Statistical Tools for Causal Inference

The SKY Community

2019-04-13

Contents

Introduction

Tools of causal inference are the basic statistical building block behind most scientific results. It is thus extremely useful to have an open source collectively aggreed upon resource presenting and assessing them, as well as listing the current unresolved issues. The content of this book covers the basic theoretical knowledge and technical skills required for implementing staistical methods of causal inference. This means:

- Understanding of the basic language to encode causality,
- Knowledge of the fundamental problems of inference and the biases of intuitive estimators,
- Understanding of how econometric methods recover treatment effects,
- Ability to compute these estimators along with an estimate of their precision using the statistical software R combined with latex using Rmarkdown.

All the notions and estimators are introduced using a numerical example and simulations.

All the code behind this book is written in Rmarkdown and is publically available on GitHub. Feel free to propose corrections and updates.

Part I

The Two Fundamental Problems of Inference

When trying to estimate the effect of a program on an outcome, we face two very important and difficult problems: the Fundamental Problem of Causal Inference (FPCI) and the Fundamental Problem of Statistical Inference (FPSI).

In its most basic form, the FPCI states that our causal parameter of interest (TT, short for Treatment on the Treated, that we will define shortly) is fundamentally unobservable, even when the sample size is infinite. The main reason for that is that one component of TT, the outcome of the treated had they not received the program, remains unobservable. We call this outcome a counterfactual outcome. The FPCI is a very dispiriting result, and is actually the basis for all of the statistical methods of causal inference. All of these methods try to find ways to estimate the counterfactual by using observable quantities that hopefully approximate it as well as possible. Most people, including us but also policymakers, generally rely on intuitive quantities in order to generate the counterfactual (the individuals without the program or the individuals before the program was implemented). Unfortunately, these approximations are generally very crude, and the resulting estimators of TT are generally biased, sometimes severely.

The Fundamental Problem of Statistical Inference (FPSI) states that, even if we have an estimator E that identifies TT in the population, we cannot observe E because we only have access to a finite sample of the population. The only thing that we can form from the sample is a sample equivalent $\hat{E}$ to the population quantity E, and $\hat{E} \neq E$. Why is $\hat{E} \neq E$? Because a finite sample is never perfectly representative of the population. What can we do to deal with the FPSI? I am going to argue that there are mainly two things that we might want to do: estimating the extent of sampling noise and decreasing sampling noise.

Chapter 1

The Fundamental Problem of Causal Inference

In order to state the FPCI, we are going to describe the basic language to encode causality set up by Rubin, and named Rubin Causal Model (RCM). RCM being about partly observed random variables, it is hard to make these notions concrete with real data. That's why we are going to use simulations from a simple model in order to make it clear how these variables are generated. The second virtue of this model is that it is going to make it clear the source of selection into the treatment. This is going to be useful when understanding biases of intuitive comparisons, but also to discuss the methods of causal inference. A third virtue of this approach is that it makes clear the connexion between the treatment effects literature and models. Finally, a fourth reason that it is useful is that it is going to give us a source of sampling variation that we are going to use to visualize and explore the properties of our estimators.

I use X_i to denote random variable X all along the notes. I assume that we have access to a sample of N observations indexed by $i \in \{1, \dots, N\}$. ''i'' will denote the basic sampling units when we are in a sample, and a basic element of the probability space when we are in populations. Introducing rigorous measure-theoretic notations for the population is feasible but is not necessary for comprehension.

When the sample size is infinite, we say that we have a population. A population is a very useful fiction for two reasons. First, in a population, there is no sampling noise: we observe an infinite amount of observations, and our estimators are infinitely precise. This is useful to study phenomena independently of sampling noise. For example, it is in general easier to prove that an estimator is equal to TT under some conditions in the population. Second, we are most of the time much more interested in estimating the values of parameters in the population rather than in the sample. The population parameter, independent of sampling noise, gives a much better idea of the causal parameter for the population of interest than the parameter in the sample. In general, the estimator for both quantities will be the same, but the estimators for the effetc of sampling noise on these estimators will differ. Sampling noise for the population parameter will generally be larger, since it is affected by another source of variability (sample choice).

1.1 Rubin Causal Model

The RCM is made of three distinct building blocks: a treatment allocation rule, that decides who receives the treatment; potential outcomes, that measure how each individual reacts to the treatment; the switching equation that relates potential outcomes to observed outcomes through the allocation rule.

1.1.1 The treatment allocation rule

The first building block of the RCM is the treatment allocation rule. Throughout this class, we are going to be interested in inferring the causal effect of only one treatment with respect to a control condition. Extensions to multi-valued treatments are in general self-explanatory.

In the RCM, treatment allocation is captured by the variable D_i. $D_i = 1$ if unit i receives the treatment and $D_i = 0$ if unit i does not receive the treatment and thus remains in the control condition.

The treatment allocation rule is critical for several reasons. First, because it switches the treatment on or off for each unit, it is going to be at the source of the FPCI. Second, the specific properties of the treatment allocatoin rule are going to matter for the feasibility and bias of the various econometric methods that we are going to study.

Let's take a few examples of allocation rules. These allocation rules are just examples. They do not cover the space of all possible allocation rules. They are especially useful as concrete devices to understand the sources of biases and the nature of the allocation rule. In reality, there exists even more complex allocation rules (awareness, eligibility, application, acceptance, active participation). Awareness seems especially important for program participation and has only been tackled recently by economists.

First, some notation. Let's imagine a treatment that is given to individuals. Whether each individual receives the treatment partly depends on the level of her outcome before receiving the treatment. Let's denote this variable Y_i^B, with B standing for "Before". It can be the health status assessed by a professional before deciding to give a drug to a patient. It can be the poverty level of a household used to assess its eligibilty to a cash transfer program.

1.1.1.1 The sharp cutoff rule

The sharp cutoff rule means that everyone below some threshold $\bar{Y}$ is going to receive the treatment. Everyone whose outcome before the treatment lies above $\bar{Y}$ does not receive the treatment. Such rules can be found in reality in a lot of situations. They might be generated by administrative rules. One very simple way to model this rule is as follows:

$$D_i = \mathbb{1}[Y_i^B \leq \bar{Y}], \tag{1.1}$$

where $\mathbb{1}[A]$ is the indicator function, taking value 1 when A is true and 0 otherwise.
Example 1.1 (Sharp cutoff rule)**.** Imagine that $Y_i^B = \exp(y_i^B)$, with $y_i^B = \mu_i + U_i^B$, $\mu_i \sim \mathcal{N}(\bar{\mu}, \sigma^2_{\mu})$ and $U_i^B \sim \mathcal{N}(0, \sigma^2_U)$. Now, let's choose some values for these parameters so that we can generate a sample of individuals and allocate the treatment among them. I'm going to switch to R for that.

```
param <- c(8,.5,.28,1500)
names(param) <- c("barmu","sigma2mu","sigma2U","barY")
param
```

```
##    barmu sigma2mu  sigma2U     barY
##     8.00     0.50     0.28  1500.00
```

Now, I have choosen values for the parameters in my model. For example, $\bar{\mu} = 8$ and $\bar{Y} = 1500$. What remains to be done is to generate Y_i^B and then D_i. For this, I have to choose a sample size ($N = 1000$) and then generate the shocks from a normal.

```
# for reproducibility, I choose a seed that will give me the same random sample each time I run the pro
set.seed(1234)
N <-1000
mu <- rnorm(N,param["barmu"],sqrt(param["sigma2mu"]))
```

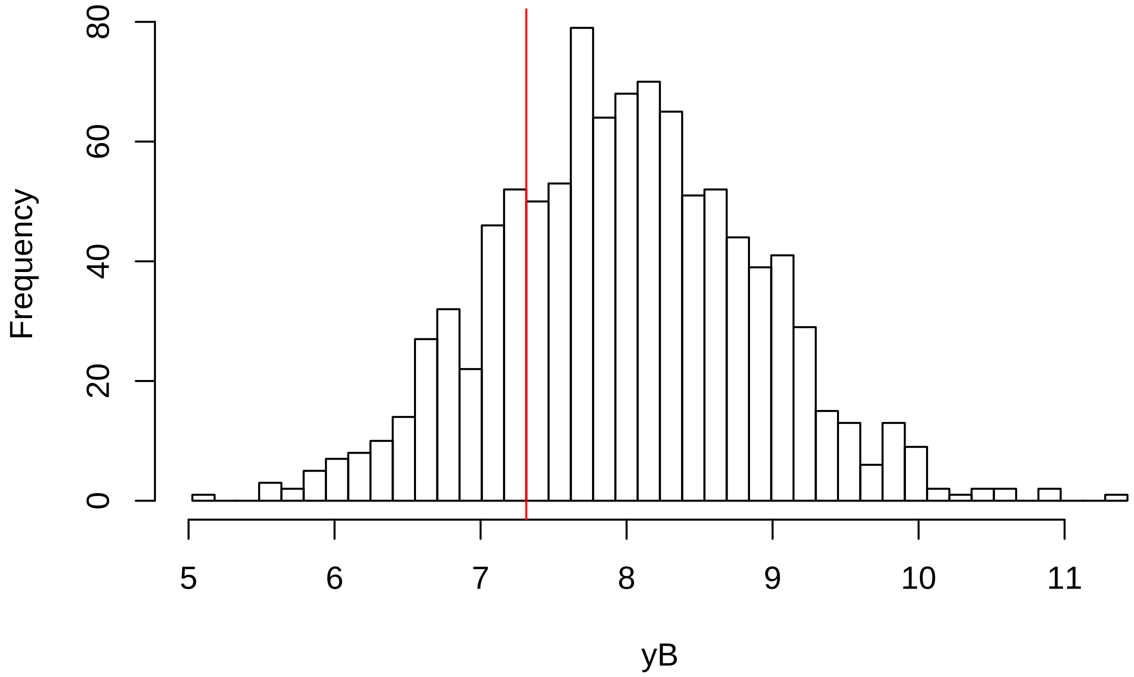

Figure 1.1: Histogram of y_B

Table 1.1: Treatment allocation with sharp cutoff rule

0	771
1	229

```
UB <- rnorm(N,0,sqrt(param["sigma2U"]))
yB <- mu + UB
YB <- exp(yB)
Ds <- ifelse(YB<=param["barY"],1,0)
```

Let's now build a histogram of the data that we have just generated.

```
# building histogram of yB with cutoff point at ybar
# Number of steps
Nsteps.1 <- 15
#step width
step.1 <- (log(param["barY"])-min(yB[Ds==1]))/Nsteps.1
Nsteps.0 <- (-log(param["barY"])+max(yB[Ds==0]))/step.1
breaks <- cumsum(c(min(yB[Ds==1]),c(rep(step.1,Nsteps.1+Nsteps.0+1))))
hist(yB,breaks=breaks,main="")
abline(v=log(param["barY"]),col="red")
```

You can see on Figure 1.1 a histogram of y_i^B with the red line indicating the cutoff point: $\bar{y} = \ln(\bar{Y}) = 7.3$. All the observations below the red line are treated according to the sharp rule while all the one located above are not. In order to see how many observations eventually receive the treatment with this allocation rule, let's build a contingency table.

```
table.D.sharp <- as.matrix(table(Ds))
knitr::kable(table.D.sharp,caption='Treatment allocation with sharp cutoff rule',booktabs=TRUE)
```

We can see on Table 1.1 that there are 229 treated observations.

1.1.1.2 The fuzzy cutoff rule

This rule is less sharp than the sharp cutoff rule. Here, other criteria than Y_i^B enter into the decision to allocate the treatment. The doctor might measure the health status of a patient following official guidelines, but he might also measure other factors that will also influence his decision of giving the drug to the patient. The officials administering a program might measure the official income level of a household, but they might also consider other features of the household situation when deciding to enroll the household into the program or not. If these additional criteria are unobserved to the econometrician, then we have the fuzzy cutoff rule. A very simple way to model this rule is as follows:

$$D_i = \mathbb{1}[Y_i^B + V_i \leq \bar{Y}], \tag{1.2}$$

where V_i is a random variable unobserved to the econometrician and standing for the other influences that might drive the allocation of the treatment. V_i is distributed according to a, for the moment, unspecified cumulative distribution function F_V. When V_i is degenerate (*i.e.* it has only one point of support: it is a constant), the fuzzy cutoff rule becomes the sharp cutoff rule.

1.1.1.3 The eligibility + self-selection rule

It is also possible that households, once they have been made eligible to the treatment, can decide whether they want to receive it or not. A patient might be able to refuse the drug that the doctor suggests she should take. A household might refuse to participate in a cash transfer program to which it has been made eligible. Not all programs have this feature, but most of them have some room for decisions by the agents themselves of whether they want to receive the treatment or not. One simple way to model this rule is as follows:

$$D_i = \mathbb{1}[D_i^* \geq 0]E_i, \tag{1.3}$$

where D_i^* is individual i's valuation of the treatment and E_i is whether or not she is deemed eligible for the treatment. E_i might be choosen according to the sharp cutoff rule of to the fuzzy cutoff rule, or to any other eligibility rule. We will be more explicit about D_i^* in what follows.

SIMULATIONS ARE MISSING FOR THESE LAST TWO RULES

1.1.2 The potential outcomes

The second main building block of the RCM are potential outcomes. Let's say that we are interested in the effect of a treatment on an outcome Y. Each unit i can thus be in two potential states: treated or non treated. Before the allocation of the treatment is decided, both of these states are feasible for each unit.

Definition 1.1 (Potential outcomes)**.** For each unit i, we define two potential outcomes:

- Y_i^1: the outcome that unit i is going to have if it receives the treatment,
- Y_i^0: the outcome that unit i is going to have if it **does not** receive the treatment.

Example 1.2. Let's choose functional forms for our potential outcomes. For simplicity, all lower case letters will denote log outcomes. $y_i^0 = \mu_i + \delta + U_i^0$, with δ a time shock common to all the observations and $U_i^0 = \rho U_i^B + \epsilon_i$, with $|\rho| < 1$. In the absence of the treatment, part of the shocks U_i^B that the individuals experienced in the previous period persist, while some part vanish. $y_i^1 = y_i^0 + \bar{\alpha} + \theta\mu_i + \eta_i$. In order to generate the potential outcomes, one has to define the laws for the shocks and to choose parameter values. Let's assume that $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$ and $\eta_i \sim \mathcal{N}(0, \sigma_\eta^2)$. Now let's choose some parameter values:

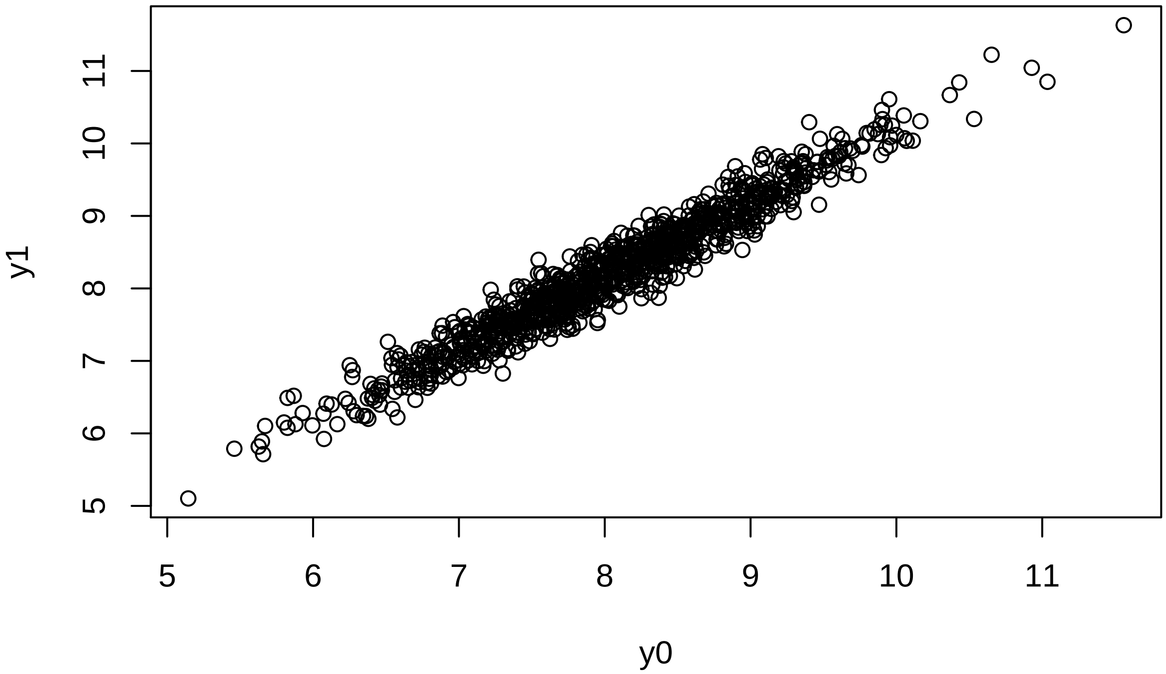

Figure 1.2: Potential outcomes

```
l <- length(param)
param <- c(param,0.9,0.01,0.05,0.05,0.05,0.1)
names(param)[(l+1):length(param)] <- c("rho","theta","sigma2epsilon","sigma2eta","delta","baralpha")
param
```

```
##         barmu      sigma2mu       sigma2U          barY           rho 
##          8.00          0.50          0.28       1500.00          0.90 
##         theta sigma2epsilon     sigma2eta         delta      baralpha 
##          0.01          0.05          0.05          0.05          0.10
```

We can finally generate the potential outcomes;

```
epsilon <- rnorm(N,0,sqrt(param["sigma2epsilon"]))
eta<- rnorm(N,0,sqrt(param["sigma2eta"]))
U0 <- param["rho"]*UB + epsilon
y0 <- mu +  U0 + param["delta"]
alpha <- param["baralpha"]+  param["theta"]*mu + eta
y1 <- y0+alpha
Y0 <- exp(y0)
Y1 <- exp(y1)
```

Now, I would like to visualize my potential outcomes:

```
plot(y0,y1)
```

You can see on the resulting Figure 1.2 that both potential outcomes are positively correlated. Those with a large potential outcome when untreated (*e.g.* in good health without the treatment) also have a positive health with the treatment. It is also true that individuals with bad health in the absence of the treatment also have bad health with the treatment.

1.1.3 The switching equation

The last building block of the RCM is the switching equation. It links the observed outcome to the potential outcomes through the allocation rule:

$$
\begin{aligned}
Y_i &= \begin{cases} Y_i^1 & \text{if } D_i = 1 \\ Y_i^0 & \text{if } D_i = 0 \end{cases} \\
&= Y_i^1 D_i + Y_i^0 (1 - D_i)
\end{aligned}
\tag{1.4}
$$

Example 1.3. In order to generate observed outcomes in our numerical example, we simply have to enforce the switching equation:

```
y <- y1*Ds+y0*(1-Ds)
Y <- Y1*Ds+Y0*(1-Ds)
```

What the switching equation (1.4) means is that, for each individual i, we get to observe only one of the two potential outcomes. When individual i belongs to the treatment group (*i.e.* $D_i = 1$), we get to observe Y_i^1. When individual i belongs to the control group (*i.e.* $D_i = 0$), we get to observe Y_i^0. Because the same individual cannot be at the same time in both groups, we can NEVER see both potential outcomes for the same individual at the same time.

For each of the individuals, one of the two potential outcomes is unobserved. We say that it is a **counterfactual**. A counterfactual quantity is a quantity that is, according to Hume's definition, contrary to the observed facts. A counterfactual cannot be observed, but it can be conceived by an effort of reason: it is the consequence of what would have happened had some action not been taken.

One very nice way of visualising the switching equation has been proposed by Jerzy Neyman in a 1923 prescient paper. Neyman proposes to imagine two urns, each one filled with N balls. One urn is the treatment urn and contains balls with the id of the unit and the value of its potential outcome Y_i^1. The other urn is the control urn, and it contains balls with the value of the potential outcome Y_i^0 for each unit i. Following the allocation rule D_i, we decide whether unit i is in the treatment or control group. When unit i is in the treatment group, we take the corresponding ball from the first urn and observe the potential outcome on it. But, at the same time, the urns are connected so that the corresponding ball with the potential outcome of unit i in the control urn disappears as soon as we draw ball i from the treatment urn.

The switching equation works a lot like Schrodinger's cat paradox. Schrodinger's cat is placed in a sealed box and receives a dose of poison when an atom emits a radiation. As long as the box is sealed, there is no way we can know whether the cat is dead or alive. When we open the box, we observe either a dead cat or a living cat, but we cannot observe the cat both alice and dead at the same time. The switching equation is like opening the box, it collapses the observed outcome into one of the two potential ones.

Example 1.4. One way to visualize the inner workings of the switching equation is to plot the potential outcomes along with the criteria driving the allocation rule. In our simple example, it simply amounts to plotting observed (y_i) and potential outcomes (y_i^1 and y_i^0) along y_i^B.

```
plot(yB[Ds==0],y0[Ds==0],pch=1,xlim=c(5,11),ylim=c(5,11),xlab="yB",ylab="Outcomes")
points(yB[Ds==1],y1[Ds==1],pch=3)
points(yB[Ds==0],y1[Ds==0],pch=3,col='red')
points(yB[Ds==1],y0[Ds==1],pch=1,col='red')
test <- 5.8
i.test <- which(abs(yB-test)==min(abs(yB-test)))
points(yB[abs(yB-test)==min(abs(yB-test))],y1[abs(yB-test)==min(abs(yB-test))],col='green',pch=3)
points(yB[abs(yB-test)==min(abs(yB-test))],y0[abs(yB-test)==min(abs(yB-test))],col='green')
abline(v=log(param["barY"]),col="red")
legend(5,11,c('y0|D=0','y1|D=1','y0|D=1','y1|D=0',paste('y0',i.test,sep=''),paste('y1',i.test,sep='')),
```

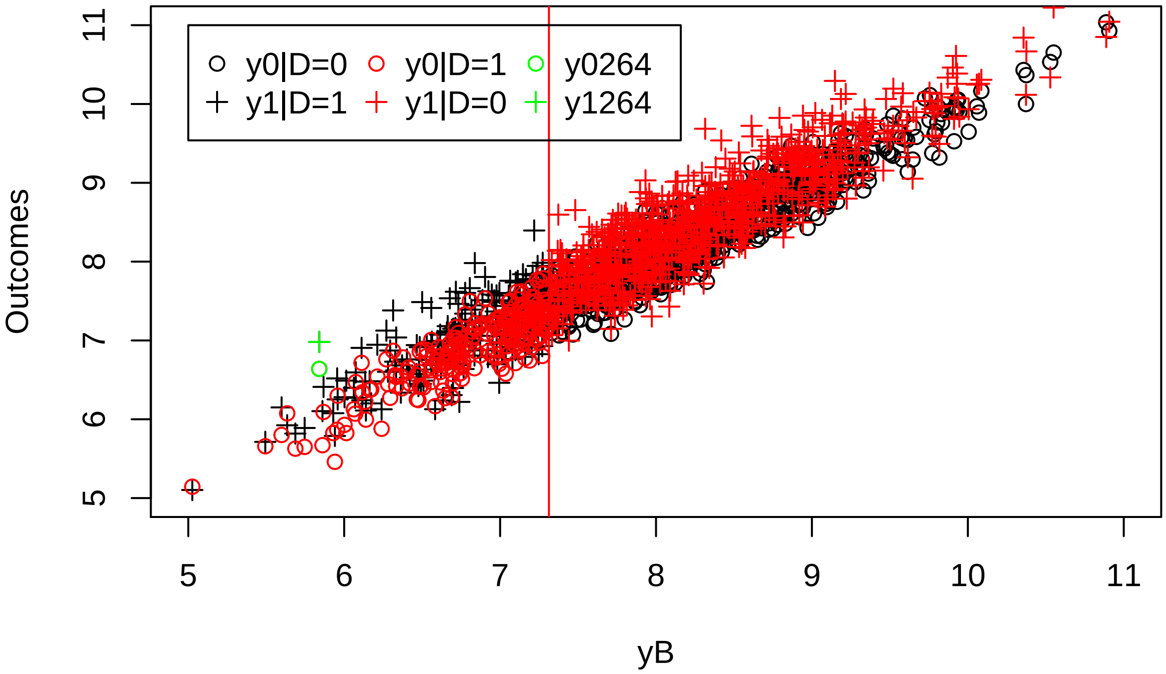

Figure 1.3: Potential outcomes

```r
plot(yB[Ds==0],y0[Ds==0],pch=1,xlim=c(5,11),ylim=c(5,11),xlab="yB",ylab="Outcomes")
points(yB[Ds==1],y1[Ds==1],pch=3)
legend(5,11,c('y|D=0','y|D=1'),pch=c(1,3))
abline(v=log(param["barY"]),col="red")
```

Figure 1.4 plots the observed outcomes y_i that results from applying the switching equation. Figure 1.4 shows that each individual in the sample is endowed with two potential outcomes, represented by a circle and a cross. Figure 1.3 plots the observed outcomes y_i along with the unobserved potential outcomes. Only one of the two potential outcomes is observed (the cross for the treated group and the circle for the untreated group) and the other is not. The observed sample in Figure 1.3 only shows observed outcomes, and is thus silent on the values of the missing potential outcomes.

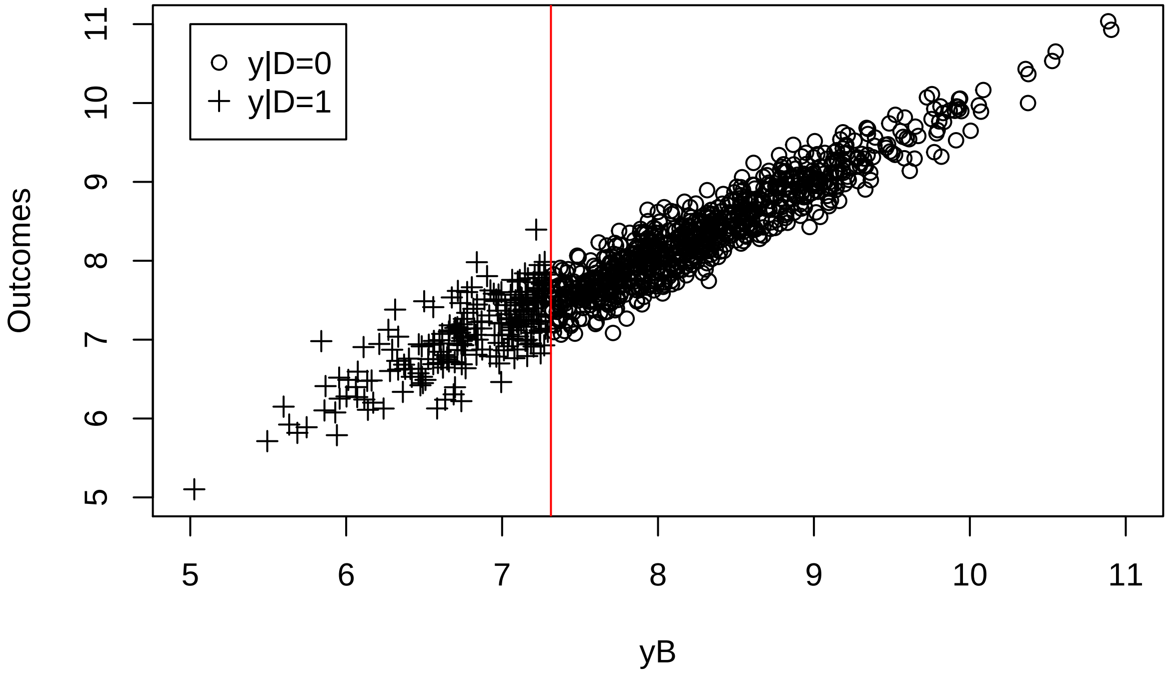

Figure 1.4: Observed outcomes