Modern Sampling Methods

Class 4: Treatment and Policy Choice

January 10, 2022

Outline

- Setup
- Example
- CES Rules and Minmax Regret
- Local Asymptotic Optimality
- Empirical Welfare Maximization

Basic Setup

Based on Manski (2004) and Dehejia (2005).

$\mathcal{T} = \{0, 1\}$: set of possible treatments.

$Y(0)$, $Y(1)$: potential outcomes

$X \in \mathcal{X}$: background characteristics.

Let θ be parameters associated with potential outcomes:

$$X \sim F_X(\cdot)$$
$$Y(0)|X = x \sim F_0(\cdot|x, \theta)$$
$$Y(1)|X = x \sim F_1(\cdot|x, \theta)$$

Treatment Assignment Rules and Social Welfare

A treatment assignment rule selects treatment based on X:

$$\delta : \mathcal{X} \to \{0, 1\}.$$

(Could also allow for randomization.)

Suppose we want to maximize average outcomes

$$W(\theta, \delta, x) = \delta(x)E_\theta[Y(1)|X = x] + (1 - \delta(x))E_\theta[Y(0)|X = x];$$

$$W(\theta, \delta) = \int W(\theta, \delta, x)dF_X(x).$$

The ideal policy is

$$\delta^*(x) = \mathbf{1}\left\{E_\theta[Y(1)|X = x] \geq E_\theta[Y(0)|X = x]\right\}.$$

Statistical Treatment Rule

This is not feasible in general b/c we do not know θ.

Suppose we have some data that is informative about θ. How to use the past data to inform the future choice of treatment rule?

<u>Statistical Treatment Rule</u>:

Before making our treatment assignment, we observe some data $Z \sim P_\theta$

(Assume Z independent of the future individual to be treated.)

We then choose δ based on the data z.

Note the timing:

1. observe Z (e.g. from a randomized controlled trial);
2. take a <u>new</u> individual, and observe their X;
3. assign this individual to treatment based on her own X as well as the data of others collected in Z.

Notation:

$$\delta(x, z)$$

indicating that the policy depends on data and on any information we have about the new individual.

Ex ante probability of assigning individuals with $X = x$ to treatment:

$$\beta(\delta, x, \theta) = E_\theta[\delta(x, Z)] = \int \delta(x, z) dP_\theta(z).$$

Ex ante expected social welfare for a given rule δ:

$$E_\theta[W(\theta, \delta(\cdot, Z))] =$$

$$\int \int \Big\{ \delta(x, z) \cdot E_\theta[Y(1)|X = x)]$$

$$+(1 - \delta(x, z)) \cdot E_\theta[Y(0)|X = x] \Big\} dF_X(x) dP_\theta(z).$$

Example: Dehejia (2005)

GAIN experiment, a randomized evaluation of a job training program in California. (Data from Alameda County.)

Tobit model for earnings of individual i in quarter t:

$$Y_{it}^* = x_{it1}'\beta_1 + T_i \cdot x_{it1}'\beta_2 + x_{it2}'\beta_3 + \epsilon_{it}, \quad \epsilon_{it} \overset{iid}{\sim} N(0, \sigma^2),$$

$$Y_{it} = 1(Y_{it}^* > 0) Y_{it}^*.$$

Parameter vector: $\theta = (\beta, \sigma^2)$. The data are:

$$Z = \{(x_{it1}, x_{it2}, T_i, , Y_{it}) : i = 1, \ldots, n, t = 1, \ldots, T\}.$$

Use Bayesian methods to simulate posterior distribution $p(\theta|Z)$.

Hypothetical decision problem: counselor is dealing with a new individual (person $n+1$), whose covariates $x_{n+1,t}$ are observed and whose earnings will follow the same Tobit model.

Can simulate outcomes for person $n+1$:

- Draw θ from posterior $p(\theta|Z)$.
- Simulate $Y_{n+1}(0)$ given $x_{n+1,t}$ and setting $T_{n+1} = 0$.
- Simulate $Y_{n+1}(1)$ given $x_{n+1,t}$ and setting $T_{n+1} = 1$.

Then choose treatment that has higher expected outcome.

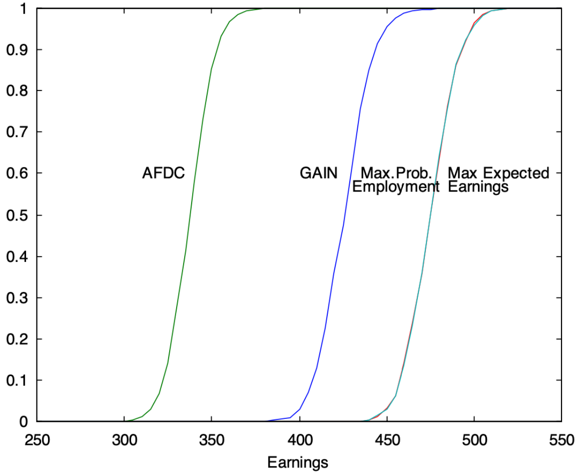

Fig. 7. Predictive distributions for average earnings.

From: Dehejia (2005)

Some other economic applications

- Job Training Programs: JTPA (Kitagawa and Tetenov 2018)
- Environmental Policy (Assuncao et al, 2019)
- Energy Incentives (Knittel and Stolper, 2019)
- Marketing (Rossi et al 1996, Dube et al 2017)

Manski (2004)

Suppose the covariate X is discrete, taking on possible values $\{x_j : j = 1, \ldots, k\}$.

Suppose data Z are obtained from a randomized experiment:

N_j units with $X = x_j$, of which
N_j^1 treated and N_j^0 controls.

Conditional Empirical Success (CES) Rule:

$$\hat{\beta}_j := \frac{1}{N_j^1} \sum_{i=1}^{N_j} T_{ji} Y_{ji} - \frac{1}{N_j^0} \sum_{i=1}^{N_j} (1 - T_{ji}) Y_{ji}.$$

Then define

$$\hat{\delta}(x_j, Z) = 1(\hat{\beta}_j > 0).$$

Manksi's CES rule is nonparametric and intuitive, but not immediately clear whether it is in some sense optimal.

Related question: if covariate X takes on many values (or is continuous), will CES work well or are there alternatives?

To analyze further, we need some measure of a statistical treatment rule's performance.

Consider expected welfare <u>regret</u>:

$$R(\theta, \delta) = E_\theta \left[W(\theta, \delta^*) - W(\theta, \delta) \right],$$

where δ^* is the ideal rule.

Maximum expected welfare regret can be used as a measure of performance for a rule δ,

$$\max_{\theta} \quad R(\theta, \delta)$$

For the class of CES rules, Manski provides bounds on maximum expected welfare regret when the randomized experiment is conducted by:

- Simple randomization
- Stratified block randomization

(see Class 6 for definitions/discussion of these random assignment mechanisms)

Stoye (2009)

Provides the minmax expected welfare regret optimal rule for some important cases.

First case

- No covariates
- Y_0 and Y_1 are binary variables
- Let $\bar{Y}_1$ and $\bar{Y}_0$ be the averages in the two groups.

Consider two different random assignment mechanisms:
(i) Suppose random assignment is by matched pairs, where n is even and we observe exactly $n/2$ treated and $n/2$ controls in our data Z.

Let

$$\hat{\delta}_{MP}(Z) = \begin{cases} 0 & \text{if } \bar{Y}_1 < \bar{Y}_0 \\ \frac{1}{2} & \text{if } \bar{Y}_1 = \bar{Y}_0 \\ 1 & \text{if } \bar{Y}_1 > \bar{Y}_0 \end{cases}$$

(Essentially the CES rule.)

(ii) Simple random assignment, where n^0 and n^1 denote the number of control and treated observations in the sample.

Let

$$\hat{\delta}_{S}(Z) = \begin{cases} 0 & \text{if } n_1\left(\bar{Y}_1 - \frac{1}{2}\right) < n_0\left(\bar{Y}_0 - \frac{1}{2}\right) \\ \frac{1}{2} & \text{if } n_1\left(\bar{Y}_1 - \frac{1}{2}\right) = n_0\left(\bar{Y}_0 - \frac{1}{2}\right) \\ 1 & \text{if } n_1\left(\bar{Y}_1 - \frac{1}{2}\right) > n_0\left(\bar{Y}_0 - \frac{1}{2}\right) \end{cases}$$

Result: Stoye shows that $\hat{\delta}_{MP}$ and $\hat{\delta}_S$ are minmax rules with respect to expected welfare regret.

Further, this minmax regret result:

- Extends to allow for bounded outcomes.
- Extends to hold with covariates: then the minmax regret rule conditions *fully* on X, even if this means very few observations per cell, or even some empty cells.

Results as sharp as Stoye's are difficult to obtain in more complex situations with

- More complex outcome distributions
- Structured/parametrized outcome distributions
- Nonexperimental (observational) data, or data from adaptive experiments
- Restrictions on the class of rules (e.g. constraints on complexity of rule)
- etc.

Then it can be useful to turn to large-sample approximations to study alternative rules.

Local Asymptotics for Treatment Assignment

Consider a setting without covariates, but where data are not necessarily from a RCT: there is just some general statistical model

$$Z^n \sim P_\theta, \quad \theta \in \Theta$$

(where n indicates sample size).

The model parameter θ is informative about average treatment effect through:

$$\text{ATE} = W(\theta, 1) - W(\theta, 0) = g(\theta)$$

for some known function g.

As $n \to \infty$, we will often be able to estimate θ consistently:

$$\hat{\theta} \xrightarrow{p} \theta \quad \Rightarrow \quad g(\hat{\theta}) \xrightarrow{p} g(\theta),$$

and therefore we can learn the optimal rule in the limit.

However, this type of analysis does not capture the finite-sample risk arising from estimation error in $\hat{\theta}$.

One useful way to better reflect finite-sample properties is to consider the behavior of rules when $g(\theta) \approx 0$: let θ_0 satisfy

$$g(\theta_0) = 0,$$

and consider parameters local to θ_0:

$$\theta = \theta_0 + \frac{h}{\sqrt{n}}.$$

Under this local parametrization:

- uncertainty about whether $g(\theta) \lesseqgtr 0$ does not vanish;
- but classic asymptotic normality theory for parametric and semiparametric statistical models largely carries through.

Hirano and Porter (2009): if P_θ is a regular parametric model, and if $\hat{\theta}$ is an asymptotically efficient estimator (e.g. MLE), then the "plug-in" rule

$$\hat{\delta} = 1(g(\hat{\theta}) > 0)$$

is locally asymptotically minmax regret.

In semiparametric settings, if $\hat{g}$ is a semiparametrically efficient estimator of the ATE, then $\hat{\delta} = 1(\hat{g} > 0)$ is locally asymptotically minmax regret.

Empirical Welfare Maximization

This suggests to replace unknown welfare with a "good" estimate.

Next consider the problem with covariates: $\delta(\cdot)$ is a function of X.

Let

$$W(\delta) = E_X\left[\delta(X)E[Y(1)|X] + (1-\delta(X))E[Y(0)|X]\right].$$

Suppose

$$\widehat{W}(\delta) = \text{estimate of } W(\delta),$$

and we set

$$\hat{\delta} = \arg\max_{\delta} \widehat{W}(\delta).$$

This is the general empirical welfare maximization principle.

Suppose we have (conditionally) randomized experimental data and X has finite support.

Then we can estimate $E[Y|T=1,X]$ and $E[Y|T=0,X]$ by empirical conditional averages $\hat{\mu}_1(X)$ and $\hat{\mu}_0(X)$.

Then set

$$\widehat{W}(\delta) = \frac{1}{n}\sum_{i=1}^{n}\big[\delta(X_i)\hat{\mu}_1(X_i) + (1-\delta(X_i))\hat{\mu}_0(X_i)\big].$$

This leads to Manski's CES rule.

Next suppose X is continuous.

Then the space of possible functions $\delta(X)$ is very large, and it is not generally possible to estimate $W(\delta)$ uniformly well.

Kitagawa and Tetenov (2018) propose to restrict the class of possible rules δ.

For example, consider only rules of the form

$$\delta(X) = 1(\alpha + \beta X > 0).$$

In applications, it may be more practical to consider simple classes of rules such as this.

For $\delta \in \mathcal{A}$ where $\mathcal{A}$ is sufficiently "small," it may be possible to construct welfare estimators $\widehat{W}(\delta)$ s.t.

$$\widehat{W}(\delta) \xrightarrow{p} W(\delta),$$

and

$$\sqrt{n}\left(\widehat{W}(\delta) - W(\delta)\right) \xrightarrow{d} N(0, V_\delta),$$

uniformly in $\delta \in \mathcal{A}$.

Then

$$\hat{\delta}(X) = \arg\max_{\delta \in \mathcal{A}} \widehat{W}(\delta)$$

will typically have good properties.

$$\widehat{W}(\delta) = \frac{1}{n}\sum_{i=1}^{n}\big[\delta(X_i)\hat{\mu}_1(X_i) + (1-\delta(X_i))\hat{\mu}_0(X_i)\big].$$

$$\widehat{\widehat{W}}(\delta) = \frac{1}{n}\sum_{i=1}^{n}\left[\delta(X_i)\frac{Y_i T_i}{p(X_i)} + (1-\delta(X_i))\frac{Y_i(1-T_i)}{1-p(X_i)}\right].$$

$$\widehat{\widehat{\widehat{W}}}(\delta) = \frac{1}{n}\sum_{i=1}^{n}\left[\delta(X_i)\frac{Y_i T_i}{\hat{p}(X_i)} + (1-\delta(X_i))\frac{Y_i(1-T_i)}{1-\hat{p}(X_i)}\right].$$