Cabras, S. & Tena Horrillo, J. D. (2016). A Bayesian non-parametric modeling to estimate student response to ICT investment. *Journal of Applied Stadistics, 43*(14), pp. 2627-2642.

DOI: 10.1080/02664763.2016.1142946

A Bayesian non-parametric modeling to estimate student response to ICT investment†

Stefano Cabras[a,b,c] and Juan de Dios Tena Horrillo[d,e,f]

[a]Department of Statistics, Universidad Carlos III, Getafe, Spain; [b]Instituto Flores de Lemus, Getafe, Spain; [c]Department of Mathematics and Informatics, Università di Cagliari, Cagliari, Italy; [d]Management School, University of Liverpool, Liverpool, UK; [e]Department of Economics, Università di Sassari, Sassari SS, Italy; [f]CRENoS, Cagliari, Italy

ARTICLE HISTORY Received 11 March 2015 Accepted 13 January 2016

ABSTRACT

This paper estimates the causal impact of investment in information and communication technologies (ICT) on student performances in mathematics as measured in the Program for International Student Assessment (PISA) 2012 for Spain. To do this we apply a new method-ology in this context known as Bayesian Additive Regression Trees that has important advantages over more standard parametric spec-ifications. Results indicate that ICT has a moderate positive effect on math scores. In addition, we analyze how this effect interacts with variables related to school features and student socioeconomic sta-tus, finding that ICT investment is especially beneficial for students from a low socioeconomic background.

KEYWORDS

Regression trees; causality; ICT; Bayesian statistics; BART

1. Introduction

Education is an important source of human capital in a country and its improvement has been linked not only to higher economic growth, Barro [3] and Hunushek and Kimko [13], but also to other non-economic benefits such as health, security and better social behavior, see Lochner [18] and references therein. Therefore, a very relevant issue concerning the decision-making process in education is to apply analytical methods that help to understand how education is produced and how it is affected by different policies.

Compared to other governmental policies aimed at improving educational levels, investment in information and communication technologies (ICT henceforth) in schools is especially appealing as it can be physically observable and is also perfectly divisible according to political discretionary criteria. However, in principle, its causal impact on students' performance can be considered ambiguous because, although it can be argued that the use of ICT can help students to a better understanding of the subjects studied, it can distract pupils from important academic obligations.

The aim of this article is to estimate the causal impact of ICT in the educational center on students' performance in mathematics using the information provided in the database

CONTACT Stefano Cabras s.cabras@unica.it

PISA 2012. A fundamental problem with this analysis, and in general with any type of analysis in social science, lies in being able to differentiate between association and causation. A satisfactory solution to this problem would only be possible with a metaphysic database where one could observe the difference between the response of a group of individuals to a given treatment and the response of the same individuals in the same circumstance without any treatment. However, Dawid [7] explains that a physical model can only identify the distribution of the model with and without treatment and not the joint distribution we are interested in making inference about.

The previous literature about the impact of ICT on student performance, which will be more extensively outlined in the next section, has dealt with this problem in two ways. The first proposed solution is to find a set of exogenous instruments that are correlated with the causal variable but uncorrelated with the error term in a parametric linear regression model. However, even if this set of instrumental variables can be found, results will hinge dramatically on the selected parametric specification and the specification of covariates. A second possible solution is to design experiments where the allocation of ICT students is random. However, even when it is economically feasible to run these experiments, in many instances they are subject to ethical problems or their results cannot be generalized to the whole population.

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), which was originally developed by Chipman *et al.* [5] and applied to analysis of causal inference by Hill [14]. This method, which mainly addresses the problem of optimal estimation of response surface, that is, PISA score, allows for causal identification in non-experimental works without being obliged to estimate two models, one to capture the potential endogeneity of the ICT variable and another to specify students' performance. Using this flexible functional form has additional advantages over other more traditional parametric specifications. Among them it can be mentioned that it does not require any previous selection procedure over the covariates to include, the possibility of dealing with a large number of regressors and estimating a large number of interactive effects between the treatment variable and other variables in the analysis, or the way this methodology copes with missing values without the necessity of dropping them from the sample.

To preview, we find that ICT has a moderately positive causal impact on students' performance. Most interestingly, the BART methodology allows us to estimate the most significant interactions between the treatment variable and other covariates related to school features and students' socioeconomic status. We find that ICT investment is especially beneficial for students coming from a low socioeconomic background.

The next section explains in more detail the previous literature about estimation of the causal impact of ICT on students' performance. Section 3 describes the BART methodology considered in this work and Section 4 shows and analyzes the main results from the application of this methodology to the PISA 2012 database. Conclusions are drawn in Section 5.

2. Literature review

The estimation of the impact that ICT exerts on students' performance has attracted a great deal of interest over the last 20 years. An early survey of this literature can be found in Kirkpatrick and Cuban [15], who indicate a lack of consensus about the effect of ICT in school performance from previous research. It is explained that most of these studies are based on simple correlations between ICT and pupils' performance, without taking into account other variables such as school characteristics, resources and quality, which are potentially linked to both ICT resources and students' outcomes.

This endogeneity issue has raised concerns about the validity of findings in early studies and the necessity of using econometric procedures that correct the potential estimation bias. Some solutions to this problem found in the literature are (1) identifying a certain education policy or practice that stems from a particular source that is not correlated with the outcome of interest, Angrist and Lavy [1], Leuven *et al.* [17], Machin *et al.* [19] and Goolsbee and Guryan [12]; (2) conducting an experiment that randomly assigns participants to treatment and control groups, Banerjee *et al.* [2]; and (3) using a database that contains information about relevant student and school characteristics that are not observed in standard analysis, Fuchs and Woessman [11] and Spiezia [24].

Most of these papers conclude that ICT has no effect on students' performance. In particular, Angrist and Lavy [1] estimated the impact of a large-scale computerization policy in elementary and middle schools in Israel. In order to estimate the causal impact of computer-aided instruction on students' performance, they control for unobservable school characteristics by exploiting the fact that funding decisions under this policy were mainly determined on the basis of the towns' ranking of applicant schools. They do not find evidence that increasing educational use of computers has a positive influence on pupil test scores. Leuven *et al.* [17] evaluated the effect of subsidies policies for computers and software on students' performance targeted at schools in the Netherlands. In principle, all schools with at least 70% disadvantaged minority pupils received the subsidy while all schools with less than 70% of such pupils did not receive the subsidy. However, there were exceptions to this rule and accordingly causal effects were estimated by considering a fuzzy regression discontinuity framework in which the probability of receiving the funds is estimated as a function of eligibility. Under this specification they find that computer subsidies in the Netherlands do not have a positive impact on students' outcomes. Goolsbee and Guryan [12] evaluate the impact on test score results of an US program to subsidize schools' investment in internet and communication. They explain that districts are allowed to decide which schools to include on their applications and that there are incentives for high-poverty districts to apply and for wealthier schools to convince higher-poverty schools in their district to add their applications, which can produce selection bias in the estimation. In order to cope with this problem they consider (1) a fuzzy regression discontinuity design by estimating the difference in students' performance for schools just below and just above the cut-off of subsidies for which each school is eligible and (2) an estimation of the change in test score using as a valid instrument the subsidy the district would have received if it had been forced to include every school in the district. In both cases they do not find that this subsidy has a significant effect on student performance.

However, other papers indicate that ICT has a positive and significant influence on students' outcome. For example, Machin *et al.* [19] evaluated the impact of ICT investment on educational output for England. In order to control for the potential endogeneity of ICT investment, they consider an instrumental variable estimation that uses as an instrument a change in the rules governing ICT investment. In this case, they find evidence of a positive causal impact of ICT investment on educational performance.

Banerjee *et al.* [2] also find a positive effect of ICT using a very different estimation approach. In particular, they conducted a randomized experiment to study the causal impact of computers on students' performance in India. They compared the change in the test scores of students who received a computer-assisted learning program with those of a control group, finding a strong effect of treatment on math scores in the short run. However, they also find that the effect declines substantially the following year.

The PISA database is particularly useful to analyze the problem of the influence of ICT on pupils' performance because it is a general survey for the whole student population that is about to finish the compulsory education period. It also includes most of the relevant information about students and schools that are typically omitted in most of the papers discussed above, and are observable in this case for a representative group of students. Fuchs and Woessman [11] and Spiezia [24] are two studies that based their analysis of ICT investment on the PISA database. However, evidence is also mixed in this case. Fuchs and Wöβmann [10] using international data from PISA 2000 find a significant and positive correlation between the availability of computers at school and school performance, but this correlation becomes insignificant when other school characteristics are taken into account. Spiezia [24] using information from PISA 2006 concludes that ICT has a positive effect on science scores.

The paper differs from the previous works discussed in this section because it is not based on a parametric specification that might not be flexible enough to capture the true relationship among the variables in the model, as they often rely on arbitrary identification assumptions. Moreover, parametric specifications can put the strong ignorability assumption at risk as it states that treatment allocation is only independent of the output variable conditional on a very specific functional form of the covariates in the model. An additional advantage of the methodology considered in this paper is that it allows for the identification and estimation of different types of nonlinear interactions between the treatment variables and the different covariates without the necessity of estimating different models for each interaction

3. 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 [14] and references therein and Chipman *et al.* [5] for details about the statistical model. Leonti *et al.* [16] provide one of the few applications of this model to the estimation of a causal effect to the use of medical plants.

3.1. Definition of causal effect

Assume that N individuals participate in the PISA test and that all the variables considered in the analysis may contain missing values. For the ith individual, $i = 1, \dots, N$, let Y_i be the score in the PISA test or a proxy value for this, as for example a draw from the posterior distribution of the PISA test (OECD [20]). Let Z be a dummy variable that indicates the state of use of computers at school, the treatment variable where $Z = 1$ if a table, laptop or a fixed computer exists and is used in the school and $Z = 0$ 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, $Y_i(1)$. However, this is impossible because only one of them can be observed while the other is non-observable and it is designated as the counterfactual result and has to be estimated with a regression model like the one described below. Such a model is mainly used in the estimation of response surfaces which is the main problem in the estimation of causal effects. In this case, it is the response Y to a 'hypothetical treatment' Z.

Once the potential causal effects have been estimated, the average total effect is defined as $\text{ATE} = E(Y(1) - Y(0))$, where the expected value is computed with respect to the probability distribution of Y for all the individuals. The causal effect for each individual is usually of no interest but it is that for a given set of individuals; for example those who have received the treatment $E(Y(1) - Y(0)|Z = 1)$, that is, individuals that have used a computer in the school. In this case, the expected value is estimated with respect to the conditional distribution of $(Y|Z = 1)$. Even more general, if we have a set of covariates X we can estimate the causal effect conditional to them, that is to $X = x$.

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. 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 belongs to a high socioeconomic status, and therefore it is the family environment (and not the use of computer) that determines a favorable score in the PISA test compared to other students with low socioeconomic status. 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. More specifically, the strong ignorability hypothesis regarding the allocation of treatment states that Y is conditionally independent of Z given X and that the probability of treatment allocation is always positive regardless of the specific value of X. In order to achieve this, it is necessary to include in X all the potential confounding factors and, because of that, matrix X typically has a very high dimensionality and is formed from different types of covariates: qualitative, quantitative and sortable variables. This situation complicates the analysis as it requires the use of sophisticated regression models in the estimation of Y.

Furthermore, the fact of considering many covariates makes it not possible for some classical approaches such as, for example, the propensity score, to be immediately applied because, given that treatment and no treatment cannot be observed for the same value of $X = x$, the estimation of the score assigned to each individual becomes difficult. This fact obliges the analyst to consider a set of variables of lower dimension, in many occasions putting the strong ignorability assumption in doubt. Finally, it is well known that the specification of regression models with many variables makes it impossible to search for all the possible models with all types of interactions. Again, this forces the analyst to consider only interactive effects among first or second order covariates or to use algorithms such as the forward or backward variable selection that provides locally optimal models. Unfortunately, there is not any theoretical evidence, but only some empirical results, which allow us to detect the scope of a local instead of global optimum.

Due to these drawbacks, in addition to others that can be found in the use of classical devices, the next section describes the BART approach, which not only frees us from model specification, because it is a non-parametric model estimated by observations, but because it also allows us to estimate with satisfactory precision the response, and with that, the counterfactual result. The proposed model belongs to the class of non-parametric Bayesian models that allow us to perform conditional inference about the available obser-vation, without the need of usual sampling hypothesis scheme, very typical in classical inference. In this case results are based on non-observed samples which are necessary for the well-known hypothesis of infinite replications under the null model of no causal effect.

3.2. 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 $\pi(\cdot|\cdot)$ the probability distribution of the left argument conditional to the right one. The aim of the analysis is to estimate the *posterior* probability distribution of the causal effect, that is
$\pi(\text{ATE}|D)$, or the distribution conditional to some covariates, $\pi(\text{ATE}|D, X = x)$. In order to do this, we use a non-parametric regression model. The novelty in these types of causal inference analysis is the use of a Bayesian regression model known as BART. As in all Bayesian models, we need a likelihood function defined for a set of parameters, $\in \Theta \notin R$, and a prior distribution $\pi(\theta), \theta \in \Theta$. The likelihood function, $(y|x, z, \theta)$, is obtained from the following additive regression model, where the mean of Y is determined by the sum of m estimated models for the response variable:

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

where $g(x, z; T_j, M_j)$ 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. Essentially g is a function that gives to each individual i its expected value in the $j$thtree, $\mu_{ij} \in M_j$. The final score estimated for the ith individual would correspond to the average of the m scores. It is well known that, in order to minimize the forecast error, classification trees tend to grow disproportionately until generating *overfitting* in the response and that in general, an estimator obtained from many simple trees is more efficient than another one obtained from a single big tree. Examples of these types of models are Boosting [23] and Random Forest [4].

In order to achieve this it is necessary to use a regularization prior on the size of the tree $\pi(T, M)$ specified in Chipman *et al.* [5]. This regularization prior precludes trees from growing too much and makes each of the μ_{ij} contribute in a marginal way to the estimation of the response function. The posterior distribution of θ is estimated in a computationally feasible way by considering a conjugate prior on σ^2, that is, an inverse-gamma that induces a conditional distribution of σ^2, $\pi(\sigma^2|T_1, \ldots, T_m, M_1, \ldots, M_m)$ that can be expressed in an analytical form, which is again an inverse-gamma. As Chipman *et al.* [5] show, the hyper parameters of all prior distributions are specified in relation to the observed sample. In particular, following Chipman *et al.* [5, Section 2.2], priors are specified for T and M independently among them and from σ^2, that is $\pi(T, M) \propto \pi(M|T)\pi(T)\pi(\sigma^2)$. The prior $\pi(\sigma^2)$ is inverse-gamma with mean $\hat{\sigma}$, the observed unconditional standard deviation of Y and scale fixed in order to guarantee that $P(\sigma < \hat{\sigma}) = q$, with $q = 0.95$. This is consistent with the fact that *a posteriori* the conditional variance σ^2 must be no larger than unconditional variance of Y because of the well-known Rao–Blackwell Theorem. Prior on split points, $\pi(T)$, is specified by three aspects: (i) the probability that a node at certain depth is non-terminal; (ii) the distribution on the splitting variable assignments at each interior node; and finally (iii) the distribution on the splitting rule assignment in each interior node, given the splitting variable. Prior (i) is proportional to $\alpha(1+d)^{-\beta}$, where d the tree depth and $\alpha = 0.95$, $\beta = 5$, such that trees with more than 5 terminal nodes have probability of 3%. Such values of $\alpha = 0.95$ and $\beta = 5$, assures that trees do not grow too much and here is where regularization prior comes from. Finally, priors (ii) and (iii) are just uniform priors on available covariates and on possible split points in each one of the covariates. Finally the prior on $\mu_{ij} \in M_j$ is the conjugate Normal with mean and variance such that the observed extremes of Y are probable a priori. Again, this guarantees that the prior is vague enough and allow the data to speak for themselves.

These priors are data dependent, which is not very orthodox from a Bayesian point of view. However, this approach, known as empirical Bayes, is very popular and have been enhanced from a theoretical point of view by a recent paper from Petrone *et al.* [21]. At any rate, as explained by Hill [14], results for this type of analysis are robust with respect to prior modifications. In particular, in the proposed analysis, the number of observations is so large that the prior is washed out by the likelihood and the posterior is highly insensitive to prior hyperameters.

3.3. Estimation of the BART model

Using the priors specified above it is possible to simulate samples from the posterior distribution with a non-excessive computational effort using Markov Chain Monte Carlo (MCMC), more specifically using Metropolis Hastings within Gibbs. This means that the simulation algorithm alternate Gibbs steps (like the one that is necessary to simulate σ^2) and Metropolis–Hastings (MH) steps when the conditional distributions for the remaining parameters are not available in a known way. In particular, the distribution used to update the values of T_j and M_j consists in adding/dropping a terminal node and changing a split variable or a split point with some probabilities specified in Chipman *et al.* [5, Section 3). In particular the Gibbs step refers to σ^2 which is drawn for an inverse-gamma distribution, while more challenging are the m draws of T_j and M_j. Such problem is actually reduced by sampling T_j and M_j given the prediction residuals $R_j = y - \sum_{k \neq j} g(x, z; T_k, M_k)$ of the tree without node j. The draw of M_j is another Gibbs step because of the conjugacy of its prior, while that of T_j is an MH step where the proposal consists of the following movements with their probabilities in parenthesis: growing a terminal node (0.25), pruning a pair of terminal nodes (0.25), changing a non-terminal rule (0.40), and swapping a rule between parent and child node (0.10). Once the posterior distribution of $\theta = (T_1, \ldots, T_m, M_1, \ldots, M_m, \sigma^2)$ has been obtained, the predictive distribution for an individual score in PISA test is:

$$m(Y_i | x_i, z_i) = \int_{\theta \in \Theta} L(Y_i; \theta) \, \mathrm{d}\pi(\theta | D), \tag{2}$$

which is practically estimated generating values of Y_i, using the normal distribution with the mean and variance for each value θ in the chain MCMC and the regression trees computed in x_i and z_i. In particular, we use $m = 500$ trees and 5000 MCMC steps after an initial burn-in of 1000 steps. Interactions among covariates do not need to be specified a priori, they are estimated from the data by the 500 trees. Although interactions are not analyzed as the main objective is prediction, these can be retrieved from the specific structures of the trees. For instance, if terminal nodes are defined upon split variables $X_1, \ldots, X_k$, this indicate the presence of an interaction between such k covariates. Interactions can be of arbitrary order and may involve many covariates at the same time.

In this way, the distribution for each individual and its corresponding counterfactual response can be estimated simply by estimating the response at $z_i = 1$ if the student does not have a computer in his/her school and at $z_i = 0$ otherwise. Once these predictive posterior distributions have been obtained, the difference between the factual and counterfactual responses are considered to obtain the distribution of the individual causal effect. Finally, $\pi(\text{ATE}|D)$ is estimated from the set of the differences for all the individuals. Then, the estimation of the conditional causal effect is required, which is obtained simply by considering the difference for the individuals that fulfill the condition $X = x$.

4. Results

4.1. Descriptive statistics

The PISA database contains information on knowledge and abilities of students who are close to the end of the compulsory education period. It is mainly used to determine the way these students are prepared for life after this period instead of focusing on evaluation of their curricular knowledge. The database contains information about students' performance in reading, math and science as well as survey answers from students, headmasters and parents. In this paper, we focus on the performance of Spanish students in mathematics, leaving analysis of other subjects and countries for future contributions. This leaves us with 25,313 observations of Spanish students. These observations must be interpreted as obtained from a random sample of the total population in a way that each observation has a relative weight according to its importance in the total population.

Regarding our dependent variable, students' performance in math, PISA shows five plausible values. Each of them represent a random observation obtained from the posterior distribution of students' performance. The population mean of these values can be obtained as a weighted average of all the plausible values that in turn has been obtained as a weighted average of the 25,313 observations that are available.

The average value for the dependent variable is 484.3 with a standard deviation 87.7. This standard deviation has been computed from a linear combination of the imputation variance, which estimates the deviation between the estimation obtained with the different plausible values, and the sample variance, which estimates the deviation due to the sample. Sample variance is equal to 7697.9 while imputation variance is 0.11, which suggests, in line with previous works, that most of the uncertainty in the population estimation corresponds to sample variability instead of to the fact of considering only one of the five plausible values. Since estimated results are almost identical regardless of the plausible variable considered, all results shown in the remainder of this paper are based only on the first plausible value.

The explanatory variable of interest that will be used to estimate causal effect 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).

It is interesting to show the features of the sample distribution for math scores. Applying the normality test proposed by Royston [22] based on an adjustment to the test of D'Agostino *et al.* [6] the null hypothesis would be rejected for all the usual significant values. Moreover, the null hypothesis is also rejected if we look at the distribution of this variable for both the values 0 and 1 of the TREAT indicator defined in the above paragraph. The absence of normality conditional and unconditional on the treatment does not exclude the possibility that the distribution of the response conditional on the considered set of variables is normal, as suggested by the residual analysis of the BART model. This is due to the possibility of using very flexible models in the definition of the linear predictor. Figure 1 shows the histogram of the math score obtained by students in both cases.

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

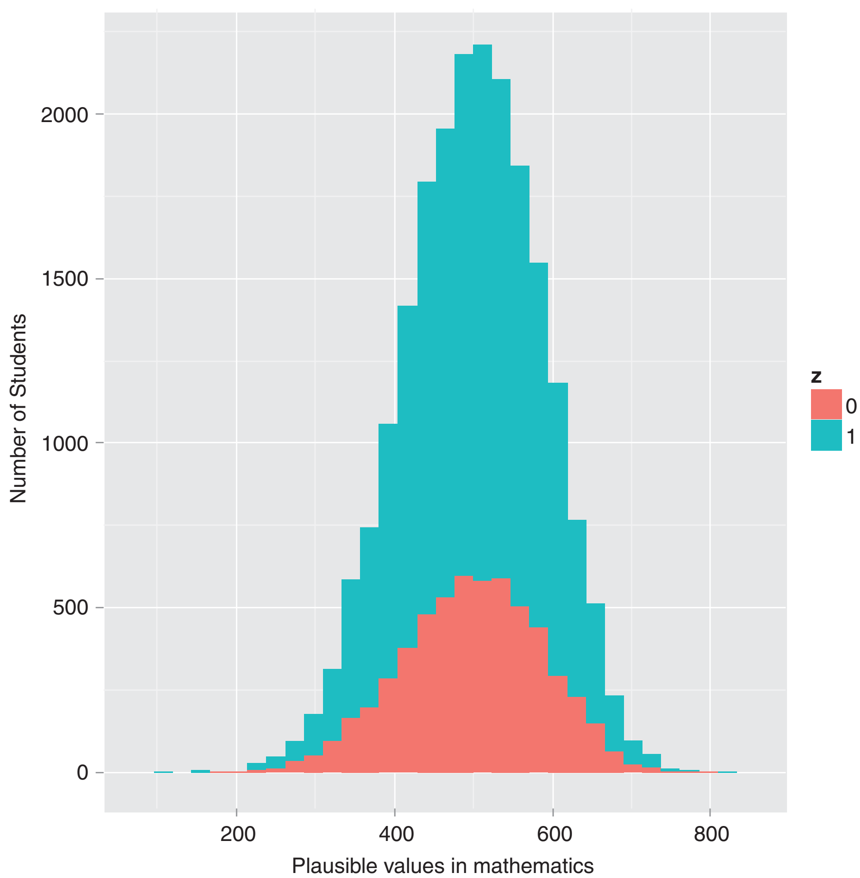

Figure 1. Sample histogram of the first plausible value of math scores for students who do not use and use a computer at the school [Colour online].

Table 1. Estadística descriptiva para las covariables consideradas en el modelo: Datos muestrales sin ponderar.

	COMPUTER = 1					COMPUTER = 0				
	Obs.	Mean	Std. Dev	Min	Max	Obs.	Mean	Std. Dev	Min	Max
RELATIVE_AGE	17850	0.50	0.50	0	1	7463	0.51	0.50	0	1
GENDER	17850	1.50	0.50	1	2	7463	1.50	0.50	1	2
EARLY_EDUCATION	17850	0.06	0.24	0	1	7463	0.06	0.25	0	1
ESCS	17812	−0.12	0.99	−3.92	2.73	7309	−0.09	1.03	−5.3	2.55
FAMSTRUC	16909	1.91	0.31	1	3	6888	1.90	0.32	1	3
HEDRES	17765	0.10	0.87	−3.93	1.12	7303	−0.02	0.91	−3.93	1.12
HISCED	17599	4.39	1.63	0	6	7219	4.46	1.66	0	6
HISEI	17658	47.58	21.42	11.01	88.96	7248	48.88	21.95	11.01	88.96
HOMEPOS	17784	0.13	0.81	−6.48	3.76	7320	0.07	0.86	−6.65	3.76
IMMIG	17227	1.18	0.57	1	3	7041	1.22	0.61	1	3
TIMEINT	17721	55.22	40.45	0	206	6596	51.57	41.67	0	206
WEALTH	17781	−0.01	0.79	−5.32	2.91	7316	−0.04	0.83	−5.32	2.91
CLSIZE	17850	33.54	24.30	13	99	7463	35.05	25.67	13	99
SCMATEDU	17648	0.11	0.91	−3.59	1.98	7361	0.04	0.90	−3.59	1.98
STRATIO	16700	11.91	6.76	1.11	139	6945	12.06	7.55	1.11	139
SMRATIO	16547	108.02	110.55	2	1820	6917	106.80	90.02	2	1820
SCHLTYPE	17309	2.58	0.58	0	8	7263	2.60	0.60	1	3

While previous features can be easily obtained from their answers to questionnaires, we use fundamental indices regarding features of their family or the school they attend that can be directly computed from students' answers, see for example OCDE [20]. More specifically, the following variables are considered in this paper:

Previous students' features: RELATIVE_AGE: binary indicator that takes value 1 if the student is born in the first 6 months of the year and 0 otherwise; GENDER: Gender indicator; EARLY_EDUCATION: binary indicator that takes value 1 if the student has attended nursery and 0 otherwise.

Socioeconomic features of the student's family: ESCS: Index of economic, social and cultural status; FAMSTRUC: Family structure; HEDRES: educational resources at home; HISCED: Educational level of parents; HISEI: Highest occupational level of parents; HOMEPOS: Possessions at home; IMMIG: Immigrant status; WEALTH: Wealth.

School features: CLSIZE: Size of the class; TIMEINT: Total time using computers (in minutes); SCMATEDU: Quality of the educational resources in the school; STRATIO: Student–teacher ratio; SMRATIO: Student–math teachers ratio; SCHLTYPE: Indicator of school ownership.

Having a computer in the school determines the expected values of many student features and this is an indicator about the necessity of controlling for the influence of all these variables in a causal analysis (Table 1).

4.2. Results

Here we describe the results of the analysis for the databases described in the previous section. The main result, that is the causal effect of computers on the PISA test for math, is shown in Figure 2 which shows the approximation, by means of MCMC draws, of the posterior distribution $\pi(\mathrm{ATE}|D)$.

According to Figure 2, a moderate positive effect of computers on scores can be observed. In particular, the probability of a positive effect is 80%, which it amounts to

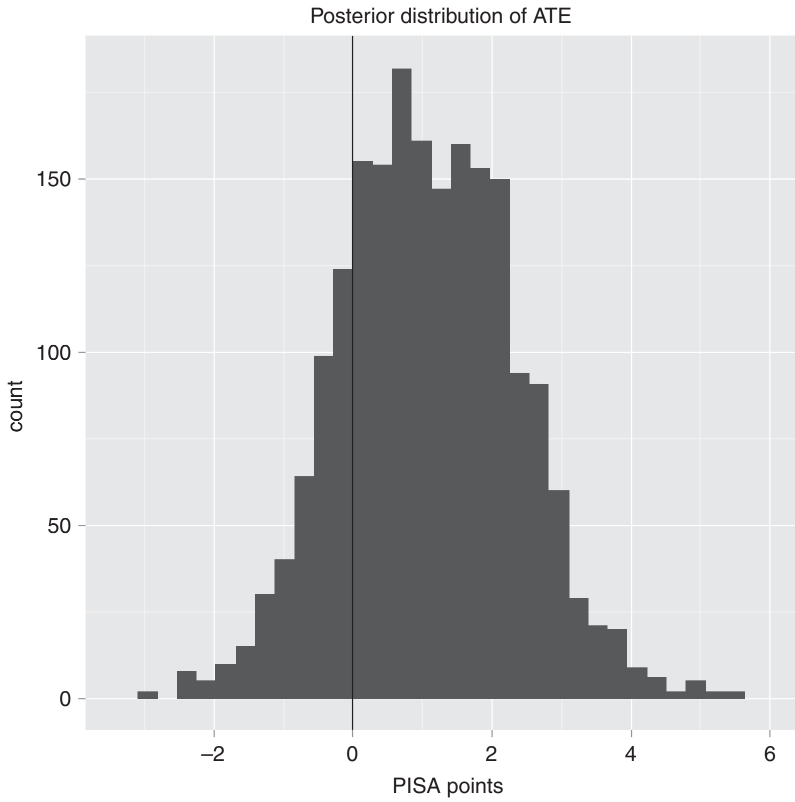

Figure 2. Approximation of the posterior distribution.

saying that it is four times more likely that the impact will be positive rather than null or negative. On average, the effect is 1 point and the 95% interval for this effect is between −1.2 and 3.4, which amounts to saying that there is evidence for strong positive or even some negative effects.

The most important variables for explaining the response can be easily estimated by counting the number of times each of the variables shows up in the classification trees. Figure 4 illustrates this point.

The most important variables are HOMEPOS, TIMEINT, WEALTH and HEDRES. However, the treatment variable (TREAT) is not among the most important ones and this explains why the causal effect of computer is not very significant.

The estimated BART model with the covariates indicated in Figure 3 explains around 29% of the variability of the response as can be seen in Figure 4. It is interesting to compare this result with a standard procedure based on a standard linear parametric regression by the weighted minimum least squares procedure of the variable of interest PV1MATH with respect to TREAT and all the other control variables already mentioned. The resulting regression has a value of the R^2 equal to 0.22 which has a value substantially lower than the BART model. In this linear regression, it can be concluded that the TREAT variable has a positive and significant effect at the usual significant levels. In particular, the value of the estimated coefficient is 3.97 with a p-value of 0.002. This could be interpreted as

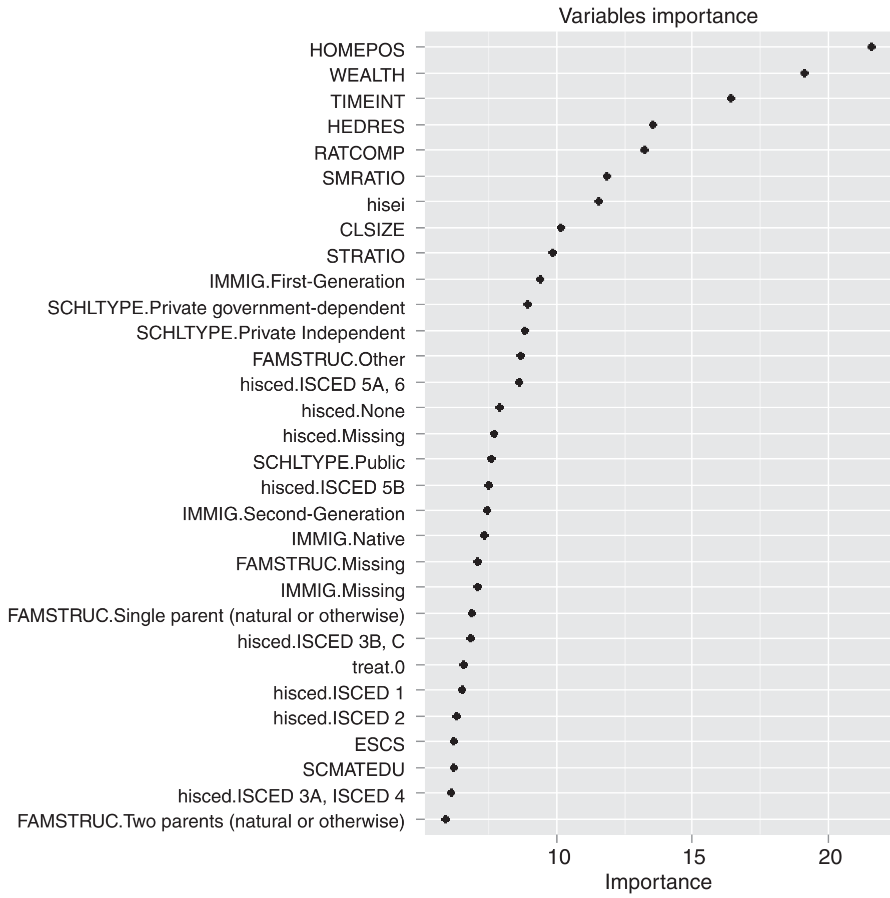

Figure 3. The most important variables according to the number of times they show up in the trees.

a causal effect estimation if all relevant covariates were introduced in the model with all the interactions. However, this is extremely difficult to do in a parametric specification. As an alternative to this estimation, we also consider a propensity score model using logistic regression before performing the matching. In this case, the estimated causal effect was 2.05 which is not significant at any of the conventional values (t-stat:1.04). However, the difficulty to include all the potential interactions across variables resulted in an unsatisfactory balance between the treatment and the control groups (see supplementary material) making the final estimation unreliable.

Based on this model, it is possible to estimate the causal effect of the use of computers conditional on the most relevant variables used in the analysis, without imposing any parametric functional form on the interactive effects. The most important ones are shown in Figure 5.

Results suggest that the impact of ICT on students' performance largely depends both on school features and indicators of the students' socioeconomic background. More specifically, for the first group of variables, private schools make more efficient use of computers than public schools. This is consistent with the previous literature; see for example Figlio

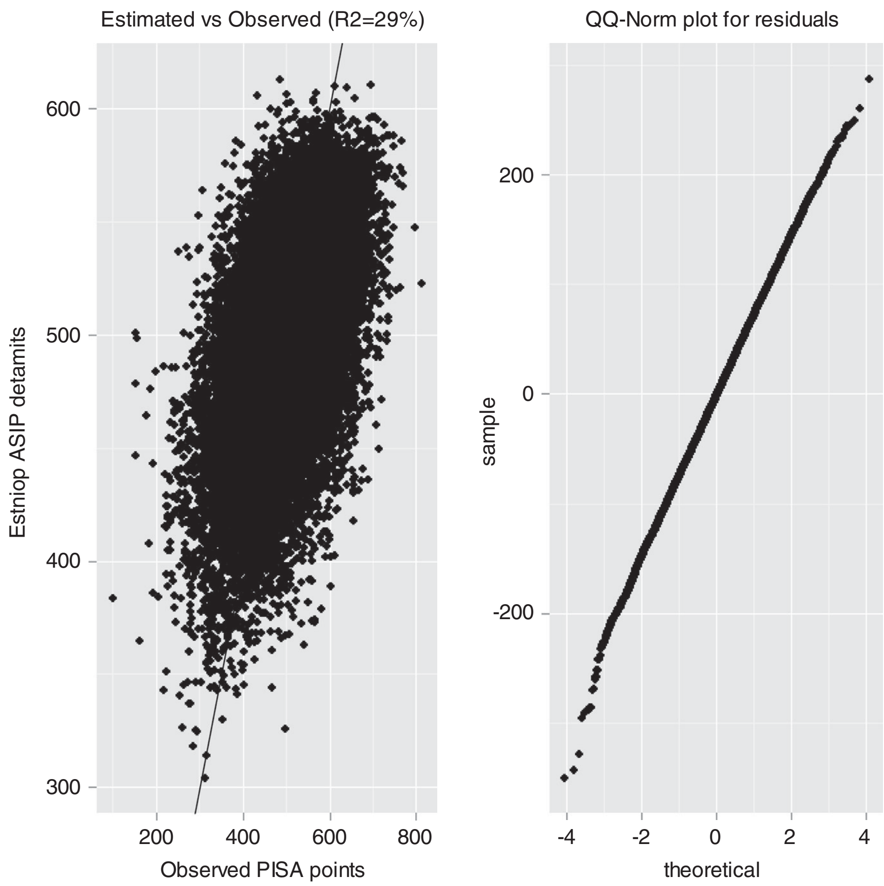

Figure 4. Observed and estimated PISA score ($R^2 = 0.29$) and QQ-plot of residuals versus normal distribution.

and Kenny [9]. However, it is important to emphasize that these results are not necessarily general for other countries and sectors; Delfgaauw and Dur [8]. Furthermore, as expected, the interactive effect of the time of computer use (TIMEINT) with the treatment variable also tends to be positive, although its slope is not very pronounced. This moderate interactive effect could be due to the fact that our treatment variable already includes information about the use of computers.

It is very interesting to analyze the interactive effect of ICT with other variables that indicate the socioeconomic position of the students. A remarkable result for decision makers is that in general, the treatment variable has a higher positive effect on students with the lowest socioeconomic status. This could be observed by the fact that ICT exerts a more positive effect on first and second generation immigrants compared to native students or by the fact that ICT effect is more positive for students with a lower level of possessions at home (HOMEPOS) and educational resources at home (HEDRES). Moreover, although the structure of the family is not necessarily linked to socioeconomic status, it can be observed that ICT school investment is more inefficient for students living with two parents compared to those living with a single parent or another familiar structure. Overall,

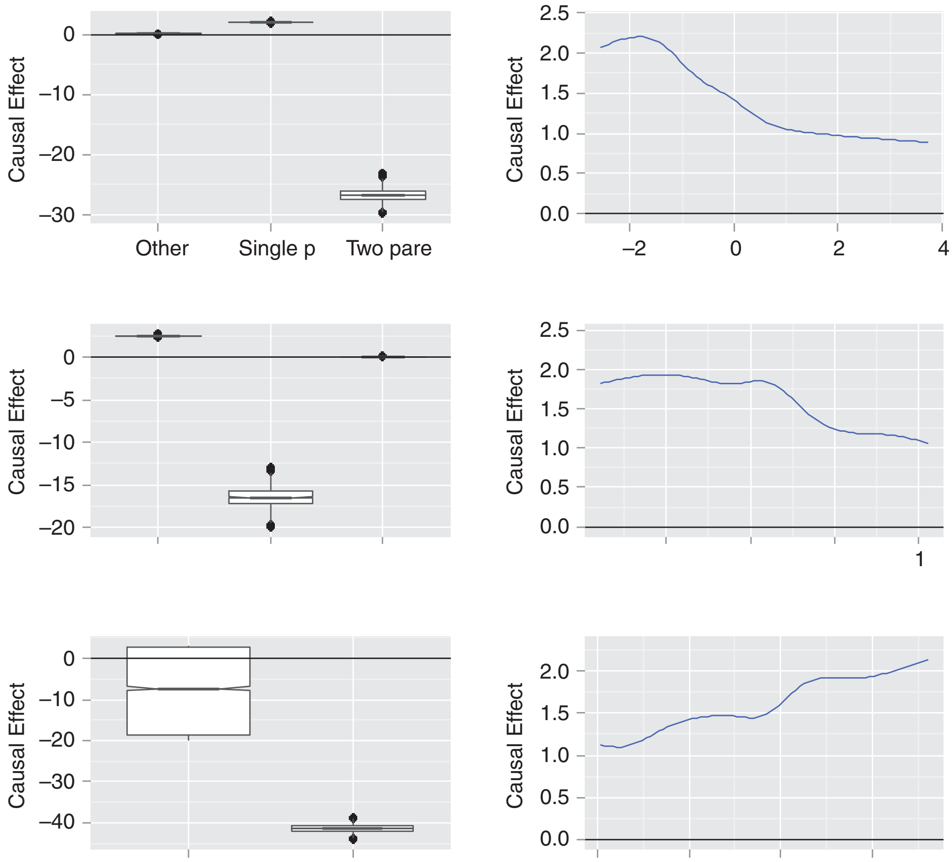

Figure 5. Distributions conditionals on the causal effects.

these results suggest that ICT in school could substitute in many cases other available ICT resources at home and, because of this, its effect is more positive for students who do not have direct access to these resources.

5. Concluding remarks

Using the PISA 2012 database, we have estimated the causal impact of computer use in the school on Spanish students' performance in math. In this estimation we have used BART models which have important advantages compared with other more standard parametric models. In particular, in order to treat endogeneity there is no need to specify two models: one for the treatment variable, use of computers at school; and another for the response variable, student's scores. Other advantages of this procedure are the lower necessity of assumptions in the model specification, the way it deals with missing values and outliers, the possibility of analyzing a large number of variables, and the interactive effects of each of them with the treatment variable.

Some evidence is found for the positive causal impact of computers on the performance of Spanish students. Especially remarkable is the fact that this positive effect is significantly higher for students belonging to economically disadvantaged environments, which reinforces the use of this intervention as a device to achieve a higher level of social equity. It has been also found that the effect of this policy depends in a non-monotonous way on how computers are used and the student–teacher ratio.

The Causal effect estimation problem is here afforded as a problem of prediction of counterfactual observations. In this sense the comparison with existing method is related to the prediction feature of the proposed model against parametric models. In particular, we highlighted that existing methods have the limitation that the structure of predictors, which and their interactions, must be specified a priori and managing arbitrary complex model is not feasible if all possible (and available) confounding effects must be taken into account. For instance, usual propensity score methods have the drawback that a specific set of covariates along with their interactions must be specified. This limitation is not present here while it is in other researches on this specific topic related to educational policy as those previously cited. Moreover, an extensive simulation study which compares the BART approach with classical ones is presented in Hill [15], showing that estimation of causal effect are more precise than standard approaches when the complicate interactions among predictor is present and unknown.

This paper is a first approximation to the analysis of the causal impact of computers on students' performance. Future research could extend this analysis to other countries and subjects. Although this is beyond the scope of this work, which is mainly focused on a specific feature of the PISA 2012 survey in Spain, we provide as Supplementary Material the R code useful in order to analyze other countries of the PISA survey and make comparison with standard classical approaches to causal inference. The code can be also modified in order to estimate the causal effect of possible other educational policies that could be evaluated by means of the PISA survey.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Stefano Cabras has been supported by Ministerio de Ciencia e Innovación grant MTM2013-42323, ECO2012-38442, RYC-2012-11455, by Ministero dell'Istruzione, dell'Univesità e della Ricerca of Italy and by Regione Autonoma della Sardegna CRP-59903. Juan de Dios Tena Horrillo has been supported by Ministerio de Educación y Ciencia, ECO2009-08100 and ECO2012-32401.

ORCID

Stefano Cabras http://orcid.org/0000-0001-6690-8378
Juan de Dios Tena Horrillo http://orcid.org/0000-0002-0036-9735

References

[1] J. Angrist, and V. Lavy, *New evidence on classroom computers and pupil learning*, Econom. J. 112 (2002), pp. 735–765.

[2] A. Banerjee, S. Cole, E. Duflo, and L. Linden, *Remedying education: Evidence from two randomized experiments in India*, Q. J. Econ. 122 (2008), pp. 1235–1264.
[3] R. Barro, *Human capital and growth*, Am. Econ. Rev. 91 (2001), pp. 12–17.
[4] L. Breiman, *Random forests*, Mach. Learn. 45 (2001), pp. 5–32.
[5] H.A. Chipman, E.I. George, and R. Mculloch, *BART: Bayesian additive regression trees*, Ann. Appl. Statist. 4 (2010), pp. 266–298.
[6] R.B. D'Agostino, A.J. Belanger, and R.B. D' Agostino, Jr., *A suggestion for using powerful and informative tests of normality*, Amer. Statist. 44 (1990), pp. 316–321.
[7] A.P. Dawid, *Causal inference without counterfactuals*, J. Amer. Statist. Assoc. 95 (2000), pp. 407–424.
[8] J. Delfgaauw and R. Dur, *Incentives and workers motivation in the public sector*, Econom. J. 18 (2008), pp. 171–191.
[9] D. Figlio and L.W. Kenny, *Individual teacher incentives and student performance*, J. Public Econ. 91(5–6) (2007), pp. 901–914.
[10] T. Fuchs and L. Wö*β*mann, *What accounts for international differences in student per-formance? A re-examination using PISA data*. Empir. Econ. 4 (2007), pp. 433–464. doi:10.1007/s00181-006-0087-0.
[11] T. Fuchs and L. Woessman, *Computers and student learning: Bivariate and multivariate evidence on the availability and use of computers at home and at schools*, Brussels Econ. Rev. 47 (2004), pp. 359–389.
[12] A. Goolsbee and J. Guryan, *The impact of internet subsidies in public schools*, Rev. Econ. Statist. 88 (2006), pp. 336–347.
[13] E. Hanushek and D. Kimko, *Schooling, labor force quality and economic growth*, Amer. Econ. Rev. 90 (2000), pp. 1184–1208.
[14] L.H. Hill, *Bayesian nonparametric modeling for causal inference*, J. Comput. Graph. Statist. 20 (2011), pp. 217–240.
[15] H. Kirkpatrick and L. Cuban, *Computers make kids smarter – right? TECHNOS Q*. Educ. Technol. 7 (1998), pp. 1–11.
[16] M. Leonti, S. Cabras, C. Weckerle, M. Solinas, and L. Casu, *The causal dependence of present plant knowledge on herbals – contemporary medicinal plant use in Campania (Italy) compared to Matthioli (1568)*, J. Ethnopharmacol. 130 (2010), pp. 379–391.
[17] E. Leuven, M. Lindahl, H. Oosterbeek, and D. Webbink, *The effect of extra funding for disadvantaged pupils on achievement*, Rev. Econ. Statist. 89 (2007), pp. 721–736.
[18] L. Lochner, *Non-production benefits of education: Crime, health, and good citizenship*, NBER Working Papers 16722, National Bureau of Economic Research, Inc., 2011.
[19] S. Machin, S. McNally, and O. Silva, *New technology in schools: Is there a payoff*? Econom. J. 117 (2007), pp. 1145–1167.
[20] OECD, *PISA 2009 technical report*, OECD, Paris.
[21] S. Petrone, J. Rousseau, and S. Scricciolo, *Bayes and empirical Bayes: Do they merge*? Biometrika 101 (2014), pp. 285–302.
[22] P. Royston, *Comment on sg3.4. and an improved D'Agostino test*, Stata Tech. Bull. 3 (1991), pp. 23–24. Reprinted in Stata Technical Bulletin Reprints, 1, 110–112. College Station, TX: Stata Press.
[23] R.E. Schapire and Y. Singer, *Improved boosting algorithms using confidence-rated predictions*, Mach. Learn. 37 (1999), pp. 297–336.
[24] V. Spiezia, *Does computer use increase educational achievements*? *Student-level evidence from PISA*, OECD J. Econ. Stud. 1 (2010), pp. 1–22.