Giacomo Petrillo, Stockholm, 2023-04-25

Opinated comments on "CAUSAL INFERENCE for Statistics, Social, and
Biomedical Sciences-An Introduction" by Imbens and Rubin, first edition (2015),
ISBN 978-0-521-88588-1

M-bias

######

In section 12.2.4, "Why Is Unconfoundedness an Important Assumption?", they never mention M-bias. Although I understand there are good reasons for almost always not worrying about M-bias in practice (see Ding and Miratrix, 2015), nevertheless it seems highly relevant that it is possible to imagine a situation where units should not be compared at equal pretreatment characteristics. I'd prefer if the book briefly mentioned this, and then explained that such context is rare.

The design phase is not safe

############################

The book often makes the point that analysis choices made in the design phase are free from predictable influences on the final causal effect estimates on the outcome. I agree that this is an important concern, and often neglected in applied work. However, I disagree that the design phase is thus safe. It is only safer than starting right away with outcomes, but not safe in absolute terms. The way the book says it strongly implies the latter interpretation. This can give a false sense of safety to readers. It reminds me of what Andrew Gelman sometimes says, that "you think you did ok because you followed the little rules in the book, but the rules were not actually sufficient, and you should not be given that mistaken impression."

Relevant quotes:

CHAPTER 12: Unconfounded Treatment Assignment, Page 270, section 12.4.2:

> One important difference between the model-based imputations and the other three (weighting, blocking, and matching methods) is that the first requires building models for the potential outcomes, whereas for the other three all decisions regarding the implementation of the estimators without covariate adjustment can be made before seeing any outcome data. This difference is important because not having outcome data prevents the researcher from adapting the model to make it fit prior notions about the treatment effects of interest. Although the researcher does have to make a number of important decisions when using weighting, blocking, and matching methods, these can be implemented in a way that does not introduce bias in the estimates for treatment effects and so have arguably more credibility.

I think this argument in favor of non-model based, so in particular non-fully bayesian, methods is valid for what people currently do in practice. However flexible Bayesian methods can be constructed that have the same property, it is "just" more difficult and less studied. See for example the success of BART. This is relevant because it makes the separate-design-phase methods less evidently a different class than model-based methods, and a fully model-based method would analyse together treatment with outcome, showing clearly the dependence.

Page 276, section 12.5:

> As a result, this analysis cannot be “contaminated” by knowledge of estimated outcome distributions, or by preferences, conscious or unconcious, for particular results.

I disagree. From a theoretical point of view, writing PS subclassification as a single model—based inference, the models for treatment and outcome must be joint. From the practical point of view, one often has strong expectations re ties between treatment and outcome distributions, e.g., for many interventions, even without specific knowledge, I'd expect a negative correlation and a positive effect. These researcher expectations are sufficient to influence the final result while making choices in the design phase.

CHAPTER 13: Estimating the Propensity Score, Page 283:

> [...] it is important to keep in mind that during this entire process, and in fact in this entire chapter, we do not use the outcome data, and there is, therefore, no way of deliberately biasing the final estimation results for the treatment effects. Consequently, there is no concern regarding the statistical properties of the ultimate estimates of the average treatment effects obtained from iterating back and forth between (i) the specification of the propensity score, and (ii) balance assesments of the estimated propensity score, until an adequate specification is found.

CHAPTER 14: Assessing Overlap in Covariate Distributions, Page 309:

> These assessments do not involve the outcome data and therefore do not introduce any systematic biases in subsequent analyses.

CHAPTER 15: Matching to Improve Balance in Covariate Distributions, Page 358, section 15.7:

> An important aspect of the analysis in this chapter is that it is entirely based on the covariate and treatment data, and never uses the outcome data. As such, it cannot intentionally introduce biases in the subsequent analyses.

CHAPTER 16: Trimming to Improve Balance in Covariate Distributions, Page 374:

> An important aspect of the analysis in this chapter, shared with the matching approach in the previous chapter, is that it is entirely based on the covariate and treatment data, and never uses the outcome data. As such it cannot intentionally introduce systematic biases in the subsequent analyses for causal effects on outcomes.

CHAPTER 26: Conclusions and Extensions, Page 589:

> We discuss the importance of the design stage of a study for causal effects where the outcome data are not yet used. At this stage a researcher can carry out preliminary analyses that make the final analyses that do involve the outcome data more credible and robust.

Here it's vague enough in saying that's more credible without implying that it's completely safe. I'd like all the other quotes sounded less stark, and only implied "additional protection" instead of "absolute protection" from influences on the final result.

Assessing unconfoundedness

##########################

I think I understand the need to use a different word, "assess" instead of "test", when dealing with uncounfoundedness, yet the way the book talks about the matter looks contradictory and as such is confusing.

CHAPTER 12: Unconfounded Treatment Assignment, Page 278:

> In Chapter 21, in Part V of the text, we discuss methods for assessing the unconfoundedness assumption. We purposely use the term “assess” here rather than “test,” because unconfoundedness has no directly testable implications. Nevertheless, there are a number of stastistical analyses that we can conduct that can shed light on its plausibility.

I don't like this wordplay with "test" and "assess." I think I understand what's meant, and I agree with it, but it should be made more explicit: it is not testable within the sample. If it wasn't testable in a general sense, we could conclude we can't know anything! Our brain does not have access to some "magic inference box" that spits out extra information you can't reach with the senses. Even if the brain had magical priors of truth, the brain comes from somewhere.

CHAPTER 21: Assessing Unconfoundedness, Page 479:

```
    This critical assumption [uncounfoundedness] is not testable. The issue is
                                                    ------------
    that the data are not directly informative about the distribution of the
    control outcome Yi(0) for those who received the active treatment (for
    those with Wi = 1, we never observe Yi(0)), nor are they directly
    informative about the distribution of the active treatment outcome given
    receipt of the control treatment (for those with Wi = 0, we never observe
    Yi(1)). Thus, the data cannot directly provide evidence on the validity of
    the unconfoundedness assumption. Nevertheless, here we consider ways to
    assess the plausibility of this assumption from the data at hand.
    ------
```

The usual bait and switch: it is not "testable"", yet we can "assess it from the data at hand." (Yes, I understand the technical meaning is different: but they never really explain it technically.)

Single model for treatment and outcome vs. ad hoc balancing criteria

##

Since the book seems to argue a preference for Bayesian inference, it then looks weird when it's forsaken without much distress for blocking and matching, in general balancing, methods. I think that a clearer exposition of the problem would recognize that it is in principle possible to write a single, coherent, model-based analysis that does as well as ad hoc methods, but that currently nobody knows how to set that up reliably.

Some specific comments:

CHAPTER 12: Unconfounded Treatment Assignment, Page 271:

```
        Y(0), Y(1) | X, θ.
```

> [...]
>
> In this approach, often there is no need to specify a parametric model for the conditional distribution of the treatment indicator given the covariates, the super-population assignment mechanism,
>
> $$p(W|X; \varphi),$$
>
> because, if φ and θ are distinct parameters, inference for causal effects is not affected by the functional form of the specification of this assignment mechanism. However, it is important for this argument that φ and θ are distinct parameters.

Using propensity score subclassification implies, from a fully model-based perspective, that the parameters θ and φ are connected. When I realized this connection, everything became quite clearer to me.

CHAPTER 13: Estimating the Propensity Score, Page 282, bottom:

> Such a criterion would always suggest that using the true propensity score is preferable to using an estimated propensity score. In contrast, for our purposes, it is often preferable to use the estimated propensity score. The reason is that using the estimated score may lead to superior covariate balance in the sample compared to that achieved when using the true super-population propensity score.

Page 306:

> A second key point is that the goal in this chapter is to obtain an estimated propensity score that balances the covariates within subclasses, rather than one that simply estimates the hypothetical true propensity score as accurately as possible. As has been noted in the literature, using the estimated propensity score often leads to better balance than using the true propensity score.

Page 307:

> The point that using the estimated propensity score rather than the true propensity score leads to better balance and better estimators for causal effects has been made in Rubin and Thomas (1992a, 1992b, 1996, 2000) and Hirano, Imbens, and Ridder (2003).

This doesn't ring true to me. If a covariate is independent, then I should not care about balancing it. If I'm caring about it, then I must be implicitly assuming there's some dependence. Tentative steelman: I assume I'm bad at stating the joint distribution (i.e., the models), and that I expect that balancing the covariates does better, even though this is a contradiction from the point of view of an ideal bayesian agent.

The point is that, even in an RCT, what I actually care about is comparing twins, not that on average I'll get similar units in the long run. This comes out naturally if I try to make a single coherent Bayesian analysis, but I have to bungle-argue it out of air if I don't.

Page 284:

> A final point to emphasize is that the primary goal is to find an adequate

> specification of the propensity score, in the sense of a specification that achieves statistical balance in the covariates. We are not directly interested in a structural, behavioral, or causal interpretation of the propensity score, ***although inspecting and assessing the strength and nature of the dependence of the propensity score on the covariates may be helpful when assessing the plausibility of the unconfoundedness assumption.*** Finding an adequate specification is, therefore, in essence, a statistical problem that relies less on subject-matter knowledge than other aspects of the modeling of causal effects. The goal is simply to find a specification for the propensity score that leads to adequate balance between covariate distributions in treatment and control groups in our sample.

The unconfoundedness assumption is not testable, and the propensity score model has no meaning, yet...

CHAPTER 14: Assessing Overlap in Covariate Distributions, Page 319:

> There was a small set of seventeen individuals who had been exposed to chemotherapy and who had experienced multiple births. These seventeen individuals were all in the control group, so we estimated the propensity score to be equal to zero for these individuals. In the calculation of the average linearized propensity score (lps) by treatment group, in the last row of Table 14.1, these seventeen individuals were excluded from further analyses.

Darn frequentists.

Matching on the linearized propensity score
###

CHAPTER 15: Matching to Improve Balance in Covariate Distributions, Page 343:

> Put differently, the potential outcomes are more likely to be approximately linear in the lps than in the propensity score. For example, if the potential outcomes are linear in the covariates, the covariates are jointly normal, and the propensity score follows a logistic form, then the potential outcomes are linear in the lps.

Are Normality of the covariates, and the logistic model, necessary for the justification?

At a conference I've argued with a guy who said that asymptotic analyses show that it's better to match on the propensity score on the original [0, 1] scale. I replied that nevertheless on finite samples I expect the lps to be better and so I'd go for that, but this book provided scarce fodder for my argument, since it is so specific. Maybe the following reasoning is a more general explanation, although not yet satisfying to me.

If a discrete stratification of the propensity score was a balancing score, it would be sufficient to match (or stratify) on that. However that can not be the case because the propensity score is the minimal balancing score.

Suspend disbelief for a moment, and imagine that the propensity score subclasses were balancing instead. This would imply that the ratio of the propensity score probability densities by treatment group is constant within the classes:

```
    E = propensity score
    C = propensity score discrete substratum
    W = treatment

    C = c if l(c) < E < r(c)

    W ⫫ X | E (by definition of the model)

    W ⫫ X | C (absurd assumption)

    P(W=w|C=c) =
    = P(W=w|C=c,X=x) =   for l(c) < e(x) < r(c)
    = P(W=w|C=c,X=x) =
    = P(W=w|E=e(x),X=x) =
    = P(W=w|E=e(x)) =
    = P(W=w|E=e(x),C=c)

    ==> W ⫫ E | C

    ==> p(E=e|W=0,C=c) / p(E=e|W=1,C=c) = 1
```

But actually, without the absurd assumption, we have

```
    p(E=e|W=0,C=c) / p(E=e|W=1,C=c) =

        (provided l(c) < e < r(c))

    = p(E=e|W=0) / p(E=e|W=1)   P(C=c|W=1) / P(C=c|W=0) =

        (apply Bayes to numerator and denominator)

    = P(W=0|E=e) P(C=c,W=1)  /  P(W=1|E=e) P(C=c,W=0) =

        (use W ⫫ X | E and E = P(W=1|X))

    = (1 - e) P(C=c,W=1)  /  e P(C=c,W=0) =

    = (1 - e)/e  p(C=c,W=1)/p(C=c,W=0)
```

So, to have balancing C, we would like (1 - e)/e (the odds) to be constant. The amount of variation of the odds within a propesity score class thus measures how much C is nonbalancing. Precisely, let us use the ratio between the maximum and minimum value of the odds within the class. Then the nonbalancing measure is evenly distributed among the classes if they are evenly spaced w.r.t. log (1 - e)/e. Analogously, the distance of units for matching is measured by the difference of linearized propensity scores.