STATISTICAL MODELLING AND CAUSALITY

Federica Russo*, Michel Mouchart**; Michel Ghins*, Guillaume Wunsch***

*Institut Supérieur de Philosophie, Université catholique de Louvain
**Institut de Statistique, Université catholique de Louvain
***Institut de Démographie, Université catholique de Louvain

Introduction

Philosophers and statisticians have been debating on causality for a long time. However, these discussions have been led quite independently from each other. An objective of this paper is to restore a fruitful dialogue between them. As is well known, at the beginning of the 20th century, some philosophers and statisticians dismissed the concept of causality altogether. It will suffice to mention Bertrand Russell (1913) and Karl Pearson (1911). Almost a hundred years later, causality still represents a central topic both in philosophy and statistics.

In the social sciences, most studies are concerned with the possible causes, determinants, factors, *etc.* of a set of observations. In particular, for planning or policy reasons, it is important to know what causes which effects. In order to attain causal knowledge, many social scientists appeal to statistical modelling to confirm or disconfirm their hypotheses about possible causal relations among the variables they consider, taking care of controlling for relevant covariates and especially for possible confounding factors.

To what extent can a statistical model say something about *causal* relations among variables? In this paper, we will attempt an answer by examining a special class of statistical models, *i.e. structural models*. The discussion, however, will not be confined to a mere examination of statistical methods, since a considerable effort will be made by considering causality from an *epistemological* perspective. This paper does not address the *nature* of causation itself; rather, we will be concerned with the question of how we come to uncover causal relations by means of statistical modelling.

I. Scientific knowledge

In order to gain cognitive access to real systems, scientists typically construct models; social scientists are no exception to this rule (Franck 2002). Broadly speaking, a model is an abstract object which may contain statements, schemes, figures and mathematical expressions designed to increase our knowledge of some *aspects* of reality. The notion of model is central to present-day philosophy of science; so far, however, no proposed account of what a model is has managed to attract universal consensus. The following remarks are thus aimed at singling out the characteristics of what we take models to be in the context of statistical modelling in social science.

A model is not a personal mental image of some reality. Models are not private psychological entities but intersubjective constructions which can typically be found in scientific textbooks and articles, and taught to students. For a social scientist in general, a model is not a set-theoretical structure that satisfies or makes true a given set of statements; this formal, mathematical, notion of model has influenced many a philosopher of science (Suppes 2002, van Fraassen 1980, Giere 1999). But social scientists take the core of a model to be a set of assumptions, *i.e.* statements, which aim at providing a simplified account of a complex reality. Of course, this view does not prevent the possible satisfaction of these assumptions by some set-theoretical structures, *i.e.* models in the mathematical sense.

Models however, as the term is used by social scientists, are not purely mathematical. Although the vast majority of scientific models contain mathematical assumptions, they usually contain non mathematical components such as explanatory statements, graphs, pictures, schemata *etc.* Giere (1999) compares models with geographical maps. But maps function essentially as pictures of some territory, whereas the models used by social scientists contain both pictorial elements, such as drawings, figures *etc.*, and components which are not figurative, such as statements. Provisionally, we will take a model to be an abstract object which permits partial cognitive access to some real systems. Systems are portions of reality which can be purely observable or also contain non observable parts.

Realists claim that at least some parts of a model correspond to some elements or aspects of a real system. This claim is notoriously controversial. Yet, the empirical success of a model, *i.e.* its ability to lead to correct, precise and even novel observational predictions, provides good grounds for the realist's belief in such partial correspondence. For many realists, a model is a – at least possible – representation of a real system. "Representation" is a hard-to-define, elusive concept. For social scientists a representation is not necessarily an image or a picture. Imaging is a particular case of representing in which there is some isomorphism or identity of form between the image and what is pictured. In a picture of a system, the organization of the elements of the picture mirrors the arrangement of the corresponding elements in the real system. Since models contain statements and since statements are not pictures of possible facts, *pace* Wittgenstein (1961), some model parts are not images. For our present purpose, it will be sufficient to say that parts of a model may correspond to some aspects of a represented system. No model pretends to capture all aspects of a real system. Some characteristics are consciously or unconsciously disregarded. Modelling consists in abstracting and results in constructing a simplified representation of a complex reality; thus, it always involves some degree of idealization. This is why some of the statements in the model are only approximately true. Modelling is approximating. Some idealizations and approximations are explicit and explain why a model does not fit the data perfectly; but some discrepancies with the observations or data are left unexplained.

Fitting the data certainly is a widely agreed upon condition of adequacy. But what does fitting the data mean? Making true predictions or statistically accurate predictions within certain limits surely is part of the story. A model itself determines what counts as a relevant observational result since those results must have corresponding counterparts in the model; typically, the observational results are denoted by the values of variables. However, a model by itself does not contain the criteria for deciding if a given observational result counts as a confirmation or a refutation. Such a decision is external to the model and hinges on the amount of error deemed acceptable for practical purposes. In some circumstances a model of geometrical optics may be considered adequate, while in other contexts we will have to use models of wave or quantum mechanical optics. Thus, good data fitting or empirical adequacy is also judged in function of the ends and purposes pursued (Giere 1999).

Minimal realists, such as van Fraassen (1980), restrict the correspondence of a model with reality to the data level. Unobservational components in the model may have counterparts in reality, but we have no means to assure ourselves that they exist because we do not have empirical access to them. As an empiricist, van Fraassen holds that the only cognitive access to reality is observation. Bolder realists, usually called scientific realists, claim that we have reasons to believe in the existence of some unobservable entities or processes as well. Some theoretical statements of our models can be asserted to be true provided they play a role in achieving empirical adequacy, especially in accounting for new observations. Some parts of schemata or figures may have a counterpart in reality and this can be attested by empirical evidence and measurements. For example, the angle between the atoms of a molecule can be measured. Usually, the correspondence of an element of a drawing *etc.* with reality can be expressed by a proposition. Here, we will espouse a moderate form of realism according to which models permit to have cognitive access to some unobservable aspects of real systems. Granted, models are falsifiable and science is not the locus of definitive and infallible truth.

This moderate realist position has a bearing on the status of causality. A social scientist doesn’t rest content with good data fitting but also attempts to construct models which provide a causal explanation of those data. A causal structure or relationship among variables is first articulated within the model. In some instances, it can be held as a working hypothesis. If a model achieves a good statistical fit with the data, especially when it succeeds in encompassing hitherto unknown data, then it is reasonable to believe that the model hits upon a real causal relationship. And we could capitalize on this in order to construct new models for different, but related, situations.

II. Data

Science tries to make sense out of observations, but the latter first have to be…observed. The selection of what one observes depends upon our underlying research questions and theoretical constructs. According to what we are looking for/at, we can use our eyes, a microscope, an electrocardiograph, *etc.* In demography for example, most data are collected by some sort of form: a census form, a birth certificate, a survey questionnaire, an inscription in a population register, and so on. The facts thus collected are however often far from perfect for the scientific enterprise. First, the data may contain voluntary or non-voluntary errors: erroneous income declarations, age-heaping in some less developed countries, sampling biases[1], *etc.*

Then, there is the issue of time. The time-ordering of events is a prerequisite for causal analysis: causes should precede their effects in time, though this criterion is disputed by some (Horwich 1987). A cross-sectional observational scheme, as one knows, does not enable us to disentangle age, period, and cohort (APC), or their durational equivalent, effects. For example, it may show that hearing decreases with age and so presently does smoking, but the former is an age effect and the second is a cohort effect. Retrospective studies give a time-dimension to the data and enable us to distinguish APC effects or to time-order events relative to their possible causes, but they are influenced by recall lapses and only those persons alive and present can obviously be interviewed. If one can afford them, prospective longitudinal studies avoid these pitfalls, but they can be affected by loss to follow-up and they thus possibly lead progressively to a selected population. Recall bias and selection bias are particularly difficult to model and correct (Freedman, 1999). In retrospective or prospective studies, one also has to choose an adequate time-frame: if the observation extends far into the future or the past, loss to follow-up and recall lapses respectively increase[2]. If the time-frame

[1] Including how the sample was drawn; for example, interviewing only hospital patients gives a biased image of the health of the whole population.

[2] As does the cost of the prospective survey!

is too short, we might miss an important lagged effect, such as the deleterious impact of a drug occurring only several years after use. Case-control studies can be of help here, though they are rarely used in the social sciences.

Another serious observational problem in demography and in the social sciences in general is the fact that many of our variables are abstract notions such as social status or intelligence, which cannot be observed directly. In structural modelling, one would call them *latent variables*. For measurement and comparisons, we must therefore agree on a common definition of these abstract concepts and on a procedure for deriving empirical measures satisfying the reliability (repeatable measures) and construct validity (accurate reflection) criteria (Babbie 2000). The problem is especially acute when contexts differ: may we use the same definition and indicators of education in Europe and in Africa, for example? Probably not. Taking into account the purpose of the study, the definition of a concept should determine a partition such that an object either is or is not subsumed under the definition, notwithstanding the possible existence of some cases of fuzzy membership. The possible various facets or dimensions of the concept should be clearly pointed out. Such a procedure is furthermore helpful in selecting the multiple indicators of the concept needed for empirical measurement. Even biological variables such as sterility are not always obvious or clearly defined; one needs in this case to distinguish between primary and secondary sterility and to avoid taking sub-fertility for infertility (Habbema *et al.* 2004). The indicators of these forms of (in)fertility will not be the same.

III. Causality and Statistical modelling

The statistical model

Formally, a statistical model **M** is a set of probability distributions; more precisely:

$$M = \left\{ S, P^{\theta} \theta \in \Theta \right\}$$

where S, called the *sample space* or observation space, is the set of all possible values of a given observable variable (or vector of variables) and for each $\theta \in \Theta$, P^{θ} is a probability distribution on the sample space, also called the *sampling distribution*; thus, θ is a characteristic, also called *parameter*, of the corresponding distribution and Θ describes the set of all possible sampling distributions belonging to the model. The basic idea is that the data can be analyzed *as if* they were a realization of one of those distributions. For example, in a univariate normal model, the sample space S is the real line and the normal distributions are characterized by a bivariate parameter, for instance the expectation (μ) and the variance (σ^2); in this case: $\theta = \left(\mu, \sigma^2\right)$.

A *statistical model* is based on a *stochastic representation of the world.* Its randomness delineates the frontier or the internal limitation of the statistical explanation, since the random component represents what is *not* explained by the model. For instance, in the simplest case of repeated measurements (x) of the weight of a given object, the statistical model, derived from the equation $x = \mu + \varepsilon$, explains the expected measurement as the "true" weight (μ) of the object to which is added an unexplained error of measurement (ε) modelled as a random variable with zero mean.

Accordingly, a *statistical model* is made of *a set of assumptions* under which the data are to be analyzed. Typical assumptions of statistical models are: the observed random variables follow or not identical distributions; the observations are, or are not, independent; the basic sampling distributions are, or are not, continuous and may pertain, or not, to a family characterized by a finite number of parameters (*e.g.* the normal distributions). When we deal with multivariate models, adequate assumptions might involve linearity (in the parameters and/or in the variables), non measurement error, or non correlation of error terms.

In particular, we often assume the model to be linear or approximately so. This is a matter of convenience, since a linear model is easy to manipulate, its parameters are easily estimated and the resulting estimators have nice properties. Often assumed, linearity may also be tested. The same holds for normal distributions. We may also assume that variables are measured without error and that the errors are not correlated with the independent variables.

If the statistical assumptions are satisfied, the statistical model correctly describes co-variations between variables, but no causal interpretation is allowed yet. In other words, it is not necessary that causal information be conveyed by the parameters, nor is it generally legitimate to give the regression coefficients a causal interpretation. It is worth noting that in specifying the assumptions typical of a statistical model, the problem is not to evaluate whether or not an assumption is true. In a sense, if a model-builder could prove that an assumption were (exactly) true, this would not be an assumption anymore, but a description of the real world. Rather, the main issue is to evaluate whether an assumption is useful, in the sense of making possible a process of *learning-by-observing* on some aspects of interest of the real world.

Statistical inference and structural models.

Statistical inference is concerned with the problem of *learning-by-observing* and is *inductive* since it implies drawing conclusions about what has not been observed from what has been observed. Therefore, statistical inference is always uncertain and the calculus of probability is the natural, and in a sense logically necessary tool (see *e.g.* de Finetti 1937, Savage 1954), for expressing the conclusions of statistical inference. Therefore, the stochastic aspect of statistical models involves a stochastic representation of the world *and* a vehicle for the learning-by-observing process.

Here, two aspects ought to be distinguished. On the one hand, learning-by-observing conveys the idea of learning about some features of interest, namely the characteristics of a distribution or the values of a future realization. On the other hand, learning-by-observing is also concerned with the problem of accumulating information as observations accumulate. These two aspects actually refer to the usefulness of the model. Structural models are precisely designed for making the process of statistical inference meaningful and operational.

To better understand the idea behind this last claim, it is worth distinguishing two families of models. In the first family we find *purely statistical* models, also called associational or descriptive models, exploratory data analysis or data mining. In these approaches, the assumptions are either not made explicit or restricted to a minimum allowing us to interpret descriptive summaries of data. Interest may accordingly focus on the distributional characteristics of one variable at a time, such as mean or variance, or on the associational characteristics among several variables, such as correlation or regression coefficients. It is worth noting that the absence or the reduced number of assumptions constituting the underlying model make these associational studies insufficient to infer any causal relations.

The second family consists in the so-called *structural* or *causal* models. "Structural" conveys the idea of a representation of the real world that is stable under a large class of interventions or of modifications of the environment. As a matter of fact, structural models incorporate not only observable, or manifest, variables but also, in many instances, unobservable, or latent, variables. The possible introduction of latent variables is motivated by the help they provide in making the observations understandable; for instance, the notion of "intelligence quotient" or of "associative imagination" might help to shape a model which explains how an agent succeeds in answering the questions of a test in mathematics. Thus a structural model may capture an underlying structure of the world. Modelling this underlying structure requires taking into account the contextual knowledge of the field of application in order to uncover the structural stability.

Structural models are also called “causal models”. Here, the concept of causality is *internal* to a model which is itself stable, in the sense of *structurally stable* (see below). The characteristics, or parameters, of a structural model are of interest, because they correspond to intrinsic properties of the observed reality and can be safely used for accumulating statistical information, precisely because of their structural stability. In this context, a structural model is opposed to a “purely statistical model”, understood as a model that accounts for observable regularities without linking those regularities to stable properties of the real world.

Besides the assumptions of stability, or of invariance, the construction of structural models typically involves other assumptions such as: covariate sufficiency, no-confounding, independence of error terms, recursivity *etc.* It is worth pointing out that the correctness and usefulness of structural or causal models also rest on a set of untested, and often untestable, assumptions, which nevertheless play a fundamental role. In particular, at the *building* stage the direction of time is usually assumed to point from the past to the future. However, this direction may be reversed at the *inference* stage; for instance, in a medical application, inference may concern a diagnosis conditional on observed symptoms, even though the pathology has been active before the symptoms appear.

Conditional models and exogeneity

By means of causal modelling, the social scientist often has to acknowledge that all the variables of interest cannot be taken into account by a unique structural model, because of too severe an environmental instability. The purpose of modelling becomes, under such circumstances, to uncover some structural, or stable, aspects of an unstable reality. In order to do that, a usual strategy consists in separating the data into two parts: $X = (Y, Z)$, along with a *marginal-conditional decomposition*, namely, in terms of density functions:

$$p(x \mid \theta) = p(z \mid \theta) \cdot p(y \mid z, \theta)$$

where the *marginal model*, constituted by $\{p(z \mid \theta), \theta \in \Theta\}$, describes how the data Z (alone) have been generated. The *conditional model*, constituted by $\{p(y \mid z, \theta), \theta \in \Theta\}$, describes the conditional distribution of $(Y \mid Z)$, that is to be interpreted as describing the data generating process of the random variable y relatively to a particular value of *Z* and, therefore, does not take into account the randomness of Z. In social science contexts, the standard practice is to gather in the marginal model the more unstable aspects of the real world, as conjectured on the basis of background and contextual knowledge. This makes plausible the assumption that only the conditional component of the model is structural. The burden of the assumptions then bears on the conditional distribution of the variables Y, leaving virtually free the marginal distributions of the conditioning variables Z.

In such a case, only the parameters describing the conditional distribution are considered of interest and, reciprocally, a variable Z is said to be *exogenous*, for a given parameter of interest, if this parameter of interest depends only on the parameters identified by the model conditional on the exogenous variables, and if the parameters identified by the marginal model and by the conditional model respectively are independent in a Bayesian sense (or variation-free in a classical sense).

This leads us to a first approach to the concept of causality: *causality is exogeneity in a structural model*. That is to say, a causal variable is an exogenous variable in a structural conditional model. A simple example, given in the Appendix, illustrates the notions just introduced.

Beyond exogeneity, towards an epistemological concept of causality

Exogeneity only provides an *operational* concept of causality. But causality is such a rich concept that to explicate it just in terms of exogeneity does not do justice to the numerous attempts both in the philosophical and scientific literature to account for it. We now examine two *epistemological* aspects of causality. On the one hand, a *single* "marginal-conditional decomposition" may not properly account for the complexity of the relationships within a large number of variables of interest. On the other hand, the concept of causality requires, at some stage, to acknowledge the role of time. This second aspect will be distinguished from the first one by labelling it "atemporal", in order to pinpoint aspects for which the role of time is not essential.

We begin with those atemporal features. The issue of determinism is definitely controversial. Structural equations may be considered functions in which probabilities come in through error terms representing some lack of knowledge (Cartwright 1989, Hausman 1998, Woodward 2003). We may otherwise adopt the statistician's viewpoint: the world of the statistician is stochastic – the error terms representing what is *not* explained. In truth, the epistemological perspective here endorsed allows us to set aside the *metaphysical* problem of determinism, *i.e.* whether or not the world is actually deterministic.

The complexity of the relationships among a large number of variables of interest requires to take further features into consideration. By excluding causal loops, *recursiveness* makes causal analysis easier; however recursive models are not always feasible because of insufficient observable data (for more details see Wold 1949 and 1954).

Another crucial assumption is called *covariate sufficiency*, namely the conditional distribution of the variable of interest (Y) only depends on the retained exogenous variables (Z) and not on other variables (W). This is often interpreted as meaning that those exogenous variables are direct causes of the dependent variables and that these are all the variables explaining the variability in the endogenous variable Y (for a similar view, see *e.g.* Stone 1993). That condition bears on the causal mechanism and raises however several practical and conceptual difficulties. Firstly, covariate sufficiency is a property of conditional independence that is Y and W are mutually independent conditionally on Z, written as ($Y \perp W \mid Z$). This property is, in principle, testable. However, a hypothesis may be statistically acceptable because of data weakness, although contextual knowledge suggests it should be rejected. Secondly, covariate sufficiency may keep an underlying causal mechanism hidden because, as a property of conditional independence, covariate sufficiency is an informational relationship rather than a structural property. For instance, if W causes Z and Z causes Y (diagrammatically: $W \to Z \to Y$), Z may be sufficient with respect to W (*i.e.* the prediction of Y knowing Z is not improved by a further knowledge of W) although W clearly is part of the causation of Y.

By ruling out other factors liable to screen-off the impact of the covariates we took into account, the assumption of *no confounding* has a complementary role to covariate sufficiency and accordingly faces similar practical and conceptual difficulties. This assumption has been discussed at length around the so-called Simpson's paradox the resolution of which is based on the fact that the two conditions $Y \perp Z \mid W$ and $Y \perp Z$ are not linked by any logical implication, neither their negations. It should however be noticed that, as pointed out by Stone (1993, 459) "good explanations (of the notion of confounding) are surprisingly rare" in spite of "appearing in most epidemiology texts and (being) ubiquitous in the quantitative social sciences".

The *invariance* condition of a structural model is actually a complex issue. This is a condition of stability not of the causal variables, but of the causal relation itself. The idea is that each variable is determined by a set of other variables through a relationship that remains *invariant* when those other variables are subject to external influence. It is in this sense that we call the model “structurally stable”. This condition allows us to predict the effects of changes and interventions. *Stability* of distributions is also assumed to ensure that the (conditional) independencies between variables will not be jeopardized by variations in the parameters (Pearl 2000 calls this condition “stability”, but is also known as “DAG-isomorphism” (Pearl 1988) or “faithfulness” (Spirtes *et al.* 1993)).

By *causal asymmetry*, we mean that the two propositions “*Z* causes *Y*” and “*Y* causes *Z*” cannot hold at the same time: either the direction of causation is known and is unique or the direction of causation is unknown and this is sometimes called a situation of simultaneity in the model.

The temporal aspect of a causal mechanism is generally accepted as an important component of the modelling effort. The first issue is evidently that of the *direction of time*. As mentioned above, we assume that the causal mechanism follows the direction of time from the past to the future. Indeed, the problem of feedback loops is solved taking into account *causal priority*, *i.e.* causes precede effects in time: $Z_t \to Y_{t'} \to Z_{t''}$, where $t < t^{'} < t^{''}$. That is, Y is an intervening variable between two temporally distinct values of *Z* and these two together guide the choice of *causal ordering*, *i.e.* the temporal order in which variables are observed.

Hypothetico-deductive methodology

Causal modelling is not concerned with the question whether the true causes of an effect are disclosed but rather with the issue of a good representation (of the world) which embodies a *causal* mechanism. For a moderate scientific realist, the process of model building involves a continuous interaction between a prior knowledge of the field and a sequence of statistical procedures for elaborating and testing the successive hypotheses, with the understanding that the true causes are attainable at least in principle. In practice, causal attribution incorporates accordingly educated guesses. Because, as explained above, causality is a property *within* a structural model, rather than a *prima facie* empirical problem, it is impossible to *deduce* causes from correlations in a purely statistical model. But this does not lead to a Humean sceptical despair: causal modelling is indeed a promising tool for causal attribution. In other words, causal inference is vindicated: what vindicates causal models is the hypothetico-deductive strategy employed in much of contemporary science.

Three remarks are in order. Firstly, a hypothetico-deductive methodology is employed in case we have at our disposal enough well confirmed theories and background knowledge to formulate a prior causal hypothesis. If this is not the case, exploratory statistical methods provide a useful tool to detect in the data a tentative structure to be further analyzed by means of the structural modelling methodology.

Secondly, from a logical point of view, a model is false if at least one of its assumptions is false. However, as mentioned above, models are deemed to be useful rather than true. Consequently, also false models can be *useful* depending on the problem at hand. In other words, although the model is not the “true” model, it can more or less faithfully represent reality and thus be useful in order to understand (at least) some aspects of the world.

Thirdly, deduction ought not to be confused with the hypothetico-deductive methodology and induction ought not to be confused with the inductive methodology. In fact, there is a sharp difference between hypothetico-deductive or inductive strategies on the one hand, and deductive or inductive inferences, on the other hand. The former are procedures for testing or formulating hypotheses, whereas the latter are types of inference. On the one hand, H-D strategies *confirm* (or disconfirm) hypotheses, while inductive strategies are employed to *discover* hypotheses. On the other hand, deduction is a *non-ampliative inference* from what is known to what is known, whereas induction is an *ampliative* inference from what is known to what is *not* known yet.

Structural models discussed above are hypothetico-deductive (H-D) models, for which empirical testing is performed through two stages:

(i) prior theorizing of out-of-sample information, including in particular the selection of variables deemed to be of interest, the formulation of a causal hypothesis (also called the conceptual hypothesis), *etc.*;

(ii) iteratively:

a. building the statistical model;

b. testing the adequacy between the model and the data to accept the empirical validity or non validity of the causal hypothesis.

Causal modelling requires accurate knowledge of the causal context: previous studies, well confirmed scientific theories or background knowledge are essential. The conceptual hypothesis states a hypothesized causal structure, *i.e.* a causal claim about the hypothesized link between conceptual variables to be put forward for empirical testing. However, the evaluation of the conceptual hypothesis cannot be done a priori but requires empirical testing. Indeed, this is what the whole statistical set-up is built for. Differently put, causality is a matter of confirmation, or borrowing the statistical vocabulary, a matter of accepting or rejecting a given hypothetical model. So, if statistical assumptions and structural stability are satisfied *and* if the model fits the data, then the hypothesized causal link is provisionally accepted.

This strategy is hypothetico-deductive because the causal claim is not inferred from the data, as in inductive methods, but confirmed or disconfirmed in the given causal context and relative to the structural model. This is, in particular, at variance from algorithms in TETRAD (Spirtes *et al.* 1993), which are supposed to allow the *inductive inference* of causal relationships from a data set regardless of any prior conceptual framework. Note that H-D strategy so described is general enough to provide a scheme of scientific practice, without commitment to a strict covering law model or to the use of any particular measure of confirmation; on this point see Williamson (2005).

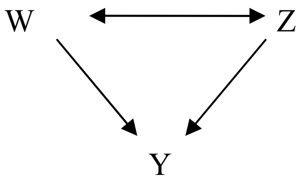
IV. The population and the individual: two levels of causation

The problem of levels of causation arises because causal conclusions drawn from statistical models concern populations as well as individuals, although probability distributions and their parameters are typically defined relative to the population. Nonetheless, populations *are* made up of individuals. Regardless of how individuals influence group behaviour and *vice versa*, how can we make sense of the following issue? Epidemiological studies establish a causal relationship between smoking and lung cancer. This conclusion, namely smoking causes lung cancer at the population level, is based on data concerning individuals of a particular population. Yet, one might be interested in Harry’s chance of getting lung cancer given that he smokes, or in the probability that his smoking caused him to contract lung cancer.

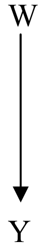
This remark leads us to the fruitful distinction between population-level causation and individual-level causation, although we are left with at least two problems. On the one hand, we face the *methodological* problem of handling the heterogeneity of individual characteristics and, on the other, the epistemological problem of relating the two levels. It is worth pointing out that to advocate two levels of causation is *not* tantamount to saying that causation operates in a different manner at the two levels.

The methodological problem can be rephrased as follows. The concept of causality that emerges from structural modelling involves two complementary aspects. Firstly, the validity of a statement such as “smoking caused Harry’s cancer” depends on the validity of a model deemed to be structural. Secondly, the concept of a structural model is a statistical concept that refers to some population of reference, and a statement such as “smoking causes lung cancer” virtually refers to each individual of a population of reference, because of the assumption of homogeneity. Also, the H-D methodology just discussed points to the issue that, as a matter of fact, the structural model on which causality is based is not justified *a priori*, but has to be uncovered by blending field knowledge and statistical methods.

When *W* causes *Y*, according to a structural model referring to a specific population of reference, we cannot be sure that this causal relation will hold for *every* individual in the population. For instance, suppose that smoking (*W*) and stress (*Z*) mutually interact and both cause cancer (*Y*):



and suppose that Z is not observable or not observed; in such a situation the statistical model might be:



Suppose now that a fourth (categorical and latent) variable *V* distinguishes the effect of cigarette smoking on stress: for some individuals ($V = 1$) cigarette smoking relieves stress, and for others ($V = 0$ it does not. Suppose also that stress increases the probability of cancer (through the weakening of the immune system). Then, if only *W* and *Y* are observed, the scientist might conclude that (i) globally smoking still causes cancer because the proportion of individuals with $V = 1$ is not very high, and (ii) for some individuals ($V = 1$) forbidding cigarette smoking increases the risk of cancer because of deterioration of the immune system. We are confronted with a *theoretically* difficult situation that also has *practical* consequences.

To begin with, some non-observed variables can be subsequently observed. For instance, in multilevel analysis, categorical variables by level can be introduced in a second step. Secondly, there is a deeper and more fundamental underlying problem, namely the heterogeneity of individuals in the same population.

Heterogeneity of individual characteristics, *i.e.* the problem raised by variables that are causal *and* non observable, has the consequence that our models, although structural, are nonetheless imperfect. The complexity of the problem of heterogeneity has led to an extremely vast body of literature. Solving the problem goes far beyond the scope of this work, and we shall be content to mention that it has been tackled and criticized from several perspectives. For instance, multilevel analysis (see Courgeau 2003) tries to get at an understanding of individual and population behaviours under the assumption that the grouping of individuals according to various levels introduces an influence of the group on its members and *vice versa*. In health sciences, *e.g.* in biostatistics, frailty models are used to model heterogeneity of populations (Vaupel and Yashin, 2001). The counterfactual approach tries to cope with the problem of heterogeneity by setting different hypothetical initial conditions. However, in spite of the intuitive appealing of counterfactual reasoning, the non observability of such different settings has been the object of several criticisms (see Dawid 2001)

Let us now come back to the practical problem. Can a physician decide whether to prescribe a treatment or not on the basis of a causal model? The answer is not straightforward. Let us see why. On the one hand, the answer ought to be positive, were the physician to be sure of covariate sufficiency. However, this is not a realistic situation. Might the physician not blindly follow what the structural model prescribe? He would also take into account the specific characteristics of his/her patients, possibly leading to the discovery of new causal variables. Progress in health sciences partly depends on the questioning of our models in particular circumstances.

Nonetheless, this practical problem hides the epistemological one. In fact, the physician's decision depends in the relationships between causality at the population level and at the individual level. The physician's incertitude about the population level causal relation is due to the methodological difficulties mentioned above: we wonder (i) what the causal (*i.e.* exogenous) variables are and whether it is possible at all to provide a sufficient list, and (ii) what mechanisms operate among the variables deemed to be causal.

In spite of this, what we discover about the average relation between smoking and lung cancer, *i.e.* at the population level, can guide causal attribution in the case of Harry through a simple tool of probabilistic reasoning, namely Bayes' theorem. In fact, Bayes' theorem allows us to calculate the posterior probability of the cause for a given individual, provided that the population risk is interpreted as a prior probability for this individual.

Conclusions

Statistical models are stochastic representations of the real world. Structural models are conditional statistical models characterized by parameters that are stable over a large class of interventions. It is *within* these structural models that we are allowed to formulate causal statements. From a statistical viewpoint, causality can be operationally defined in terms of exogeneity in a structural model. Nonetheless, it is not hard to see that exogeneity is not enough if we consider causality from an *epistemological* perspective.

A more complex and rich concept of causality may be worked out once we consider the fundamental role of assumptions made in structural models, of background and contextual knowledge and of the hypothetico-deductive methodology. These considerations, however, do not enable us to attain a unique and consensual *definition* of causality. But at least this allows us to attain a concept of causality that is *internal* or relative to the structural model itself. This is not to deny the existence of causation. Rather, this is to emphasize that our knowledge of causal relations – at least in the social science – depends on structural models that mediate the epistemic access to causal relations.

Appendix

In this appendix we illustrate through a very simple example some basic notions regarding structural modelling and exogeneity. We also take this opportunity to show the progressive nature of model specification.

Let us consider the following economic situation. We start with annual data on the price $p(t)$ and the quantity sold $q(t)$ of a well-defined holiday destination; imagine, for instance, a full-board stay in a three star hotel during the first two weeks of August on a well-known beach. Without further specification, a simple model would consist in assuming that the pair of variables $\left(p(t),q(t)\right)$is generated by a bivariate normal distribution with expectations $(\mu_{p(t)},\mu_{q(t)})$, variances $\left(\sigma^2_{p(t)},\sigma^2_{q(t)}\right)$ and covariance $\sigma_{p(t),q(t)}$. The argument "t" in the parameters indicates that the characteristics of the process change over time: in the statistical jargon the parameters are purely "incidental". Although acceptable from an economic point of view, this parametric instability leads to a not operational model: each new observation $\left(p(t),q(t)\right)$ introduces 5 new parameters and no observation might possibly put the model into question, let alone refute it. We should add more structure, based on contextual, rather than empirical, knowledge.

Suppose now that the market operates as follows. In January, each year, the tour operator prints a catalogue in which the price $p(t)$ is announced. From January to July the buyers enter their orders, compounding the aggregated quantity $q(t)$ at the announced price $p(t)$. This suggests to decompose the joint distribution, generating $(p(t),q(t))$, into a *marginal distribution* generating $p(t)$, with parameters $\left(\mu_{p(t)},\sigma^2_{p(t)}\right)$ and a *conditional distribution* generating $(q(t)\,|\,p(t))$: $P[p(t),q(t)]=P[p(t)]\cdot P[q(t)\,|\,p(t)]$.

The motivation for this decomposition is that the economic context suggests that the *marginal process* captures the behaviour of the tour operator (the supply side) where the expectations $\mu_{p(t)}$ expresses, in particular, the (changing) expectations on the evolution of the costs (cost of kerosene for air flight, local cost of accommodation *etc.*) and the variances $\sigma^2_{p(t)}$ reflects the indeterminateness of the behaviour, in particular due to changing levels of uncertainty. Similarly, the economic context suggests that the *conditional process* generating $(q(t)\,|\,p(t))$ captures the consumer behaviour (demand side), taking in particular into account that in our societies once a catalogue is printed the buyer does not try to bargain on the price: (*s*) he either accepts the price, and buys, or rejects the price, and does not buy (and possibly looks for a cheaper offer…) and therefore behaves *"as if"* the price were fixed, *i.e.* is not random.

Suppose now that we are only interested in the supply side behaviour. As shown above the parameters $\left(\mu_{p(t)},\sigma^2_{p(t)}\right)$ cannot be estimated because of their incidental nature. Thus we should add more structure, typically in making these expectations and variances known functions of past observations and of a finite number of parameters (for instance α and β in: $\mu_{p(t)}=\alpha+\beta\,p(t-1)$). By doing so, we aim at capturing a parameter constancy: this is one of the major objectives of statistical modelling.

Suppose now that we are only interested in the demand side. Again for the same reason as before we should endow the conditional expectation $E[q(t)\,|\,p(t)]$ and the conditional variance $V[q(t)\,|\,p(t)]$ with some structure. For the sake of simplicity, let us assume that $E[q(t)\,|\,p(t)]=\gamma+\delta p(t)$ and $V[q(t)\,|\,p(t)]=\sigma^2$, which is independent of t (an homoscedasticity assumption). Apart from the so-introduced structural stability, we should now face the issue of exogeneity; this is the question whether it is “admissible” (*i.e.* without loss of information) to only consider the conditional model, *i.e.* to treat the price $p(t)$ *“as if”* the price was not random. The exogeneity of the price involves two different aspects. Firstly, do we grant that the conditional distribution actually captures the behaviour we are interested in? Thus, do we grant that the demand really behaves under “given” price, that only the (current) price is relevant for the buying behaviour, that the conditional variance really is constant etc?

Quite a different question is the following: do we accept that the randomness of the price gives no information on the parameters of the conditional model? In other words, do we accept that the unknown value of the parameters of the conditional model is independent of the values of the parameters of the marginal model? If so, the statistical jargon would say that the parameters of the marginal models and the parameters of the conditional model are “variation-free”, in a sampling theory approach, or are “a priori independent” in a Bayesian approach. Thus the exogeneity problem involves two aspects. The first one regards the parameters of interest: are we really - *i.e.* contextually - interested in the parameters of the conditional model? The second aspect regards the statistical efficiency: may we ignore the random character of the conditioning variable?

Does the exogeneity of the price in the conditional model generating $\left(q(t)\,|\,p(t)\right)$ imply that “the price causes the quantity”? The answer is “yes” in quite a specific sense, *i.e.* under several provisos. Firstly, it is a concept of causality internal to a specific model: the price $p(t)$ “causes” the quantity in a particular model which is assumed to represent a demand behaviour but the price does not cause the quantity “in general”. Secondly, this concept of causality is relative to a particular family of models, namely models that are both conditional and structural: the price causes the quantity because the conditional model, generating $\left(q(t)\,|\,p(t)\right)$ is assumedly structural. In other words, to the best of the scientist’s knowledge, analyzing the data on the price alone, by means of a marginal model generating $p(t)$ only, would give no information about the characteristics of the conditional distribution *and* this conditional distribution is by assumption stable under a large class of interventions.

References

BABBIE E. (2001), *The Practice of Social Research*, Wadsworth, Belmont.

CARTWRIGHT N. (1989), *Nature's Capacities and their Measurement*, Clarendon Press, Oxford.

- (1999), *The Dappled World. A Study of the Boundaries of Science*, Cambridge University Press, Cambridge.

DAWID A.P. (2001), “Causal Inference Without Counterfactuals”, in CORFIELD D. and WILLIAMSON J. (2001) (eds), *Foundations of Bayesianism*, Kluwer Applied Logic Series, Kluwer Academic Publisher, Dordrecht, 37-74.

DE FINETTI B. (1937), “La prévision, ses lois logiques, ses sources subjectives”, *Annales de l'Institut Henri Poincaré*, 7, 1-68.

DUCHENE J. and WUNSCH G. (1989) (eds), *L'explication en sciences sociales : la recherche des causes en démographie*, Ciaco, Louvain-la-Neuve. (Chaire Quetelet 1987)

EDWARDS A.W.F. (1972), *Likelihood. An Account of the Statistical Concept of Likelihood and its Application to Scientific Inference*, Cambridge University Press, Cambridge.

ELLETT F.S. and ERICSON D.P. (1983), "The Logic of Causal Methods in Social Science", *Synthese*, 57, 67-82.

- (1984), "Probabilistic Causal Systems and the Conditional Probability Approach to Causal Analysis", *Quality and Quantity*, 18, 247-259.

- (1986), "Correlation, Partial Correlation, and Causation", *Synthese* 67, 157-173.

- (1986), "An Analysis of Probabilistic Causation in Dichotomous Structures", *Synthese* 67, 175-193.

- (1989), "Causal Modelling and Theories of Causation", in DUCHÊNE J. and WUNSCH G. (1989) (eds.), 397-424.

FISHER R.A. (1922), “On the Mathematical Foundations of Theoretical Statistics”, *Philosophical Transactions of the Royal Society*, 222, 309-368.

- (1930), “Inverse Probability”, *Proceeding of the Cambridge Philosophical Society*, 26, 528-535.

- (1935), “The Logic of Inductive Inference”, *Journal of the Royal Statistical Society*, XCVIII, 39-54.

FRANCK R. (2002) (ed), *The Explanatory Power of Models*, Kluwer, Dordrecht.

FREEDMAN D. (1999). “From association to causation: some remarks on the history of statistics”, *Statistical Science*, 14(3), 243-258.

GIERE, R. (1999), *Science without Laws*, University of Chicago Press, Chicago.

HABBEMA J.D., COLLINS J., LERIDON H., EVERS J., LUNENFELD B., and TE VELDE E. (2004), “Towards Less Confusing Terminology in Reproductive Medicine: A Proposal”, *Fertility and Sterility*, 82(1), 36-40.

HAUSMAN D. (1998), *Causal Asymmetries*, Cambridge University Press, Cambridge.

HOLLAND P. W. (1986), “Statistics and Causal Inference”, *Journal of the American Statistical Association*, 81, 945-970.

HORWICH P. (1987), *Asymmetries in Time. Problems in the Philosophy of Science*, The MIT Press, Cambridge, Massachusetts.

PEARL J. (1988), “Graphs, Causality, and Structural Equation Models”, in *Sociological Methods and Research*, 27(2), 226-284.

- (2000), *Causality. Models, Reasoning, and Inference*, Cambridge University Press, Cambridge.

PEARSON K. (1911), *The Grammar of Science*, Dent London, London.

RUSSELL B. (1913), “On the Notion of Cause”, *Proceeding of the Aristotelian Society*, 13, 1-26.

SAVAGE L.J. (1954), *The Foundations of Statistics*, John Wiley, New York.

SOBER, E. (1986), "Causal Factors, Causal Influence, Causal Explanation", *Proceedings of Aristotelian Society*, 60, 97-136.

SPIRTES P., GLYMOUR C. and SCHEINES R. (1993). *Causation, Prediction, and Search*, Springer-Verlag. 2nd Edition, MIT Press (2001), New York, N.Y.

STONE R. (1993), “The Assumptions on which Causal Inferences Rest”, *Journal of the Royal Statistical Society, Series B (Methodological),* 55 (2), pp. 455-466.

SUPPES P. (1970), *A Probabilistic Theory of Causality*, North Holland Publishing Company, Amsterdam.

- (1962), “Models of data”, in E. NAGEL, P. SUPPES & A. TARSKI (eds.), *Logic, Methodology and Philosophy of Science*, Stanford University Press, Standford, 252-261.

VAN FRAASSEN, B. (1980), *The Scientific Image.* Oxford University Press, Oxford.

VAUPEL J. and YASHIN A. (2001). “L’hétérogénéité cachée des populations”, in: G. CASELLI, J.

VALLIN and G. WUNSCH, *Démographie: analyse et synthèse*, Volume 1 *La dynamique des populations*, Editions de l’INED, Paris, 463-478.

WILLIAMSON J. (2005), *Bayesian Nets and Causality. Philosophical and Computational Foundations*, Clarendon Press, Oxford.

WITTGENSTEIN L.(1961), *Tractatus Logico*-Philosophicus, translated by D. F. Pears and B. F. McGuinness, (First German edition in 1921), Routledge & Kegan, London.

WOLD H. (1949), “Statistical Estimation of Economic Relationships”, *Econometrica*, 17 (Supplement), 1-22.

- (1954), “Causality and Econometrics”, *Econometrica*, 22, 162-177.

WOODWARD J. (2003), *Making Things Happen: A Theory of Causal Explanation*, Oxford University Press, Oxford.