UNIVERSIDAD CARLOS III DE MADRID working papers

UC3M Working Papers
Statistics and Econometrics
15-15
ISSN 2387-0303
July 2015

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-48

A BAYESIAN MODEL TO ESTIMATE CAUSALITY IN PISA SCORES: A TUTORIAL WITH APPLICATION TO ICT

S. Cabras[ab], J.D. Tena[ab]

Abstract

This paper presents a step-by-step tutorial to estimate causal effects in PISA 2012 by means of a nonparametric Bayesian modeling approach known as Bayesian Additive Regression Trees (BART), with an illustration of the causal impact of ICT on Spanish students' performance. The R code is explained in a way that can be easily applied to other similar studies. The application shows that, compared to more traditional methodologies, the BART approach is particularly useful when a high-dimensional set of confounding variables is considered as its results are not based on a sampling hypothesis. BART allows for the estimation of different interactive effects between the treatment variable and other covariates. BART models do not require the analyst to make explicit subjective decisions in which covariates must be included in the final models. This makes it an easy procedure to guide policy makers' decisions in different contexts.

Keywords: *Causality in education, BART models, propensity score.*

[a] Department of Statistics, Universidad Carlos III de Madrid.
[b] Instituto Flores de Lemus, Universidad Carlos III de Madrid.

Running head: TUTORIAL ON BART FOR CAUSAL INFERENCE

A Bayesian model to estimate causality in PISA scores: a tutorial with application to ICT

Stefano Cabras[(1)] and Juan de Dios Tena Horrillo[(2)]

(1) Universidad Carlos III and Instituto Flores de Lemus, Madrid (Spain) and UniversitÃă di Cagliari, Cagliari (Italy).

(2) Universidad Carlos III and Instituto Flores de Lemus, Madrid (Spain) and UniversitÃă di Sassari and CRENoS, Sassari (Italy)

Abstract

This paper presents a step-by-step tutorial to estimate causal effects in PISA 2012 by means of a nonparametric Bayesian modeling approach known as Bayesian Additive Regression Trees (BART), with an illustration of the causal impact of ICT on Spanish students' performance. The R code is explained in a way that can be easily applied to other similar studies. The application shows that, compared to more traditional methodologies, the BART approach is particularly useful when a high-dimensional set of confounding variables is considered as its results are not based on a sampling hypothesis. BART allows for the estimation of different interactive effects between the treatment variable and other covariates. BART models do not require the analyst to make explicit subjective decisions in which covariates must be included in the final models. This makes it an easy procedure to guide policy makers' decisions in different contexts.

A Bayesian model to estimate causality in PISA scores: a tutorial with application to ICT

Introduction

The evaluation of the causal impact of discrete decisions on students' performance is an issue of obvious interest for a variety of stakeholders, such as researchers, school managers and/or politicians. However, developing an experimental design for this task is not always possible due to a potential range of ethical and economic considerations. Therefore, the use of observational studies may be regarded as an easy alternative for the estimation of causal effects in education.

The Program for International Student Assessment (PISA) provides comprehensive and internationally comparable information on students' performance, as well as on family and institutional factors. For this reason, it may be considered a fertile field in which to analyze the causal effects of different educational variables in a non-experimental setting. Nevertheless, this task necessarily requires the consideration of a large number of confounding variables to render the treated and the control samples comparable. This is a potential problem when considering traditional matching estimation techniques, such as propensity score, given that the two samples cannot be observed for the same values of confounding variables.

This paper presents a step-by-step tutorial to estimate causal effects in PISA 2012 by means of a nonparametric Bayesian modeling approach known as Bayesian Additive Regression Trees (BART). Originally developed in Chipman et al. (2010), is a very flexible nonparametric model also used in Leonti et al. (2010) for a similar causal analysis. This method, which addresses mainly the optimal estimation of response surface, i.e. the PISA score, allows for causal estimation in non-experimental works without being obliged to estimate two models, one to capture the potential endogeneity of the treatment variable and another to specify students' performance. Moreover, this flexible approach allows for both the inclusion of a large number of covariates, as it does not require a sampling hypothesis, and the estimation of a large number of interactive effects between the treatment and other variables in the analysis. The fact that it does not require any subjective decision by the analysist, except for deciding the response and the treated variable, makes it an easy procedure for decision makers to implement in different contexts.

We illustrate the use of BART models for causal analysis with an application to the estimation of the causal effect of information and communication technologies (ICT) on performance of Spanish students in mathematics as measured in PISA 2012. This tutorial extends from the collection of the database to the comparison of the estimated causal effect obtained with BART models and those obtained with other more traditional approaches, such as linear regression and matching.

The remainder of this paper is structured as follows. We explain the insight of causal estimation with non-experimental data in Section 2. Sections 3 and 4 briefly describe the estimation of causal effects using some traditional methodologies and the BART approach, respectively. We present the PISA database and provide some guidelines about how to download it in Section 5. Sections 6 and 7 show how to perform causal analysis with some traditional methodologies and with the BART approach in R, with an illustration of the estimated causal impact of ICT on Spanish students' performance in mathematics. We draw some conclusions in Section 8.

The causal estimation problem in brief

Assume that N individuals participate in the PISA test. For the ith 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 draw from the posterior distribution of the PISA test (OECD, 2009). Let z be a dummy variable that indicates the state of use of computers at school, the treatment variable where $z = 1$ if a tabletop, laptop or 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, $Y_i(1)$, and not under the use of computers, $Y_i(0)$. However, this is impossible because only one of these can be observed, while the other is unobservable and it is designated as the counterfactual result that has to be estimated with a regression model like the BART model described below. Such a model is used mainly 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 outcomes have been estimated the average total effect is defined as $ATE = E(Y(1) - Y(0))$, where the expected value is computed with respect to the probability distribution of Y for all the individuals, observed and potential outcomes. The causal effect for each individual is of no interest. Instead we are interested on the causal effect for a given set of individuals; for example those who have received the treatment $E(Y(1) - Y(0)|z = 1)$, that is, the set of individuals who 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 generally, if we have a set of covariates X, we can estimate the causal effect conditional on them, that is on $X = x$. 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 assigned 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 computers) that determines a favorable score on the PISA test compared to 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 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; 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, considering many covariates makes it impossible for some classical approaches, such as, the propensity score, to be immediately applied because treated and not-treated individuals cannot all be observed for the same value of $X = x$, and, thus, 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 cases 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 which may provide only locally optimal models. Unfortunately, there are no any theoretical ways to assess whether a local or global optimum has been reached, unless all possible models are fitted.

Due to these drawbacks in the use of classical devices as well as others that will be explained later, BART models not only free us from model specification, because they are nonparametric models estimated by observations, but also allow us to estimate the response and, thus, the counterfactual result with satisfactory precision. This model belongs to the class of nonparametric Bayesian models that allow us to perform conditional inference from the observed data about the causal effect, without considering resampling arguments that are necessary to interpret classical inference.

Traditional approaches for estimating causal effects.

Although a detailed discussion of the different methodologies to estimate causal effects is beyond the scope of this paper, linear regression models are, perhaps, the simplest and most common method of evaluating alternative explanations for a given outcome of interest. If we are interested in the estimation of causal effects, the basic strategy should be to avoid the potential omitted variable bias by including all the possible confounding factors that could affect both the probability of treatment and the response variable. The standard linear regression model takes the following form:

$$Y = \alpha + \delta z + \beta X + \epsilon, \tag{1}$$

where X is the design matrix of confounding factors supposed to render z uncorrelated with the error term ϵ.

Given that in the PISA database, individuals are weighted according to their importance in the sample and under the key assumption that covariates X in (1) contains all the relevant information to explain z and Y, we can consider the weighted least squares estimate of δ as an unbiased estimated of ATE.

However, the regression approach can be subject to at least two important pitfalls; see Morgan & Winship (2014). First, it rests on the assumption that the causal effect is weighted over students according to the PISA weights. In this case, the estimated causal effect represents a conditional variance weighted estimate of causal effects of individuals, and the causal estimation is unbiased and consistent only with this particularly weighted average that is not usually the parameter of interest. The second problem with the linear regression approach is that the strong ignorability condition does not necessarily imply that treatment is uncorrelated with the error term net of adjustment for X, as this error term depends on the specification of covariates in X. Therefore, in order to interpret the estimation of a regression strategy as an actual causal effect, we require the full inclusion in X of all covariates.

In contrast to this approach, matching estimation is a method of strategic subsampling across treated and control cases in such a way that the researcher selects a nontreated control case for each treated case based on the set of covariates X and nonmatched control cases are discarded. For this comparison, an estimation of the probability of treatment it is typically needed, i.e. *propensity score*. Subsequently, the average differences in the observed responses for the treated and matched cases are considered as the treatment effect estimate for individuals given the treatment. Matching estimators can be seen in many cases as weighted regressions, where the weights are functions of the estimated propensity scores (PS); see Imbens (2004).

Here, some of the most common problems with PS will be considered along with recent improvements that partially mitigate them. Such problems are absent in the BART model approach, beginning with the unconditional interpretation of the results; that is, we do not need to resort to hypothetical resampling schemes in order to interpret the significance of the estimated ATE effect. The next section contains a description of this approach.

The BART model: likelihood and priors

Let $\mathcal{D}$ be the available data that is, the set y,z and X observed for the N students and $\pi(\cdot|\cdot)$ be the probability distribution of the left argument conditional on the right one. The aim of the analysis is to estimate the posterior probability distribution of the causal effect, $\pi(ATE|\mathcal{D})$, or even more some conditional distribution to a suitable set of covariates, $\pi(ATE|\mathcal{D}, X = x)$. In order to do this, we use a nonparametric regression model. The novelty in this type 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, $\theta \in \Theta \notin \Re$, and a prior distribution $\pi(\theta)$, $\theta \in \Theta$. The likelihood function, $L(y; X, z, \theta)$, is obtained from the following additive regression model, where the conditional mean of Y is determined from the sum of estimated models for the response variable:

$$Y = \sum_{[j=1]}^{m} g(x, z; T_j, M_j) + \epsilon,\ \epsilon \sim Normal(0, \sigma^2), \tag{2}$$

where $g(x, z; T_j, M_j)$ is a regression binary tree (or classification tree if Y is a categorical variable) with its splitting variables and splitting 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 jth tree, $\mu_{ij} \in M_j$. 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 one obtained from a single big tree. Examples of these types of models are Boosting (Friedman, 2001) and Random Forest (Breiman, 2001). 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. (2010). 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. (2010) shows, the hyper parameters of all prior distributions are specified in relation to the observed sample. It produces priors that are dependent on the sample. This procedure, which is not very orthodox from a Bayesian point of view, is part of the approaches known as empirical Bayes methods, which are very popular and have been enhanced from a theoretical point of view (Petrone et al., 2014). As explained in Hill (2011), results for this type of analysis are robust with respect to prior modifications.

Using the priors specified above it is possible to simulate samples of the posterior distribution with a non-excessive computational effort using Markov Chain Monte Carlo (MCMC) and 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 steps when the conditional distributions for the remaining parameters are not available in a closed form expression. In particular, the distribution used to update the values of T_j and M_j consists of adding/dropping a terminal node and changing a split variable or a split point with some probabilities specified in Chipman et al. (2010). 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:

$$m(Y_i|x_i, z_i) = \int_{\theta\in\Theta} L_(Y_i; \theta) d\pi(\theta|\mathcal{D}) \tag{3}$$

which is practically estimated generating values of Y_i, using the normal distribution with the mean and variance for each value in the chain MCMC and the regression tress computed in x_i and z_i. In particular, we use $m = 500$ trees and 10000 MCMC steps after an initial burn-in of 1000 steps. In this way, the distribution for each individual and his corresponding counterfactual response can be estimated simply by estimating the response in $z_i = 1$ if the student does not have a computer in his/her school and in $z_i = 0$ otherwise. Once we obtain the predictive posterior distributions, we consider the difference between the factual and counterfactual responses to obtain the distribution of the individual causal effect. Finally, we estimate $\pi(ATE|\mathcal{D})$ 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 $X = x$.

In what follows, we illustrate how to estimate the causal effect of ICT on the performance of Spanish students in mathematics using the PISA database. We compare this procedure to other traditional alternative procedures used to perform causal analysis, such as the linear regression and propensity score models.

The PISA database

The PISA database contains information on the knowledge and abilities of students who are close to the end of their compulsory education. It is used mainly to determine how well these students are prepared for life after compulsory education instead of focusing on the evaluation of their curricular knowledge. The PISA surveys take place every three years. The last one took place in 2012, and the database can be downloaded from the webpage `http://pisa2012.acer.edu.au/downloads.php`. Here, we focus on the performance of Spanish students in mathematics. Observations must be taken as an outcome from a weighted survey from the total population of students in such a way that each observation has a relative weight according to its importance in the total population. This weight is included in the analysis that follows.

In this section, we provide a guide to creating a R dataframe to be used for our analysis. Raw data were first transformed into SPSS format files, as explained in `http://edutechwiki.unige.ch/en/PISA#Setting_up_SPSS_files`, and then read in R by means of `read.spss()`. The final `dat` dataframe contains:

Y the response variable, in this case is the first plausible value for math score `PV1MATH`. Note that Sample variance is equal to 7697.9 while imputation variance is 0.11, which indicates, in line with previous works, that most of the uncertainty in the population estimation corresponds to sample variability rather than 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 work are based only on the first plausible value.

z the treatment variable, in this case, the set `IC02Q01, IC02Q02, IC02Q03` is used to define the use of IT at schools, which is a one-dimensional variable, as explained below;

X possible covariates that act as confoudings factors for students `v.student.conf` and their schools `v.school.conf`. In this analysis, we included:

- Home information: 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; TIMEINT: total time using computers (in minutes).
- School information: CLSIZE: size of the class; SCMATEDU: quality of the educational resources in the school; STRATIO: students-teachers ratio; SMRATIO: math students-teachers ratio; SCHLTYPE: indicator of school ownership; RATCMP15: the index of computer availability.

More information about the different variables can be obtained from the PISA codebooks at `http://pisa2012.acer.edu.au/downloads.php`.

```
> rm(list=ls())
> #############################
> # Name of variables to be extracted
> v.resp=c("PV1MATH") # Response Variable
> v.treat=c("IC02Q01","IC02Q02","IC02Q03") # Treatment variable(s)
> # Student Confoundings
> v.student.conf=c("IMMIG", "HEDRES", "WEALTH", "ESCS","FAMSTRUC",
+  "HISCED","HISEI","HOMEPOS", "TIMEINT")
> # School Confoundings
> v.school.conf=c("CLSIZE","SCMATEDU","STRATIO",
+  "SMRATIO","SCHLTYPE","RATCMP15")
```

Here, we read the SPSS files, created according to `http://edutechwiki.unige.ch/en/PISA#Setting_up_SPSS_files` and merge everything into one unique data frame `dat`:

```
> library(intsvy)
>  dat <- pisa.select.merge(folder="mySAVfolder/",
+          school.file="INT_SCQ12_DEC03.SAV",
+          student.file="INT_STU12_DEC03.SAV",
+          parent.file="INT_PAQ12_DEC03.SAV",
+          student= c(v.resp,v.treat,v.student.conf),
+          parent =c(),
+          school = v.school.conf,
+          countries = "ESP")
>  dim(dat)

[1] 25313   155
```

We have 25313 students and 155 variables, which is not a random sample but, rather, a weighted sample with weights in the w vector:

```
> w=dat$W_FSTUWT
```

Such a vector enters in the computation of all regressions needed for classical approaches and in approximating the posterior distribution of the causal effect.

Consider the database with only the variables that we need and a subset of only complete cases that can be used to build the prediction model for the response. We end up with 16869 students and 19 variables:

```
> dat=dat[c(v.resp,v.treat,v.student.conf,v.school.conf)]
> names(dat)[names(dat)==v.resp]="y"
> w=w[complete.cases(dat)]
> w=w/sum(w)
> nw=function(w) w/sum(w)
> dat=dat[complete.cases(dat),] # Remove NAs
> dim(dat)

[1] 16869    19
```

Let's define the treatment status as treated if one of the following is used at school: desktop, laptop computer or tablet,

```
> z=factor(0+apply(dat[v.treat],1,function(xx) any(xx=="Yes, and I use it")))
> dat$z=z
> dat=dat[!(names(dat)%in%v.treat)]
> table(z)

z
    0     1
 4505 12364
```

Therefore, 12364 out of the 16869 students used ITC. This is a quiet unbalanced sample that poses problems in classical approaches.

Once we collected the database, we were able to obtain some descriptive statistics of the variables under analysis by the following command:

```
> by(dat, dat$z,summary)

dat$z: 0
       y                        IMMIG          HEDRES
 Min.   :186.4   Native           :4089   Min.   :-3.93000
 1st Qu.:442.8   Second-Generation:  44   1st Qu.:-0.69000
 Median :503.6   First-Generation : 372   Median : 0.04000
 Mean   :502.5                            Mean   : 0.02507
 3rd Qu.:565.0                            3rd Qu.: 1.12000
 Max.   :794.7                            Max.   : 1.12000

     WEALTH               ESCS
 Min.   :-2.890000   Min.   :-2.88000
 1st Qu.:-0.590000   1st Qu.:-0.82000
 Median :-0.110000   Median :-0.07000
 Mean   : 0.004444   Mean   :-0.05014
 3rd Qu.: 0.450000   3rd Qu.: 0.81000
 Max.   : 2.910000   Max.   : 2.55000

                                 FAMSTRUC                  HISCED
 Single parent (natural or otherwise): 466   None             :  32
 Two parents (natural or otherwise)  :4013   ISCED 1          : 215
 Other                               :  26   ISCED 2          : 650
                                             ISCED 3B, C      :  71
                                             ISCED 3A, ISCED 4:1039
                                             ISCED 5B         : 666
                                             ISCED 5A, 6      :1832
     HISEI          HOMEPOS            TIMEINT          CLSIZE    
 Min.   :11.01   Min.   :-3.7400   Min.   :  0.0   Min.   :13.0  
 1st Qu.:27.91   1st Qu.:-0.3300   1st Qu.: 69.0   1st Qu.:23.0  
 Median :44.94   Median : 0.1700   Median :122.0   Median :23.0  
 Mean   :49.12   Mean   : 0.1333   Mean   :175.6   Mean   :25.2  
 3rd Qu.:68.88   3rd Qu.: 0.6200   3rd Qu.:214.0   3rd Qu.:28.0  
 Max.   :88.96   Max.   : 3.7600   Max.   :823.0   Max.   :48.0  

    SCMATEDU           STRATIO           SMRATIO      
 Min.   :-3.59200   Min.   :  1.111   Min.   :   2.0  
 1st Qu.:-0.52140   1st Qu.:  8.682   1st Qu.:  65.0  
 Median : 0.01800   Median : 10.785   Median :  85.5  
 Mean   : 0.04361   Mean   : 12.034   Mean   : 106.0  
 3rd Qu.: 0.46060   3rd Qu.: 14.681   3rd Qu.: 128.5  
 Max.   : 1.97600   Max.   :139.000   Max.   :1820.0  

                         SCHLTYPE       RATCMP15      z       
 Private Independent         : 272   Min.   :0.0000   0:4505  
 Private government-dependent:1286   1st Qu.:0.3760   1:   0  
 Public                      :2947   Median :0.5360           
                                     Mean   :0.6292           
                                     3rd Qu.:0.7890           
                                     Max.   :6.2860           

------------------------------------------------------------ 
dat$z: 1
```

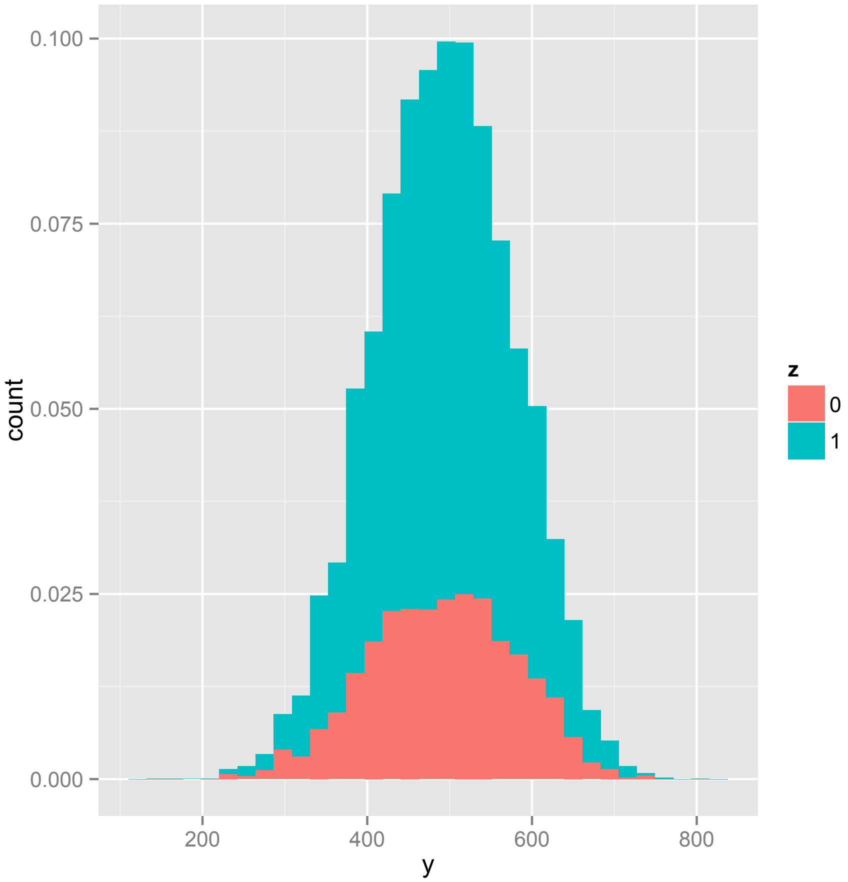
```
      y                       IMMIG            HEDRES             WEALTH
Min.   :150.6   Native           :11255   Min.   :-3.9300   Min.   :-5.3200
1st Qu.:444.2   Second-Generation:  122   1st Qu.:-0.6900   1st Qu.:-0.5900
Median :504.1   First-Generation :  987   Median : 0.0400   Median :-0.1100
Mean   :502.2                             Mean   : 0.1401   Mean   : 0.0296
3rd Qu.:561.8                             3rd Qu.: 1.1200   3rd Qu.: 0.4500
Max.   :811.8                             Max.   : 1.1200   Max.   : 2.9100

     ESCS                                            FAMSTRUC
Min.   :-3.1200   Single parent (natural or otherwise): 1152
1st Qu.:-0.8300   Two parents (natural or otherwise)  :11134
Median :-0.1400   Other                               :   78
Mean   :-0.1056
3rd Qu.: 0.6900
Max.   : 2.6000

              HISCED          HISEI           HOMEPOS            TIMEINT
None             : 114   Min.   :11.01   Min.   :-5.3300   Min.   :  0.0
ISCED 1          : 586   1st Qu.:27.91   1st Qu.:-0.3300   1st Qu.: 90.0
ISCED 2          :1820   Median :43.33   Median : 0.2000   Median :148.0
ISCED 3B, C      : 233   Mean   :47.48   Mean   : 0.1739   Mean   :194.2
ISCED 3A, ISCED 4:3184   3rd Qu.:67.04   3rd Qu.: 0.6200   3rd Qu.:246.0
ISCED 5B         :2033   Max.   :88.96   Max.   : 3.7600   Max.   :823.0
ISCED 5A, 6      :4394
    CLSIZE         SCMATEDU           STRATIO            SMRATIO
Min.   :13.00   Min.   :-3.5920   Min.   :  1.111   Min.   :   2.00
```

```
1st Qu.:23.00   1st Qu.:-0.5214   1st Qu.:  8.250   1st Qu.:  62.57
Median :23.00   Median : 0.0180   Median : 10.538   Median :  83.27
Mean   :25.02   Mean   : 0.1075   Mean   : 11.736   Mean   : 105.27
3rd Qu.:28.00   3rd Qu.: 0.7524   3rd Qu.: 14.649   3rd Qu.: 121.78
Max.   :48.00   Max.   : 1.9760   Max.   :139.000   Max.   :1820.00

                        SCHLTYPE       RATCMP15      z
Private Independent         : 558   Min.   :0.0000   0:    0
Private government-dependent:3972   1st Qu.:0.4440   1:12364
Public                      :7834   Median :0.6000
                                    Mean   :0.7451
                                    3rd Qu.:0.9330
                                    Max.   :8.0000
```

A glance at the above statistics makes clear that the two samples are different. For example, students with z=1 show on average higher values of HEDRES, WEALTH or HOMEPOS than students with z=0. This suggests the necessity to control for the influence of all these variables in a causal analysis.

The following histograms illustrate the conditional empirical distribution of the PISA scores (the first plausible value for math), y, conditionally on z. We can see that, marginal to all other student characteristics, these conditional distributions are very similar. However, this is due to the presence of confounding factors that hide the effect of ICT at school on PISA scores.

```
> library(ggplot2)
> library(Hmisc)

> ggplot(dat, aes(y,weights=w,fill=z))+geom_histogram()
```



Some standard causal estimators in R

The well-known `glm()` command can be used to estimate a simple linear regression or, depending on the nature of the response variable, a logistic regression, just by changing the `family()` argument. What matters here is the that sample weights are automatically included in the calculus using the argument `weight`, which is general to all regression approaches, as illustrated below:

```
> linear.reg=glm(y~.,data=data.frame(y=dat$y,X),weight=w)

> summary(linear.reg)
```

```
Call:
glm(formula = y ˜ ., data = data.frame(y = dat$y, X), weights = w)


Deviance Residuals:
    Min       1Q   Median       3Q      Max  
-5.6991  -0.2480   0.0316   0.2901   3.7963  


Coefficients:
                                             Estimate Std. Error t value
(Intercept)                                -9.284e+02  4.783e+02  -1.941
IMMIGSecond-Generation                     -2.239e+01  4.849e+00  -4.616
IMMIGFirst-Generation                      -3.033e+01  2.191e+00 -13.844
HEDRES                                     -1.194e+01  9.749e-01 -12.245
WEALTH                                     -4.570e+01  1.679e+00 -27.220
ESCS                                       -5.393e+02  1.887e+02  -2.857
FAMSTRUCTwo parents (natural or otherwise) -4.272e-01  1.948e+00  -0.219
FAMSTRUCOther                              -3.724e+01  8.307e+00  -4.484
HISCEDISCED 1                               1.787e+02  5.179e+01   3.450
HISCEDISCED 2                               4.052e+02  1.286e+02   3.151
HISCEDISCED 3B, C                           5.444e+02  1.800e+02   3.024
HISCEDISCED 3A, ISCED 4                     7.109e+02  2.313e+02   3.073
HISCEDISCED 5B                              7.727e+02  2.570e+02   3.006
HISCEDISCED 5A, 6                           1.044e+03  3.469e+02   3.008
HISEI                                       1.249e+01  4.145e+00   3.014
HOMEPOS                                     2.525e+02  6.756e+01   3.738
TIMEINT                                    -4.976e-02  3.821e-03 -13.025
```

```
CLSIZE                                     -1.143e-01  1.099e-01  -1.041
SCMATEDU                                    2.231e+00  6.660e-01   3.351
STRATIO                                    -3.754e-02  7.190e-02  -0.522
SMRATIO                                     1.853e-02  4.904e-03   3.778
SCHLTYPEPrivate government-dependent       -4.062e+00  2.586e+00  -1.571
SCHLTYPEPublic                             -1.848e+01  2.486e+00  -7.434
RATCMP15                                   -1.457e+00  1.290e+00  -1.130
z1                                          3.970e+00  1.305e+00   3.043
                                           Pr(>|t|)
(Intercept)                                0.052252 .
IMMIGSecond-Generation                     3.94e-06 ***
IMMIGFirst-Generation                       < 2e-16 ***
HEDRES                                      < 2e-16 ***
WEALTH                                      < 2e-16 ***
ESCS                                       0.004278 **
FAMSTRUCTwo parents (natural or otherwise) 0.826442
FAMSTRUCOther                              7.39e-06 ***
HISCEDISCED 1                              0.000563 ***
HISCEDISCED 2                              0.001632 **
HISCEDISCED 3B, C                          0.002495 **
HISCEDISCED 3A, ISCED 4                    0.002120 **
HISCEDISCED 5B                             0.002650 **
HISCEDISCED 5A, 6                          0.002634 **
HISEI                                      0.002584 **
HOMEPOS                                    0.000186 ***
TIMEINT                                     < 2e-16 ***
```

```
CLSIZE                                    0.298084
SCMATEDU                                  0.000808 ***
STRATIO                                   0.601617
SMRATIO                                   0.000159 ***
SCHLTYPEPrivate government-dependent      0.116285
SCHLTYPEPublic                            1.11e-13 ***
RATCMP15                                  0.258652
z1                                        0.002350 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1


(Dispersion parameter for gaussian family taken to be 0.3283823)


    Null deviance: 7126.5  on 16868  degrees of freedom
Residual deviance: 5531.3  on 16844  degrees of freedom
AIC: 204115


Number of Fisher Scoring iterations: 2

> par(mfrow=c(2,2))
> plot(linear.reg)
```

Thus, a simple estimation of the causal effect of z on student performance is given by the estimated coefficient 3.97, which is significant at some conventional values. The coefficient of $z = 1$ could be interpreted as causal effect if all relevant covariates were introduced in the model with all their interactions. However, this is extremely difficult, and, as discussed in Section 3, a matching estimator can be used to overcome this pitfall of the linear regression technique, when used to conduct a causal analysis.

A typical approach consists of estimating the propensity score using logistic regression before perfoming the matching. These two steps can be easily run in R by means of the package `Matching`. In our example, we could use the commands:

```
> library(Matching)
> score=glm(z~.,data=X,weight=w,family=quasibinomial())
> summary(score)
```

```
Call:
glm(formula = z ~ ., family = quasibinomial(), data = X, weights = w)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.041383  -0.004815   0.002989   0.004998   0.021682

Coefficients:
                                             Estimate Std. Error t value
(Intercept)                                 1.4242258 14.6649492   0.097
IMMIGSecond-Generation                     -0.0252920  0.1454695  -0.174
IMMIGFirst-Generation                      -0.0424445  0.0661806  -0.641
HEDRES                                      0.2755387  0.0301859   9.128
WEALTH                                      0.2110849  0.0522649   4.039
ESCS                                        0.1257630  5.7875723   0.022
FAMSTRUCTwo parents (natural or otherwise)  0.1367852  0.0581225   2.353
FAMSTRUCOther                               0.3405603  0.2670018   1.275
```

```
HISCEDISCED 1                              -0.7672763  1.5897365  -0.483
HISCEDISCED 2                              -0.7125142  3.9435331  -0.181
HISCEDISCED 3B, C                          -0.5595017  5.5187300  -0.101
HISCEDISCED 3A, ISCED 4                    -0.7720550  7.0923226  -0.109
HISCEDISCED 5B                             -0.7532584  7.8815151  -0.096
HISCEDISCED 5A, 6                          -0.9838771 10.6372991  -0.092
HISEI                                      -0.0022556  0.1270942  -0.018
HOMEPOS                                    -0.2817867  2.0717910  -0.136
TIMEINT                                     0.0011462  0.0001246   9.197
CLSIZE                                     -0.0073291  0.0033697  -2.175
SCMATEDU                                    0.0313542  0.0205633   1.525
STRATIO                                    -0.0043420  0.0020363  -2.132
SMRATIO                                     0.0005761  0.0001740   3.310
SCHLTYPEPrivate government-dependent        0.3494151  0.0772320   4.524
SCHLTYPEPublic                              0.0846907  0.0729966   1.160
RATCMP15                                    0.3450919  0.0475195   7.262
                                           Pr(>|t|)
(Intercept)                                0.922634
IMMIGSecond-Generation                     0.861974
IMMIGFirst-Generation                      0.521309
HEDRES                                      < 2e-16 ***
WEALTH                                     5.40e-05 ***
ESCS                                       0.982664
FAMSTRUCTwo parents (natural or otherwise) 0.018614 *
FAMSTRUCOther                              0.202151
HISCEDISCED 1                              0.629355
HISCEDISCED 2                              0.856622
HISCEDISCED 3B, C                          0.919248
HISCEDISCED 3A, ISCED 4                    0.913317
HISCEDISCED 5B                             0.923861
HISCEDISCED 5A, 6                          0.926307
HISEI                                      0.985840
HOMEPOS                                    0.891814
TIMEINT                                     < 2e-16 ***
CLSIZE                                     0.029644 *
SCMATEDU                                   0.127337
STRATIO                                    0.032995 *
SMRATIO                                    0.000934 ***
SCHLTYPEPrivate government-dependent       6.10e-06 ***
SCHLTYPEPublic                             0.245984
RATCMP15                                   3.98e-13 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for quasibinomial family taken to be 5.95755e-05)

    Null deviance: 1.1678  on 16868  degrees of freedom
Residual deviance: 1.1438  on 16845  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

```
> #match=Match(Y=dat$y, Tr=dat$z==1, X=score$fitted.values,estimand="ATE",weights=nrow(dat)

> #save(match,file="match.RData")

> load(file="match.RData")

> summary(match)
```

```
Estimate...  2.0546

AI SE......  1.9683

T-stat.....  1.0438

p.val......  0.29658


Original number of observations..............  16869

Original number of treated obs (weighted)....  12303.74

Original number of treated obs...............  12364

Matched number of observations...............  16869

Matched number of observations  (unweighted).  207697
```

Before seeing the results, it is interesting to check whether we have achieved a satisfactory balance between the treatment and control groups. Generally, one requests balance statistics on an ad-hoc selection of higher order terms and interactions that were included in the propensity score matching Dehejia & Wahba (1999). However, given the large number of covariates in our example, we refer to the `MatchBalance()` function using as an example the school type variable, which is an important confounding effect that should be taken into account.

```
> MatchBalance((z==1)~SCHLTYPE,data=X, match.out=match,

+                 nboots=500,weights=nrow(dat)*w)
```

```
***** (V1) SCHLTYPEPrivate government-dependent *****

                       Before Matching                  After Matching
```

```
mean treatment........      0.255                       0.24135
mean control..........    0.20446                         0.249
std mean diff.........     11.595                       -1.7864

mean raw eQQ diff.....   0.035738                      0.023804
med  raw eQQ diff.....          0                             0
max  raw eQQ diff.....          1                             1

mean eCDF diff........   0.017897                      0.011902
med  eCDF diff........   0.017897                      0.011902
max  eCDF diff........   0.035795                      0.023804

var ratio (Tr/Co).....     1.1678                       0.97917
T-test p-value........ 1.6525e-12                      0.079496


***** (V2) SCHLTYPEPublic *****
                       Before Matching                   After Matching
mean treatment........    0.68136                       0.69065
mean control..........    0.71976                        0.6796
std mean diff.........    -8.2409                        2.3898

mean raw eQQ diff.....   0.020644                      0.032316
med  raw eQQ diff.....          0                             0
max  raw eQQ diff.....          1                             1
```

```
mean eCDF diff........   0.010274                     0.016158

med  eCDF diff........   0.010274                     0.016158

max  eCDF diff........   0.020548                     0.032316


var ratio (Tr/Co).....     1.0762                      0.98122

T-test p-value........ 1.0635e-06                     0.022685




Before Matching Minimum p.value: 1.6525e-12

Variable Name(s): SCHLTYPEPrivate government-dependent  Number(s): 1


After Matching Minimum p.value: 0.022685

Variable Name(s): SCHLTYPEPublic  Number(s): 2
```

The school type is a very important confounding factor, and we can see that it cannot be matched with this approach, as it is significantly different between the treated and control groups ($p < 0.05$).

Recently, however, more sophisticated procedures have been used to find an optimal balance for the data; see the examples regarding the use of the `GenMatch()` function in Sekhon (2008).

```
> #gmatch=GenMatch(Tr=dat$z==1, X=score$fitted.values,estimand="ATE",weights=nrow(dat)*w)

> #save(gmatch,file="gmatch.Rdata")

> load("gmatch.Rdata")

> match2=Match(Y=dat$y, Tr=dat$z==1, X=score$fitted.values,

+              estimand="ATE",weights=nrow(dat)*w,Weight.matrix=gmatch)

> MatchBalance((z==1)~SCHLTYPE,data=X, match.out=match2,
+                 nboots=500,weights=nrow(dat)*w)

***** (V1) SCHLTYPEPrivate government-dependent *****
                       Before Matching                   After Matching
mean treatment........      0.255                      0.24175
mean control..........    0.20446                      0.24836
std mean diff.........     11.595                      -1.5428

mean raw eQQ diff.....   0.035738                     0.023657
med  raw eQQ diff.....          0                            0
max  raw eQQ diff.....          1                            1

mean eCDF diff........   0.017897                     0.011829
med  eCDF diff........   0.017897                     0.011829
max  eCDF diff........   0.035795                     0.023657

var ratio (Tr/Co).....     1.1678                      0.98196
T-test p-value........ 1.6525e-12                      0.12948


***** (V2) SCHLTYPEPublic *****
                       Before Matching                   After Matching
mean treatment........    0.68136                      0.69014
mean control..........    0.71976                      0.68037
std mean diff.........    -8.2409                       2.1133

mean raw eQQ diff.....   0.020644                     0.032265
```

```
med  raw eQQ diff.....          0                            0
max  raw eQQ diff.....          1                            1


mean eCDF diff........   0.010274                     0.016132
med  eCDF diff........   0.010274                     0.016132
max  eCDF diff........   0.020548                     0.032265


var ratio (Tr/Co).....     1.0762                      0.98335
T-test p-value........ 1.0635e-06                     0.044007




Before Matching Minimum p.value: 1.6525e-12
Variable Name(s): SCHLTYPEPrivate government-dependent  Number(s): 1


After Matching Minimum p.value: 0.044007
Variable Name(s): SCHLTYPEPublic  Number(s): 2
```

We can see that the school type is still not matched between cases and controls ($p < 0.05$, $p = 0.044$). If such an important variable is not matched, then the interaction of this with the parents' education-level variable `HISCED` surely is not matched.

Therefore, in general, matching estimation provides a more reliable estimation of causal effects than simple regression models as the former uses weights that make the treatment and control groups comparable. However, when there are a large number or regressors, as in our case, strong hypothesis about the covariates to be included in the model is required in order to obtain a balance match between the treatment and control groups. Moreover, consideration of all possible type interactions between the treatment indicator and other covariates of the model can be impractical in many cases. Again, this forces the analyst to consider only interactive effects among first- or second-order covariates or to use algorithms such as forward or backward variable selection that provide locally optimal models. Unfortunately, there is no theoretical justification to guide us in assessing the scope of a local instead of a global optimum.

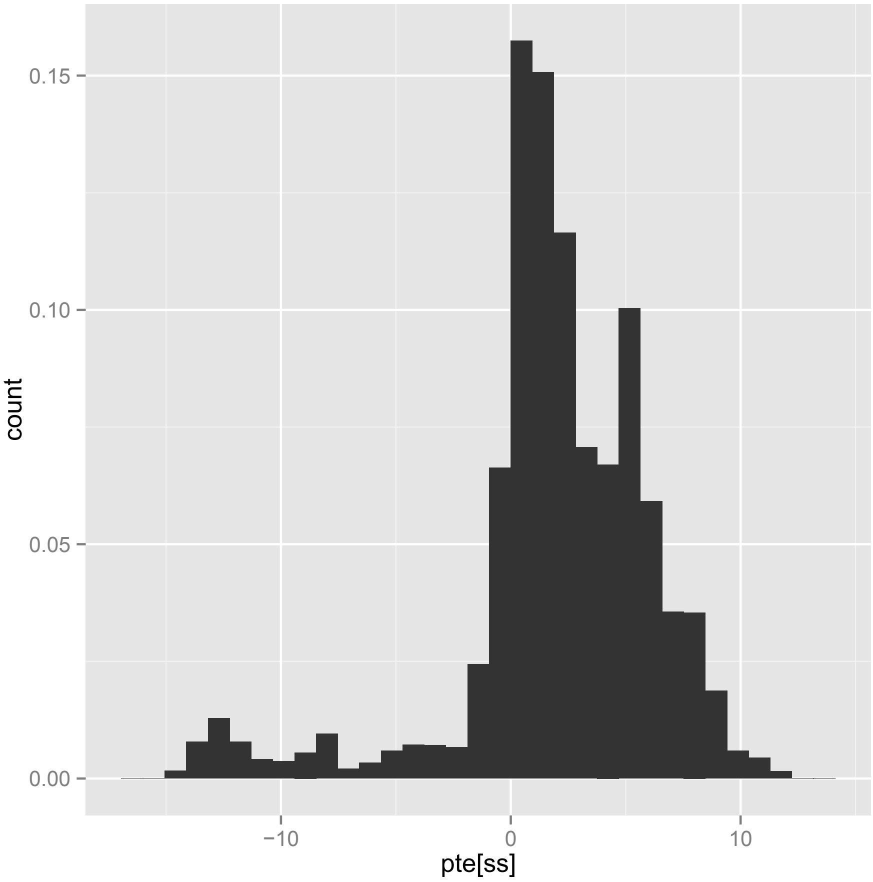
How to estimate causal effects with BART

In order to estimate the prediction model and the prediction for the counterfactuals using BART, we start by defining the prediction matrix `X` with observed `z` along with that of counterfactuals `Xc`. The train data are `X` and `dat$y`, while the prediction is made for `Xc`, which is the same set of covariates, except for the switched treatment status. We do not care to define the possible interactions between covariates, as these will be estimated in the BART model. All BART tuning parameters are left at their default values.

```
> Xc=X=dat[,-1] # The first column is the response
> Xc$z=factor((1:0)[X$z])
> library(BayesTree)
> # This may take a while ...
> bartFit = bart(X,dat$y,Xc)
> pte=apply(bartFit$yhat.train-bartFit$yhat.test,2,mean)
```

The posterior distribution of the marginal causal effect on treated (i.e. ATE) is obtained from the simulated differences between the mean of the posterior predictive distribution for actual students that used the ITC at school and the mean of the posterior predictive distribution for the same students, assuming that they do not use ITC at school. This distribution is reported in the following histogram:

```
> ss=dat$z==1
> ggplot()+geom_histogram(aes(pte[ss],weights=nw(w[ss])))
```



Using the following commands, we can get information about the posterior probability of a positive effect and its magnitude, respectively:

```
> weighted.mean(pte[ss]>0,w[ss])

[1] 0.8236357

> weighted.mean(pte[ss],w[ss])

[1] 1.995664
```

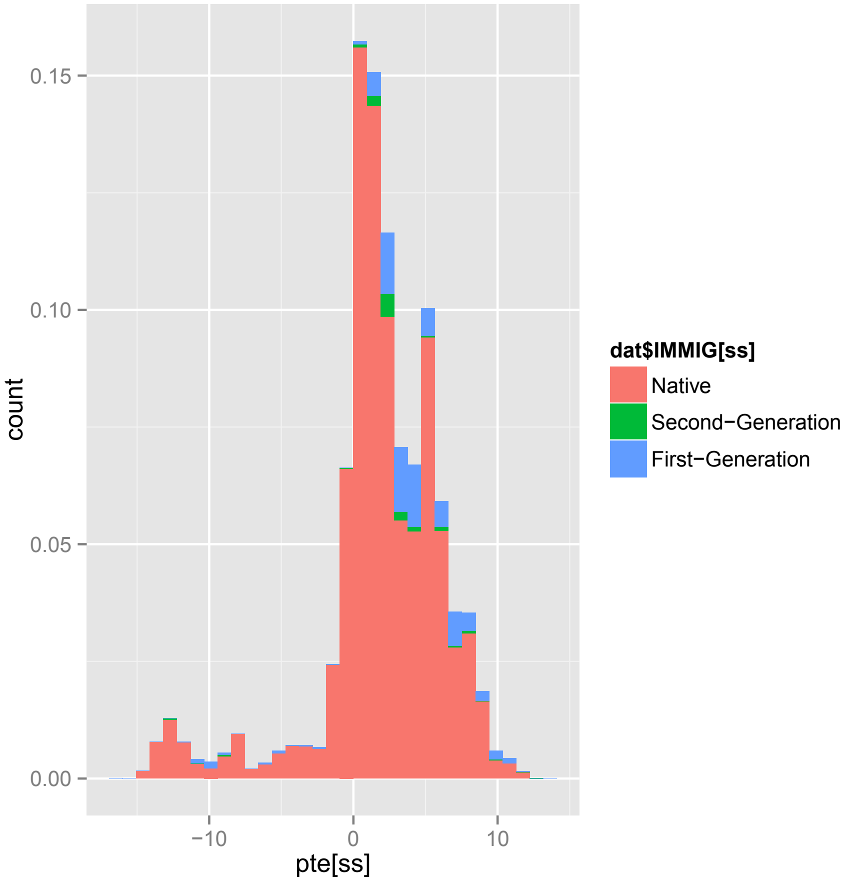
It is also straightforward to obtain credible interval values of the estimated causal effects:

```
> wtd.quantile(pte[ss], weights=w[ss], probs=c(0.025, .975),normwt=TRUE)
```

```
      2.5%       97.5%
-11.853324   8.648826

> wtd.quantile(pte[ss], weights=w[ss], probs=c(0.05, .95),normwt=TRUE)

       5%       95%
-7.834857  7.951276
```

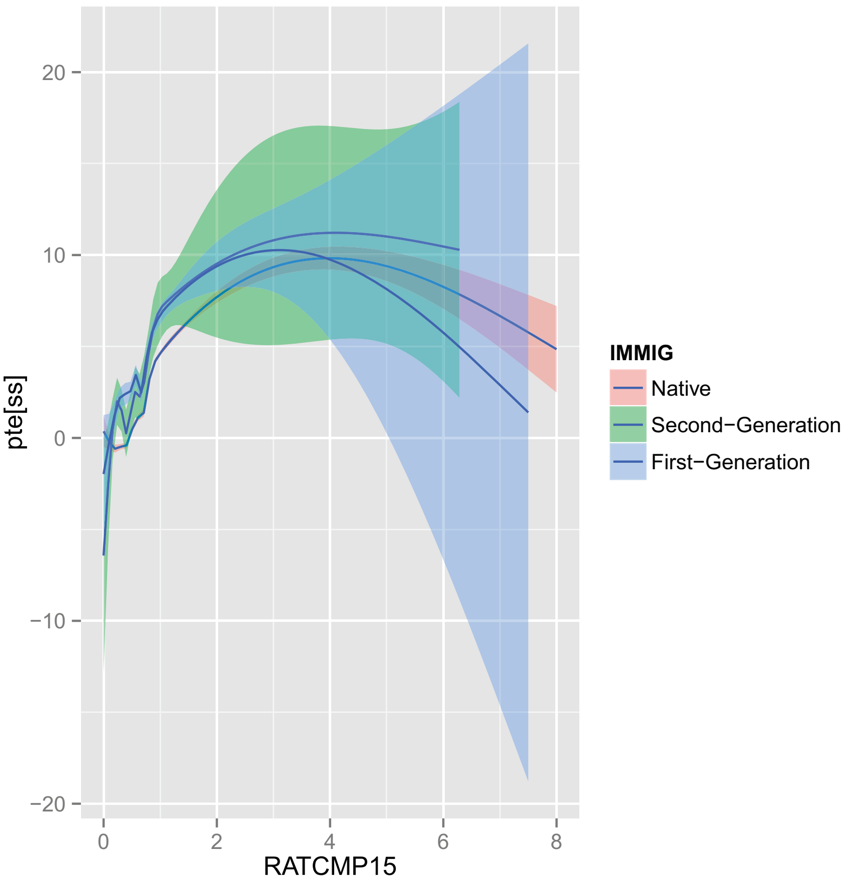
One important advantage of the BART methodology in this context is that it allows for the analysis of conditional causal effects. For instance, it is possible to estimate the ATE on non-native students and compare it with native ones. For non-native the posterior probability of a positive effect is around 94%, and it reduces to 81% for native students. We can also compare the estimated ATE for native, first-generation and second-generation students.

```
> cond.nn=(dat$z==1)&(dat$IMMIG!="Native")
> weighted.mean(pte[cond.nn]>0,w[cond.nn])

[1] 0.9357242

> cond.n=(dat$z==1)&(dat$IMMIG=="Native")
> weighted.mean(pte[cond.n]>0,w[cond.n])

[1] 0.8121684

> ggplot()+geom_histogram(aes(pte[ss],weights=nw(w[ss]),fill=dat$IMMIG[ss]))
```



Another interesting exercise is to analyze the interaction of the three groups of students with the ratio of computers to students The following graph illustrates the conditional regression functions, along with the 95% credible intervals for the posterior distribution of ATE conditional on immigration status and the ratio of computers to students.

```
> ggplot(dat[ss,])+geom_smooth(aes(y=pte[ss],x=RATCMP15,fill=IMMIG,weight=w[ss]))
```



The picture allows us to identify the optimal number of computers to students for each group. Thus, an increase in the number of computers is more effective when they are more scarce for non-native than for native students. However, when there are many computers per students, i.e., more than 4 every 15 students, increasing the number of computers does not improve mathematics performance.

Concluding remarks

We have illustrated how to estimate the effect of ICT on the performance of Spanish students in mathematics by means of the BART model in R, as well as its main advantages over other more traditional approaches. In particular, we have shown how difficult is to obtain a balanced sample for the treatment and control groups under classical methods even when more sophisticated automated process to search the data for the best matches, such as `GenMatch`, are used. In general, this is likely to be a problem when the number of potential confounding variables is large, as is typically the case with the PISA database. BART models are a way to circumvent this issue. In addition, BART models provide a result that it is not based on hypothetical resampling arguments which are very difficult to justify in causal analysis.

This tutorial also explains how to estimate conditional causal effects for different types of students and to obtain implications for policy makers as such as finding the optimal level of ICT investment for a target group of students (e.g., native versus non-native students). In principle, the fact that it does not require any subjective decision by the analyst, apart from defining the response and the treated variable, makes it an easy procedure for decision makers to implement in different contexts.

References

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5-32.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266-298. Retrieved from `doi:10.1214/09-AOAS285`

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, *94*(448), 1053–1062.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189-1232.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 1-24. Retrieved from http://pubs.amstat.org/doi/abs/10.1198/jcgs.2010.08162 doi: 10.1198/jcgs.2010.08162

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, *86*(1), 4–29.

Leonti, M., Cabras, S., Weckerle, C., Solinas, M., & L., C. (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.

Morgan, S. L., & Winship, C. (2014). *Counterfactuals and causal inference.* Cambridge University Press.

OECD. (2009). *Pisa 2009 technical report* (Tech. Rep.). OECD,Paris.

Petrone, S., Rousseau, J., & Scricciolo, C. (2014). Bayes and empirical bayes: do they merge? *Biometrika*, *to appear*(–).

Sekhon, J. S. (2008). Multivariate and propensity score matching software with automated balance optimization: the matching package for r. *Journal of Statistical Software*.

Author Note

Corresponding Author: Stefano Cabras, Universidad Carlos III, Department of Statistics, C/Madrid 126. 28903 Getafe (Madrid). Office: 10.1.02. Phone: (+34) 916249849, e-mail:stefano.cabras@uc3m.es.

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 EducaciÃşn y Ciencia, ECO2009-08100 and ECO2012-32401.