A Bayesian nonparametric modelling to estimate students' response to ICT investment

Un approccio bayesiano non parametrico per la stima del profitto scolastico dovuto all'investimento in ICT

Stefano Cabras[1] and J. D. Tena[2]

Abstract This paper estimates the causal impact of Information Communication Technology (ICT) investment on student performances as measured in the Program for International Student Assessment (PISA) 2012 for Spain. We employ a new methodology in this context known as Bayesian Additive Regression (BART) that has important advantages over more standard parametric specifications, namely, (1) it is far simpler and more flexible than many recent competitors as it does not require ad-hoc assumptions about model fitting, number of covariates and how they interact, (2) it easily deals with missing data, and (3) in order to treat endogeneity it does not require the estimation of two equations one for ICT investment and one for the score. Our preliminary results suggest that, contrary to most previous studies, ICT have a moderate effect on math scores.

Abstract *In questo lavoro si stima l'effetto causale dell'investimento in Informazione, Comunicazione e Tecnologia (ICT) sul profitto scolastico misurato attraverso il Programma forma International Student Assessment (PISA) 2012 per la Spagna. Per far ciò, applichiamo una metodologia relativamente nuova in questo contesto, nota come Bayesian Additive Regressione Trees (BART). Tale tecnica fornisce vantaggi rilevanti rispetto ai modelli parametrici standard. In particolare: (1) è più semplice da implementare, più flessibile incluso rispetto alle tecniche econometriche più recenti poiché non richiede assunzioni specifiche sul modello di regressione come il l'insieme di covariate; (2) considera la presenza di dati mancanti e (3) per il trattamento delle variabili endogene non richiede la stima di due equazioni separate.*

Key words: Regression trees, Causality, ICT, Bayesian Statistics, BART models.

---

[1]Università di Cagliari (Italy) and Universidad Carlos III (Spain). Phone: (+34) 916249314. E-Mail: stefano.cabras@uc3m.es.
[2]Università di Sassari (Italy) and Universidad Carlos III (Spain). Phone: (+39) 079207323. E-Mail: juande@uniss.it.

1 Introduction

This paper estimates the causal impact of ICT on students' results in PISA 2012 using a non parametric Bayesian modeling approach known as Bayesian Additive Regression Trees (BART henceforth) that was originally developed by Chipman et al. (2010) and applied to the analysis of causal inference by Hill (2013).
A fundamental issue in this context lies in being able to differentiate between association and causation. Some solutions to this endogeneity problem are (1) to identify an exogenous education policy that is not correlated with the outcome of interest, Leuven et al. (2007); and (2) to conduct an experiment that randomly assign participants to the treatment and control group, Banerjee et al. (2008).
The proposed method allows estimation of causal effects in non-experimental works without being obliged, as usual in the econometric literature to estimate 2 models: one to capture the potential endogeneity of the ICT variable and another to specify students' performance. Moreover, using this flexible functional form has additional advantages over other more traditional parametric specifications as those usually required for the Difference in Difference method or the propensity score method. The BART method does not require any sort of hypothesis over the set of the covariates that influence the response and it can deal with a large number of regressors estimating thus a possible infinite number of interaction effects between covariates that are relevant to predict the response or treatment variable.

The remaining of this article is structured as follows: next section describes the BART methodology, Section 3 shows and analyzes the main results from the application of this methodology to the PISA 2012 database. Conclusions are drawn in Section 4.

2 Methodology

In this section we define the causal effects of interests as well as the estimation model. Regarding the estimation of causal effects we will mainly follow the notation in Hill (2011) and references therein and Chipman et al. (2010) for details about the statistical model. Leonti et al. (2011) provides one of the few applications of this model to the estimation of a causal effect to the use of medical plants.

Assume that N individuals participate in the PISA test and that all the variables considered in the analysis may contain missing values. For the i th individual, $i=1,\ldots,N$, let Y_i be the score in the PISA test or a proxy value for this as for example a drawn from the posterior distribution of the PISA test (OECD, 2005). Let Z_i be the intervention or treatment variable, that is a dummy variable that indicates the state of use of computers at school, where $Z_i=1$ if a table, laptop or a fixed computer exists and it is used in the school and $Z_i=0$ otherwise. In order to compute the causal effect of Z on the response variable Y

we should know, in principle, the potential results of the value of the test for the same individual under the use, $Y_i(0)$, and not use of computers, $Y_i(1)$. Assuming that the potential outcome is estimated, then average causal effects is $ATE = E(Y(1) - Y(0))$, where the expected values is computed with respect to the probability distribution of Y for all the individuals. The conditional causal effect as for instance the effect on students that used the computer is estimated by means of a conditional expectation: $E(Y(1) - Y(0) | Z = 1)$.
In observational studies, such as the PISA test, potential results are not typically independent of the treatment. This is known in the literature as the endogeneity problem. For instance, in the case of the PISA test, it is more likely that a student is allocated to a school with computers when his/her family has a high socioeconomic status and therefore it is the family environment (and not the use of computer) what may determines a favorable score in the PISA test. In order to assume that there exists independence in the treatment it is necessary to include in the analysis all the possible confounding factors represented, in this case, by the covariates X. This situation complicates the analysis as it requires the use of sophisticated regression models in the estimation of Y as the BART model.

Next section describes the BART approach, which not only frees us from specifying a set of covariates, as such set is estimated with the model. BART is a non parametric model that allows us to estimate with a satisfactory precision the response, and with that, the counterfactual result. This model belongs to the class of non-parametric Bayesian models that allows us to perform inference conditionally on the available observations, without considering the hypothesis of infinite resampling, typical in classical inference.

2.1 *The BART model: likelihood and priors.*

Let D be the available data, that is the set y, x, z observed for the N students and $\pi(.|.)$ the probability distribution of the left argument conditional to the right argument. The aim of the analysis is to estimate the *posterior* distribution of the causal effect, that is $\pi(ATE|D)$, or the distribution conditional to some covariates, $\pi(ATE|D, X = x)$. In order to do this we use a non parametric regression model. The novelty in this paper is the use of a Bayesian regression model known as BART for causal inference. Like in all Bayesian model we need a likelihood function defined for a set of parameters, $\theta \in \Theta \not\subset R$, and a prior distribution $\pi(\theta)$, $\theta \in \Theta$. The likelihood function, $L(D;\theta)$, is obtained from the following additive regression model, where Y is assumed to be normal with mean determined by the sum of estimated regression trees:

$$Y=\sum_{j=1}^{m} g\left(x,z;T_j,M_j\right)+\varepsilon,\varepsilon \quad N\left(0,\sigma^2\right)(1) \ ,$$

where $g\left(x,z;T_j,M_j\right)$ is a regression binary tree (or classification tree if Y were a categorical variable) with their variables and split points represented by T_j and their terminal nodes denoted by M_j and computed with respect to the values x,z that belong to the individual whose response is Y. The aim of this model is mainly to predict the outcome rather than the interpretation of the relations between variables. For this reason the model must be complex and in fact we do not have interpretation tools as regression coefficients because the model is made of trees that represent partitions of the space spanned by the covariates.
In order to estimate the model we use the regularization prior, specified in Chipman et al. (2010), to avoid that trees grow too much. The posterior distribution of θ is estimated in a computationally feasible using Markov Chain Monte Carlo (MCMC) as detailed in Chipman et al. (2010). Once the posterior distribution of $\theta=\left(T_1,\ldots,T_m,M_1,\ldots,M_m,\sigma^2\right)$ has been obtained, causal effects for the individual and the population can be computed by using the posterior predictive distribution of for the individuals whose Z value has been switched from 0 to 1 or from 1 to 0.

3 The PISA2012 Database

The PISA database contains information on knowledge and abilities of students who are closed to the end of the compulsory education period. Here we will only focus on the performance of Spanish students in math using the PISA 2012 database leaving for a further version of the paper a more complete analysis that includes more countries and skills. This leave us with 25,313 observations of Spanish students.

Regarding our dependent variable, students' performance in math, PISA show 5 plausible values. For the sake of simplicity we only compute the effect using the first of the five plausible values in math (PV1MATH) given that this analysis is almost identical to the one obtained with the remaining plausible values.

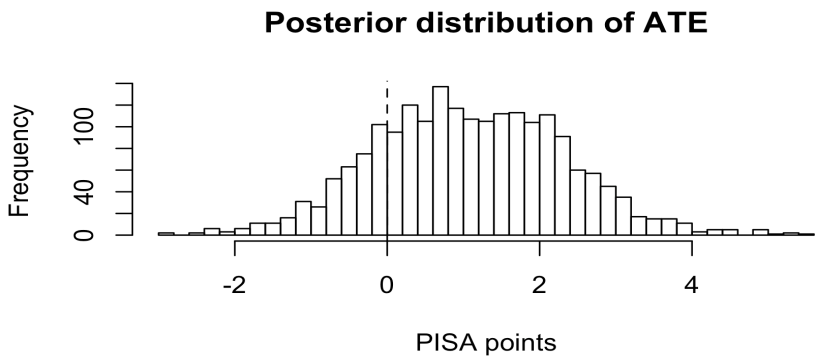
The explanatory variable of interest that will be used to estimate causal inference is the use of computers in the school. This information has been obtained by the specification of a dummy variable TREAT that takes value 1 when the student answers "Yes, and I use it" to at least one of these three questions: In the center when you study, do you have the possibility to use one of the following devices ? A fixed computer (IC02Q01), a laptop (IC02Q02), or a tablet (IC02Q03).

The additional explanatory variables, or covariates, contain information about previous students features, socioeconomic position of their family and information about the school.

4 Results

Here we report some results of the analysis for the databases described in the previous section using as response variable the first plausible value in math PV1MATH. The main result, that is the causal effect of computers on the PISA test for math is shown in Figure 1.

Figure 1: Draws of the posterior distribution



According to Figure 1, it can be seen some positive effect of computers on score, with probability 80% that amounts to saying that it is four times more probable that the impact is positive rather than null or negative. On average, the effect is 1 point, however its 95% credible interval is between -1.2 and 3.4. Although it is not correct to use credible intervals to make hypothesis testing (as they are not invariant under parameter transformation) the obtained credible interval indicates that we cannot be strongly convinced of a positive effect as there is also certain evidence also for negative effects. A parametric regression provides a poorer fit with estimated coefficients and p-values of 1.9 and 0.13 respectively. The estimated causal impacts are neither significant for the majority of papers in the previous literature discussed in the introduction.

We can also estimate the causal effect of the use of computers conditionally to the most relevant variables used in the analysis. The computation of these effects would be very complex under a parametric approach but it can be easily obtained by BART models. Analytical results are not shown for the sake of simplicity but suggest in first place that the treatment effect is conditional to the number of computers by class and the time these computers are used. In particular, the interaction between the treatment variable with the index ratio of the number of

computers in the school by students has a threshold effect indicating that investment in computers only has a positive effect if the number of computers in the school is higher than a certain value. Another very remarkable result from a political point of view is that in general the treatment variable has a higher positive effect on students that belong to the lowest socioeconomic environment.

5 Concluding remarks

Using the PISA 2012 database we have estimated the causal impact of ICT on Spanish students' performance in math. It is found some evidence of the positive causal impact of computers on the performance of Spanish students. It is especially remarkable the fact that this positive effect is significantly higher for students belonging to economically disadvantaged environments what suggests that the use of this intervention is a device to achieve higher level of social equity. It has been also found that the effect of this policy depends on a non monotonous way on how computers are used and the students teacher ratio.

References.

1. Banerjee, A., S. Cole, E. Duflo and L. Linden (2008). "Remedying Education: Evidence from Two Randomized Experiments in India", *Quarterly Journal of Economics*, 122(3), 1235-1264.
2. Chipman, H.A., George, E.I., and Mculloch R. (2010). "BART: Bayesian Additive Regression Trees", *The Annals of Applied Statistics*, **4**(1), 266–298.
3. Hill, L.H. (2013). "Bayesian Nonparametric Modeling for Causal Inference", *Journal of Computational and Graphical Statistics*, **20**(1), 217-240.
4. Leonti, M., S. Cabras, C. Weckerle, M. Solinas, and C. L. (2010). "The causal dependence of present plant knowledge on herbals—contemporary medicinal plant use in campania (italy) compared to matthioli (1568)", *Journal of Ethnopharmacology* **130** (2), 379–391.
5. Leuven, E., M. Lindahl, H. Oosterbeek and D. Webbink (2007). "The effect of extra funding for disadvantaged pupils on achievement", The Review of Economics and Statistics, 89(4), 721–736.
6. OECD, PISA 2009 Technical Report, OECD, Paris.