered sequence of sub-mechanisms, reflecting the causal ordering of the variables underlying the putative mechanism and making causal assessment feasible. Following Pearl (2000), this recursive decomposition can be represented by a *Directed Acyclic Graph* (DAG).

(ii) Congruence with background information: causal ordering of the variables is usually based on prior knowledge, often but not necessarily limited to economic theory, including information about the temporal ordering of variables and on the context. Background knowledge can also include preliminary analysis of data.

(iii) Invariance or stability of the recursive decomposition for a specified population of interest and historical time, in opposition to the idea of a "universal" theory or model.

6.2 The recursive decomposition

More specifically, once the vector of variables X is decomposed into an ordered sequence of p components, namely $X = (X_1, X_2, \cdots X_p)$ (with p typically much larger than 2), a recursive decomposition is a systematic marginal-conditional decomposition of the joint distribution of X, namely:

$$\begin{aligned} p_X(x \mid \theta) \quad = \quad & p_{X_p|X_1,X_2,\cdots X_{p-1}}(x_p \mid x_1, x_2, \cdots x_{p-1}, \theta_{p|1,\cdots p-1}) \\ & \cdot p_{X_{p-1}|X_1,X_2,\cdots X_{p-2}}(x_{p-1} \mid x_1, x_2, \cdots x_{p-2}, \theta_{p-1|1,\cdots p-2}) \cdots \\ & \cdot p_{X_j|X_1,X_2,\cdots X_{j-1}}(x_j \mid x_1, x_2, \cdots x_{j-1}, \theta_{j|1,\cdots j-1}) \cdots p_{X_1}(x_1 \mid \theta_1) \end{aligned} \tag{8}$$

where each $\theta_{j|1,\cdots j-1}$ stands for the parameters characterizing the corresponding conditional distribution $p_{X_j|X_1,X_2,\cdots X_{j-1}}$.

Once the number p of components increases, background knowledge, substantiated by analysis of the data and statistical tests, can provide a simplification of the factors in the form of conditional independence properties. More specifically, it is typically the case that the distribution of $(X_j \mid X_1, \cdots, X_{j-1})$ is known not to depend on some of the conditioning variables. Thus there is a subset $\mathcal{I}_j \subset \{X_1, \cdots, X_{j-1}\}$ of variables[1] whose actual relevance for the conditional process generating $X_j \mid X_1, \cdots, X_{j-1}$ is defined by the property

$$X_j \perp\!\!\!\perp X_1, \cdots, X_{j-1} \mid \mathcal{I}_j, \theta. \tag{9}$$

This property implies that the factor $p_{X_j|X_1,X_2,\cdots X_{j-1}}$ in (8) is actually simplified into $p_{X_j|\mathcal{I}_j}$ and $\mathcal{I}_j$ may be called the *relevant information of the j-th sub-mechanism*. Once $\mathcal{I}_j$ has been specified for each factor, (8) is condensed into

$$p_{X_1,X_2,\cdots X_p|\theta} = \prod_{1 \le j \le p} p_{X_j|\mathcal{I}_j,\theta_{j|1,\cdots j-1}} \tag{10}$$

This form is called a *condensed recursive decomposition*.

The condensed recursive decomposition is interpreted as a global mechanism decomposed into an ordered sequence of acting sub-mechanisms. For the sub-mechanisms to act autonomously, one should also add a condition of mutual

[1] More formally, rather than a subset of variables, $\mathcal{I}_j$ is a sub-σ-algebra of the σ-algebra generated by $(X_1, \cdots, X_{j-1})$.

independence among the parameters $\theta_{j|1,\cdots j-1}$, *i.e.* these parameters should be variation-free in a sampling-theory approach or a priori independent in a bayesian approach.

6.3 Causal assessment

Structural causal models

The recursive decomposition is the cornerstone of the explanatory power of a structural causal model because it endows the distribution P_X^θ with the interpretation that each component of the decomposition, *i.e.* the distribution of an outcome variable conditional on its immediate (or direct) putative causes, stands for one of the sub-mechanisms that compose the joint DGP of X. The recursive decomposition is built in such a way that the identified sub-mechanisms are interpretable from background knowledge. The order of the decomposition of X is crucial for the interpretability of the components as sub-mechanisms. Finally, invariance or stability of the model is required, as a major aim of SCM is to distinguish structural from incidental components of a data generating process.

Parameterization, finite-dimensional or infinite-dimensional (*i.e.* non-parametric or semi-parametric models), is an issue for measuring the effect of a causing variable on an outcome but is independent of the identification of the sub-mechanisms, this being achieved by the recursive decomposition. Note also that the "*ceteris paribus*" clause and the reliance on "stylized facts" are two techniques for isolating a mechanism from the context. These approaches may be useful for expositing an abstract theory but jeopardize the validity of a structural model that makes use of these techniques, because causal models in the social sciences are context-dependent. Econometric models used for empirical analysis are not always structural causal models, in the sense of this section. It can happen, for example, that forecasting is the main purpose for which the model is built, *e.g.* a statistical model which relies mainly on time series, or measurement studies that focus on constructing and summarizing economic macro data like unemployment, inflation, GDP and so on. Or again descriptive models which rely on economic theory and serve to describe and study the dynamic characteristics of a macroeconomic variable, without specifying the key aspects of the statistical model on which these descriptions are based. In this spirit, for example, is the contribution of Phillips (1958) who documented the inverse relationship between unemployment and inflation, relying on the analysis of observed data for the two variables. Models of this type have certainly their value; the main thing that unites them is that they have no claim to define and analyze a causal relationship. As a result, they cannot properly be used for evaluations and comparisons of alternative economic policies, as well as for out-of-sample counterfactual scenarios.

Structural and non-structural models

Structural and non-structural models are fundamentally different since they

make a very different use of economic theory and statistical methods. In other words, the two types of models differ in their theoretical properties rather than in their empirical performance. The difference becomes clear when referring to the interpretation of the empirical results. Results based on a structural causal model can be interpreted according to the postulated causal mechanisms, while the same cannot be said for those obtained by way of a non-structural model. Thus, classifying a model as structural or non-structural should be based on the modeling strategy, rather than on the empirical performance.

Can we speak of causality when the empirical analysis has been performed by means of a non-structural model? Strictly speaking, the answer is *no*. A structural causal model is supported *inter alia* by an economic theory, decomposes the DGP into an ordered sequence of sub-mechanisms, and is subjected to an empirical control of its stability viewed as an essential structural feature, in a particular context and during a specific period of time. A non-structural model can possibly also rely on economic intuition and/or on an economic theory but is not adequately controlled for its structural stability and, by definition, does not identify the time- and context-dependent causal sub-mechanisms as data generating processes.

6.4 Endogeneity and Causal Assessment

Choosing variables

Any variable one may consider to enter into a model has a story behind it and there must be a reason to include it in the model. This is typically provided by economic theory and background knowledge. In other words, model specification should rely on knowledge of the domain and on the theoretical framework which is at the origin of the empirical research and that can guide the search for relevant explanatory variables.

A thorough reflection on the reasons or motivations for the inclusion of variables at the right hand side of the model and their meaning, may help in identifying sources of endogeneity at the outset. More explicitly, the endogenous character of some variables of the right-hand side (rhs) arises under different contexts. A first situation is due to the presence of neglected confounding variables associated both with the left-hand variable (lhs) and with some of the rhs variables. A solution, data permitting, is to condition on these neglected variables on the rhs. When the neglected variables are not observable (*i.e.* , latent), one might make use of proxy variables or have resort to an instrumental variable approach. The interpretation of the empirical results may be assisted by background knowledge but, in any case, should remain cautious. A second situation arises in the case of simultaneity when the data of the lhs and rhs variables are aggegated on a same period of time. This can be due either to incomplete information or to a real simultaneity of the relationship between the lhs and rhs variables, within the period of time. In such a case, outside information can sometimes give the ordering of the variables and a consequent recursive decomposition. Otherwise, the absence of a recursive decomposition makes causal

assessment impossible. For example, consider a simultaneous equations model with two equations representing the demand and the supply of a specific good, to be estimated on the basis of price and quantity at market equilibria. When the model is not recursive, the equations do not represent the mechanism underlying the data generating process but may be interpreted as representing a hypothetical behaviour where the variable on the rhs, namely demand or supply, is exogenous.

Endogeneity once again

A frequent dilemma of simultaneity arises when, according to the theory considered, X causes Y and, roughly at the same time, Y causes X. In general, the problem of endogeneity cannot be solved by adding new control variables or increasing the sample size. In such a case, direct estimates will remain biased and inconsistent: the problem is of a structural nature. This suggests that a way of getting consistent parameter estimates is to consider, in the theory, a recursive decomposition of the joint distribution of X and Y, that would specify, if adequate data are available, the relevant sub-mechanisms producing X and Y.

Faced with this kind of problem, it becomes necessary to think more carefully about the processes that underpin the phenomena under study and examine in more depth the specification of the complete model. The simultaneity problem is widely recognized in economics, at a theoretical level, but does not always lead to the right solution in empirical analysis. Very often single-equation models continue to be proposed even when the problem of simultaneity (endogeneity) seems pretty obvious.

Often the reason that endogeneity is overlooked is due to the fact that data have been taken for granted, ignoring the underlying data generating processes. Any attempt to address endogeneity cannot leave aside the understanding of the DGP, since for the purposes of modeling it is important to know how the data are generated and what information is embedded in the data. In economics, as in other social sciences, data collected by others are often used, and there is no control over the process that generated these data.

Many advances in econometric theory were made to deal with problems related to simultaneity, the so-called simultaneity bias. In the past, Sims (1980) pointed out that there was a great discrepancy between the results obtained from current macroeconomic models and those obtained from descriptive statistical models, a-theoretical and solely data-based, a typical example being the so-called vector autoregressive (VAR) models.

Haavelmo's contribution

In his fundamental contribution, Haavelmo (1944) distinguishes between effects on outputs resulting from variation in inputs and explained by a theoretical economic model underlying a statistical model, from discovering causal relationships between variables based on empirical associations, derived from the analysis of data. According to Haavelmo, a causal effect must be based on a theoretical model that does not necessarily coincide with the empirical data

generating process. This is mainly due to the fact that economic theories, as is well known, have recourse to assumptions of *ceteris paribus* that do not always correspond to the actual DGP. Actually, in modeling a theory, one tries to make use of the minimum *ceteris paribus* conditions, but in most cases they cannot be eliminated completely since the theories model very complex systems that can hardly be written in all their particular details. In any empirical analysis one can never know for certain that the potential gap between the theoretical model and the DGP is bridged. What Haavelmo suggests is to model the residual component in a probabilistic approach. In this way one ends up obtaining a statistical model of the theory which, on the one hand, has coherent and interpretable implications in economics, and, on the other hand, explains the data sufficiently, with the characteristic that the unexplained component is made up of random and independent errors.

Such a statistical model provides a reliable basis for learning from data about the economic phenomenon of interest. It enables probing the adequacy of specific theories by way of substantively relevant questions. Modeling a real economic phenomenon requires representing a highly complex system whose properties should be derived from data that reflect a single non-replicable realization of a multivariate process. Obviously a DGP can only be approximated by simpler relationships, which characterize the data sufficiently for the purpose of analysis. A statistical model links economic theory to data when it reinforces an understanding of both the DGP and the theoretical model.

The Cowles Commission followed Haavelmo and, in the field of applied econometrics, more and more models began to appear with several equations in which endogenous variables appear among the regressors, that is models with simultaneous equations. The contribution of Haavelmo, many years after its publication, remains a cornerstone of econometrics.

A simultaneous equations model without recursivity does not specify however the sub-mechanisms of the DGP and therefore does not allow a justified causal assessment. When feasible, the recursive decomposition presented in Section 6.2 is probably the most efficient and statistically optimal way to solve the problem of simultaneity.

7 Discussion and Conclusions

Without intending to be exhaustive, this paper has presented some approaches aiming at causal inference in econometrics. Moreover, a structural causal modeling approach has been proposed, based on the concepts of mechanism and sub-mechanisms, and of recursive decomposition of the global mechanism.

The Need for Theory.
A policy maker must decide on the implementation of the best economic policies and knowledge of the causal link between variables is undoubtedly important.

This knowledge is also useful for understanding the functioning of the economic mechanism representing the structure within which this link exists. Understanding this mechanism, which can sometimes be very complex, requires theory but also other elements such as preliminary data analysis, intuition and reflexion, and the experience of those who have been working in the field, along with a relevant statistical model. The field information must also involve a thorough knowledge of the context which may be very different according to time and space.

Much has already been written, in the social sciences, on the process of theory building and testing (see for example the thorough discussion in Gérard, 2006). In Section 2 of the present paper, we distinguished two important facets of this process: the specification of the theoretical model and that of the statistical model. Following the work of Hubert Blalock and others, Duchêne and Wunsch (1985) have identified two components in the theoretical model, the main model elaborated in terms of concepts and the operational or auxiliary one based on available indicators of the concepts. This distinction takes into account the fact that, in many cases, not all concepts of the main model have a corresponding indicator in the data at hand. The choice of indicators depends not only on the availability of data but also on the definitions given to the concepts in the main theory. As Gérard (2006) has stressed, it is essential to give a precise meaning to each concept in the theory, that is to define the necessary characteristics of the concept. An applied econometric research should therefore always differentiate these three classes of models, the conceptual, the operational, and the statistical one.

In Section 3, the distinction in observational studies between model-based and design-based approaches was recalled. In particular, the Neyman-Rubin counterfactual average treatment effect (ATE) model, replicating a randomized experimental design, was criticized by Heckman (2008) due to the fact that it is a black-box that ignores the mechanisms producing the effect. Other critical comments addressed to the Neyman-Rubin model and more generally to randomized designs can be found in Russo *et al.*(2011) and in Deaton and Cartwright (2018).

Structural Causal Modeling

A specific definition of structural modeling was given in Section 6 as a basis for causal inference. The three main features of this structural causal modeling (SCM) approach are respectively: (a) The recursive decomposition of the joint distribution interpretable as an ordered sequence of meaningful sub-mechanisms (entities and activities), reflecting the causal ordering of the variables in the putative mechanism representing the DGP. This decomposition can be represented by a directed acyclic graph (DAG). (b) Congruence with background information, not necessarily limited to economic theory, concerning the context, the causal ordering and the role-function of the variables. (c) Invariance of the recursive decomposition for a specified population and context.

The advantage of the SCM mechanistic explanation approach is that it 'opens the black-box' in terms of the variables and their order responsible for

the data generating process (see in particular Russo *et al.*, 2019). SCM is well suited for drawing causal inferences from observational studies, in the absence of experimental designs. The approach has also been applied in demography (Gourbin *et al.*, 2017) and in epidemiology (Mouchart *et al.*, 2019). If the decomposition is fully developed, it avoids - to the best of one's knowledge - loss of exogeneity, due to confounding or simultaneity, that can hamper many other approaches as we have seen. SCM can specify the direct and indirect paths leading from causes to outcomes, distinguishing between mediators, moderators, and confounding variables. It can also take into account causal priority and reverse causation if the variables are ordered in time.

Of course, SCM is dependent on a good knowledge of the putative causes of the outcomes considered and of their order, *i.e.* of the various sub-mechanisms involved in the DGP. For this purpose, in addition to economic theory, a thorough review of the empirical literature on the topic of interest is mandatory.

The decisional process leading to the actual behaviors of the economic agents, in interaction with others and constrained by the institutional context, will unfortunately often remain unknown. Furthermore, as Herbert Simon (1959) has pointed out a long time ago, the actual decision-making process can be quite complex. The requirements of SCM are demanding in terms of causal knowledge and data availability, and many empirical works cannot meet these requirements. Even in these cases, developing an incomplete SCM on the basis of the information and data available, can nevertheless be a fruitful theoretical exercise.

Implications

Several practical implications can now be drawn. If putative causes of outcomes and their ordering can be assumed on the basis of background knowledge, and if relevant data are available, it can be recommended in observational studies to have recourse to structural models, such as SCM, in order to infer causality from the mechanism and sub-mechanisms considered to represent the DGP. Causality assessments, and therefore the measurement of the effects of causes, derived from non-structural models are, on the other hand, questionable. This might explain the failure of implications drawn from many such models.

This fact does not however imply that recourse to non-structural models be useless. They might have descriptive or predictive value, such as Granger causality and transfer entropy, or they can be used to study the possible effects of various scenarios, such as in history-friendly simulation and in agent - based modeling. Exploratory methods using *Big Data*, such as data mining, machine learning, and other techniques for extracting information from these data, can also come up with novel associations among variables and suggest new theoretical insights, especially when evidence is thin. They are especially useful for finding patterns in the data. Bareinboim and Pearl (2016) have examined how to take into account control of confounding, sampling bias, and generalization across populations in data-fusion from multiple sources. It should be pointed out, in this regard, that low quality *Big Data* can lead to wrong causal conclusions (see Brodie *et al.*, 2018). The evaluation of the quality of the data, big or

small, is always a prerequisite for sound causal modeling.

One of the problems with *Big Data* analysis consists in the 'Big'. In the field of public health and health economics, for example, (too?) many variables are now collected in a variety of forms, both structured and non-structured. Heinis and Ailamaki (2015) have even argued that 'forgetting' or shedding information should be part of today's data management. Methods exist to analyze these data in their various forms and to possibly link them together. But the sheer amount of variables available can be an obstacle for policy evaluation, taking into account the fact that a policy intervention should be based on cause-outcome relations. A structural modeling approach can often yield, for this purpose, the proper or strict subset of variables that are the most relevant from a causal viewpoint, *i.e.* the subgroup of variables that one should focus on.

For example, suppose that one is interested in evaluating the impact of increasing the tax T on cigarettes, from $T = t$ to $T = t + \alpha\,(\alpha > 0)$, in order to decrease smoking and eventually lung cancer and cardiovascular diseases. If C stands for consumption of cigarettes, it is not sufficient to compare C before and after this policy intervention, *i.e.* $C = c$ and $C = c + \beta$ with $\beta < 0$. In addition to tax T, consumption C is associated with a vector of variables Z. A change in some of these variables could also be a cause of the decrease observed in C. A structural analysis could tell us what are the most relevant causes, say X among the Z, that have to be controlled for, in order to decide that the decrease in smoking is due to the tax increase rather than to other causes.

All models are context-dependent

Finally, it should be stressed once again that no model is 'true', as models are always simplified representations of the real world. And no model in the social and economic sciences can be deemed universal, as models in these sciences are context-dependent and imbedded in historical time. The same may be said about theory.

References

Angrist, J.D. and Pischke, J.-S. (2010), The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics, *The Journal of Economic Perspectives*, **24**(2), 3-30.

Arthur, W. B. (1994), *Increasing Returns and Path Dependence in the Economy*, Ann Arbor, University of Michigan Press.

Athey, S. and Imbens, G. (2017). The state of applied econometrics: causality and policy evaluation, *The Journal of Economic Perspectives*, **31**(2), 3-32.

Bareinboim E. and Pearl J. (2016). Causal inference and the data-fusion problem, *Proceedings of the National Academy of Sciences*, 113(27), 7345-7352.

Barnett L., Barret A.B. and Seth K. (2009), Granger causality and transfer entropy are equivalent for gaussian variables, *Physical Review Letters* , **103**(23), 238701-1-238701-4.

Becker G. (1976) , *The Economic Approach to Human Behavior*, The University of Chicago Press (Paperback Edition, 1978).

Blossfeld H.-P. (2009). Causation as a generative process, Chapter 5 in: Engelhardt, H., Kohler, H.-P. and Fürnkranz-Prskawetz, A. (eds.), *Causal Analysis in Population Studies*, Springer, 83-109.

Brenner T. and Murmann J.P. (2003) The Use of simulations in developing robust knowledge about causal processes: methodological considerations and an application to industrial evolution, *Papers on Economics and Evolution* #0303, Max Planck Institute, Jena.

Brodie M.A. *et al.*(2018), Big data *vs.* accurate data in health research, *Medical Hypotheses*, 119, 32-36.

Collier J.(1999), Causation is the transfer of information, in H. Sankey (ed.), *Causation and Laws of Nature*, Dordrecht: Kluwer, 215-245.

Cox D.R. (1992), Causality: some statistical aspects, *Journal of the Royal Statistical Society*, Series A, **155**(2), 291-301.

Davidson R. (2015), Computing, the bootstrap and economics, *Canadian Journal of Economics*, **48**(4), 1195-1214.

Deaton A. and Cartwright N. (2018), Reflections on randomized control trials, *Social Science & Medicine*, 210, 86-90, and the whole issue of this journal focused on randomized controlled trials, pp. 1-90.

Duan P., Yang F., Chen T. and Shah S.L. (2013), Direct causality detection via the transfer entropy approach, *IEEE Transactions on Control Systems Technology*, **21**(6), 2052-2066.

Duchêne J. and Wunsch G. (1985), From theory to statistical model, in IUSSP: *International Population Conference, Florence 1985*, Vol. 2, IUSSP, Liège, 209-224.

Florens J.-P., Mouchart M. and Rolin J.-M. (1990), *Elements of Bayesian Statistics*, New York: Marcel Dekker.

Gérard H. (2006), Theory building in demography, Chapter 129 in G. Caselli, J. Vallin and G. Wunsch (Eds.): *Demography. Analysis and Synthesis*, Academic Press, 647-659.

Gourbin C., Wunsch G., Moreau L. and Guillaume A. (2017), Direct and indirect paths leading to contraceptive use in urban Africa. An application to Burkina Faso, Ghana, Morocco and Senegal,*Revue Quetelet / Quetelet Journal*, **5**(1), 33-70.

Granger C.W.J. (1969), Investigating causal relations by econometric models and cross-spectral methods, *Econometrica*, **37**, 424-438.

Haavelmo T. (1944), The Probability approach in econometrics, *Econometrica*, **12**, Supplement, iii-vi+1-115.

Heckman J. J. (2008), Econometric causality, *International Statistical Review* **76**(1), 1-27.
doi:10.1111/j.1751-5823.2007.00024.x

Heinis T. and Ailamaki A. (2015), Reconsolidating data structures, *Proceedings of the 18th International Conference on Extending Database Technology* (EDBT), 665-670.

Hicks J. (1980), *Causality in economics*, Australian National University Press, Canberra.

Holland P. W. (1986), Statistics and causal inference, *Journal of the American Statistical Association*, **81**(396), 945-960.

Hoover K.D. (2008), Causality in economics and econometrics, in Steven N. Durlauf and Lawrence E. Blume, eds, *The New Palgrave Dictionary of Economics*, 2nd edition, Palgrave Macmillan, The New Palgrave Dictionary of Economics Online. <http://www.dict C000569>
doi:10.1057/9780230226203.0209.

Johnston J. (1972), *Econometric Methods*, 2nd edition, McGraw-Hill.

Koch G. G. and Gillings D. B. (2004), Inference, design-based vs. model-based. In *Encyclopedia of Statistical Sciences*. Wiley.
doi:10.1002/0471667196.ess1235.pub2

Koopmans T.C. (1953), Identification problems in econometric model construction, in *Studies in Econometric Methods*, W.C. Hood and T.C. Koopmans (eds.), 27-48, New York, Wiley.

Leroy S.F. (2006), Causality in economics, University of California, Santa Barbara.
`http://econ.ucsb.edu/~sleroy/downloads/causal10-3.pdf`

Little D. (2011), Causal mechanisms in the social realm , chap. 13 in P. McKay Illari, F. Russo, and J. Williamson (eds), *Causality in the Sciences*, Oxford University Press, 273- 295.

Malerba F. and Orsenigo L. (2002.), Innovation and market structure in the dynamics of the pharmaceutical industry and biotechnology: toward a history-friendly model, *Industrial and Corporate Change*, **11**, 667-703.

Marschak C. F. (1953), Economic measurements for policy and prediction, in *Studies in Econometric Methods*, W.C. Hood and T.C. Koopmans (eds.), New York, Wiley. 1-26.

Moneta A. and Russo F. (2014), Causal models and evidential pluralism in econometrics, *Journal of Economic Methodology*, **21**(1), 54-76.

Mouchart M., Bouckaert A. and Wunsch G. (2019), Pharmacological and residual effects in randomized placebo-controlled trials. A structural causal modelling approach, *Revue d'Epidémiologie et de Santé Publique*, **67**(4), 267-274.

Mouchart M. and Orsi R. (2016), Building a structural model: Parameterization and structurality, *Econometrics*, **4**(23), 1-16. doi:10.3390/econometrics4010023,www.mdpi.com/journal/econometrics.

Mouchart M., Russo F. and Wunsch G. (2010), Inferring causal relations by modelling structures, *Statistica*, LXX(4), 411-432.

Mouchart M., Wunsch G and Russo F. (2016, a), Controlling variables in social systems - A structural modelling approach, *Bulletin of Sociological Methodology/ Bulletin de Méthodologie Sociologique*, **132**, 5-25.

Mouchart M., Wunsch G and Russo F. (2016, b), The issue of control in multivariate systems: A contribution of structural modelling, *Revista de la Sociedad Argentina de Estadística*, **13**, `https://revistas.unc.edu.ar/index.php/ReSAE/article/view/16355`

Pearl J. (2000), *Causality. Models, Reasoning, and Inference*, Cambridge University Press, Cambridge, revised and enlarged in 2009.

Phillips A. W. (1958), The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861-1957, *Economica* , **25**(100), 283-299.

Rubin D.B.(1974), Estimating causal effects of treatments in randomized and non randomized studies, *Journal of Educational Psychology*, **66**(5), 688-701.

Russo F., Wunsch, G. and Mouchart M. (2011), Inferring causality through counterfactuals in observational studies. Some epistemological issues, *Bulletin of Sociological Methodology*, **111**, 43-64.

Russo F., Wunsch G. and Mouchart M. (2019), Causality in the social sciences: a structural modelling framework, *Quality & Quantity*, doi 10.1007/s11135-019-00872-y.

Salmon W. (1984), *Scientific Explanation and the Causal Structure of the World* , Princeton: Princeton University Press.

Schreiber T. (2000), Measuring information transfer, *Physical Review Letters* , **85**(2), 461-464.

Simon H. A. (1959), Theories of decision-making in economics and behavioral science, *The American Economic Review*, **49**(3), 253-283.

Sims C. A. (1980), Macroeconomics and reality, *Econometrica*, **48**(1), 1-48.

Spanos A. (2006), Where do statistical models come from? Revisiting the problem of specification,*IMS Lecture Notes-Monograph Series*, 2nd Lehmann Symposium-Optimality, **49**, 98-119.

Sterba S. K. (2009), Alternative model-based and design-based frameworks for inference from samples to populations: From polarization to integration, *Multivariate Behavioral Research*, **44**(6), 711-740.
doi:10.1080/00273170903333574

Stock J.H. and Watson M.W. (2003), *Introduction to Econometrics*, Addison-Wesley.

Verbeek M. (2004), *A Guide to Modern Econometrics*, 2nd edition, Wiley.

Wold H. (1954), Causality and econometrics, *Econometrica*, **22**(2), 162-177.

Wooldridge J. M. (2013), *Introductory Econometrics: A Modern Approach*, South-Western, 5th edition.

Wunsch G., Mouchart M. and Russo F.(2018), Causal attribution in block-recursive social systems - A structural modeling perspective, *Methodological Innovations*,**11**(1), 1-11.
doi: 10.1177/2059799118768415.