Notes on Kasy: Why Experimenters Might Not Always Want to Randomize, and What They Could Do Instead

notes David Reinstein 20 Jul 2018

cf, Barrios 'This paper shows that stratifying on the conditional expectation of the outcome given baseline variables is optimal in matched-pair randomized experiments.'

Intro

Page 1

> If, for instance, the experimenter is interested in estimating the average treatment effect and evaluates an estimate in terms of the squared error, then she should minimize the expected mean squared error (MSE) through choice of a treatment assignment. We provide explicit expressions for the expected MSE that lead to easily implementable procedures for experimental design

> following situation (cf. Morgan and Rubin 2012). They have selected a random sample from some population and have conducted a baseline survey for the individuals in this sample. Then a discrete treatment is assigned to the individuals in this sample, usually based on some randomization scheme. Finally, outcomes are realized, and the data are used to perform inference on some average treatment effect. A key question for experimenters is how to use covariates from the baseline survey in the assignment of treatments. Intuition and the literature suggest to use stratified randomization conditional on covariates, also known as blocking

(p2)

> Conditional independence only requires a controlled trial (CT), not a randomized controlled trial (RCT)

- *DR: We may be able to do better than what he is proposing as we have outcomes for previous comparable observations.*

-The experimenter has to choose a treatment assignment and an estimator, given knowledge of the covariate distribution in the sample. Her objective is to minimise risk based on a loss function such as the mean squared error of a point estimator. The decision criteria considered are Bayesian average risk and (conditional) minimax risk-

trade off balance across various dimensions of covariates … picking a prior for condnl expectation of potential outcomes allows one to calculate an obj fncn (Bayesian risk)… to do this in a principled way.

Sec 2. Motivating Example

For experimental units. Continuous covariate for each, assign each to binary treatment. Want to estimate conditional average treatment affect across these units.

DR: I would think you care about a population affect not about an affect for the actual units sampled

- Plans to estimate simple difference in means across treatment
- 16 possible treatment assignments (considered in table 1)

Potential outcomes are determined as $Y_i^d = x_i + d + \epsilon_i^d$, latter independent|x, zero mean, variance 1.

(This assumes the true ATE=1)

Simple expressions for the variance of the estimator: $var(\hat{(\beta)}) = 1/n_1 + 1/n_0$ … simply a function of the number receiving each treatment

The (ex-post) bias = $E[\hat{(\beta)}] - \beta$.. is (?) the simple difference in the average covariates between treatment and control units..

As always, MSE=Bias^2 + Var

Next suppose we lost/ignored covariates. Here -there is no bias, even ex-post- (?), and the variance is simply the $1/n_1 + 1/n_0$ term times var(X_i)+ 1. (The variance in the X-s gets folded in, and this is only scaled down by larger numbers in each treatment.) This then argues for (randomly) assigning half the units to the treatment.

DR: This discussion of ex-post and ex-ante bias could use some clarification

2.2 Randomised designs, continuing the example

Can be characterised by the probability it assigns to each of the rows of table 1, and the MSE will be the weighted avg of the MSE for each row (weighted by the assigned probability)

Consider pure randomisation, a coin flip each time … but throwing out -all the same- assignments -> MSE=3.2

Next, Consider only balanced (2 treatments) assignment -> MSE=2.7

Pairwise randomization: groups into one pairs with X=0 or X=1 and another with X=2 or X=3, and then randomise within. ->MSE=1.5

Removing the assignments 0,1,0,1 and 1,0,1,0, retaining only 0,1,1,0 and 1,0,0,1 reduces bias further, leading to MSE=1.0

- *DR: should give some intuition for this… depends on functional form*

2.3: other DGP

We noted the MSE of *any* randomised procedure is the weighted avg of MSEs it chooses among. However, we *assumed* we knew the model. -Model 2- flips the sign of the effect of x and makes it nonlinear. Result: The bias in each assignment changes, but the ranking of MSE-s across designs is the same!

DR: Why? Any intuition? Strange to give an example where this occurs if it is not a general property.

But in general we cannot assert this, and need to consider performance across alternative data generating processes. -> use a nonparametric Bayesian prior.

Decision theory

evidence X, decision $a = delta(x)$. State of world θ.

Loss $L = L(a, \theta)$,

E.g., MSE of estimate $\hat{(\beta)}$

Frequentist risk function averages losses over randomness f a treatment allocation scheme, for a given θ: $R(\delta, \theta) = E[L(\delta(X), \theta|\theta]$.

If loss is squared error of estimate, then this is the same MSE. … previously calculated for two possible values of θ.

How to trade off across states θ?

- can focus on state of the world with the worst value of the risk function -> minimax criterion.
- Can assign weights π (prior probabilities) to each sotw and minimise the weighted avg -> Bayesian

He gives his main intuition here. As a randomised procedure has an expected risk that is simply the average over deterministic procedures, it cannot be better. It can be as good as is the optimal deterministic procedure if it randomised Between the set of assignments that are all equally the best. This can only happen if there is a space where the relevant observable variables are the same across observations, but not Where we have these continuous variables.

Section 4: our proposed procedure

-the key object that we need prior to the conditional expectation of potential outcomes given covariates…-

DR: this could be estimated using prior data on similar units . In our case, at least we can make this prediction for the -nontreated- observations, As we know how do use covariates, As we know how do use covariates relate to the amount a page raises overall, and how it relates to the amount a page raises overall, and how it relates to the amount of the next five contributions.

He simply assumes there is some function of the covariates and treatment that yields this expectation, and this is an expectation about variances and covariances (between the treatment and control values). He assumes a prior over the conditional variance of the outcome, and begins by assuming it is homoscedastic (an important practical issue - the treatment may affect variance in the rw).

Another set of prior moments builds on linear models… Here the MSE 'corresponds to' balance measured by difference in covariate means.

Prop 2… MSE and expected loss for posterior blp estimator

DR: Are we talking about something like a controlled regression here? … in fn 7, how can a range of standard estimators all be 'best'?

Prop 3… MSE and expected loss for difference in

In the case of a linear separable true model, the bias is proportional to the square the difference in covariate means … so minimise this

DR: But this is far from clear if we have multiple covariates!

With MSE and expected loss, we can compute and implement the assignment procedure to minimise this, it would seem

4.3 Discrete optimisation

How to choose these things in practice? He suggests considering the machine learning literature and -squared exponential priors-.

But with many observations there are too many assignments to find this optimum easily. He recommends a 'random search'… rerandomising and computing until it performs reasonably well.

DR: I don't understand the iterative procedure he proposes

5. Arguments for randomisation

Arguments against these:

- not necessary for the equivalent of 'randomization inference'

DR: why do I want to do randomization inference anyways?

- identification based on conditional independence *is* still valid
- Bayesian choice of prior issue

DR: What about the choice of functional form in many x's?… seems difficult

Great point:

> This does not imply, however, that the interpretation of the data obtained from the experiment has to rely on the same prior as the one used for designing the experiment, or on any prior at all. From the perspective of a researcher analyzing data produced by an experiment where treatment was assigned based on observables as well as a randomization device, it is immaterial how treat- ment was assigned, as long as the conditional independence condition… holds.

Conclusion

Second, we can consider nonparametric priors that lead to very tractable estimators and expressions for Bayesian risk (MSE). The general form of the expected MSE for such priors is …, where $\bar{C}$ and C are the appropriate covariance vector and matrix from the prior distribution, cf. Section 4.

DR: I'm still a bit intuitively confused as to where we get or how we can make assumptions over these elements… and if we have some prior data on outcomes, how that can be used and prioritized

---

DR: Some further comments:

- A more descriptive version, expanding on things, would be helpful
- It's unclear why we focus on the *sample* ATE; we care about the super-population
- It is hard to see how this works; give intuition for how we trade off among balancing the x's if there are several, and we do not have outcome data. What exactly does the procedure do with merely variation in the observed X's
    - What if we observe outcomes from a *similar* untreated population ?
- This might actually be useful for *lab* and survey-based experiments (Prolific, Qualtrics, Mturk even) where we have, or could collect, a rich set of predetermined observables