Statistical Tools for Causal Inference

The SKY Community

2019-04-09

Contents

Chapter 1

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.

Chapter 2

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.

2.1 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 $i \in \{1, \ldots, N\}$. "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).

2.1.1 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.

2.1.1.1 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. First, 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.

2.1.1.1.1 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:

$$D_i = \mathbb{1}[Y_i^B \leq \bar{Y}], \tag{2.1}$$

where $\mathbb{1}[A]$ is the indicator function, taking value 1 when A is true and 0 otherwise.

Table 2.1: Treatment allocation with sharp cutoff rule

0	771
1	229

Example 2.1 (Sharp cutoff rule)**.** Imagine that $Y_i^B = \exp(y_i^B)$, with $y_i^B = \mu_i + U_i^B$, $\mu_i \sim \mathcal{N}(\bar{\mu}, \sigma_\mu^2)$ and $U_i^B \sim \mathcal{N}(0, \sigma_U^2)$. 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"]))
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 2.1 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 2.1 that there are 229 treated observations.

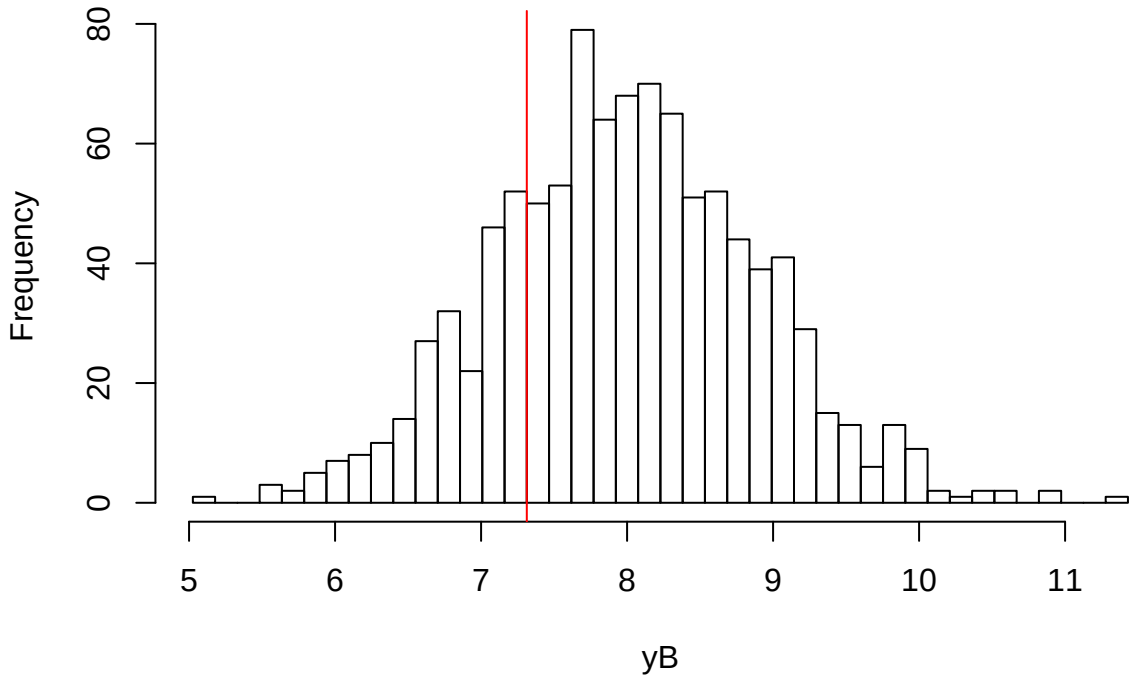

Figure 2.1: Histogram of y_B