Why experimenters might not always want to randomize, and what they could do instead

Maximilian Kasy
Department of Economics, Harvard University

Setup

1. *Sampling:*
   random sample of n units
   baseline survey $\Rightarrow$ vector of covariates X_i
2. *Treatment assignment:*
   binary treatment assigned by $D_i = d_i(X, U)$
   X matrix of covariates; U randomization device
3. *Realization of outcomes:*
   $Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$
4. *Estimation:*
   estimator $\widehat{\beta}$ of the (conditional) average treatment effect, $\beta = \frac{1}{n} \sum_i E[Y_i^1 - Y_i^0 | X_i, \theta]$

Questions

- How should we assign treatment?
- In particular, if X has continuous or many discrete components?
- How should we estimate β?
- What is the role of prior information?

Some intuition

- "Compare apples with apples"
  $\Rightarrow$ balance covariate distribution.
- Not just balance of means!
- We don't add random noise to estimators
  – why add random noise to experimental designs?
- Identification requires controlled trials (CTs),
  but not randomized controlled trials (RCTs).

Framework

1. *Decision theoretic:*
   $\mathbf{d}$ and $\widehat{\beta}$ minimize risk $R(\mathbf{d}, \widehat{\beta} | X)$
   (e.g., expected squared error)
2. *Nonparametric:*
   no functional form assumptions
3. *Bayesian or minimax*

Main results

1. The unique optimal treatment assignment does not involve randomization.
2. Identification using conditional independence is still guaranteed without randomization.
3. Tractable nonparametric priors
   $\Rightarrow$ Explicit expressions for risk as a function of treatment assignment
   $\Rightarrow$ choose $\mathbf{d}$ to minimize these

A formal decision problem

- Risk function of treatment assignment $\mathbf{d}(X, U)$, estimator $\widehat{\beta}$, for loss L, data generating process θ:
  $$R(\mathbf{d}, \widehat{\beta} | X, U, \theta) := E[L(\widehat{\beta}, \beta) | X, U, \theta]$$
- (Conditional) Bayesian risk:
  $$R^B(\mathbf{d}, \widehat{\beta} | X, U) := \int R(\mathbf{d}, \widehat{\beta} | X, U, \theta) dP(\theta)$$
- Conditional minimax risk:
  $$R^{mm}(\mathbf{d}, \widehat{\beta} | X, U) := \max_{\theta} R(\mathbf{d}, \widehat{\beta} | X, U, \theta)$$
- Objective: $\min R^B$ or $\min R^{mm}$

Theorem (Optimality of deterministic decisions)

Consider a general decision problem.
Let R^* equal R^B or R^{mm}. Then:

1. *The optimal risk $R^*(\delta^*)$, when considering only deterministic procedures $\delta(X)$, is no larger than the optimal risk when allowing for randomized procedures $\delta(X, U)$.*
2. *If the optimal deterministic procedure δ^* is unique, then it has strictly lower risk than any non-trivial randomized procedure.*

Numerical example

- Covariates: $(x_1, \ldots, x_4) = (0, 1, 2, 3)$
- Model 1: $Y_i^d = x_i + d + \epsilon_i^d$
- Model 2: $Y_i^d = -x_i^2 + d + \epsilon_i^d$
- Prior over models (covariance kernel for f):
  $C((x_1, d_1), (x_2, d_2)) =$
  $10 * \exp\left(-\left(\|x_1 - x_2\|^2 - (d_1 - d_2)^2\right)/10\right)$

Table:

- One row for each non-random treatment assignment $(d_1, \ldots, d_4)$
- Randomized designs: Assign equal probability to all assignments $(d_1, \ldots, d_4)$ marked by "1"
- Bias, Variance, and MSE for difference-in-means estimator, under either model, and integrating over prior
- MSE for random designs:
  average of non-random MSEs

assignment				designs					model 1			model 2			EMSE
d_1	d_2	d_3	d_4	1	2	3	4	5	bias	var	MSE	bias	var	MSE	
0	1	1	0	1	1	1	1	1	0.0	1.0	1.0	2.0	1.0	5.0	1.77
1	0	0	1	1	1	1	1	1	0.0	1.0	1.0	-2.0	1.0	5.0	1.77
1	0	1	0	1	1	1	1	0	-1.0	1.0	2.0	3.0	1.0	10.0	2.05
0	1	0	1	1	1	1	1	0	1.0	1.0	2.0	-3.0	1.0	10.0	2.05
1	1	0	0	1	1	1	0	0	-2.0	1.0	5.0	6.0	1.0	37.0	6.51
0	0	1	1	1	1	1	0	0	2.0	1.0	5.0	-6.0	1.0	37.0	6.51
0	1	0	0	1	1	0	0	0	-0.7	1.3	1.8	3.3	1.3	12.4	2.49
0	0	1	0	1	1	0	0	0	0.7	1.3	1.8	-0.7	1.3	1.8	2.49
1	1	0	1	1	1	0	0	0	-0.7	1.3	1.8	0.7	1.3	1.8	2.49
1	0	1	1	1	1	0	0	0	0.7	1.3	1.8	-3.3	1.3	12.4	2.49
1	0	0	0	1	1	0	0	0	-2.0	1.3	5.3	4.7	1.3	23.1	6.77
1	1	1	0	1	1	0	0	0	-2.0	1.3	5.3	7.3	1.3	55.1	6.77
0	0	0	1	1	1	0	0	0	2.0	1.3	5.3	-7.3	1.3	55.1	6.77
0	1	1	1	1	1	0	0	0	2.0	1.3	5.3	-4.7	1.3	23.1	6.77
0	0	0	0	1	0	0	0	0	-	-	∞	-	-	∞	-
1	1	1	1	1	0	0	0	0	-	-	∞	-	-	∞	-
MSE model 1:				∞	3.2	2.7	1.5	1.0							
MSE model 2:				∞	20.6	17.3	7.5	5.0							

Nonparametric Bayesian setup

- Prior moments:
  $f(x, d) := E[Y_i | X_i = x, D_i = d]$
  $\mathrm{Cov}(f(x_1, d_1), f(x_2, d_2)) = C((x_1, d_1), (x_2, d_2))$
  $\mu_i = E[Y_i | X, D], \quad \mu_\beta = E[\beta | X, D]$
  $\Sigma = \mathrm{Var}(Y | X, D, \theta)$
  $C_{i,j} = C((X_i, D_i), (X_j, D_j))$ and $\overline{C}_i = \mathrm{Cov}(Y_i, \beta | X, D)$.
- Mean squared error objective:
  Loss $L(\widehat{\beta}, \beta) = (\widehat{\beta} - \beta)^2$,
  Bayes risk $R^B(\mathbf{d}, \widehat{\beta} | X) = E[(\widehat{\beta} - \beta)^2 | X]$.
- Linear estimators:
  $\widehat{\beta} = w_0 + \sum_i w_i Y_i$.

Theorem (Best linear predictor, posterior variance)

The optimal estimator is
$$\widehat{\beta} = \mu_\beta + \overline{C}' \cdot (C + \Sigma)^{-1} \cdot (Y - \mu),$$
and the corresponding expected loss (risk) is
$$R^B(\mathbf{d}, \widehat{\beta} | X) = \mathrm{Var}(\beta | X) - \overline{C}' \cdot (C + \Sigma)^{-1} \cdot \overline{C}.$$

Discrete optimization

- The optimal design solves
  $$\max_{\mathbf{d}} \overline{C}' \cdot (C + \Sigma)^{-1} \cdot \overline{C}.$$
- Possible minimization algorithms: Search over random $\mathbf{d}$; greedy algorithm; simulated annealing

Contact Information

- Web: http://www.scholar.harvard.edu/kasy
- Email: maximiliankasy@fas.harvard.edu
- Phone: +1 (310) 666 8071

ASCINA
Austrian Scientists & Scholars in North America