Econometrics Math Camp

Day 2: Moments, Distributions, Estimators, Asymptotics

Michael Droste

August 2021

Introduction

- Welcome back!

- Materials (slides, notes, and problems) are available on GitHub:
  https://github.com/mdroste/metrics-mathcamp-2021/

- We have three main objectives for the today's half of econometrics math camp:
  1. Review the most important take-aways from Friday
  2. Review some additional material
  3. Tackle two important simulation-based exercises

Today's Outline

- Review from Day 1
- Statistical Models and Estimators
- Asymptotics
  - Convergent Sequences of Random Variables
  - Slutsky's Theorem
  - (A) Law of Large Numbers
  - (A) Central Limit Theorem
- Ordinary Least Squares

Review: Continuous Random Variables

- Let X be a random variable. We say that X is a continuous random variable if (and only if) F_X can be written as:

$$F_X(x) = \int_{-\infty}^{\infty} f_X(t)dt$$

where f_X satisfies $f_X(x) \geq 0$ and $\int_{-\infty}^{\infty} f_X(t)dt = 1$.

- At the points where F_X is continuous, we have:

$$f_X(x) = \frac{dF_x(x)}{dx}$$

- We call $f_X(x)$ the probability density function (or pdf) of X.

- The support of X is $S_X = \{x : f_X(x) > 0\}$

Review: Cumulative Distribution Function

- Let X be a random variable. The cumulative distribution function (or cdf) or X, $F : \mathbb{R} \to [0, 1]$, is defined as:

$$F_X(x) = P(X^{-1}(x)) = P(\{\omega \in \Omega : X(\omega) \leq x\})$$

  We often write the cdf of X as:

$$F_x(x) = P(X \leq x)$$

- The CDF of a random variable X is a function that tells us: for any given value of x, what is the probability that the random variable X takes on a value less than or equal to x?

Review: Quantiles

- The quantiles of a random variable X are closely related to the CDF.

- The quantile function is:

$$Q(u) = \inf \left\{ x : F_X(x) \geq u \right\}$$

If F_X is invertible, then:

$$Q(u) = F_X^{-1}(u)$$

Review: Expectations of Continuous Random Variables

- Let X be a continuous random variable. Its expectation (or expected value) is defined as:

$$E[X] = \int_{S_x} x f_X(x) dx$$

  if $\int_{S_x} |x| f_X(x) dx < \infty$. Otherwise, the expectation does not exist.

- Note that expectations (still) play nicely with transformations of (continuous) random variables. For instance, let $g : \mathbb{R} \to \mathbb{R}$. Then:

$$E[g(X)] = \int_{S_x} g(x) f_X(x) dx$$

Review: Expectations as Linear Operators

- Expectations are a linear operator. What does this mean? Let X be a random variable, $a \in \mathbb{R}$ a constant, and $g_1(\cdot), g_2(\cdot)$ be real-valued functions. Then:
  1. $E[a] = a$
  2. $E[ag_1(X)] = aE[g_1(X)]$
  3. $E[g_1(X) + g_2(X)] = E[g_1(X)] + E[g_2(X)]$

Review: Conditional Expectations

- Let X and Y be random variables with a joint density $f_{X,Y}(x, y)$. The conditional expectation of Y given X=x is:

$$E[Y|X = x] = \int_{S_Y} y f_{Y|X}(y|x) dy$$

- Note that this is a function of x, and is sometimes called the conditional expectation function or regression function. It is sometimes useful to denote the CEF of a variable Y as a function of x as $\mu_Y(x)$.

Review: Law of Iterated Expectations

- The law of iterated expectations is a really, really useful law for manipulating conditional expectations. It will show up all the time in your homework in a variety of settings.

- One form of the law of iterated expectations can be stated as:

$$E_Y[Y] = E_X E_{Y|X}[Y]$$

  where E_X denotes the expectation taken with respect to the marginal density of X and $E_{Y|X}$ denotes the expectation taken with respect to the conditional density of Y given X.

- The way I think about this rule is as follows. Suppose I have the expectation of the conditional density of Y given X. If I take this expression and apply an expectation operator with respect to X, then I am left with the expectation of the marginal density of Y.

Review: Law of Total Probability

- Consider K disjoint events C_k that partition the sample space Ω; that is, $C_i \cap C_j = \emptyset$ for all $i \neq j$ and $\cup_{i=1}^{K} C_i = \Omega$. Let A be some event.

- The law of total probability states that we can write $P(A)$ in terms of $P(A|C_i)$ and $P(C_i)$ in a way that 'adds up'.

$$P(A) = \sum_{i=1}^{K} P(A|C_i)P(C_i)$$

Review: Bayes' Rule

- Given two events A, B, Bayes' rule (sometimes seen as Bayes' law) relates the conditional probabilities $P(A|B)$, $P(B|A)$ and the marginal probabilities $P(A)$, $P(B)$.

- One simple formulation of Bayes law can be expressed as:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Break-Out Session #3

- Let's warm up our brains with one quick pencil-and-paper problem. Please work with your classmates to prove the law of total variance. See Worksheet #3 on GitHub.

- We will reconvene in approx. 30 mins to discuss.

Statistical Models

- You might notice that any discussion of a statistical model, an estimator, or even a dataset has been conspicuously absent for metrics math camp. Let's fix that.

- Let our data be $D = (D_1, ..., D_n)$. Each D_i consists of a $1 \times K$ vector of variables.

- A statistical model is a set of assumptions regarding the joint distribution of the data. Most generally, a model can be thought of as saying that $D \sim F(\theta)$, for some distribution F and parameter θ. We sometimes write the model as $\{F(\theta) : \theta \in \Theta\}$, where Θ is the "parameter space".

- Note that θ can be a single parameter, a function, a distribution, or all of these. If θ is finite-dimensional, our model is said to be parametric. Otherwise, it is non-parametric.

Estimators

- OLS is one particular example of an estimator, which is simply a function from the data to the parameters of interest in your statistical model.

- When the parameter is θ, we often represent an estimator as $\hat{\theta}$.

- It is useful to characterize estimators by their properties. Estimators can be categorized in many ways, but you will spend a lot of the first year talking about consistency, efficiency, unbiasedness, and asymptotic normality.

Properties of Estimators

- An estimator $\hat{\theta}$ is consistent if it converges in probability to the true parameter θ as the sample size N (or some other relevant index) grows large. Consistent estimators are sometimes thought of as "asymptotically unbiased".

- An estimator is asymptotically normal if the distribution of $\hat{\theta}$ converges in distribution to a normal distribution with standard deviation proportional to $1/\sqrt{n}$ as the sample grows large. What does it mean for an estimator to have a distribution? We'll explore this in the second half of today's class!

- Much more on this to come in 2120 and 2140! But what are these terms, "converges in probability" and "converges in distribution"? We need some way to think about what it means for sequences of random variables to converge.

Asymptotics

- How do we say something about the behavior of our estimators without strong parametric assumptions (e.g. assuming a distribution for the errors)?

- Answer: Use limiting behavior of estimators in large samples, where the behavior of most estimators becomes much simpler, thanks to a set of really powerful results (like the law of large numbers and the central limit theorem).

- Of course, the asymptotic (limiting) behavior of an estimator as the sample grows infinitely large is only an approximation for the estimator's behavior in finite samples - and potentially a poor approximation.

Stochastic Convergence

- The idea behind a lot of asymptotics is to consider “sequences” of estimators, usually indexed by the population size (N). We need to extend the idea of converging sequences of real numbers to random variables.

- It will be helpful to start by remembering the idea of convergence for a non-stochastic sequence of real numbers. Let $\{x_n\}$ be a sequence of real numbers. We say:

$$\lim_{n \to \infty} x_n = x$$

  if for all $\epsilon > 0$, there exists some N such that for all $n > N$, $|x_n - x| < \epsilon$.

- We will now think about several different ways to generalize this definition to sequences of random variables.

Almost Sure Convergence

- Let $\{X_n\}$ denote a sequence of random variables. We say that the sequence $\{X_n\}$ converges to the random variable X almost surely if:

$$P(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1$$

- This definition is a little bit unwieldy, so to make life simpler, we often just write:

$$X_n \xrightarrow{\text{a.s.}} X$$

Almost Sure Convergence: In English

- For a given outcome ω in the sample space Ω, we can ask whether:

$$\lim_{n \to \infty} X_n(\omega) = X(\omega)$$

  holds using the definition of non-stochastic convergence.

- If the set of outcomes for which this holds has probability = 1, then:

$$X_n \xrightarrow{\text{a.s.}} X$$

Convergence in Probability

- Let $\{X_n\}$ be a sequence of random variables and X be a random variable. The sequence of random variables $\{X_n\}$ converges in probability to the random variable X if for all $\epsilon > 0$,

$$\lim_{n \to \infty} P(|X_n - X| > \epsilon) \to 0$$

- As with almost sure convergence, it is helpful and very common to express this in the shorthand:

$$X_n \xrightarrow{\text{p}} X$$

Convergence in Probability: In English

- Fix some $\epsilon > 0$. We can then compute:

$$P_n(\epsilon) = P(|X_n - X| > \epsilon)$$

- This is just a number, o we can check whether $P_n(\epsilon) \to 0$ using the standard definition of non-stochastic convergence.

- If $P_n(\epsilon) \to 0$ for all values $\epsilon > 0$, then we write $X_n \xrightarrow{\text{p}} X$

Convergence in Distribution

- Let $\{X_n\}$ be a sequence of random variables and $F_n(\cdot)$ be the cdf of X_n. Let X be a random variable with cdf $F(\cdot)$. The sequence $\{X_n\}$ converges in distribution (or weakly converges, or converges in law) to X if:

$$\lim_{n\to\infty} F_n(x) = F(x)$$

  for all points x at which $F(x)$ is continuous.

- We often write $X_n \xrightarrow{\text{d}} X$.

Convergence in Distribution: In English

- Convergence in distribution describes the convergence of cdfs of variables. It does not mean that individual realizations of variables get close to each other.

- Recall that the cdf of a variable is defined as:

$$F(x) = P(X \leq x) = P(\{\omega \in \Omega : X(\omega) \leq x\})$$

- As a result, $F_n(x) \to F(x)$ does not inform us about whether $X_n(\omega)$ is getting close to $X(\omega)$ for any $\omega \in \Omega$

Multivariate Convergence

- We can extend each of these definitions to random vectors very easily.
- We say that a sequence of random vectors $\{X_n\}$ converges almost surely to X if each element of X_n converges almost surely to X. Likewise, the same element-wise extension applies for convergence in probability.
- A sequence of random vectors converges in distribution to a random vector if we apply the definition above to the joint cdf.
- In a full-semester statistics course, you would likely spend time being more formal about why this is the case. Essentially, the fact that our random variables are defined on a metric space is crucial. Distance metrics often generalize to n dimensions quite easily.

Relationship Between Types of Convergence

- We discussed three types of convergence for sequences of random variables: almost sure convergence, convergence in probability, and convergence in distribution.

- It turns out that the ordering of these topics was not accidental, and we moved from 'strong' notions of convergence to 'weak' notions.

- In particular, almost sure convergence implies convergence in probability, and convergence in probability implies convergence in distribution.

- The reverse direction does not hold: convergence in distribution does not imply convergence in probability.

Asymptotic Tools

- Now that we have defined three different types of convergence, we can think about four distinct theorems that are going to come up repeatedly when we talk about asymptotics.

- These theorems are statements involving the types of convergence we just discussed, and in practice, using them means that we (often) don't have to explicitly show that a sequence converges using the definition of convergence.

- These tools are Slutsky's Theorem; the Continuous Mapping Theorem; the Law of Large Numbers; and the Central Limit Theorem. All four of these are worth memorizing.

Slutsky's Theorem

- Let c be a constant, X_n and Y_n denote sequences of random variables indexed by n, and X and Y denote random variables. Suppose that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$. Then:
  1. $X_n + Y_n \xrightarrow{d} X + c$
  2. $X_n Y_n \xrightarrow{d} Xc$
  3. $X_n / Y_n \xrightarrow{d} X/c$ if $c \neq 0$.

- Slutsky's theorem is useful because it gives us extra 'tools' for establishing the convergence of sequences of random variables defined by adding, multiplying, or dividing sequences of random variables.

Continuous Mapping Theorem

- Let g be a continuous function. Then:
  1. If $X_n \xrightarrow{\text{d}} X$, then $g(X_n) \xrightarrow{\text{d}} g(X)$.
  2. If $X_n \xrightarrow{\text{p}} X$, then $g(X_n) \xrightarrow{\text{p}} g(X)$

- The continuous mapping theorem is easy to remember. Intuitively, it tells us that convergence in probability and distribution are preserved over a continuous transformation g.

The Law(s) of Large Numbers

- This is a big one! The law of large numbers (there are actually several different flavors) provides conditions under which sample averages converge to expectations.

- One form is the weak law of large numbers (weak LLN or WLLN for short). Let $X_1, ..., X_n$ be a sequence of random variables with $E[X_i] = \mu$, $Var[X_i] = \sigma^2 < \infty$, $Cov(X_i, X_j) = 0$ for all $i \neq j$. Then:

$$\bar{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{p}} \mu$$

- Another form is the strong law of large numbers (strong LLN or SLLN for short). Let $X_1, X_2, ...$ be i.i.d with $E[X_i] = \mu < \infty$. Then:

$$\bar{X_n} \xrightarrow{\text{a.s.}} \mu$$

The Central Limit Theorem(s)

- Another important class of theorems for asymptotics are central limit theorems, which provide conditions under which "properly-centered" sample averages converge in distribution to normal random variables.

- There are several different flavors, but I'll discuss the most useful one for you.

- Let $X_1, X_2, \ldots$ be an i.i.d. sequence of random variables with mean μ and variance σ^2. Then:

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\text{d}} N(0, \sigma^2)$$

- This generalizes to random vectors, where you will often see this applied. If $X_1, X_2, \ldots$ are random vectors with mean vector μ and covariance matrix Σ, then we have:

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

Motivating the Least Squares Estimator

- Now, let's switch gears and talk about one particular (class of) estimator(s) - the ordinary least squares estimator.

- OLS can be motivated in multiple ways. In a few weeks, you'll walk through Gary Chamberlain's view of ordinary least squares through the lens of an inner product space and the projection theorem.

- I want to give a simpler, more familiar presentation so that we can work with this estimator in our breakout problems today.

Ordinary Least Squares: Model

- We observe data $(X_1, ..., X_N)$ and $(Y_1, ..., Y_N)$, where X_i is a $1 \times k$ (row) vector and each Y_i is a scalar.

- We are interested in estimating the model:

$$Y = X\hat{\beta} + e$$

- We want to choose $\hat{\beta}$ such that the sum of squared residuals, $e'e$, is minimized.

Ordinary Least Squares: Solution

- We can write the sum of squared residuals:

$$e'e = (X\hat{\beta} - Y)'(X\hat{\beta} - Y) = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

- The first order condition $\partial e'e/\partial\hat{\beta}$ yields $-2X'Y + 2X'X\hat{\beta} = 0$. Solving for $\hat{\beta}$ yields the OLS estimator in matrix form:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

- Note that estimates are functions of the data (random variables); therefore, estimates are random variables.

- We have been careful with our notation to write that our OLS estimates $\hat{\beta}$ are not β; that will be the subject of next week's final math camp session on statistical inference.

Break-Out Session #4

- This concludes today's new materials. Tomorrow, we'll consider statistical inference.

- We'll now move to break-out rooms once again to consider two problems that you should solve with computers.