Lecture I: Basic Concepts in Empirical Analysis

Ye Wang
University of North Carolina at Chapel Hill

Linear Methods in Causal Inference
POLI784

Estimand, estimator, and estimate

- In social science studies, we are often interested in the value of some quantity.
- E.g., approval rating for Mr. Trump; the effect of democracy on development.
- Such a quantity is referred to as the estimand or the target parameter (τ).
- Whether the estimand should be studied can only be justified by substantive arguments.
- We do not observe the estimand directly.
- We collect data generated under this target parameter.

Estimand, estimator, and estimate

- E.g., τ is the average approval rating for Trump in the US, and we observe $Y_i = \tau + \varepsilon_i$ for each individual i in a survey.
- The relationship between the data and the estimand is known as the data-generating process (DGP).
- Our goal: to infer the value of τ from data under assumptions on the DGP.
- Assumptions on the DGP should also be justified by substantive knowledge.
- In a more general case, $Y_i = f(\tau, \mathbf{X}_i, \varepsilon_i)$, with $\mathbf{X}_i$ being some covariates.
- Is $f(\cdot)$ smooth? Could it be linear in $\mathbf{X}_i$?
- The relationship $Y_i = \tau + \varepsilon_i$ is built upon multiple assumptions that can be wrong in practice.

Estimand, estimator, and estimate

- We attempt to achieve this goal using an estimator.
- An estimator is a mapping from you data to a number (or several numbers).
- You can think it as an algorithm (e.g., sample average $\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N} Y_i.$).
- Also known as a functional (a function of the distribution function, or the DGP).
- The number we obtain is called an estimate.
- We hope the estimator has good properties: the estimate it generates should be close to the estimand τ we care about.

Linear estimators

- We focus on linear estimators in this course.
- Suppose we have a sample of N units and observe the outcome Y_i, the treatment D_i, and the covariates $\mathbf{X}_i$.
- A linear estimator $\hat{\tau}$ takes the form

$$\hat{\tau} = \sum_{i=1}^{N} w(D_i, \mathbf{X}_i) * Y_i.$$

- A linear combination of Y_i.

Linear estimators

- For example, if Y_i and D_i are mean-zero and there are no covariates, the regression coefficient equals

$$\hat{\tau} = \frac{\sum_{i=1}^{N} D_i Y_i}{\sum_{i=1}^{N} D_i^2}$$

- Here $w(D_i, \mathbf{X}_i) = \frac{D_i}{\sum_{i=1}^{N} D_i^2}$.
- It can be more complicated and covers most methods we have for causal inference.
- Another example: the nearest-neighbor matching estimator:

$$\hat{\tau} = \frac{1}{N_1} \sum_{i:D_i=1} (Y_i - Y_{\mathcal{N}_i}),$$

where $Y_{\mathcal{N}_i}$ is i's nearest neighbor from the control group.

Identification

- If the estimate generated by the estimator equals the estimand τ when N is infinite, we say τ can be identified.
- Identification means whether we can infer the value of the target parameter at least in theory.
- In the previous example, it means we can find an estimator $\hat{\tau}$ such that $\tau = E[Y_i] = E[\hat{\tau}]$.
- Whether this is possible depends on assumptions we have imposed.

Properties of an estimator

- If $E[\hat{\tau}] = \tau$, we say the estimator is unbiased for τ.
- If there exists an unbiased estimator for τ, then τ can be identified.
- If $\lim_{N\to\infty} \hat{\tau} = \tau$, we say the estimator is consistent.
- Consistency holds when the variance of the estimator declines to zero:

$$P(|\hat{\tau} - \tau| > \varepsilon) \leq \frac{Var(\hat{\tau} - \tau)}{\varepsilon^2}. \text{ (Markov's inequality)}$$

- It is essentially the proof of the law of large numbers.

An example: sample average

- What are the properties of the sample average estimator?
- Suppose
  1. $Y_i \sim F(y)$, $E[Y_i] = \mu$,
  2. $Var[Y_i] = \sigma^2 < \infty$, and
  3. data are i.i.d. (independent and identically distributed)
- Remember that $\sigma^2 = E[Y_i^2] - \mu^2$.
- It is unbiased: $E[\hat{\tau}] = \frac{1}{N} \sum_{i=1}^{N} E[Y_i] = \frac{1}{N} \sum_{i=1}^{N} \mu = \mu$.

An example: sample average (*)

- The variance of the estimator is

$$
\begin{aligned}
Var[\hat{\tau}] = E[\hat{\tau}^2] - (E[\hat{\tau}])^2 &= E\left[\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} Y_i Y_j\right] - \mu^2 \\
&= \frac{1}{N^2}\sum_{i=1}^{N} E\left[Y_i^2\right] + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i} E\left[Y_i Y_j\right] - \mu^2 \\
&= \frac{1}{N^2}\sum_{i=1}^{N}(\sigma^2 + \mu^2) + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}\mu^2 - \mu^2 \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\sigma^2 = \frac{\sigma^2}{N} \to 0.
\end{aligned}
$$

- It is thus consistent.

Unbiasedness vs. consistency

- Consistency and unbiasedness do not imply each other.
- What estimator is consistent but biased?
- What estimator is unbiased yet inconsistent?
- Usually consistency is more important as we focus on large samples in social science.
- Unbiasedness matters more if the sample size is smaller.

From identification to estimation

- If an estimand can be identified, usually it can be estimated in finite sample.
- A common principle is to rely on the sample analogue.
- Suppose the estimand can be written as the expectation over a functional of the data: $\tau = E[f(Y_i, D_i, \mathbf{X}_i)]$.
- We replace the expectation sign $E[\cdot]$ with sample average $\frac{1}{N}\sum_{i=1}^{N} \cdot$.
- Identification is hard while estimation is easier.
- The estimate is the first number you are going to report in your quantitative analysis.
- It is important to discuss the magnitude of the estimate!
- Sometimes this is referred to as the economic significance of your estimate.
- It has welfare implications.

From estimation to inference

- But we also want to let our readers know how confident we are in the estimate.
- We want to construct confidence intervals for the estimate (often 95%).
- This process is called statistical inference.
- We can replace confidence intervals with confidence sets when the estimand is multi-dimensional.

Statistical inference

- First, we want to derive the theoretical variance of $\hat{\tau}$, $Var(\hat{\tau})$.
- If possible, we hope that $Var(\hat{\tau})$ is as small as possible (efficiency).
- $Var(\hat{\tau}) = E[\hat{\tau} - \tau]^2$ when $\hat{\tau}$ is unbiased.
- We have seen that if $Var(\hat{\tau}) \to 0$ when $N \to \infty$, $\hat{\tau}$ is consistent.
- It is often essential to know how fast $Var(\hat{\tau})$ declines to zero.

Statistical inference

- For most estimators, $N * Var(\hat{\tau})$ converges to a constant.
- Then, we have that $\sqrt{N}(\hat{\tau} - \tau)$ converges to a fixed distribution.
- We say $\hat{\tau}$ is root-N consistent.
- As we will see, most nonparametric estimators are not root-N consistent.
- For example, if $\hat{\tau}$ is based on kernel regression, then $N^{2/5}(\hat{\tau} - \tau)$ converges to a fixed distribution (under regularity conditions).

Statistical inference

- The variance's value often hinges on unknown parameters.
- We also need to find an estimate for $N * Var(\hat{\tau})$, denoted as $\hat{\sigma}^2$.
- We call $\frac{\hat{\sigma}}{\sqrt{N}}$ the standard error of $\hat{\tau}$.
- This becomes another estimation problem.
- We hope our variance estimate to be unbiased and consistent.
- At least, it should be conservative: $\hat{\sigma}^2 \geq N * Var(\hat{\tau})$ when $N \to \infty$.
- This is usually the second number you report in your analysis.

Statistical inference

- To construct confidence intervals, we need to know the distribution of $\hat{\tau}$, $F_N(\hat{\tau})$ even when $\hat{\tau}$ is root-N consistent.
- When N is finite, it is often impossible to know the answer.
- But as N is sufficiently large, the distribution is often close to the normal distribution: $\mathcal{N}(\tau, N * Var(\hat{\tau}))$.
- This is justified by the central limit theorem (CLT):

$$\sqrt{N}(\hat{\tau} - \tau) \to \mathcal{N}(0, N * Var(\hat{\tau})).$$

- Remember that our estimators have the linear form, hence they often converge to normality.

Statistical inference

- Another approach is to approximate $F_N(\hat{\tau})$ with resampling techiniques.
- Common choices: jackknife and bootstrap.
- If we can approximate $F_N(\hat{\tau})$, we can construct the confidence intervals as

$$\hat{\mathcal{C}} = \left[\hat{\tau} - z_{1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{N}}, \hat{\tau} - z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{N}}\right]$$

- What is the interpretation of the confidence interval?
- Remember that $\hat{\mathcal{C}}$ is an approximation!

Statistical inference

- Confidence intervals are closely connected with hypothesis testing.
- Under the null hypothesis $H_0 : \tau = 0$, we know that

$$\frac{\hat{\tau}}{\sqrt{Var(\hat{\tau})}} \rightarrow \mathcal{N}(0, 1)$$

- We reject the null if $\frac{\hat{\tau}}{\sqrt{Var(\hat{\tau})}}$ is larger (smaller) than the $100 * (1 - \alpha/2)$th ($100 * (\alpha/2)$th) percentile of the normal distribution.
- α is called the level of the test.
- A critical property of the confidence interval is the coverage rate, defined as

$$P(\tau \in \hat{\mathcal{C}}).$$

- We hope the coverage rate is at least $(1 - \alpha)\%$ when $N \rightarrow \infty$:

$$\lim_{N \rightarrow \infty} P(\tau \in \hat{\mathcal{C}}) \geq (1 - \alpha).$$

An example: sample average (continued)

- We can prove that the sample average is efficient.
- We can estimate its variance via either $\frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$ or $\frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$.
- Both variance estimators are consistent but only the latter is unbiased.
- We can show that $\sqrt{N}(\hat{\tau} - \tau) \rightarrow \mathcal{N}(0, \sigma^2)$ using the CLT.
- The 95% confidence interval of $\hat{\tau}$ can be conducted using critical values from the normal distribution.

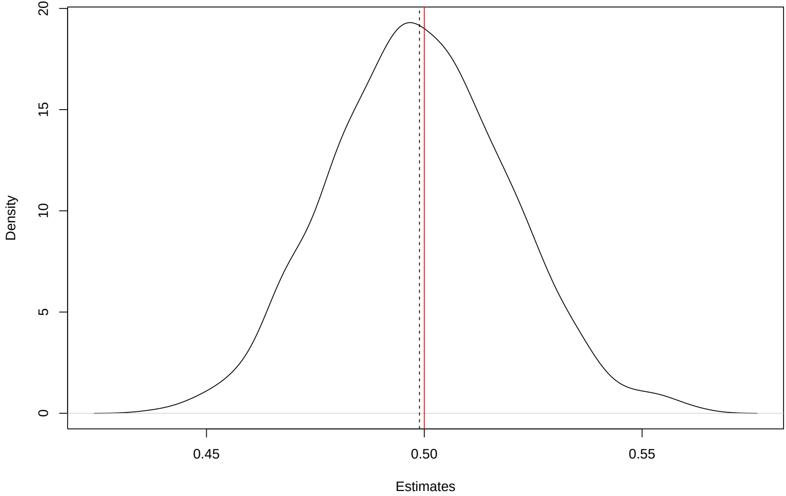
Monte Carlo experiment

- With real data, we never know what the true DGP or the estimand is.
- But we can specify them in simulation, or Monte Carlo experiments.
- It is thus important to examine the performance of any method via simulation.
- We generate the data from a distribution that satisfies the requirement of the method.
- We apply the method to the data, obtaining all the quantities we need (the estimate, the variance estimate, the confidence interval, etc.).
- Remember that we can do this repeatedly and allow N to increase.

Monte Carlo experiment: sample average

```r
N <- 200
Nboots <- 1000
ests <- matrix(NA, Nboots, 3)
covered <- rep(NA, Nboots)
for (b in 1:Nboots){
  Y <- runif(N) # population mean: 0.5
  # true variance is 1/12 = 0.0833
  Y_bar <- mean(Y)
  Y_var1 <- var(Y)
  Y_var2 <- var(Y) * ((N - 1) / N)
  ests[b, 1] <- Y_bar
  ests[b, 2] <- Y_var1
  ests[b, 3] <- Y_var2
  CI <- c(Y_bar - 1.96 * sqrt(Y_var1 / N),
          Y_bar + 1.96 * sqrt(Y_var1 / N))
  covered[b] <- CI[1] <= 0.5 & CI[2] >= 0.5
}
```

Monte Carlo experiment: sample average



Monte Carlo experiment: sample average

```
mean(ests[, 1]) - 0.5 # bias
```

```
## [1] -0.001103303
```

```
N*var(ests[, 1]) # true variance (simulated)
```

```
## [1] 0.08312325
```

```
mean(ests[, 2]) # avg. of estimated variance
```

```
## [1] 0.08348342
```

```
mean(ests[, 3]) # avg. of estimated variance
```

```
## [1] 0.083066
```

```
mean(covered) # coverage rate
```

```
## [1] 0.953
```

References I