Lectures 2 & 3

Properties of Estimators

Oscar Volpe

9/29/2021 & 10/4/2021

Drawing Data

When we collect data, we are observing the realizations of random vectors.

Definition (Sample)

A *sample* of size n, denoted by $\{X_i\}_{i=1}^n$, is a collection of random vectors.

- *Note:* the *sampling process* may be characterized in a variety of ways.
- When we take independent draws from a population, the resulting sample will be (on average) representative of the sample space.

Definition (Independent and Identically Distributed)

A sample $\{X_i\}_{i=1}^n$ is *independent and identically distributed (i.i.d.)* if elements of the $\{X_i\}_{i=1}^n$ are mutually independent and are all distributed according to the same distribution, i.e. $X_i \perp X_j$ and $F_{X_i} = F_{X_j}$ for all $i \neq j$.

Defining an Estimator

Our goal is to use data to say something about the *true* features of the wider population. To accomplish this task, we construct *estimators*.

Definition (Estimator, Estimate)

Given a sample $\{X_i\}_{i=1}^n$ and an unknown parameter θ in the population, an *estimator* for θ, denoted by $\hat{\theta}_n$, is a function of $\{X_i\}_{i=1}^n$ used to learn about θ. We call the realization of $\hat{\theta}_n$ an *estimate* of θ.

- The *target parameter* (or *estimand*) is object we wish to estimate.
- Given data $\{X_i\}_{i=1}^n$, we might want to estimate the population mean, the population variance, or even the entire distribution function.
- *Important Distinction:* the target parameter θ versus the estimator $\hat{\theta}_n$.

Sample Analogue Principle

Suppose we know some properties that are satisfied for the "true parameter" in the population. If we can find a parameter value in the sample that causes the sample to mimic the properties of the population, we might use this parameter value to estimate the true parameter.

Suppose we have a sample $\{X_i\}_{i=1}^n$ drawn from distribution F. The *sample analogue principle* tells us we can estimate $\theta(F)$ using $\hat{\theta}_n = \theta(\hat{F}_n)$, where:

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leq t)$$

We call $\hat{F}_n(t)$ the *empirical distribution function*, as it approximates $F(t)$ by computing the proportion of draws that satisfy the condition $X_i \leq t$.

- The *sample analogue principle* gives us a "natural" estimator of θ.
- We compute the estimator by acting as if $\hat{F}_n = F$.

Method of Moments

In practice, the *sample analogue principle* suggests estimating parameters using sample averages. It leads to what we call the "Method of Moments".

- *Step 1:* write θ in terms of population moments: $E(X)$, $E(X^2)$, etc.
- *Step 2:* replace the population moments with sample averages

The Sample Mean

Given a sample $\{X_i\}_{i=1}^n$, a natural estimator for $E(X)$ is $\bar{X}_n$, where:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

We call this quantity the *sample mean* of X. Similarly, a natural estimator for $E(X^k)$ is $\frac{1}{n}\sum_{i=1}^{n} X_i^k$. These are all *method of moments* estimators.

Estimating the Variance

Any parameter θ that can be expressed in terms of population moments has a *method of moments* estimator, e.g. $g(\bar{X}_n)$ approximates $g(E(X))$.

- This gives us an estimator for a wide variety of target parameters.
- *Example:* you have a sample $\{X_i, Y_i, Z_i\}_{i=1}^n$ and $\theta = E(XY^2Z^3)$ is your target parameter. An estimator is $\hat{\theta}_n^{\text{MoM}} = \frac{1}{n}\sum_{i=1}^n X_i Y_i^2 Z_i^3$.

What is *method of moments* estimator for $\text{Var}(X) = E((X - E(X))^2)$?

$$\hat{\theta}_n^{\text{MoM}} = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}_n^2$$

Is $\hat{\theta}_n^{\text{MoM}}$ the "best" estimator for $\text{Var}(X)$? What do we mean by "best"?

Finite-Sample Properties

Definition (Bias)

Let θ be some target parameter. The *bias* of an estimator $\hat{\theta}_n$ for θ equals:

$$\text{Bias}(\hat{\theta}_n) = E(\hat{\theta}_n) - \theta$$

- We say that $\hat{\theta}_n$ is *unbiased* if $\text{Bias}(\hat{\theta}_n) = 0$.
- We call the sign of $\text{Bias}(\hat{\theta}_n)$ the "direction of bias".

Definition (Precision)

The *variance* of an estimator $\hat{\theta}_n$ is $\text{Var}(\hat{\theta}_n)$, and the *precision* of an estimator is the reciprocal of its variance, i.e. $\text{Precision}(\hat{\theta}_n) = 1/\text{Var}(\hat{\theta}_n)$.

- Even for small samples, it is often desirable to have *precise* estimators.
- Let $\hat{\theta}_n$ and $\tilde{\theta}_n$ be unbiased estimators. We say that $\hat{\theta}_n$ is *more efficient* than $\tilde{\theta}_n$ if it has higher precision, i.e. lower variance, than $\tilde{\theta}_n$.

The Sample Variance

We can show that $\hat{\theta}_n^{\text{MoM}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ is *biased downward*. Why?

The Sample Variance

An unbiased estimate of the $\text{Var}(X)$ is the *sample variance* s_n^2, where:

$$s_n^2 = \frac{1}{n-1}\sum_{i-1}^{n}(X_i - \bar{X}_n)^2$$

In this case, the *method of moments* estimator is not most desirable.

- Dividing by $n-1$ instead of n is called "Bessel's correction".
- Similarly, the *sample covariance* $s_{XY} = \frac{1}{n-1}\sum_{i-1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)$ will give you an unbiased estimator for the covariance $\theta = \text{Cov}(X, Y)$.

Mean Squared Error

What properties would it be "nice" for an estimator to have?

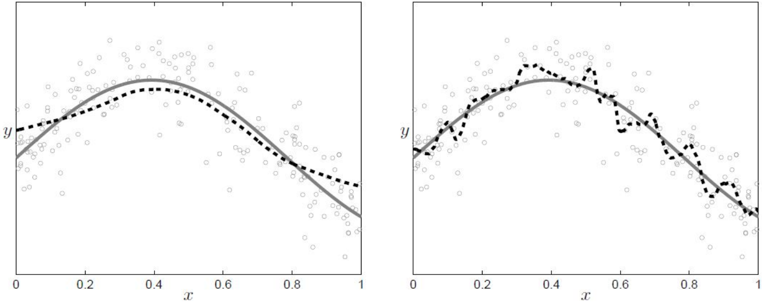
1. Low Bias: $\mathbb{E}[\hat{\theta}_n] \approx \theta$
2. High Precision/Low Variance: $E[(\hat{\theta}_n - E[\hat{\theta}_n])^2]$ "small"

In practice, when fitting models, we often encounter trade-offs.

$$\begin{aligned} MSE(\hat{\theta}_n) &= \mathbb{E}[(\hat{\theta}_n - \theta)^2] \\ &= \mathbb{E}[\hat{\theta}_n^2] + \theta^2 - 2\mathbb{E}[\hat{\theta}_n]\theta \\ &= \underbrace{\mathbb{E}[\hat{\theta}_n^2] - \mathbb{E}[\hat{\theta}_n]^2}_{\text{Variance}} + \underbrace{\theta^2 - 2\mathbb{E}[\hat{\theta}_n]\theta + \mathbb{E}[\hat{\theta}_n]^2}_{\text{Bias}^2} \end{aligned}$$

Visualizing the Trade-off

Visualizing the Bias-Variance Trade-off



- The left is **oversmoothed** — high bias, low variance
- The right is **undersmoothed** — low bias, high variance
- (What does "too much/little" mean? Here we use "eyeball optimality")

Diagram from A. Torgovitsky.

Convergence in Probability

Definition (Convergence in Probability)

A sequence of random vectors $\{X_i\}_{i=1}^n$ *converges in probability* to X, denoted by $X_n \xrightarrow{p} X$, if, for all $\varepsilon > 0$, $P(|X_n - X| > \varepsilon) \to 0$ as $n \to \infty$.

If an estimator $\hat{\theta}_n$ converges in probability to θ, i.e. if $\hat{\theta}_n \xrightarrow{p} \theta$, then we say that $\hat{\theta}_n$ is a *consistent* estimator of θ. This is an *asymptotic property*.

- Intuitively, $\hat{\theta}_n$ is *consistent* if it gets "closer" (in a $\xrightarrow{p}$ sense) to θ when the sample size n becomes larger. For large samples, this is desirable.
- $\xrightarrow{p}$ differs from other types of convergence, such as *a.s. convergence*, *convergence in q_{th} moment*, and *convergence in distribution*.

Markov's Inequality

Theorem (Markov's Inequality)

For any random variable X, $P(|X| > \varepsilon) \leq \frac{E(|X|^q)}{\varepsilon^q}$ for all $q, \varepsilon > 0$.

- Notice that *Markov's inequality* places an upper bound on the probability that $|X| > \varepsilon$ in terms of the moments of $|X|$.
- Use it constructing confidence regions for $\hat{\theta}_n$ or to show $\hat{\theta}_n \xrightarrow{p} \theta$.

Application (WLLN)

Let $\{X_i\}_{i=1}^n$ be i.i.d. random variables with mean μ and variance σ^2. Then:

$$\begin{aligned}
P(|\bar{X}_n - \mu| > \varepsilon) &\leq \frac{E(|\bar{X}_n - \mu|^2)}{\varepsilon^2} \\
&= \frac{E\left(\sum_{i=1}^n (X_i - \mu)^2 + \sum_{i=1}^n \sum_{j \neq i} (X_i - \mu)(X_j - \mu)\right)}{n^2 \varepsilon^2} \\
&= \frac{n\text{Var}(X_i)}{n^2 \varepsilon^2} = \frac{\sigma^2}{n \varepsilon^2} \to 0, \quad \text{as } n \to \infty
\end{aligned}$$

Weak Law of Large Numbers

Theorem (Weak Law of Large Numbers)

Let $\{X_i\}_{i=1}^{n}$ be a sample of i.i.d. random variables. If $E(X)$ exists, then the sample mean $\bar{X}_n$ is a consistent estimator for $E(X)$, i.e. $\bar{X}_n \xrightarrow{p} X$.

- *Important Result:* as long as the sample is i.i.d., the sample mean will tend toward the *true* mean as the sample size becomes larger.
- There is also a *Strong Law of Large Numbers*, stating: $\bar{X}_n \xrightarrow{a.s.} E(X)$.
- The WLLN implies that $\frac{1}{n}\sum_{i=1}^{n} g(X_i) \xrightarrow{p} E(g(X_i))$ if $E(g(X_i))$ exists, since functions of i.i.d. random variables are also going to be i.i.d..
- The WLLN is even more powerful when combined with the Continuous Mapping Theorem (see the next section).

Theorem Statement

Theorem (CMT for $\xrightarrow{p}$)

Let $\theta_1, \ldots, \theta_k$ be unknown parameters in the population. Let $\{X_i\}_{i=1}^n$ be a sample, and let $\hat{\theta}_n^{(1)}, \ldots, \hat{\theta}_n^{(k)}$ be estimators for $\theta_1, \ldots, \theta_k$ (respectively). If the function g is continuous over the support of $(\theta_1, \ldots, \theta_k)$, then:

$$\hat{\theta}_n^{(1)} \xrightarrow{p} \theta_1, \ldots, \hat{\theta}_n^{(k)} \xrightarrow{p} \theta_k \quad \Longrightarrow \quad g(\hat{\theta}_n^{(1)}, \ldots, \hat{\theta}_n^{(k)}) \xrightarrow{p} g(\theta_1, \ldots, \theta_k)$$

Example

We can show $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ is consistent for $\text{Var}(X)$ using the CMT.

$$\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}_n^2, \quad \text{where:} \quad \begin{aligned} \frac{1}{n}\sum_{i=1}^n X_i^2 &\xrightarrow{p} E(X^2) \\ \bar{X}_n &\xrightarrow{p} E(X) \end{aligned}$$

Since $g(y, z) = y - z^2$ is a continuous function, the continuous mapping theorem guarantees that $\frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}_n^2 \xrightarrow{p} E(X^2) - E(X)^2 = \text{Var}(X)$.

Finite-Sample vs. Asymptotic Properties

Suppose $\{X_i\}_{i=1}^{n}$ *are iid random variables generated from* F.

- *Q1.* Is $\bar{X}_n$ unbiased for $\mathbb{E}[X]$? Is it consistent?
- *Q2.* Is $\frac{X_1+X_2}{2}$ unbiased for $\mathbb{E}[X]$? Is it consistent?
- *Q3.* Is $\bar{X}_n^{-1}$ unbiased for $\mathbb{E}[X]^{-1}$? Is it consistent?
- *Q4.* Is $g(\bar{X}_n)$ unbiased for $g(E(X))$? Is it consistent?
- *Q5.* Is $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ unbiased for $\mathrm{Var}[X]$? Is it consistent?

Limiting Distributions

Definition (Convergence in Distribution)

A sequence of random vectors $\{X_i\}_{i=1}^n$ *converges in distribution* to X, denoted by $X_n \xrightarrow{d} X$, if, for all x at which $P(X \leq x)$ is continuous:

$$P(X_n \leq x) \rightarrow P(X \leq x) \quad \text{as } n \rightarrow \infty$$

- *Important Note:* $\xrightarrow{p}$ implies $\xrightarrow{d}$, but $\xrightarrow{d}$ *does not* imply $\xrightarrow{p}$.
  - As a counterexample, let $X \sim N(0,1)$ and $X_n = -X$.
- $\xrightarrow{d}$ is useful for deriving the asymptotic distributions of estimators.

Useful Properties

- $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} X_n$ implies $Y_n \xrightarrow{d} X$.
- $X_n \xrightarrow{d} c$ implies $X_n \xrightarrow{p} c$ *if* c is a constant.

Theorem Statement: Univariate Case

Theorem (Central Limit Theorem)

Let $\{X_i\}_{i=1}^n$ be a sample of i.i.d. random variables. If $Var(X) < \infty$, then:

$$\sqrt{n}(\bar{X}_n - E(X)) \xrightarrow{d} N(0, Var(X))$$

- For "large" samples, $\sqrt{n}(\bar{X}_n - E(X))$ is approximately normally distributed, regardless of what the initial distribution of X_i is.
- Extremely useful for deriving the limiting distributions of estimators.
- Even more powerful when used with *Slutsky's theorem* (next slide).
- We say that $\hat{\theta}_n$ is a $\sqrt{n}$*-consistent* estimator for θ if:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2),$$

for some σ^2, which we call the *asymptotic variance* of $\hat{\theta}_n$.

Continuous Mapping Theorem for $\xrightarrow{d}$

Theorem (CMT for $\xrightarrow{d}$)

Let θ be an unknown parameter in the population. Let $\{X_i\}_{i=1}^n$ be a sample, and let $\hat{\theta}_n$ be an estimator for θ. If the function g is continuous over the support of θ, then $\hat{\theta}_n \xrightarrow{d} \theta$ implies that $g(\hat{\theta}_n) \xrightarrow{d} g(\theta)$.

- Importantly, note that marginal $\xrightarrow{d}$ does not imply joint $\xrightarrow{d}$.

An important special case of this theorem is *Slutsky's theorem:*

Theorem (Slutsky's Theorem)

Suppose $\hat{\theta}_n^{(1)} \xrightarrow{d} X$ and $\hat{\theta}_n^{(2)} \xrightarrow{d} c$ for some constant $c \neq 0$. Then:

$$\hat{\theta}_n^{(1)} + \hat{\theta}_n^{(2)} \xrightarrow{d} X + c, \quad \hat{\theta}_n^{(1)}\hat{\theta}_n^{(2)} \xrightarrow{d} Xc, \quad \hat{\theta}_n^{(1)}/\hat{\theta}_n^{(2)} \xrightarrow{d} X/c$$

Theorem Statement: Multivariate Case

Theorem (Multivariate Central Limit Theorem)

Let $\{X_i\}_{i=1}^n$ be a sample of i.i.d. random vectors in $\mathbb{R}^k$. Suppose that the variance-covariance matrix $\Sigma \in \mathbb{R}^{k \times k}$ exists. Then:

$$\sqrt{n}(\bar{X}_n - E(X)) \xrightarrow{d} N(\mathbf{0}, \Sigma)$$

Note. This theorem is particularly useful when we look at linear models:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + U_i,$$

where each coefficient β_j is estimated by an estimator $\hat{\beta}_j$. The multivariate Central Limit Theorem allows us to derive the limiting distribution:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(\mathbf{0}, V),$$

where $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)' \in \mathbb{R}^{k+1}$ and $\beta = (\beta_0, \beta_1, \ldots, \beta_k)' \in \mathbb{R}^{k+1}$.

Bessel's Correction

We show that $E(\hat{\theta}_n^{\text{MoM}}) = \left(\frac{n-1}{n}\right)\text{Var}(X_i)$, so $\hat{\theta}_n^{\text{MoM}}$ is downward biased.

Setting $s_n^2 = \frac{1}{n-1}\sum_{i-1}^n (X_i - \bar{X}_n)^2 = \frac{n}{n-1}\hat{\theta}_n^{\text{MoM}}$, we have $E(s_n^2) = \text{Var}(X_i)$.

$$
\begin{aligned}
E(\hat{\theta}_n^{\text{MoM}}) &= E\Big(\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2\Big) = \frac{1}{n}\sum_{i=1}^n E\Big((X_i - \bar{X}_n)^2\Big) \\
&= \frac{1}{n}\sum_{i=1}^n E\Big((X_i - E(X_i) - (\bar{X}_n - E(X_i)))^2\Big) \\
&= \frac{1}{n}\sum_{i=1}^n E\Big((X_i - E(X_i))^2 - 2(X_i - E(X_i))(\bar{X}_n - E(X_i)) + (\bar{X}_n - E(X_i))^2\Big) \\
&= \frac{1}{n}\sum_{i=1}^n \text{Var}(X_i) - \frac{2}{n}\sum_{i=1}^n E((X_i - E(X_i))(\bar{X}_n - E(X_i))) + \frac{1}{n}\sum_{i=1}^n E((\bar{X}_n - E(X_i))^2) \\
&= \frac{n}{n}\text{Var}(X_i) - E\Big((\bar{X}_n - E(X_i))\frac{2}{n}\sum_{i=1}^n (X_i - E(X_i))\Big) + \frac{n}{n}E((\bar{X}_n - E(X_i))^2) \\
&= \text{Var}(X_i) - E((\bar{X}_n - E(X_i))^2) = \text{Var}(X_i) - \text{Var}(\bar{X}_n) = \text{Var}(X_i) - \frac{1}{n}\text{Var}(X_i)
\end{aligned}
$$

Go Back