Introduction to causal inference without counterfactuals

A. Forcina

Dipartimento di Economia, Finanza e Statistica Università di Perugia

Outline

- Dependence and causality
- A formal language for causality: directed acyclic graphs
- Identifiability with unobservable variables: Back and Front door
- partial compliance and instrumental variables
- Marginal Structural mean models
- Parametric latent class model

Some references

The first part of these notes are based on
Pearl, J. (1995), ”Causal diagrams for empirical research,” Biometrika, **82**, 669-688.
For an updated presentation which covers recent developments, see
I. Shpitser, J. Pearl, (2008), ”Complete Identification Methods for the Causal Hierarchy”, Journal of Machine Learning, 9, 1941-1979;
A critique of counterfactuals in Causal Inference is in:
P. Dawid (JASA, 2000), see also his notes for a course on casual inference at
http://www.ucl.ac.uk/Stats/research/Resrprts/psfiles/rr279.pdf
For an instance of the Rubin's school approache see:
Angrist, J., Imbens, G. W. and Rubin, D. B. (1996), ”Identification of causal effects using instrumental variables,” JASA **91**, 444-472, or Holland P. W. (1086) ”Statistics and causal inference”, JASA **81**, 946-960
For an outline of the approach inspired by J. M. Robins see
Vansteelandt, S, and Goetghebeur, E. (2003), ”Causal inference with generalized structural mean models,” Journal of the Royal Statistical Society, **65**, 817-835, or the draft of his book with M. Hernan at
http://www.hsph.harvard.edu/faculty/miguel-hernan/causal-inference-book/

1 Objectives

The purpose of these notes is to provide a soft introduction to one of the approaches to causal inference from a personal perspective. The objective is not a philosophical disquisition into causality but a formal framework for handling causal question.

2 Counterfactuals, why not ?

They belong to what I like to call ”scientific fiction”, that is mental constructions which some find inspiring. I only disagree with Rubin and his school who have often claimed that causal inference cannot do without counterfactuals.

In most of Pearl’s construction, with which I am more comfortable, counterfactuals are an optional and, at least for the problema I am familiar with, I claim one can do without.

The reason I do not like counterfactual’s ”fiction” may be a matter of taste. In short, I claim that:

- my condition (whether unemployed) if I had not gone to university (contrary to what I did) is undefined as it requires the so called ”most closed world” to the actual one, a notion which I find hard to define and confusing;
- perhaps many people reason in terms of counterfactuals because they had been led to believe they could not do without and never asked themselves what they exactly mean;
- we could not do without counterfactuals if we really were interested in determining individual causal effects (which, however, are never identifiable and hardly of scientific interest).

2 Why a ”grammar” for reasoning on causation ?

Consider the following examples:

- In a study of income and height for a sample of adults somebody finds that income increases with height;
- It has been established that the probability of being unemployed is lower among young people who took a course relative to those who did not, though offered a chance to do so.

Has height a causal effect on income ? Probably, had we recorded sex of participants, we could have observed that, conditionally on sex, height has no effect on income.

In the second instance one could argue that, if we could have recorded individual attitudes, we could have established that subjects with certain attitudes were more likely to take the course and also to get a job. If this was the case, intuitively, in assessing the causal effect of taking the course we would have to compare individuals with similar attitudes.

3 The Simpson's paradox

It is well known that if, say, X, Y, Z are qualitative ordered variables, the marginal distribution of X, Y can exhibit independence or positive association, irrespectively of the fact that in the conditional distribution $X, Y \mid Z$ there is positive or even negative association. Thus the marginal association is not even an ”average” of the conditional association.

Thus, if X is the possible ”cause” and Y the ”effect”, it is crucial to know whether or not we should control for Z. The message that, hopefully, will emerge from these notes is that there is not a simple answer and that we should describe in more detail our understanding of the system.

4 Structural equations and DAGs

In many contexts we can formalize our knowledge (or hypothesis) by a system of non parametric structural equations

$$X_i = f_i[\mathcal{G}(X_i), \epsilon_i]$$

where X_i is one of the variables that describe a given problem, $\mathcal{G}(X_i)$ is a subset of $X_1, \ldots, X_{i-1}$, called the ”parents” of X_i, which, together with ϵ_i, are expected to determine X_i. The system is *recursive* because only ”earlier variables can cause ”later” ones.

From a counterfactual point of view, the left hand side is the value that X_i would take for a given subject if the ”causes” were set to $\mathcal{G}(X_i)$. A structural equation may also be interpreted as defining the conditional distribution of X_i within an exchangeable population of subjects with the same value of $\mathcal{G}(X_i)$.

A system of structural equations with such recursive property may be represented by a directed acyclic graph (DAG). However, the contrary is not true: a DAG in itself defines only a possible factorization of a joint distribution satisfying certain constraints of conditional independence and the arrows do not have necessarily a causal interpretation.

5 Conditional independence and DAGs

Intuitively, whenever an earlier variable does not appear in the structural equation of a later one, there must be some conditional independence. The problem is whether we can detect any possible independence constraint without algebraic manipulations but simply by inspecting the DAG. The answer is yes and there are several methods for doing this; the d-separation rule derived by Pearl is perhaps the most direct; however it requires some preliminary terminology.

a path is a sequence of arrows joining two nodes (variables), regardless of their direction;

a collider is an internal node contained in a path with converging arrows;

descendant: a node V is a descendant of U if there is a directed path from U to V;

Examples: in the DAG below V is a collider, but W and T are not;

$$X \longrightarrow V \longleftarrow Y \qquad X \longrightarrow W \longrightarrow Y \qquad X \longleftarrow T \longrightarrow Y$$

in the following DAG Y is a descendant of U, but T and W are not

$$\begin{tikzcd}
U \arrow[r] \arrow[d] & V & T \arrow[l] \\
X \arrow[r] & Y & W \arrow[ul] \arrow[u]
\end{tikzcd}$$

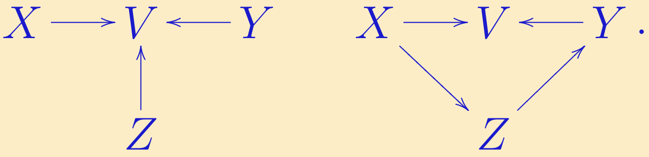
6 *d*-separation

A path from X to Y is said to be intercepted by a set $\mathcal{Z}$:

by conditioning: the path includes at least a node $V \in \mathcal{Z}$ which is not a collider, as in the following example:

$$X \longrightarrow V \longrightarrow W \longrightarrow Y$$

by marginalization: the path includes at least a node V such that: (i) V is a collider (ii) $V \notin \mathcal{Z}$, (iii) no element of $\mathcal{Z}$ is a descendant of V, as in the example below:



d-**separation**: $X \perp\!\!\!\perp Y \mid \mathcal{Z}$ if and only if $\mathcal{Z}$ intercepts all possible paths from X to Y.

7 Causal effect and the ”do” operator

Following Pearl (1995), we say that, if the causal structure of a problem can be represented by a DAG, to compute the causal effect of X on a variable Y, we should first remove from the DAG all arrows pointing to X and then marginalize all other variables. This is equivalent to remove the structural equation in which X is determined by its ”parents”.

This definition may be motivated as follows: in order to measure the effect of a given cause, we must take full control of its variations by preventing other variables to change the value of X.

The distribution of Y induced by an hypothetical intervention on X will be denoted with $P(Y\backslash X)$, a notation which is similar to that of conditioning though, as we will see, intervention is equivalent to conditioning only in special cases. Once the distribution has been constructed, the causal effect may be measured in different ways, a rather general formulation is outlined below

$$h[P(Y \in \mathcal{A}\backslash X = x_1)] - h[P(Y \in \mathcal{A}\backslash X = x_0)],$$

where $h(\cdot)$ is an arbitrary link function and $\mathcal{A}$ is a set of interest.

8 The intervention distribution

Formally, apart from the marginalization which follow the usual rules of probability, the operation of removing the arrows pointing to X is equivalent to divide the joint distribution associated with the DAG by $P(X \mid \mathcal{G}(X))$. The resulting distribution is, in all respects, a proper probability distribution for any given value of X which is determined by the original joint distribution and by the causal question of interest.

If we could run an experiment where the values of X for each unit were fixed from the outside while the remaining variables of the system were left free, there would be no arrows pointing to X in the corresponding DAG. Thus, to compute the intervention distribution is equivalent to perform an ideally randomized experiment in a context where we were unable or did not care to do so.

9 Intervention equals conditioning ?

The intuitive notion that to compute the causal effect we should ”control for covariates” is true is the following example

$$\begin{tikzcd}
U \arrow[d] \arrow[dr] & \\
X \arrow[r] & Y
\end{tikzcd}$$

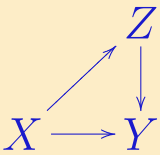
the intervention distribution gives

$$P(Y \backslash X) = \sum_u \frac{P(u)P(X \mid u)P(Y \mid X,u)}{P(X \mid u)} = \sum_u P(u)P(Y \mid X,u)$$

thus, if U was a known covariate, we should first compute the effect conditionally on U and then average across possible values of U. However, if U was unobservable, both operations could not be performed from knowledge of the distribution of the observable variables and we would say that the effect of X on Y is ”confounded” by U.

10 Direct and indirect effects

Now consider a DAG only slightly different



and suppose we want to compute the causal effect on the linear scale

$$D(y; x_0, x_1) = P(Y > y \backslash x_1) - P(Y > y \backslash x_0).$$

because there are no arrows pointing to X, the intervention distribution is now equal to $\sum_z P(z \mid X)P(Y \mid X, z)$ = $P(Y \mid X$, so here we should not adjust for covariates. However, by expanding on Z and some manipulations we can obtain

$$D(y; x_0, x_1) = \sum_z D(y; x_0, x_1 \mid z)P(z \mid x_1) + \sum_{z > z_0} D(z; x_0, x_1)D(y; z, z_0 \mid x_0)$$

where the first component may be interpreted as the direct effect of X on Y averaged across Z while the second may be seen as the product of the effect of X on Z (having selected a reference value z_0) times the effect of Z on Y and is usually called ”indirect”.
It is worth noting that above decomposition is not unique, except when conditional expectations are linear functions.

11 Identifiability of causal effects

Now the question is: can the observed distribution answer the causal query of interest or we should have organized a different experiment or collected different data ? Clearly, the problem arise only when the causal DAG contains nodes which were not observed or could not be observed.

When there are unobserved nodes and we are not prepared to make parametric assumptions concerning the joint distribution, in addition to those which are encoded in the causal DAG, the question is whether the intervention distribution of interest can be marginalized or, in other words, whether the quantities which are required to do so can be estimated from the observables.

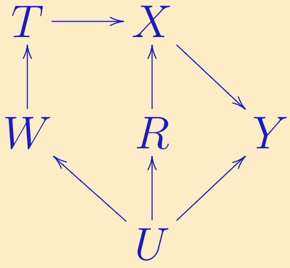
12 The ”back-door” rule

This is a result due to Pearl stating that, under certain conditions based on the DAG, certain causal effects are identifiable. Special cases of the same result are known as ”Average Partial Effect”, see for instance the book by Wooldrige.

A set of nodes $\mathcal{Z}$ satisfy the back-door condition for computing the causal effect of X on Y if:

- no element of $\mathcal{Z}$ is a descendant of X,
- any path between X and Y which contains an arrow pointing to X is intercepted by $\mathcal{Z}$

In the example below where U is latent, W, R or T, R satisfy the back-door conditions for the effect of X on Y



13 Computing the effect with back-door variables

To compute the causal effect in the previous DAG, first compute $P(Y, U, R, W, T\backslash X$ and marginalize W

$$P(Y\backslash X) = \sum_u \sum_t \sum_r P(u)P(t \mid u)P(r \mid u)P(Y \mid X, u)$$

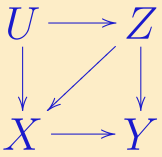
with arrow into X removed, note that $U \perp\!\!\!\perp X \mid R, T$ and $Y \perp\!\!\!\perp R, T \mid X, U$ and that $P(U, R, T) = P(R, T)P(U \mid R, T)$, then by standard manipulation

$$\begin{aligned} P(Y\backslash X) &= \sum_u \sum_t \sum_r P(u \mid r, t, X)P(r, t)P(Y \mid X, u, r, t) \\ & \sum_u \sum_t \sum_r P(r, t)P(Y, u \mid X, r, t) \\ &= \sum_t \sum_r P(Y \mid X, t, r)P(t, r). \end{aligned}$$

Rule Condition on back-door variables and then average.

14 Other examples

To familiarize with the notion, consider the following example where U is latent and Z satisfy the back-door conditions



write the intervention distribution

$$P(Y\backslash X) = \sum_z \sum_u P(u)P(z \mid u)P(Y \mid X, z)$$

and note that, because $\sum_u P(u)P(z \mid u) = P(z)$, the latent U can be marginalized in a more direct way and the resulting expression may be interpreting again as ”condition on the back-door and average”.

15 Identification by indicator variables

Consider the two DAGs below

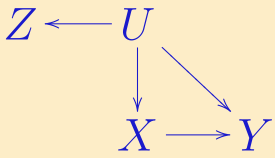
$$\begin{tikzcd}
W \arrow[r] \arrow[d] & U \arrow[d] & & W \arrow[d] & U \arrow[l] \arrow[d] \\
X \arrow[r] & Y & & X \arrow[r] & Y
\end{tikzcd}$$

first nota that in both cases W is a back-door variable and that, once the intervention distribution is written down, U is easily marginalized and the final expression is the same for the two cases

$$P(Y \backslash X) = \sum_{w} P(Y \mid X, w) P(w).$$

16 Models with a proxy

Some would call the variabile Z in the DAG below a proxy because, being determined by U, it provides, intuitively, an observable substitute for the latent U



however it can be verified that Z does not satisfy the back-door conditions and we get

$$P(Y\backslash X) = \sum_u \sum_z P(u)P(z \mid u)P(Y \mid u, X) = \sum_u P(u)P(Y \mid u, X)$$

which is the basic formula for a confounded causal effect. However, intuition suggests that, if Z was strongly correlated with U, we could approximate the causal effect by replacing U with Z. This is true and can be verified easily by simulation.

However, if U is unobservable, there is no way to check how strong is the association between U and Z.

17 The front-door rule

Though the following DAG is apparently similar to those seen so far, here the effect of X carries through Z and is confounded with the effect of U

$$\begin{tikzcd} X \arrow[d] & U \arrow[l] \arrow[d] . \\ Z \arrow[r] & Y \end{tikzcd}$$

Apparently, there is no obvious way to marginalize U from the expression below

$$P_(Y \backslash X) = \sum_u \sum_z P(u) P(z \mid X) P(Y \mid u, z).$$

However, Z satisfies the conditions for the ”front-door” formula

- Z intercepts all directed paths from X to Y,
- there is no back-door path between X and Z.
- X intercepts all paths from Y with an arrow into Z.

A result due to Pearl shows that under these conditions the causal effect of X on Y.

18 The front-door formula

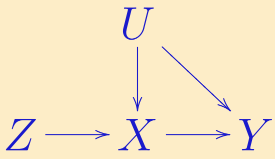
For an algebraic derivation use the identity $P(U) = \sum_x P(U \mid x)P(x)$ and exploit the following conditional independencies: $Y \perp\!\!\!\perp X \mid Z, U$ and $U \perp\!\!\!\perp Z \mid X$, reorganizing terms and marginalize to obtain

$$\begin{aligned}
P_{(}Y\backslash X) &= \sum_u \sum_z \sum_x P(u \mid x)P(x)P(z \mid X)P(Y \mid u, z) \\
&= \sum_z P(z \mid X) \sum_x P(x) \sum_u P(Y \mid x, u, z)P(u \mid x, z) \\
&= \sum_z P(z \mid X) \sum_x P(x) \sum_u P(Y, u \mid x, z) \\
&= \sum_z P(z \mid X) \sum_x P(x)P(Y \mid x, z).
\end{aligned}$$

The final formula is, essentially, the average causal effect of Z on Y with X as a back-door, weighted with the distribution of Z conditionally on X.

19 Models with an instrumental variable

The DAG below corresponds to a basic model with an instrumental variable Z



where U represents individual heterogeneity and is marginally independent from the ”instrument” Z and Y is independent from Z given X, U.
It is well known that the effect of X on Y is not identifiable without parametric restrictions of the model, for instance, in the case of binary variables, Balke and Pearl(1997) or Dawid (2002) have determined appropriate upper and lower bounds for the effect.
Observe also that the DAG contains the confounded component U, X, Y and that in the binary case the observable distribution is determined by 7 independent parameters while the latent requires at least 8: 2 for the marginals of Z, U and 3 for each conditional distribution of $X, Y \mid U$.

20 An IV model of partial compliance

A context to which the instrumental variable model can be applied is one where a treatment Z is assigned at random so that assignment is independent of individual heterogeneity. If X denotes received treatment, when experimental units are not aware of which treatment they have been assigned to, it seems reasonable to assume that Z is irrelevant to the response once X, U are given.

To describe the full latent structure, suppose that the assigned treatment Z is binary and that the received treatment X can vary between 0 (no compliance) to 1 (perfect compliance). For simplicity we also assume that U is discrete, meaning that individual heterogeneity can be represented with c distinct latent types.

The latent distribution is determined by the following vectors with one entry for each latent class: $\boldsymbol{\pi}$, the marginal distribution of U, $\boldsymbol{\tau}_{xz} = P(X = x \mid \boldsymbol{u}, z)$ and $\boldsymbol{\theta}_x = E(Y \mid \boldsymbol{u}, x)$. When we marginalize with respect to U, let $p_{xz} = P(X = x \mid z) = \boldsymbol{\tau}'_{xz}\boldsymbol{\pi}$.

21 Effect of treatment on the treated

Let also $\boldsymbol{\pi}_{xz}$ be the vector containing the posterior probabilities $P(U = u \mid X = x, Z = z)$. These can be interpreted as the weight of the different latent classes within the population of those who would self select a given level of compliance. If X was binary, there would be a vector of causal effects $\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0$, with 1 entry for each latent class. The effect of treatment on the treated is the average across the latent, when the posterior (rather than the marginal) probabilities of the latent are used

$$\Delta_1(1;1) = (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)'\boldsymbol{\pi}_{11} = E_u[E(Y \mid X = 1, u) - E(Y \mid X = 0, u) \mid X = 1],$$

where $\Delta_z(t, x)$ is the effect of an amount t of treatment among those assigned to $Z = z$ who would have self selected $X = x$. Had we averaged with the marginal of U, we would have obtained the overall causal effect which, however, is not identifiable.

22 The IV estimaand and partial compliance

When Z is binary, the IV estimand can be written as

$$\delta_{IV} = \frac{Cov(Y, Z)}{Cov(X, Z)} = \frac{E(Y \mid Z = 1) - E(Y \mid Z = 0)}{E(X \mid Z = 1) - E.(X \mid Z = 0)}.$$

To clarify the relation between the IV estimand and causal effects when X can take any value between 0 and 1, let

$$\Delta_z(t; x) = (\boldsymbol{\theta}_t - \boldsymbol{\theta}_0)' \boldsymbol{\pi}_{xz} = E_u[E(Y \mid X = t, u) - E(Y \mid X = 0, u) \mid X = x, Z =$$

which may be interpreted as the effect of an amount of treatment equal to t among those who would self select $X = x$ when assigned to $Z = z$..

23 Decomposing the IV numerator

By using the fact that $Z \perp\!\!\!\perp U$ and $Y \perp\!\!\!\perp Z \mid T, U$ and the identity $\boldsymbol{\tau}_{0z} = \mathbf{1}_c - \sum_{t>0} \boldsymbol{\tau}_{tz}$,

$$\begin{aligned} E(Y \mid Z) &= \sum_t \sum_u E(Y \mid u, t, Z) P(t \mid u, Z) P(u \mid Z) \\ &= \boldsymbol{\theta}_0' \operatorname{diag}(\mathbf{1} - \sum_{t>0} \boldsymbol{\tau}_{tz}) + \sum_{t>0} \boldsymbol{\theta}_t)' \operatorname{diag}(\boldsymbol{\tau}_{tz}) \boldsymbol{\pi} \\ &= \theta_0' \boldsymbol{\pi} + \sum_{t>0} (\boldsymbol{\theta}_t - \boldsymbol{\theta}_0)' \operatorname{diag}(\boldsymbol{\tau}_{tx}) \boldsymbol{\pi}. \end{aligned}$$

The numerator of the instrumental variable estimator $E(Y \mid Z = 1) - E(Y \mid Z = 0)$ may then be expanded as

$$\begin{aligned} &= \sum_{t>0} (\boldsymbol{\theta}_t - \boldsymbol{\theta}_0)' \operatorname{diag}(\boldsymbol{\tau}_{t1} - \boldsymbol{\tau}_{t0}) \boldsymbol{\pi} \\ &= \sum_{t>0} (\boldsymbol{\theta}_t - \boldsymbol{\theta}_0)' (\boldsymbol{\pi}_{t1} p_{t1} - \boldsymbol{\pi}_{t0} p_{t0}) \\ &= \sum_{t>0} [\Delta_1(t;t)(p_{t1} - p_{t0}) + p_{t0}(\Delta_1(t;t) - \Delta_0(t;t))]. \end{aligned}$$

24 The IV estimand as an average causal effect

A sufficient condition for interpreting the IV estimand as a causal effect is that

$$\sum_{t>0} p_{t0}[\Delta_1(t;t) - \Delta_0(t;t)] = 0, \tag{1}$$

when (1) holds, the IV numerator reduces to $\sum_{t>0} \Delta_1(t;t)(p_{t1} - p_{t0})$. This may happen in the following cases:

- treatment not available to controls, that is $p_{t0} = 0$;
- no effect modification by assignment, meaning that $\Delta_1(t;t) = \Delta_0(t;t)$ for all t;
- compensation of modifications, when (1) holds while neither of the two above conditions hold.

25 .. continuation

From the decomposition of the numerator of the IV estimand it is clear that, if (1) holds, by direct calculations

$$\delta_{IV} = \frac{\sum_{x>0}[\Delta_1(x;x)/x]x(p_{x1} - p_{x1})}{\sum_{x>0} x(p_{x1} - p_{x0})}.$$

This may be interpreted as a weighted average of $\Delta_1(x,x)/x$; this is equivalent to assume that $\Delta_1(x,x) = x\Delta_1(1,x)$, the effect is proportional to x and, because $x \leq 1$, it must be inflated to recover what the effect would have been under full compliance.

It is worth noting that while an average of $\Delta_1(1,x)$ across different values of x is equivalent to average the effect of full compliance across different sub-populations, $\Delta_1(x,x)$ is an hybrid function of the causal effect of x across different populations who also take a different amount of tretment.

26 An extended IV estimand

The interpretation of the IV estimand somehow relies on the assumption that, within a certain sub-populations, the effect of received treatment is proportional to the amount of compliance. This could be relaxed by assuming that $\Delta_1(t;t) = g(t)\Delta_1(1;t)$, where $g(t)$ is increasing in t, $g(0) = 0$, $g(1) = 1$ and $\Delta_1(1;t)$ is the effect of full compliance among those who would self select an amount t if given the opportunity to choose.

Then, simply replace X with $g(x)$ in the denominator of the IV estimand.

The resulting estimand will have an interpretation as an average causal effect whenever equation (1) holds and, in addition, the corresponding model for $\Delta_1(t;x)$ is assumed.

27 Linear models

It is worth noting that, when $E(Y \mid u, x) = h(x; \boldsymbol{\beta}) + k(u)$, that is the effect of individual heterogeneity and that of treatment are assumed to be additive, the causal effect conditional on the latent, that is the vector $\boldsymbol{\theta}_x - \boldsymbol{\theta}_0$ will be a constant vector with entry equal to $h(x; \boldsymbol{\beta}) - h(0; \boldsymbol{\beta}) = \Delta_z(x; x)$.
This implies that:

1. the effect of treatment on the treated and the overall average effect coincide because no averaging is necessary;
2. there is no effect modification by Z, thus the IV estimand is an average causal effect irrespective of whether treatment is available to controls.

28 Structural marginal mean models

This is a class of models which somehow extend the IV estimand, were introduced by J.M. Robins (1994). Let t denote the amount of received treatment and $\boldsymbol{x}$ be a vector of individual covariates and assume a linear model for the effect of receiving an amount t of treatment conditionally on covariates

$$\Delta_1(t; t \mid \boldsymbol{x}) = \mu_{tt,\,\boldsymbol{x}} - \mu_{0t,\,\boldsymbol{x}} = \boldsymbol{\psi}' \boldsymbol{g}(t, \boldsymbol{x}) \tag{2}$$

where $\boldsymbol{\psi}$ is a vector of unknown parameters, $g(t, \boldsymbol{x})$ is a known function of t and covariates such that $g(0, \boldsymbol{x}) = 0$ and $g(1, \boldsymbol{x}) = 1$.

It is important to note that while a consistent estimate of $\mu_{tt,\,\boldsymbol{x}}$ would be provided by individuals with covariates equal to $\boldsymbol{x}$ who self selected $T = t$, there is no unbiased estimate of $\mu_{0t,\,\boldsymbol{x}}$, the average response of those who would have selected t but were not allowed to reveal their preferences.

29 .. continuation

However, under the assumption that treatment is not available among controls, $E_T(\mu_{0T,\,\boldsymbol{x}})$, being the expected response for any subject with covariates equal to $\boldsymbol{x}$, can be estimated consistently by the corresponding sample average.
In the following let $\bar{y}_{1,i}$ denote the sample average among subjects with covariate $\boldsymbol{x}_i$ who were assigned to treatment, $\bar{y}_{0,i}$ the overall average among controls and $\bar{\boldsymbol{g}}_i$ the average of $g(t, \boldsymbol{x}_i)$, then define

$$d_i = \bar{y}_{1,i} - \boldsymbol{\psi}'\bar{\boldsymbol{g}}_i - \bar{y}_{0,i}$$

and note that, under the above assumptions, $E(d_i) = 0$. Suppose that $\boldsymbol{\psi}$ has size k and that there are $n \geq k$ distinct covariate configurations, let $\boldsymbol{d}$ be the vector with elements d_i; then a $k \times n$ matrix $\boldsymbol{S}$ exists which makes the generalized method of moments applied to the equation $\boldsymbol{S}'\boldsymbol{d} = \boldsymbol{0}$ most efficient.

30 Generalized marginal mean models

Let $h(\mu)$ be a suitable link function; in principle the above formulation may be extended by assuming

$$h(\mu_{tt,i}) - h(\mu_{0t,i}) = \boldsymbol{\psi}'\boldsymbol{g}_i(t)$$

which implies

$$h^{-1}[h(\mu_{tt,i}) - \boldsymbol{\psi}'\boldsymbol{g}_i(t)] - \mu_{0t,i} = 0.$$

After averaging with respect to t, a set of estimating equations would have the form

$$d_i = h^{-1}[h(\bar{y}_{1,i}) - \boldsymbol{\psi}'\bar{\boldsymbol{g}}_i] - \bar{y}_{0,i};$$

unfortunately these are not applicable to the logit link for two reasons: (i) adjustments are required when observations are 0 or 1, (ii) $E(d_i)$ is not 0 unless $\boldsymbol{\psi}$ is 0. Thus estimation will be biased.

31 Latent class models

Traditional latent class models try to explain association among observed responses by individual heterogeneity, modeled as a collection of discrete types so that, conditionally on the latent, responses are independent.

Such a model is determined by the marginal distribution of the latent and by the conditional distribution of each response given the latent. It is the structure of these conditional distributions which can characterize the latent type of each class.

Ordinary finite mixture models are latent class models when a single categorical response and covariates are available.

32 Extended latent class models

Recent extensions of latent class models allow

- the marginal distribution of the latent to depend on covariates,
- certain pairs of responses to be associated even conditionally on the latent;

the only restrictions being determined by the need for the model to be identifiable.

33 Latent class models and causal inference

Suppose that we can describe a causal model with a DAG with one or more modes representing unobserved individual heterogeneity. In this DAG any variables having an unobserved parent would be treated as endogenous and one could try to model the joint distribution of the latent and the endogenous variables conditionally on covariates.

The joint distribution could be factorized recursively and the model estimated. An early outline of this approach was described by Hagenaars (2002).

34 Causal interpretation of parameters

To be specific, consider the following example: let U denote a latent affecting treatment received and response, T is the amount of received treatment, Y is the response and $\boldsymbol{x}$ is a vector of covariates. In an observational study, the causal effect of T on Y would be confounded and not identifiable.

However, if a parametric model describing the conditional distributions of $U \mid \boldsymbol{x}$, $T \mid U, \boldsymbol{x}$, $Y \mid T, U, \boldsymbol{x}$ were identifiable, this model would have parameters specifying the dependence of Y on T for given $U, \boldsymbol{x}$, thus, adjustment for the unobservable confounder would be provided and these parameters would have a causal interpretation.

In addition, the parameters describing the dependence of Y on U for given $T, \boldsymbol{x}$ would allow to measure the effect of individual heterogeneity.

35 Identifiability of latent class models

Obviously all of this does not come for free: without important parametric restrictions, most of these models would not be identifiable. In addition, it will not be possible to submit to empirical test the assumed restrictions.

In principle, assessing local identifiability requires checking that the jacobian of the parameters for the unrestricted observed distribution with respect to the parameters of the latent class model is of full rank.

Except for very special cases where general results are available: though the jacobian can be written down, its rank is very hard to determine algebraically. However, algorithms for numerical computation of the jacobian are simple and fast. This suggests a practical solution: to determine identifiability of a complex model, sample the parameter space and determine the rank of the jacobian, if this is of full rank for a reasonably sample of parameter values, the model is very likely to be identifiable because simulations suggest that, if a model is not identifiable, the jacobian will be singular on most sample points.