Causal Models for Monitoring University Ordinary Financing Fund

Salvatore Marcantonio, Antonella Plaia

Abstract Recently iterated decreasing government transfers and an increasing proportion of budget allotted basing on competitive performances, took Italian Universities started struggling with competition for funds, in particular for the University Ordinary Financing Fund (FFO). Aim of this paper is monitoring variables responsible for FFO indicators, where monitoring means: describing, analysing retrospectively, predicting and intervening on variables responsible for indicators. All this aims can be achieved by statistical techniques that should be theoretically equipped with the distinction between predicting under observation and predicting under intervention, in order to provide correct answers to the distinct tasks of pure out of sample extrapolation and policy making.

Key words: bayesian inference, causal modelling, counterfactuals, University Ordinary Financing Fund

1 Introduction to Causal Modelling: Structural Causal models, DAGS and Counterfactuals

Recently iterated decreasing government transfers and an increasing proportion of budget allotted basing on competitive performances, took Italian Universities started struggling with competition for funds, in particular for the University Ordinary Financing Fund (FFO). Aim of this paper is to use causal machineries to monitor FFO indicators, in order to know which are the variables responsible for them, finding also what strategies can be taken to increase the value of the indicators.

Knowing that two events are associated each other (symmetric relationship) sometimes is not sufficient to ask questions we are interested in. This is especially the case when one event is interpreted as a cause and the other as the effect (asymmetric relationship), echoing the popular slogan ”association does not imply causation”.

Salvatore Marcantonio, Antonella Plaia
Department of Statistical and Mathematical Sciences, University of Palermo
Viale delle Scienze - Building 13, 90128 Palermo, Italy
e-mail: {salvatore.marcantonio,antonella.plaia}@unipa.it

The methodology followed in this paper, due to Pearl [3], formalizes such difference using the standard notation of conditional probability for describing observational relationships ("given that you see"), while using a new notation of do(X=x) operator at the right side of the conditional bar (|) for describing causal effects ("given that you do") or simple counterfactual sentence as "the value that Y would assume in unit u, had X been x" (Pearl [3] Ch 7). All such features can be implemented within a probabilistic structural causal model (PSCM), the nonparametric version of SEM, defined as a tuple M = $\langle U, V, F, P(u)\rangle$, where:

- $U = (U_1, ..., U_m)$ is a set of exogenous variables, namely determined by variables outside the model;
- $V = (V_1, ..., V_n)$ is a set of endogenous variables. These variables are functionally dependent on a subset of $U \cup V$. These are the variables to be analysed, usually empirically observable even though some of them can be unobserved or defined pure mathematically;
- F is a set of deterministic functions such that each f_i maps a subset of $U \cup V \setminus \{V_i\}$ in V_i, $v_i = f_i(pa(v_i), u_i)$, and such that V is a function of U through F;
- $P(u)$ is a joint probability distribution of U.

A PSCM defines three type of V_i variable: those who depend only on some U_k are completely unexplained by the model; those who depend only on other V_j which are deterministic, and those who depend both on V_j and U_k which are regular random variables ready to be regressed (whenever U_k are independent each other, usually denoted with ε_k).

The do(X=x) operator can be thought as encoding the simplest counterfactual sentence "the value that Y would assume in unit u, had X been x", $P(Y_u = y|do(X = x))$, i.e. the solution for Y in the submodel M_x where the function for X, $X = f(Pa[X], u_X)$ is replaced by the constant x, $X = x$, being $U = u$ the status of the exogenous variables, formally $Y(x, u) \triangleq Y_{M_x}(u)$. Thorough PSCM different types of causal effects can be defined and identified, such as direct and indirect [2], mediation [4], and confounding (Pearl [3], chap 6).

Every PSCM induces a Direct Acyclic Graph (DAG) (see fig. 1), where every node represents a variable and every function is represented by a set of arcs going from independent variables to the dependent one. DAGs are an intuitive and powerful tools for reading, by simple inspection, (in)dependencies implied by the model (via d-separation rule Pearl [3]). Graphically, the do(X=x) operator resolves in removing the arcs pointing towards X.

2 Analysis: Application to Universities

A rewarding part of FFO is allotted proportional to a set of indicators based on empirical data. Indicator A1 is concerned with educational offer and encompasses three different aspects: regularity of studies (ARS_x) is measured by the number of active and regular students enrolled on Degree Courses belongings to group $x \in G =$

$\{A,B,C,D\}$. Educational offer sustainability (KA) is measured by I_{KA}, namely the ratio between the number of regular teachers who cover "base" and "characteristic", courses (TEA) and the total teoric number of Degree Courses (TDC), normalized over its national median value (MKA). TDC is the sum, for all bachelors and master courses, of the ratio between new enrolments (NE_i) and a predetermined threshold depending on the Degree Course (T_i). Local context (KT) is measured by a variable which is in inverse proportion of the regional family income. The indicator I (see (1)) is then normalized with a national value S.

There are some other relationships not explicitly mentioned by government definition that may help to show how the indicator works. For example, ARS_A and ARS_B are essentially the same variable measured in different degree courses, therefore, it appears quite plausible to assume they could have a common exogenous cause U_{ARS} (the same with ARS_C and ARS_D) due to possible central policies, so they are correlated. Moreover there must be a probabilistic effect of NE on ARS because an "active" new enrolment can be seen as the event "success" ($ANE \sim Bin(\theta,NE)$), and active new enrolments in turn are a part of ARS. Focusing on a single University, we will call I^*_{KA} the set of the values of all the universities but this one, and S^* the sum of I over all the universities but this one. There could be a correlation between S^* and I or between I_{KA} and I^*_{KA} due to national policies common effect. The key idea is to interpret the Ministerial indicators defining functions as the following PSCM and utilize it for asking causal and counterfactual queries:

$$A1 \begin{cases} TARS = 4ARS_A + 3ARS_B + 2ARS_C + ARS_D \\ KT = f(\text{regional family income}) \\ TDC = \sum_i max\{\frac{NE_i}{T_i},1\} \qquad i = 1,2,...,\#(DC) \\ I_{KA} = \frac{TEA}{TDC} \\ I^*_{KA} = \{I_{(1,KA)},...,I_{(53,KA)}\} \\ MR_X = Median(\{I_{KA},I^*_{KA}\}) \\ KA = \frac{I_{KA}}{MR_X} \\ I = (KA+KT)TARS \\ S = \sum_{i=1}^{54} I_i = I + \sum_{i=1}^{53} I_i = I + S^* \\ A1 = 100\frac{I}{S} \\ ARS_X = f(NE_X,\varepsilon) \\ Corr(ARS_X,ARS_Y) \neq 0, Corr(S^*,I) \neq 0, Corr(I^*_{KA},I_{KA}) \neq 0 \end{cases} \tag{1}$$

$A1$ is a deterministic function of a set $U = \{S^*,KT,ARS_X,NE_i,TEA,I^*_{KA}\}$, but as PSCM prescribes giving probability to such set, every $A1$ becomes a probabilistic function of U. This acquires an empirical meaning by parametrising U with time and forecasting, leading to $A1_{t+1} = F(\hat{U}_{t+1})$, where $\hat{U}_{t+1} = E(U_{t+1}|U_t)$.

PSCM (1) and the induced DAG (Fig. 1) are also effective in evaluating the effect of new policies implemented by university governments. For example, considering the university of Palermo, we found [1] that new enrolments, NE_i, has a double and opposite effect on $A1$, positive through $TARS$ and negative through KA, and we asked whether an increasing new enrolments policy is useful or harmful for $A1$: since such effect influences two subsequent years, it is not possible to answer the question only looking at the difference between two subsequent $A1$ values, because they depend on other time varying quantities.

We solved the problem using counterfactuals and rephrased the question: what $A1$ would be had NE 10% more than its observed value (direct effect) and had active NE_t equal to active NE_{t+1} (indirect effect)? If that counterfactual value is higher than the observed $A1$, then the increasing policy will be advantageous. To the best of our knowledge no other systematic statistical models aiming to predict FFO have been published.

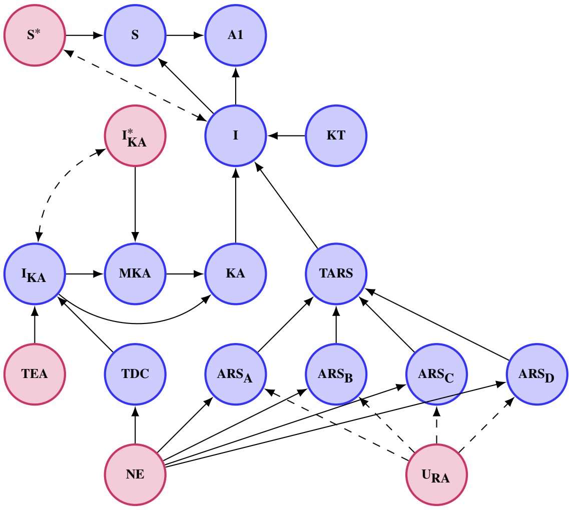

Fig. 1 *A*1 induced DAG

Acknowledgements This paper has been supported from Italian Ministerial grant PRIN 2008 ”Measures, statistical models and indicators for the assessment of the University System”, n. 2008WXSLH

References

1. Marcantonio, S.: Causal Models for Monitoring University of Palermo Ordinary Financing Fund. PhD Thesis (2012)
2. Pearl, J.: Direct and indirect effects. In Proceedings of the seventeenth conference on uncertainty in artificial intelligence (2001)
3. Pearl, J.: Causality: Models, Reasoning and Inference 2ed. Cambridge University Press (2009)
4. Pearl, J.: The Causal Mediation Formula – A Guide to the Assessment of Pathways and Mechanisms. Prevention Science, DOI: 10.1007/s11121-011-0270-1, March 2012. Available in `http://ftp.cs.ucla.edu/pub/stat_ser/r379.pdf`