Section 5: Conditional expectations and convergence

ARE 210

September 26, 2017

The section notes are available on the section Github at github.com/johnloeser/are210 in the "section5" folder.

1 Definitions

- **Conditional expectation**: Let $(\Omega, \mathbf{F}, P)$ be a probability space, $Y : \Omega \to \mathbf{R}$ a random variable, and $\mathbf{G} \subset \mathbf{F}$ a σ-algebra.[1] Then $\mathbf{E}[Y|\mathbf{G}]$ is 1) $\mathbf{G}$-measurable, and 2) solves $\int Y(\omega)\mathbf{1}_B(\omega)P[d\omega] = \int \mathbf{E}[Y|\mathbf{G}](\omega)\mathbf{1}_B(\omega)P[d\omega]$ for all $B \in \mathbf{G}$.
  - $\mathbf{E}[Y|\mathbf{G}]$ is $\mathbf{G}$-measurable $\Leftrightarrow \mathbf{E}[Y|\mathbf{G}]^{-1}(B) \in \mathbf{G}$ for all $B \in \mathbf{B}$
    - More intuitively, this will mean that $\mathbf{E}[Y|\mathbf{G}]$ will be constant over the smallest non-empty sets in $\mathbf{G}$.
    - For example, $\mathbf{E}[Y] = \mathbf{E}[Y|\{\emptyset, \Omega\}]$, and $Y = \mathbf{E}[Y|\mathbf{F}]$
  - Let $\sigma(X) = \{A \in \mathbf{F} : \exists\ B \in \mathbf{B}$ such that $A = X^{-1}(B)\}$ for some random variable X, then $\mathbf{E}[Y|X] \equiv \mathbf{E}[Y|\sigma(X)]$.
    - More intuitively, $\mathbf{E}[Y|X]$ is taking the average of Y over the set of ω that correspond to any one value of X
  - If W is $\mathbf{G}$-measurable, $\mathbf{E}[YW|\mathbf{G}] = W\mathbf{E}[Y|\mathbf{G}]$
    - More intuitively and equivalently, $\mathbf{E}[Yf(X)|X] = f(X)\mathbf{E}[Y|X]$
  - **Law of iterated expectations**: $\mathbf{E}[\mathbf{E}[Y|\mathbf{G}]] = \mathbf{E}[Y] = \mathbf{E}[\mathbf{E}[Y]|\mathbf{G}]$
  - $X \perp Y|Z \Leftrightarrow \mathbf{E}[f(Y)|X, Z] = \mathbf{E}[f(Y)|Z]$ and $\mathbf{E}[f(X)|Y, Z] = \mathbf{E}[f(X)|Z]$.
  - Define the **potential outcomes framework** to be 1) potential outcomes (Y_1, Y_0), 2) treatment status D, 3) observed outcome $Y = Y_1 D + Y_0(1 - D)$, where (Y_1, Y_0, D, Y) is a random variable
    - $(Y_1, Y_0) \perp D$ is random assignment of D
    - Let X be another random variable, then $(Y_1, Y_0) \perp D|X$ is the selection on observables assumption
    - $P[D = 1|X]$ is the propensity score
      - Selection on observables $\Rightarrow (Y_1, Y_0) \perp D|P[D = 1|X]$

[1] We also make the technical assumption that $\mathbf{E}[|Y|] < \infty$.

- $\cdot$ Selection on observables $\Rightarrow \mathbf{E}\left[\frac{DY}{P[D=1|X]}\middle| X\right] = \mathbf{E}[Y_1|X]$
- Let $L^p(\mathbf{G}) \equiv \{X : \Omega \to \mathbf{R} \mid X \text{ is } \mathbf{G}\text{-measurable}, \int |X(\omega)|^p P[d\omega] < \infty\}$, and assume $Y \in L^2(\mathbf{F})$.
  - $\widehat{Y} = \mathbf{E}[Y|\mathbf{G}]$ is the unique (almost everywhere) random variable satisfying $\mathbf{E}[(Y-\widehat{Y})^2] = \min_{X \in L^2(\mathbf{G})} \mathbf{E}[(Y-X)^2]$, or equivalently, $\mathbf{E}[(Y-\widehat{Y})X] = 0 \; \forall \; X \in L^2(\mathbf{G})$.
  - More intuitively, the conditional expectation function "projects" Y onto X.
  - $Y = \widehat{Y} + (Y - \widehat{Y})$ is a useful decomposition of Y into two mean independent random variables

- **Convergence**: Let $(\Omega, \mathbf{F}, P)$ be a probability space, $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables
  - $X_n \xrightarrow{\text{a.s.}} X \Leftrightarrow X_n(\omega) \to X(\omega)$ almost everywhere
  - Assume $X_n \in L^p$.[2] $X_n \xrightarrow{L^p} X \Leftrightarrow \mathbf{E}[|X_n - X|^p] \to 0$
  - $X_n \xrightarrow{\text{p}} X \Leftrightarrow P[|X_n(\omega) - X(\omega)| > \epsilon] \to 0$
    - $X_n \to X$, $Y_n \to Y$ implies 1) $X_n + Y_n \xrightarrow{\text{p}} X + Y$, 2) $X_nY_n \xrightarrow{\text{p}} XY$, 3) $f(X_n) \to f(X)$ for sufficiently nice f
    - **Weak law of large numbers**: $\{X_n\}_{n=1}^{\infty}$ iid random variables, let $\overline{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i$, and assume $X \in L^2$. Then $\overline{X}_n \xrightarrow{\text{p}} \mathbf{E}[X_i]$.
    - $\xrightarrow{\text{a.s.}} \Rightarrow \xrightarrow{L^p} \Rightarrow \xrightarrow{\text{p}}$
  - Let $\{\mu_n\}_{n=1}^{\infty}$ be a sequence of probability measures, and assume X_i has probability measure μ_i, and let μ be the probability measure of X. Then $X_n \xrightarrow{\text{d}} X \Leftrightarrow \mu_n \to \mu \Leftrightarrow \int f(x)\mu_n[dx] \to \int f(x)\mu[dx]$ for all bounded, continuous f
    - Probability measures converging is equivalent to CDFs converging which is equivalent to characteristic functions converging (Levy's Theorem)
    - $X_n \xrightarrow{\text{d}} X \Leftrightarrow r'X_n \xrightarrow{\text{d}} r'X$ for any matrix r (joint distributions being the same is equivalent to arbitrary linear combinations having the same distribution)
    - **Lindeberg-Levy CLT**: $\{X_i\}_{i=1}^{\infty}$ iid random variables, $X_i \in L^2$, $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, $Y_n = \frac{\overline{X}_n - \mathbf{E}[\overline{X}_n]}{\sqrt{V[\overline{X}_n]}}$, then $Y_n \xrightarrow{\text{d}} N(0,1)$.

[2] Using the definition above, here $L^p \equiv L^p(\mathbf{F})$.

2 Some helpful tips

- The potential outcomes framework is one of the most important recent developments in econometrics, if not the most. Generally, think about D as a proxy for some policy, and Y as some outcome of interest, where we're interested in estimating the effect of the policy on that outcome.
- Try constructing an example that matters to anything we do of $\xrightarrow{\text{p}}$ but not $\xrightarrow{\text{a.s.}}$ ...
- Chebyshev's inequality ($P[|X - \mathbf{E}[X]| > \epsilon] \leq \frac{V[X]}{\epsilon^2}$) got a lot of use here, almost always to show a sequence of random variables with variance going to 0 is converging in probability to the limit of their expectation

3 Practice questions

1) **Sum of Bernoulli**: Let $X_{n1}, \dots, X_{nn}$ be iid Bern$[p_n]$, and let $S_n = \sum_{j=1}^{n} X_{nj}$. Suppose $np_n \to \lambda$. Show that $S_n \xrightarrow{\text{d}}$ Pois$[\lambda]$.

Hint: the characteristic function of a Pois$[\lambda]$ RV is $\Psi(t) = \exp(\lambda(\exp(it) - 1))$.

2) **Law of iterated expectations**: Show $\mathbf{E}[\mathbf{E}[Y|X]] = \mathbf{E}[Y]$.

3) **LATE**: Let (Y_1, Y_0, D_1, D_0, Z) be a vector of random variables, where D_1, D_0, and Z are binary. Assume $(Y_1, Y_0, D_1, D_0) \perp Z$ (independence), $\mathbf{E}[D_1] > \mathbf{E}[D_0]$ (first stage), and $D_1 \geq D_0$ almost everywhere (monotonicity). Assume $D = D_1 Z + D_0(1 - Z)$, $Y = Y_1 D + Y_0(1 - D)$, and Z are observed.

a) What is $\mathbf{E}[Y|D = 1] - \mathbf{E}[Y|D = 0]$? What are necessary and sufficient conditions for it to equal $\mathbf{E}[Y_1 - Y_0|D = 1]$?

b) Show that $\frac{\mathbf{E}[Y|Z=1]-\mathbf{E}[Y|Z=0]}{\mathbf{E}[D|Z=1]-\mathbf{E}[D|Z=0]} = \mathbf{E}[Y_1 - Y_0|D_1 > D_0]$.

c) Assume $D = Z$, what does this reduce to?

4) **Distribution of Y_d**: Assume the potential outcomes framework with selection on observables, and assume common support ($\mathbf{E}[D|X] \equiv p(X) \in (0, 1)$).

a) Temporarily consider the case where $(Y_1, Y_0) \perp D$. How can the marginal distributions of Y_1 and Y_0 be calculated using (Y, D)? Can the joint distribution be calculated? Intuitively, why not?

b) Now return to the selection on observables assumption. How can the marginal distributions of Y_1 and Y_0 be calculated using (Y, D, X)?