ECMA31000: Introduction to Empirical Analysis

Estimation

Joe Hardwick

University of Chicago

Autumn 2021

Outline

- This week:
  - Delta Method + Examples
  - Estimation: Definitions, properties of estimators.
  - Method of Moments Estimation.

Delta Method

Theorem

Let $\{X_n\}_{n\geq 1}$ be a sequence of $(K \times 1)$ random vectors and suppose that

$$n^r (X_n - c) \xrightarrow{d} X$$

for some $r > 0$ and constant vector c. Let $g : \mathbb{R}^k \to \mathbb{R}^d$ be differentiable at the point c. Let $Dg(c)$ be the $d \times k$ matrix of partial derivatives evaluated at c. Then

$$n^r (g(X_n) - g(c)) \xrightarrow{d} Dg(c) X.$$

In particular, if $X \sim \mathcal{N}(0, \Sigma)$, then

$$n^r (g(X_n) - g(c)) \xrightarrow{d} \mathcal{N}\left(0, Dg(c) \Sigma Dg(c)'\right).$$

Delta Method

- $Dg\left(c\right)$ is the following matrix of partial derivatives:

$$Dg\left(c\right)=\begin{bmatrix}\frac{\partial g_1}{\partial x_1}\left(c\right) & \frac{\partial g_1}{\partial x_2}\left(c\right) & \cdots & \frac{\partial g_1}{\partial x_k}\left(c\right)\\ \frac{\partial g_2}{\partial x_1}\left(c\right) & \frac{\partial g_2}{\partial x_2}\left(c\right) & \cdots & \frac{\partial g_2}{\partial x_k}\left(c\right)\\ \vdots & \vdots & \ddots & \vdots\\ \frac{\partial g_d}{\partial x_1}\left(c\right) & \frac{\partial g_d}{\partial x_2}\left(c\right) & \cdots & \frac{\partial g_d}{\partial x_k}\left(c\right)\end{bmatrix}.$$

Delta Method

Proof.

By Taylor's theorem:

$$g\left(x\right)=g\left(c\right)+Dg\left(c\right)\left(x-c\right)+h_1\left(x\right)\left(x-c\right),$$

for some function $h_1\left(x\right)$ with $\lim_{x\to c}h_1\left(x\right)=h_1\left(c\right)=0$. It follows that

$$n^r\left(g\left(X_n\right)-g\left(c\right)\right)=Dg\left(c\right)n^r\left(X_n-c\right)+h_1\left(X_n\right)n^r\left(X_n-c\right).$$

Since $n^r\left(X_n-c\right)\overset{d}{\to}X$, by Slutsky's Theorem:

$$Dg\left(c\right)n^r\left(X_n-c\right)\overset{d}{\to}Dg\left(c\right)X.$$

Now show that $h_1\left(X_n\right)n^r\left(X_n-c\right)=o_p\left(1\right)$:

Delta Method

Proof.

Since $n^r (X_n - c) \xrightarrow{d} X$, we have $n^r (X_n - c) = O_p(1)$ and $X_n \xrightarrow{p} c$ (See Problem Set 3). Since h_1 is continuous at c by construction,

$$h_1(X_n) \xrightarrow{p} h_1(c) = 0$$

by the CMT. Therefore,

$$h_1(X_n)\, n^r (X_n - c) = o_p(1) \cdot O_p(1) = o_p(1),$$

so

$$\begin{aligned} n^r (g(X_n) - g(c)) &= Dg(c)\, n^r (X_n - c) + o_p(1) \\ &\xrightarrow{d} Dg(c)\, X. \end{aligned}$$

□

Delta Method

- Note that if $Dg\left(c\right)=0$,

$$n^r\left(g\left(X_n\right)-g\left(c\right)\right)\overset{d}{\to}0,$$

  which is a degenerate limiting distribution, (so $\overset{p}{\to}0$ also).

- If g has higher order derivatives, we can derive an alternate form of the Delta Method when $Dg\left(c\right)=0$.
- Suppose $\left\{X_n\right\}_{n\geq 1}$ is a sequence of random variables and $g:\mathbb{R}\to\mathbb{R}$ has 2 derivatives.
- Taylor's theorem implies

$$g\left(x\right)=g\left(c\right)+g'\left(c\right)\left(x-c\right)+\frac{g''\left(c\right)}{2}\left(x-c\right)^2+h_2\left(x\right)\left(x-c\right)^2,$$

  where $\lim_{x\to c}h_2\left(x\right)=h_2\left(c\right)=0$.

Delta Method

- Repeating the argument in the proof of the original Delta Method:

$$n^{2r}\left(g\left(X_n\right)-g\left(c\right)\right)=g'\left(c\right)n^{2r}\left(X_n-c\right)+\frac{g''\left(c\right)}{2}n^{2r}\left(X_n-c\right)^2 \\ +h_2\left(X_n\right)n^{2r}\left(X_n-c\right)^2.$$

- We have

$$\begin{aligned} g'\left(c\right)n^{2r}\left(X_n-c\right)&=0 \\ \frac{g''\left(c\right)}{2}n^{2r}\left(X_n-c\right)^2&\xrightarrow{d}\frac{g''\left(c\right)}{2}X^2 \\ h_2\left(X_n\right)n^{2r}\left(X_n-c\right)^2&=o_p\left(1\right)O_p\left(1\right)=o_p\left(1\right). \end{aligned}$$

- In summary, if $g'\left(c\right)=0$,

$$n^{2r}\left(g\left(X_n\right)-g\left(c\right)\right)\xrightarrow{d}\frac{g''\left(c\right)}{2}X^2.$$

Example: Sample variance

- We will tie together the concepts we have learned so far to find the asymptotic distribution of

$$S_n^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 .$$

- Let X_i be iid random variables with $\mathrm{E}\left(X_i\right) = \mu$, $Var\left(X_i\right) = \sigma^2$ and $\mathrm{E}\left(X_i - \mu\right)^4 = \kappa$.
- We are looking for an $r > 0$, constant c and random variable X such that

$$n^r \left(S_n^2 - c\right) \xrightarrow{d} X,$$

for some non-degenerate X.

Example: Sample variance

- We have already shown that $S_n^2 \xrightarrow{p} \sigma^2$, so we must take $c = \sigma^2$.
- Unfortunately, S_n^2 is not in a form where we can apply the CLT directly:

$$\sqrt{n}\left(S_n^2 - \sigma^2\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\left(X_i - \bar{X}_n\right)^2 - \sigma^2\right].$$

- If we could replace $\bar{X}_n$ with μ, would consider

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[(X_i - \mu)^2 - \sigma^2\right].$$

Example: Sample variance

- Note that $\left\{(X_i-\mu)^2\right\}_{i\geq 1}$ is an iid sequence with $\mathrm{E}\left(X_i-\mu\right)^2=\sigma^2$ and

$$\begin{aligned} Var\left((X_i-\mu)^2\right) &= \mathrm{E}\left[(X_i-\mu)^4\right]-\left[\mathrm{E}\left(X_i-\mu\right)^2\right]^2 \\ &= \kappa-\sigma^4. \end{aligned}$$

- Therefore, by the CLT:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[(X_i-\mu)^2-\sigma^2\right]\xrightarrow{d}\mathcal{N}\left(0,\kappa-\sigma^4\right).$$

- It remains to show that

$$\sqrt{n}\left(S_n^2-\sigma^2\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[(X_i-\mu)^2-\sigma^2\right]+o_p\left(1\right).$$

Example: Sample variance

- Note that

$$
\begin{aligned}
\sqrt{n}\left(S_n^2-\sigma^2\right) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\left(X_i-\bar{X}_n\right)^2-\sigma^2\right] \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\left(X_i-\mu-\left(\bar{X}_n-\mu\right)\right)^2-\sigma^2\right] \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\left(X_i-\mu\right)^2-\sigma^2\right]+\sqrt{n}\left(\bar{X}_n-\mu\right)^2 \\
&-2\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\left(X_i-\mu\right)\left(\bar{X}_n-\mu\right)\right] \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\left(X_i-\mu\right)^2-\sigma^2\right]-\sqrt{n}\left(\bar{X}_n-\mu\right)^2 .
\end{aligned}
$$

Example: Sample variance

- Finally, since

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right),$$
$$\left(\bar{X}_n - \mu\right) = o_p\left(1\right),$$

  we get

$$\sqrt{n}\left(\bar{X}_n - \mu\right)^2 = o_p\left(1\right) O_p\left(1\right) = o_p\left(1\right).$$

- In summary:

$$\sqrt{n}\left(S_n^2 - \sigma^2\right) \xrightarrow{d} \mathcal{N}\left(0, \kappa - \sigma^4\right).$$

Example: Sample variance

- We are not done yet: The limiting distribution is non-degenerate iff
$$\kappa - \sigma^4 > 0.$$
- Jensen's inequality gives
$$\kappa = \mathrm{E}\left[(X_i - \mu)^4\right] \geq \left[\mathrm{E}\,(X_i - \mu)^2\right]^2 = \sigma^4,$$
with equality if and only if the random variable $(X_i - \mu)^2$ is constant almost surely.
- In this case, this does NOT imply X_i is constant, since
$$(X_i - \mu)^2 = \sigma^2$$
will have two solutions: $X_i = \mu \pm \sigma$ when $\sigma > 0$.

Example: Sample variance

- First, since $\mathrm{E}\left(X_i\right) = \mu$, we must have

$$X_i = \begin{cases} \mu + \sigma & \text{with probability } \frac{1}{2} \\ \mu - \sigma & \text{with probability } \frac{1}{2} \end{cases}.$$

- First, note that:

$$\begin{aligned} S_n^2 &= \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 \\ &= \frac{\sigma^2}{n} \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} - \left(\frac{\bar{X} - \mu}{\sigma} \right) \right)^2. \end{aligned}$$

- The behaviour of S_n^2 does not depend on μ and σ is just a scale factor.

Example: Sample variance

- Therefore, without loss of generality, let $\mu = \frac{1}{2}, \sigma = \frac{1}{2}$ for simplicity. Then:

$$X_i = \begin{cases} 1 & \text{with probability } \frac{1}{2} \\ 0 & \text{with probability } \frac{1}{2} \end{cases},$$

where $\mathrm{E}\left(X_i\right) = \mu$ and $Var\left(X_i\right) = \frac{1}{2}\left(1 - \frac{1}{2}\right) = \frac{1}{4}$.

- In this case:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \bar{X} - \bar{X}^2 = \bar{X}\left(1 - \bar{X}\right),$$

since $X_i^2 = X_i$.

Example: Sample variance

- The CLT provides

$$\sqrt{n}\left(\bar{X}_n - \frac{1}{2}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{4}\right).$$

- We might try to use the Delta Method to find the limiting distribution of $\bar{X}\left(1-\bar{X}\right)$, but we already know this will fail.
- Let $g\left(x\right) = x\left(1-x\right)$. Indeed,

$$g\left(\bar{X}_n\right) - g\left(\frac{1}{2}\right) = \bar{X}\left(1-\bar{X}\right) - \frac{1}{4},$$

but the function g has a unique maximum at $x = \frac{1}{2}$, giving $g\left(\frac{1}{2}\right) = \frac{1}{4}$. Therefore,

$$g\left(\bar{X}_n\right) - g\left(\frac{1}{2}\right) \leq 0$$

almost surely, so it is hopeless to ask for a normal limiting distribution.

Example: Sample variance

- Note that $g'(x) = 1 - 2x$ and $g''(x) = -2$.
- The Delta Method tells us

$$\sqrt{n}\left(g\left(\bar{X}_n\right) - g\left(\frac{1}{2}\right)\right) \xrightarrow{d} \mathcal{N}\left(0, g'\left(\frac{1}{2}\right)\frac{1}{4}\right),$$

  but $g'\left(\frac{1}{2}\right) = 0$!
- Since $g'\left(\frac{1}{2}\right) = 0$, but $g''\left(\frac{1}{2}\right) = -2$,

$$\begin{aligned}\left(\sqrt{n}\right)^2\left(g\left(\bar{X}_n\right) - g\left(\frac{1}{2}\right)\right) &\xrightarrow{d} \frac{g''\left(\frac{1}{2}\right)}{2}\mathcal{N}\left(0, \frac{1}{4}\right)^2 \\ &\overset{d}{=} -\frac{1}{4}\chi_1^2.\end{aligned}$$

Example: Sample variance

- In summary,

$$n^{1/2}\left(S_n^2-\sigma^2\right)\xrightarrow{d}\mathcal{N}\left(0,\kappa-\sigma^4\right),$$

unless $\kappa-\sigma^4=0$, in which case $X_i=\mu\pm\sigma$, where each outcome has probability $\frac{1}{2}$. In that case,

$$n\left(S_n^2-\sigma^2\right)\xrightarrow{d}-\sigma^2\chi_1^2.$$

- Note: Could have derived this result without the details simply by noting that if $(X_i-\mu)^2$ is constant, then

$$\begin{aligned}n\left(S_n^2-\sigma^2\right)&=\sum_{i=1}^{n}\left[(X_i-\mu)^2-\sigma^2\right]-n\left(\bar{X}_n-\mu\right)^2\\&=-n\left(\bar{X}_n-\mu\right)^2.\end{aligned}$$

Example: Sample Standard Deviation

- Suppose $\kappa > \sigma^4$ so that:

$$n^{1/2}\left(S_n^2 - \sigma^2\right) \xrightarrow{d} \mathcal{N}\left(0, \kappa - \sigma^4\right).$$

- By the delta method, with $g(x) = \sqrt{x}$, we obtain

$$\begin{aligned} n^{1/2}\left(S_n - \sigma\right) &\xrightarrow{d} \mathcal{N}\left(0, g'\left(\sigma^2\right)^2\left(\kappa - \sigma^4\right)\right) \\ &\overset{d}{=} \mathcal{N}\left(0, g'\left(\sigma^2\right)^2\left(\kappa - \sigma^4\right)\right) \\ &\overset{d}{=} \mathcal{N}\left(0, \frac{\kappa - \sigma^4}{4\sigma^2}\right) \end{aligned}$$

unless $\sigma^2 = 0$, (though this corresponds to a constant random variable).

Example: Sample Standard Deviation

- What to do if $\kappa^4 - \sigma^4 = 0$? Note that

$$n\left(S_n^2 - \sigma^2\right) \xrightarrow{d} -\sigma^2\chi_1^2,$$

which is a non-normal limiting distribution. We could apply the delta method again.. But instead, note

$$n\left(S_n - \sigma\right) = \frac{n\left(S_n^2 - \sigma^2\right)}{S_n + \sigma} \xrightarrow{d} \frac{1}{2\sigma} \times \left(-\sigma^2\chi_1^2\right)$$

by Slutsky's Theorem. Therefore, we obtain

$$n\left(S_n - \sigma\right) \xrightarrow{d} -\frac{\sigma}{2}\chi_1^2.$$

Questions?

Estimation: Introduction

- Suppose there exists a population distribution with CDF $F \in \mathcal{F} = \{F_s : s \in \mathcal{S}\}$, where $\mathcal{S}$ is an index set.
  - $\mathcal{F}$ is called an indexed family of distributions.
- We want to learn some feature $\theta(F)$ of the distribution, e.g.

$$\theta_1^k(F) = \int_{\mathbb{R}} x^k \mathrm{d}F(x),$$
$$\theta_2(F) = \inf\left\{x : F(x) \geq \frac{1}{2}\right\},$$
$$\theta_3(F) = F,$$

- If F is a multivariate distribution, and $(Y, X_1, \ldots, X_k) \sim F$, may want to estimate conditional mean:

$$\theta_4(F) = \mathrm{E}(Y|X_1, \ldots, X_k).$$

Estimation: Classes of Distributions

- When the index set $\mathcal{S}$ is a subset of $\mathbb{R}^d$, for some $d > 0$, $\mathcal{F}$ is called a <u>parametric class</u>.
- Example: The family of univariate normal distributions is given by

$$\mathcal{F} = \{F_{\mu,\sigma} : \mu \in \mathbb{R}, \sigma > 0\}$$

where

$$F_{\mu,\sigma}(x) = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

is a parametric class represented by (μ, σ).

Estimators

- Given a sample $\{X_i\}_{i=1}^n$ of draws from a distribution F, a statistic is a function T_n mapping observed data to some set V:

$$T_n : (X_1, \ldots, X_n) \to V,$$

  where V could be a subset of $\mathbb{R}^d$or a function space.
- An estimator is a statistic used to learn about some feature of F, $\theta(F)$.
- We have seen various estimators already: Sample mean, Sample correlation coefficient etc. and studied their properties.
- We have also seen an estimator of the feature $\theta(F) = F$, given by the empirical distribution function

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

Point Identification

- Intuitively, a parameter is identified if it can be deduced from the joint distribution of observables.
- Formally, a class of distributions $\{F_s : s \in \mathcal{S}\}$ is identifiable if the map $s \rightarrow F_s$ is one-to-one.
- In other words, $\{F_s : s \in \mathcal{S}\}$ is identifiable if each distinct parameter value produces a different distribution of observables.

- Example: The class of normal distributions is identifiable, since if $(\mu_1, \sigma_1) \neq (\mu_2, \sigma_2)$, then $F_{(\mu_1,\sigma_1)} \neq F_{(\mu_2,\sigma_2)}$.
- We say a parameter is point identified if no other parameter value gives rise to the same joint distribution of observables.

Partial Identification

- We may not always be able to deduce a parameter value from observables, but the distribution may restrict the possible values the parameter can take.
- Example: Wish to learn the population mean wage (or earnings potential), but observations only come from employed individuals.
- Let W_i denote the wage of individual i, and $D_i \in \{0,1\}$ their employment status. We observe W_i iff $D_i = 1$.
- We write:

$$\mathrm{E}\left(W_i\right) = \mathrm{E}\left(W_i | D_i = 1\right) \mathrm{P}\left(D_i = 1\right) + \mathrm{E}\left(W_i | D_i = 0\right) \mathrm{P}\left(D_i = 0\right).$$

- We do not observe draws from the joint distribution F_{WD}. Instead, we observe draws from F_{ZD}, where

$$Z = W\mathbf{1}\left(D = 1\right).$$

Partial Identification

- If we assume nothing more than $W_i \geq 0$, what can we say about $\mathrm{E}(W_i)$? Lower bound

$$\begin{aligned}\mathrm{E}(W_i) = & \underbrace{\mathrm{E}(W_i|D_i=1)}_{F_{Y|D=1}}\underbrace{\mathrm{P}(D_i=1)}_{F_D}+\underbrace{\mathrm{E}(W_i|D_i=0)}_{\textit{unobserved}!}\underbrace{\mathrm{P}(D_i=0)}_{F_D}. \\ & \geq \mathrm{E}(W_i|D_i=1)\,\mathrm{P}(D_i=1)+0.\end{aligned}$$

- What if we restrict the joint distribution of F_{WD} by assuming $\mathrm{E}(W_i|D_i=1) \geq \mathrm{E}(W_i|D_i=0)$? (Monotonicity)
- Then

$$\begin{aligned}\mathrm{E}(W_i) = & \mathrm{E}(W_i|D_i=1)\,\mathrm{P}(D_i=1)+\mathrm{E}(W_i|D_i=0)\,\mathrm{P}(D_i=0) \\ \leq & \mathrm{E}(W_i|D_i=1)\end{aligned}$$

so

$$\mathrm{E}(W_i|D_i=1)\,\mathrm{P}(D_i=1) \leq \mathrm{E}(W_i) \leq \mathrm{E}(W_i|D_i=1).$$

Partial Identification

- Still haven't point identified $\mathrm{E}\left(W_i\right)$, but now we have upper and lower bounds.
- What if we restrict F_{WD} even further?
  $\mathrm{E}\left(W_i|D_i=1\right)=\mathrm{E}\left(W_i|D_i=0\right)$ (Mean independence):

$$\begin{aligned}\mathrm{E}\left(W_i\right) &= \mathrm{E}\left(W_i|D_i=1\right)\mathrm{P}\left(D_i=1\right)+\mathrm{E}\left(W_i|D_i=0\right)\mathrm{P}\left(D_i=0\right)\\ &= \mathrm{E}\left(W_i|D_i=1\right)\mathrm{P}\left(D_i=1\right)+\mathrm{E}\left(W_i|D_i=1\right)\mathrm{P}\left(D_i=0\right)\\ &= \mathrm{E}\left(W_i|D_i=1\right).\end{aligned}$$

- The mean wage in the population is point identified if we assume wages are mean independent of employment status.
- Stronger assumptions on joint distribution of (un)observables lead to sharper conclusions, but are less credible.

Sample Analogue Principle

- Let $\hat{F}_n$ denote the empirical distribution of data $\{X_i\}_{i=1}^n$, drawn from F:
$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \leq t).$$
- The sample analogue principle suggests we estimate $\theta(F)$ using $\theta\left(\hat{F}_n\right)$, so
$$\hat{\theta}_n = \theta\left(\hat{F}_n\right).$$

Sample Analogue Principle

- Let $\{X_i\}_{i=1}^n$ be iid draws and suppose we want to estimate $\theta(F) = \int_{\mathbb{R}} x^k \mathrm{d}F(x)$.
- Sample analogue principle suggests

$$\hat{\theta}_n = \int_{\mathbb{R}} x^k \mathrm{d}\hat{F}_n(x).$$

- Since $\hat{F}_n(x)$ is a step function, this integral gives uniform weight to each sample point:

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} X_i^k.$$

Sample Analogue Principle

- Variance represented by

$$\sigma^2 = \int_{\mathbb{R}} \left[x - \int_{\mathbb{R}} x \mathrm{d}F(x) \right]^2 \mathrm{d}F(x).$$

- Sample analogue estimator given by

$$\begin{aligned} \hat{\theta}_n &= \int_{\mathbb{R}} \left[x - \int_{\mathbb{R}} x \mathrm{d}\hat{F}_n(x) \right]^2 \mathrm{d}\hat{F}_n(x) \\ &= \int_{\mathbb{R}} \left[x - \bar{X}_n \right]^2 \mathrm{d}\hat{F}_n(x) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}_n \right)^2. \end{aligned}$$

Sample Analogue Principle

- Sample analogue principle often works well, but fails with conditional mean/density estimation.
- Recall

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

  is a step function for every n.
- The density $\theta(F) = f$ is given by $f(x) = F'(x)$, but

$$\hat{F}_n'(x) = 0$$

  almost everywhere, except at the points of discontinuity, where the derivative is undefined.
- Without restricting the estimator to be a density we end up placing point masses at each of the sample points.

Sample Analogue Principle

- To avoid this problem, we need to reduce the class of possible estimators somehow to prevent the result being too specific to the particular sample we draw, while being general enough to reflect the true density in a large sample.

Example: Best Linear Predictor

- Suppose (Y, X) have joint distribution F_{YX} and $\theta(F) = \mathrm{E}(Y|X = x)$.
- In Lecture 2, we showed that $\mathrm{E}(Y|X = x)$ solves

$$g^*(x) = \min_{g \in L^2(X)} \mathrm{E}(Y - g(X))^2$$

where $L^2(X) = \left\{ f : \mathrm{E}\left[f(X)^2 \right] < \infty \right\}$.

- We can restrict $L^2(X)$ to a smaller subset and ask for the best linear predictor of Y given X, which solves

$$b^* = \min_{b \in \mathbb{R}^N} \mathrm{E}\left(Y - X'b \right)^2 .$$

Example: Best Linear Predictor

- Showed, provided $\mathrm{E}(XX')$ is full rank, that:

$$b = \mathrm{E}\left(XX'\right)^{-1}\mathrm{E}\left(XY\right).$$

- Let $\hat{F}_{YX}$ be the empirical distribution of (Y, X). The sample analog minimization problem is:

$$\min_{b \in \mathbb{R}^N} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_i' b\right)^2.$$

- Unsurprisingly, the minimizer is the OLS estimator:

$$\hat{b} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n} X_i Y_i.$$

Questions?

Properties of Estimators: Bias

- If $\{X_i\}_{i=1}^n$ is an iid sample from F with mean μ, the sample mean is an unbiased estimate of μ:

$$\mathrm{E}\left(\bar{X}_n\right) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left(X_i\right) = \mu.$$

- If $\{X_i\}_{i=1}^n$ is an iid sample from F with mean μ and variance σ^2, $S_n^2 = \frac{1}{n}\sum\left(X_i - \bar{X}_n\right)^2$ is $\underline{\text{biased downward}}$.
- Intuitively, this happens because $\bar{X}_n$ solves

$$\min_{a\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} \left(X_i - a\right)^2 \leq \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \mu\right)^2,$$

where the latter is an unbiased (though infeasible) estimator of σ^2.

Properties of Estimators: Bias

- To show this, recall

$$S_n^2 - \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left[(X_i-\mu)^2-\sigma^2\right] - \left(\bar{X}_n-\mu\right)^2.$$

- Taking expectations yields

$$\begin{aligned}\mathrm{E}\left(S_n^2-\sigma^2\right) &= -\mathrm{E}\left[\left(\bar{X}_n-\mu\right)^2\right] \\ &= -\frac{\mathrm{E}\left(\sum_{i=1}^{n}(X_i-\mu)^2+\sum\sum_{j\neq i}(X_i-\mu)(X_j-\mu)\right)}{n^2} \\ &= -\frac{\sigma^2}{n}.\end{aligned}$$

- Therefore, $\mathrm{E}\left(S_n^2\right)=\frac{n-1}{n}\sigma^2$. An unbiased estimator of σ^2 is

$$\tilde{S}_n^2 = \frac{1}{n-1}\sum\left(X_i-\bar{X}_n\right)^2.$$

Properties of Estimators: Variance

- Low variance and low bias are desirable so that the estimator concentrates most of its probability mass around the true parameter value.
- $\tilde{S}_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$ is unbiased, so

$$Var\left(\tilde{S}_n^2\right) = \mathrm{E}\left(\left(\tilde{S}_n^2 - \sigma^2\right)^2\right).$$

- Some algebra gives:

$$Var\left(\tilde{S}_n^2\right) = \frac{\mathrm{E}\left(X_i - \mu\right)^4}{n} - \frac{\sigma^4\left(n-3\right)}{n\left(n-1\right)}.$$

- If $X_i \sim iid\mathcal{N}\left(\mu, \sigma^2\right)$:

$$Var\left(\tilde{S}_n^2\right) = \frac{2\sigma^4}{n-1}.$$

Properties of Estimators: Variance

- "Low" variance is relative: It depends on the class of estimators we use as a comparison.
- For example, we may ask for the minimum variance unbiased estimator, or the minimum variance linear unbiased estimator.
- For an example of the latter, let $\{X_i\}_{i=1}^n$ be iid with mean $\mu \neq 0$ and variance σ^2.
- The set of linear unbiased estimators of μ is the class of all estimators

  $$\hat{\mu} : (X_1, \ldots, X_n) \to \mathbb{R}$$

  such that $\hat{\mu} = \sum_{i=1}^n a_i X_i$ for some $\{a_i\}_{i=1}^n \in \mathbb{R}^n$ and $\mathrm{E}(\hat{\mu}) = \mu$.

Properties of Estimators: Variance

- The condition $\mathrm{E}\left(\hat{\mu}\right) = \mu$ implies $\sum_{i=1}^{n} a_i \mu = \mu$, so $\sum_{i=1}^{n} a_i = 1$.
- The variance of $\hat{\mu}$ is given by

$$Var\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} Var\left(a_i X_i\right) = \sum_{i=1}^{n} a_i^2 Var\left(X_i\right) = \sigma^2 \sum_{i=1}^{n} a_i^2 .$$

- Minimizing this objective over $\{a_i\}_{i=1}^{n} \in \mathbb{R}^n$ subject to $\sum_{i=1}^{n} a_i = 1$ gives $a_i = \frac{1}{n}$ for all n.
- Therefore $\bar{X}_n$ is the best linear unbiased estimator of μ.

Variance vs. Bias

- The mean squared error of a scalar estimator $\hat{\theta}_n$ is

$$\begin{aligned} MSE\left(\hat{\theta}_n\right) &= \mathrm{E}\left[\left(\hat{\theta}_n - \theta\right)^2\right] \\ &= Var\left(\hat{\theta}_n\right) + \text{Bias}\left(\hat{\theta}_n\right)^2 \end{aligned}$$

- Minimizing MSE takes into account both the bias and variance of an estimator. There is often a tradeoff between bias and variance when minimizing MSE.
- Accepting some bias can reduce the variance significantly. If $\hat{\theta}_n$ is an unbiased estimator of θ with variance Δ, shrinking $\hat{\theta}_n$ lowers its MSE:

Variance vs. Bias

- Consider the class of estimators $\left\{ k\hat{\theta}_n : k \in \mathbb{R} \right\}$.
- Note that

$$\begin{aligned} MSE\left(k\hat{\theta}_n\right) &= Var\left(k\hat{\theta}_n\right) + \left(\theta - \mathrm{E}\left(k\hat{\theta}_n\right)\right)^2 \\ &= k^2 Var\left(\hat{\theta}_n\right) + (1-k)^2\,\theta^2 \\ &= k^2\left(\Delta + \theta^2\right) - 2k\theta^2 + \theta^2 \end{aligned}$$

- Minimizing wrt. k using strict convexity and the FOC gives

$$k = \frac{\theta^2}{\Delta + \theta^2}.$$

Properties of Estimators: Consistency

- An estimator $\hat{\theta}_n : (X_1, \ldots, X_n) \to \mathbb{R}$ of θ is consistent if

$$\hat{\theta}_n \overset{p}{\to} \theta.$$

- $\hat{\theta}_n$ is strongly consistent if

$$\hat{\theta}_n \overset{a.s.}{\to} \theta.$$

- We have seen $\bar{X}_n \overset{a.s.}{\to} \mu$, $S_n^2 \overset{a.s.}{\to} \sigma^2$, and strong consistency of covariance and correlation estimators.

Properties of Estimators: Consistency

- You were asked to show in PSET 2 that $\mathrm{E}\left(\hat{\theta}_n\right) \to \theta$ and $Var\left(\hat{\theta}_n\right) \to 0$ implies $\hat{\theta}_n \xrightarrow{p} \theta$.
- Holds because $\mathrm{E}\left(\hat{\theta}_n - \theta\right)^2 \to 0$ iff $\mathrm{E}\left(\hat{\theta}_n\right) \to \theta$ and $Var\left(\hat{\theta}_n\right) \to 0$, and convergence in $r-$th mean implies convergence in probability.
- In general, (asymptotically) unbiased estimators are not consistent:

Properties of Estimators: Consistency

- Suppose $\theta = \mathrm{E}(X_i)$. If we ignore all but the first observation of X, and set
$$\hat{\theta}_n(X_1, \ldots, X_n) = X_1$$
for all n, then clearly $\hat{\theta}_n$ is unbiased but not consistent, since
$$X_1 \xrightarrow{p} X_1; \qquad X_1 \overset{p}{\not\to} \mathrm{E}(X_i).$$
- The sufficient condition for consistency fails because $Var\left(\hat{\theta}_n\right) \not\to 0$.

Asymptotic distribution of an estimator

- We think of the asymptotic distribution of an estimator as a non-degenerate distribution.
- Suppose that $\hat{\theta}_n$ is a consistent estimate of θ. If for some sequence $r > 0$ and non-degenerate random variable X,

  $$n^r \left(\hat{\theta}_n - \theta \right) \xrightarrow{d} X,$$

  we say that F_X is the asymptotic distribution of $\hat{\theta}_n$, where F_X is the distribution function of X.
- Often the CLT can be applied and the asymptotic distribution is normal. If

  $$\sqrt{n} \left(\hat{\theta}_n - \theta \right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma\right),$$

  we say $\hat{\theta}_n$ is asymptotically normal with asymptotic variance covariance matrix Σ.

Questions?

Method of Moments

- To illustrate these concepts we will study MoM and ML estimators.
- Suppose we wish to estimate $\theta \in \mathbb{R}^d$ that solves the equation

$$g(\theta) = \mathrm{E}(h(X)).$$

- $g : \mathbb{R}^d \to \mathbb{R}^K$ and h are known functions.
- Given a sample $\{X_i\}_{i=1}^n$, we apply the sample analogue principle

$$g\left(\hat{\theta}_n\right) = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$$

- A solution to this equation is a <u>method of moments estimator</u>.

Method of Moments: Example

- Let $\{X_i\}_{i\geq 1}$ be iid draws from $U[0,\theta]$.
- To find the functions g and h it helps to write the moments of X in terms of θ:

$$\begin{aligned}
\mathrm{E}(X) &= \int_0^\theta \frac{x}{\theta}\mathrm{d}x = \left[\frac{x^2}{2\theta}\right]_0^\theta = \frac{\theta}{2} \\
\mathrm{E}\left(X^2\right) &= \int_0^\theta \frac{x^2}{\theta}\mathrm{d}x = \left[\frac{x^3}{3\theta}\right]_0^\theta = \frac{\theta^2}{3} \\
&\vdots
\end{aligned}$$

- In this case, the first moment of X will suffice, but we can pick any of them.

Method of Moments: Example

- Let $g(\theta) = \frac{\theta}{2}$ and $h(x) = x$. We see that

$$g(\theta) = \mathrm{E}(h(X)).$$

- Next we use the sample analogue principle:

$$\frac{\hat{\theta}_{1n}}{2} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

which gives

$$\hat{\theta}_{1n} = 2\bar{X}_n.$$

Method of Moments: Example

- We could also have chosen $g(\theta) = \frac{\theta^2}{3}$ and $h(x) = x^2$, yielding

$$\frac{\hat{\theta}_{2n}^2}{3} = \frac{1}{n} \sum_{i=1}^{n} X_i^2,$$

  or

$$\hat{\theta}_{2n} = \sqrt{\frac{3}{n} \sum_{i=1}^{n} X_i^2}.$$

- Both $\hat{\theta}_{1n}, \hat{\theta}_{2n}$ are consistent, which can be shown with the SLLN and CMT.
- However, while $\hat{\theta}_{1n}$ is unbiased, $\hat{\theta}_{2n}$ is biased downward (by Jensen's inequality).

Bias

- The bias of $\hat{\theta}_{2n}$ us given by Bias $\left(\hat{\theta}_{2n}\right) = \mathrm{E}\left(\hat{\theta}_{2n}\right) - \theta$.
- By Jensen's inequality, since $g(x) = \sqrt{x}$ is a strictly concave function:

$$\mathrm{E}\left(\hat{\theta}_{2n}\right) = \mathrm{E}\left(\sqrt{\frac{3}{n}\sum_{i=1}^{n} X_i^2}\right) < \sqrt{\mathrm{E}\left[\frac{3}{n}\sum_{i=1}^{n} X_i^2\right]} = \sqrt{3 \cdot \frac{\theta^2}{3}} = \theta,$$

so $\hat{\theta}_{2n}$ is <u>biased downward</u>.

Consistency

- Recall: An estimator $\hat{\theta}_n : (X_1, \ldots, X_n) \to \mathbb{R}$ of θ is consistent if

$$\hat{\theta}_n \xrightarrow{p} \theta.$$

- Step 1: See that $\hat{\theta}_{2n}$ contains a sample average (we can apply SLLN!):

$$\hat{\theta}_{2n} = \sqrt{\frac{3}{n} \sum_{i=1}^{n} X_i^2}.$$

Conclude that

$$\frac{1}{n} \sum_{i=1}^{n} X_i^2 \xrightarrow{a.s.} \mathrm{E}\left(X^2\right) = \frac{\theta^2}{3}.$$

Consistency

- Step 2: Use the continuous mapping theorem. The condition we have to check is that $g(x) = \sqrt{3x}$ is continuous at the limit $\frac{\theta^2}{3}$ with probability 1. Since the limit random variable is a constant, and $\theta > 0$, this is satisfied, so

$$g\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) \overset{a.s.}{\to} g\left(\frac{\theta^2}{3}\right) = \theta.$$

- Since $\overset{a.s.}{\to}$ implies $\overset{p}{\to}$, we conclude that $\hat{\theta}_{2n}$ is a consistent estimator of λ.

Asymptotic distribution

- We are searching for constants $r \geq 0$ and c such that

$$n^r \left(\hat{\theta}_{2n} - c\right) \xrightarrow{d} Y$$

  for some non-degenerate Y.

- Step 1: Check $r = 0$: We previously established $\hat{\theta}_{2n} \xrightarrow{p} \theta$. If $r = 0$,

$$\hat{\theta}_{2n} - \theta \xrightarrow{p} 0,$$

  which is degenerate.

Example: Asymptotic distribution

- Step 2: Find c given that $r > 0$. In PSET 3 we showed that if

$$n^r \left(\hat{\theta}_{2n} - c\right) \xrightarrow{d} Y$$

  then $\hat{\theta}_{2n} \xrightarrow{p} \theta$. So, if there exist constants $r > 0$ and c such that this holds, $c = \theta$.

- Step 3: Find $r > 0$ and Y. Looking for $r > 0$ such that

$$n^r \left(\hat{\theta}_{2n} - \theta\right) \xrightarrow{d} Y.$$

Example: Asymptotic distribution

- Step 3a: How did we establish consistency? Was there a sample average? (Yes! Use the CLT):
- Note that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{\theta^2}{3}\right) \xrightarrow{d} \mathcal{N}\left(0, Var\left(X_i^2\right)\right)$$

by the CLT. Can show that

$$Var\left(X_i^2\right) = \mathrm{E}\left(X_i^4\right) - \mathrm{E}\left(X_i^2\right)^2 = \frac{\theta^4}{5} - \left(\frac{\theta^2}{3}\right)^2 = \frac{4\theta^4}{45}.$$

yielding

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{\theta^2}{3}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{4\theta^4}{45}\right).$$

Example: Asymptotic distribution

- Step 3b: Compare CLT result with what we actually want:
- Note that

$$\sqrt{n}\left(\hat{\theta}_{2n}-\theta\right)=\sqrt{n}\left(\sqrt{\frac{3}{n}\sum_{i=1}^{n}X_i^2}-\sqrt{3\cdot\frac{\theta^2}{3}}\right),$$

so we try the Delta Method with $g\left(x\right)=\sqrt{3x}$ and $g'\left(x\right)=\frac{\sqrt{3}}{2\sqrt{x}}$: This yields

$$\sqrt{n}\left(g\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right)-g\left(\frac{\theta^2}{3}\right)\right)\xrightarrow{d}\mathcal{N}\left(0,g'\left(\frac{\theta^2}{3}\right)^2\times\frac{4\theta^4}{45}\right)$$
$$\overset{d}{=}\mathcal{N}\left(0,\frac{\theta^2}{5}\right).$$

Questions?