

ECMA31000: Introduction to Empirical Analysis

Introduction; Probability; Distributions

Joe Hardwick

University of Chicago

Autumn 2021

Logistics

Lecture: TuTh 2:00PM - 3:20PM in Kent Chem Lab 101.

Discussion Section: W 4:30PM - 5:20PM in Rosenwald Hall 015.

Instructor: Joseph Hardwick

Office: All meetings will be held via Zoom.

Email: hardwick@uchicago.edu

Office Hours: Via Zoom on Tuesdays 3:30PM-5:00PM and
Wednesdays 9:30AM-11:00AM.

Teaching Assistant: Tanya Rajan

Email: tanyar@uchicago.edu

Office Hours: Via Zoom on Mondays and Tuesdays
10AM-11:30AM.

Readings

My lecture slides and annotations, which will be on Canvas.

There is no required textbook, but parts of the following are useful references:

- *Probability and Statistics for Economists* (2021) and *Econometrics* (2021) by Bruce Hansen. Available for free on Bruce's website.
- *Econometric Analysis of Cross Section and Panel Data* (2010) by Jeffrey M. Wooldridge. Available online through the university library.
- *A Primer in Econometric Theory* (2016) by John Stachurski. Many solved exercises on material covered in this class. Slides/Sample Chapters/Code available online.

Evaluation

Final grade determined by assignments, Midterm and Final. 2

Exams:

- Exam 1: October 26 in class 2PM-3:20PM; 80 mins in length.
- Exam 2: TBA, 120 mins in length. Weighting:

Problem Sets: 20% Submitted weekly on Canvas.

Lowest score dropped.

Exam 1: 30% (Covers lectures 1-8)

Exam 2: 50% (Covers lectures 1-16)

| Letter Grade | A | A- | B+ | B | B- | C+ | C | C- |
|---------------|----|----|----|----|----|----|----|----|
| Overall Score | 90 | 85 | 80 | 75 | 70 | 65 | 60 | 55 |

Any student scoring higher than the cutoff given above will earn at least that grade in the course.

Expectations

- All lectures/sections are in person unless we must go remote.
- Please wear masks at all times. No food and drink in classrooms.
- Please ask questions in class!
- Use the discussion boards for clarifications/questions on material.
- Any questions/concerns, please reach out!

Objective 1: Description

- Learn about some feature of the population from a finite sample.
 - Example: Rate of unemployment in the US, households sampled in Current Population Survey.
 - Example: Mean height of population of UChicago undergraduates.
- Can learn about unemployment rate/heights using sample employment/height data x_1, \dots, x_n .
 - Data are sampled from an unknown population distribution.
 - We estimate the quantity of interest e.g. Mean/Expected height, μ , estimated by sample mean $\frac{1}{n} \sum_{i=1}^n x_i$.
 - Test whether mean height is 175cm. Can assess whether hypothesis is reasonable based on sample data.
- Key difference: Actual average height is constant, but sample average is random.

Objective 1: Description

- Often interested in measuring differences in means.
- e.g. Difference in apartment prices with 3 bedrooms vs. 2 bedrooms.
- Quantity of interest:

$$E(\text{Price} | \text{Beds} = 3) - E(\text{Price} | \text{Beds} = 2).$$

- If floor area is fixed, then quantity of interest is

$$E(\text{Price} | \text{Beds} = 3, \text{Area} = 1000\text{sqft}) \\ - E(\text{Price} | \text{Beds} = 2, \text{Area} = 1000\text{sqft}).$$

Objective 2: Prediction

- Given a student's ACT and high school GPA, what is your best prediction of their college GPA?
- We don't think higher test scores during high school cause a high GPA, but may predict it.
- We can consider linear predictors of the form

$$\widehat{colGPA} = \beta_0 + \beta_1 ACT + \beta_2 hsGPA$$

where the β_j are specially chosen to give the “best” predictor of college GPA, using *ACT* and *hsGPA* as predictors.

- We use the data available to draw line/plane of best fit.

Objective 3: Causality

- Learn the effect of changing an observable characteristic on an individual's outcome.
 - Example: The effect of an extra year of schooling on earnings
 - Example: The effect of a job training grant on firm productivity
 - Example: Effect of increasing statewide alcohol tax on wine sales.
- This is different from description:
 - Example: On average, people with 4 years of high school education earn \$X more than those with 3 years.
 - Example: On average, firms receiving job training grants have a X% higher scrap rate than those which don't.
 - Example: On average, wine sales increased by X% in year following tax increase.

Linear Regression

- In this class we consider linear models for the purposes of description, prediction and inferring causality.
- The observed variables are the explanatory variables x_1, \dots, x_k and outcome y .
- For an individual i drawn from the population, the model/data generating process is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + u_i,$$

- “Linear” means linear in parameters β_j .
- e.g. if $x_1 = r$ and $x_2 = r^2$ for some variable r , the model $y = \beta_0 + \beta_1 r + \beta_2 r^2 + u$ is linear in parameters even though it contains a quadratic term in r .

Linear Regression

- Model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i$$

- u_i is the error term/idiosyncratic shock. In a causal model, u_i contains unobserved determinants of y_i .
- Parameters β_j are assumed to be unknown constants.
- Want to get a “good estimate” of the β_j from a sample of individuals from the population.

Why large sample theory?

- Suppose we observe a random sample of X_i , distributed as $X_i \sim U[0, \theta]$.
- We wish to estimate θ . Consider the Method of Moments estimator

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

What is the exact distribution of $\hat{\theta}$? For n small, this is somewhat straightforward. For $n = 100$, it isn't! But, scaled and centered, it is approximated well by the Standard Normal Distribution, due to the central limit theorem.

- There are other ways to estimate θ , e.g. maximum likelihood. The asymptotic distribution is non-normal in this case!

Why large sample theory?

- In the regression example, we will generally not assume the distribution of u is known.
- Implies distribution of estimator $\hat{\beta}$ is unknown, even if we could compute it
 - Properly scaled and centered version of $\hat{\beta}$ still has an approximately normal distribution!
- The central limit theorem provides this result, and allows us to conduct hypothesis tests on the unknown parameter β .

Overview: Part I

- Probability:
 - Sample space, Events, Probability.
 - Random variables, distributions, conditional distributions.
- Large Sample Theory
 - Modes of Convergence
 - Law of Large Numbers, CLT
 - Continuous mapping theorem, Delta Method
- Estimation
 - Definitions, finite sample and asymptotic properties
 - Maximum likelihood , (Generalized) Method of Moments

Overview: Part II

- Hypothesis testing
- Linear Regression
 - Review of Linear Algebra, Projections
 - Interpretations and Estimation of Linear Model
 - Properties of OLS
 - Testing with Normal errors, non-normal errors
- Instrumental Variables
 - Reasons best fit line/plane doesn't accurately measure causal effect of x on y (endogeneity)
 - How IV methods can help
 - Properties of IV estimators
 - GMM
- Topic (if time permits)

Questions?

Probability Spaces

- Probability spaces provide a formal model for uncertainty.

They consist of three parts:

- ① Ω , the sample space, is a non-empty set containing all possible outcomes ω of an experiment.
 - A subset of $E \subset \Omega$ is called an event.
 - When the uncertainty is resolved, if the outcome $\omega \in E$, we say event E occurred.
- ② \mathcal{F} , a sigma algebra, is a collection of events to which we can assign a probability satisfying:
 - ① $E \in \mathcal{F} \implies E^c \in \mathcal{F}$;
 - ② $E_1, E_2, \dots \in \mathcal{F} \implies \bigcup_{n=1}^{\infty} E_n \in \mathcal{F}$;
 - ③ $\Omega \in \mathcal{F}$.
- ③ P , a probability measure which is a mapping $P : \mathcal{F} \rightarrow [0, 1]$ such that
 - ① $P(\Omega) = 1$;
 - ② $P(\bigcup_{n=1}^{\infty} E_n) = \sum_{n=1}^{\infty} P(E_n)$ for disjoint events $E_1, E_2, \dots \in \mathcal{F}$.

Probability Spaces

- Probability spaces provide a formal model for uncertainty.

They consist of three parts:

- ① Ω , the sample space, is a non-empty set containing all possible outcomes ω of an experiment.
 - A subset of $E \subset \Omega$ is called an event.
 - When the uncertainty is resolved, if the outcome $\omega \in E$, we say event E occurred.
- ② \mathcal{F} , a sigma algebra, is a collection of events to which we can assign a probability satisfying:
 - ① $E \in \mathcal{F} \implies E^c \in \mathcal{F}$;
 - ② $E_1, E_2, \dots \in \mathcal{F} \implies \bigcup_{n=1}^{\infty} E_n \in \mathcal{F}$;
 - ③ $\Omega \in \mathcal{F}$.
- ③ P , a probability measure which is a mapping $P : \mathcal{F} \rightarrow [0, 1]$ such that
 - ① $P(\Omega) = 1$;
 - ② $P(\bigcup_{n=1}^{\infty} E_n) = \sum_{n=1}^{\infty} P(E_n)$ for disjoint events $E_1, E_2, \dots \in \mathcal{F}$.

Probability Spaces

- Probability spaces provide a formal model for uncertainty.

They consist of three parts:

- ① Ω , the sample space, is a non-empty set containing all possible outcomes ω of an experiment.
 - A subset of $E \subset \Omega$ is called an event.
 - When the uncertainty is resolved, if the outcome $\omega \in E$, we say event E occurred.
- ② \mathcal{F} , a sigma algebra, is a collection of events to which we can assign a probability satisfying:
 - ① $E \in \mathcal{F} \implies E^c \in \mathcal{F}$;
 - ② $E_1, E_2, \dots \in \mathcal{F} \implies \bigcup_{n=1}^{\infty} E_n \in \mathcal{F}$;
 - ③ $\Omega \in \mathcal{F}$.
- ③ P , a probability measure which is a mapping $P : \mathcal{F} \rightarrow [0, 1]$ such that
 - ① $P(\Omega) = 1$;
 - ② $P(\bigcup_{n=1}^{\infty} E_n) = \sum_{n=1}^{\infty} P(E_n)$ for disjoint events $E_1, E_2, \dots \in \mathcal{F}$.

Properties of P

- \mathcal{F} keeps track of which events we can assign a probability to.
 P tells us how likely these events are to occur.
- Let $A, B \in \mathcal{F}$ with $A \subset B$. Then:
 - ① $P(B \setminus A) = P(B) - P(A)$, where $B \setminus A = B \cap A^c$
 - ② $P(A) \leq P(B)$
 - ③ $P(A^c) = 1 - P(A)$
 - ④ $P(\emptyset) = 0$
- Subadditivity: For any $A, B \in \mathcal{F}$,
 $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (Exercise!)

Properties of P : Proofs

- 1: Note that A and $B \cap A^c$ are disjoint, and $A \cup (B \cap A^c) = B$. By property 2 of probability measure, $P(B) = P(A) + P(B \setminus A)$.
- 2: Since P is nonnegative, $P(B) = P(A) + P(B \setminus A) \geq P(A)$
- 3: Follows from 1 with $B = \Omega$.
- 4: Follows from 3 with $A = \Omega$.

(In)dependence

- For any events $A, B \in \mathcal{F}$, the conditional probability of A given B is

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

- Events $A_1, \dots, A_N \in \mathcal{F}$ are independent if

$$P\left(\bigcap_{n=1}^N A_n\right) = \prod_{n=1}^N P(A_n).$$

- If A and B are independent events, $P(A|B) = P(A)$.

Law of total probability

- A partition of Ω is a countable collection of events $\{B_n\}_{n=1}^{\infty}$ such that
 - ① $B_n \cap B_m = \emptyset$ when $n \neq m$ (events are disjoint)
 - ② $\cup_{n=1}^{\infty} B_n = \Omega$ (events cover the sample space)
- (LOTP) Let $\{B_n\}_{n=1}^{\infty}$ be a partition of Ω and A be any event. If $P(B_m) > 0$ for all m , then

$$P(A) = \sum_{m \geq 1} P(A|B_m) \cdot P(B_m)$$

Bayes' Law

- (Bayes' Law) By definition of the conditional probabilities $P(A|B)$, $P(B|A)$:

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

$$(\text{LOTP}) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}.$$

Example: Screening for health conditions

- A test for a certain condition indicates either +, a signal the patient has a condition (C), or -, a signal they don't have the condition (NC).
- Base rate of the condition in the population is

$$P(C) = 0.001.$$

- Test has a false positive rate of 0.01. There are no false negatives:

$$P(+|NC) = 0.01,$$

$$P(-|C) = 0.$$

Example: Screening for health conditions

- The test does not often produce false positives or false negatives. How likely is it that a patient has the condition given a positive result?

$$\begin{aligned}P(C|+) &= \frac{P(+|C)P(C)}{P(+|C)P(C) + P(+|NC)P(NC)} \\&= \frac{1 \times 0.001}{1 \times 0.001 + 0.01 \times 0.999} \approx \frac{1}{10}.\end{aligned}$$

- Although false positives are rare, the base rate of the condition is so low that even if a positive result is found it is still unlikely the patient has it.

Random Variables

- A random variable X maps outcomes from a sample space Ω to real numbers:

$$X : \Omega \rightarrow \mathbb{R}$$

e.g. We observe the outcome of 2 coin flips, want to know how many heads observed:

$$\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$$

$$\mathcal{F} = \mathcal{P}(\Omega)$$

$$P(\{\omega\}) = 1/4 \text{ for all } \omega \in \Omega.$$

For $\omega \in \Omega$, define

$$X_1(\omega) = \begin{cases} 0 & \omega \in \{(T, T)\} \\ 1 & \omega \in \{(H, T), (T, H)\} \\ 2 & \omega \in \{(H, H)\} \end{cases}.$$

Indicator Functions

- A binary random variable which indicates that some event A has occurred is called an indicator function:

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}$$

- A look ahead: Hypothesis tests are indicator functions! We reject the null hypothesis if a statistic T is larger than some critical value c .
 - The event $A = \{\omega : T(\omega) > c\}$ is the set of sample outcomes that produce a test stat larger than c .
 - Write $\phi(\omega) = \mathbf{1}_A(\omega)$. We reject if and only if we observe $\phi = 1$.

Random Vectors

- If X maps Ω onto \mathbb{R}^N for $N > 1$ we call X a random vector.
- E.g. If I collect a random sample of heights, the collection of observations is a random vector.
- E.g. (Coin flips) I care about number of heads X_1 AND binary rv indicating whether the second flip is a tail (X_2):

$$X(\omega) = \begin{pmatrix} X_1(\omega) \\ X_2(\omega) \end{pmatrix} = \begin{cases} \begin{pmatrix} 0 & 1 \end{pmatrix}' & \omega = (T, T) \\ \begin{pmatrix} 1 & 1 \end{pmatrix}' & \omega = (H, T) \\ \begin{pmatrix} 1 & 0 \end{pmatrix}' & \omega = (T, H) \\ \begin{pmatrix} 2 & 0 \end{pmatrix}' & \omega = (H, H) \end{cases}$$

- In this case: knowledge of both X_1 and X_2 implies knowledge of the actual outcome.

Transformations of Random Variables

- If $X : \Omega \rightarrow \mathbb{R}$ is a random variable, what about $g(X)$ for some function g ? For our purposes, yes:

$$g(X(\omega)) : \Omega \rightarrow \mathbb{R}$$

is still a map from the sample space to \mathbb{R} .

- Requires: X is a random variable, g is “measurable”.
- Note: All continuous functions are “measurable”, as are indicator functions of most sets of interest.
- Hypothesis tests are usually transformations of random variables/vectors:
 - We observe X , and $T(\omega)$ is in fact $T(X(\omega))$. Then $\phi(\omega)$ is in fact $\phi(T(X(\omega)))$.

Questions?

Distributions

- With coin flips, we observe the sample outcome ω and compute $X(\omega)$.. but:
- In many cases we only ever observe X , so we never fully specify experiment (Ω, \mathcal{F}, P) .
 - Not a problem if X is object of interest.
 - Goal is to learn about features of the distribution P_X of X .
- For $B \subset \mathbb{R}$, we define

$$P_X(X \in B) := P(\omega : X(\omega) \in B).$$

The probability measure P_X is the distribution of X .

Distribution Function

- The distribution function of X , F_X , is found by evaluating $P_X(B)$ on all sets B of the form

$$B = (-\infty, c]; \quad c \in \mathbb{R},$$

so that

$$F_X(c) = P_X((-\infty, c]).$$

Notice that the event $\{X \in (-\infty, c]\}$ is just $\{X \leq c\}$.

- Writing down F_X much easier than specifying P_X , since far fewer sets B to consider.
- Good news: P_X is completely determined by F_X !

Properties of distribution functions

- Properties of F_X :
 - Weakly increasing: $F_X(x) \leq F_X(y)$ if $x \leq y$.
 - Right continuous: $F_X(x_n) \rightarrow F_X(x)$ if $x_n \geq x$ for all n and $x_n \rightarrow x$.
 - $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow \infty} F_X(x) = 1$.

Example (Coin Flips)

- Recall 'number of heads' in 2 consecutive coin flips described by

$$X_1(\omega) = \begin{cases} 0 & \omega \in \{(T, T)\}, \\ 1 & \omega \in \{(H, T), (T, H)\}, \\ 2 & \omega \in \{(H, H)\}. \end{cases}$$

- Suppose the coin is fair. Draw F_{X_1} :

Expectation

- Recall $X : \Omega \rightarrow \mathbb{R}$. We say X induces a distribution on \mathbb{R} , because P_X , the distribution of X , is determined by:
 - P , the prob. measure on the underlying space (Ω, \mathcal{F}, P) ,
 - The random variable X :

$$P_X(B) = P(\omega : X(\omega) \in B).$$

- Expectation of X can be computed directly from P_X . e.g. If X is continuous:

$$\begin{aligned} E(X) &= \int x dP_X \\ &= \int x f_X(x) dx. \end{aligned}$$

- This means $E(X)$ is a feature of P_X : Can “learn” $E(X)$ by observing draws from P_X .

Discrete Distributions

- A random variable X is discrete if there is a countable set of values $\{x_j\}_{j \geq 1}$ such that $P_X(X \in \{x_j\}_{j \geq 1}) = 1$.
- For each j , denote $P_X(x_j) = p_j$. The probability mass function is

$$p_X(x) = \begin{cases} p_j & x = x_j, \\ 0 & \text{otherwise.} \end{cases}$$

- The CDF F_X is determined by summing probability masses:

$$F_X(x) = \sum_{j \geq 1} p_j \mathbf{1}(x_j \leq x).$$

and for a transformation $h(X)$:

$$E(X) = \sum_{j \geq 1} p_j h(x_j).$$

Continuous distributions

- A density function is a nonnegative function on \mathbb{R} which integrates to 1.
- X is (absolutely) continuous if there exists a density function f_X such that

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

for all $x \in \mathbb{R}$.

- In contrast to discrete distributions, absolutely continuous distributions assign probability 0 to any countable set.
- For any transformation $h(X)$:

$$E(h(X)) = \int_{\mathbb{R}} h(x) f_X(x) dx.$$

Simulating a continuous distribution

- Suppose we want to simulate draws from continuous F_X , but our software only produces independent draws $Y_i \sim U[0, 1]$.
- Can turn these into draws from F_X simply by applying F_X^{-1} to the outcome:

Distribution of a Random Vector

- Distribution of $X = (X_1, \dots, X_n)$ characterized by the joint distribution function F_X , defined by:

$$F_X(x_1, \dots, x_n) := P_X(X_1 \leq x_1, \dots, X_n \leq x_n).$$

- As in scalar case, X is continuous if there exists a density function $f_X : \mathbb{R}^N \rightarrow \mathbb{R}_+$ such that

$$F_X(x_1, \dots, x_n) = \int_{s_1=-\infty}^{x_1} \cdots \int_{s_n=-\infty}^{x_n} f_X(s_1, \dots, s_n) ds_1 \dots ds_n.$$

- X is discrete if there is a countable set of $(n \times 1)$ vectors $\{x_j\}_{j \geq 1}$ such that $P_X(X \in \{x_j\}_{j \geq 1}) = 1$.
- If P_{X_1}, \dots, P_{X_n} are the same univariate distribution, we say X_1, \dots, X_n are identically distributed.

Questions?

Properties of Expectation

- Expectation is Linear: If $E(X)$ and $E(Y)$ exist then for any constants a, b ,

$$E(aX + bY) = aE(X) + bE(Y).$$

- If $P(\omega : X(\omega) \leq Y(\omega)) = 1$, then $E(X) \leq E(Y)$.
- (Cauchy-Schwarz Inequality) If $E(X^2)$ and $E(Y^2)$ exist:

$$|E(XY)|^2 \leq E(X^2) E(Y^2),$$

with equality iff $X = aY$ for some constant a .

Jensen's Inequality

- (Jensen's inequality) If $E(X)$ and $E(g(X))$ exist and g is convex:

$$E(g(X)) \geq g(E(X)).$$

- The inequality is strict if X is not almost surely constant and g is strictly convex.

Moments

- If $E(X^k)$ exists, then:
 - $E(X^k)$ is the k -th moment of X .
 - $E[(X - E(X))^k]$ is the k -th central moment of X .
 - $k = 2$ gives the Variance of X .
- The covariance of X and Y is given by

$$\text{Cov}(X, Y) := E[(X - E(X))(Y - E(Y))].$$

- The correlation between X and Y is given by

$$\text{Corr}(X, Y) := \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}}.$$

Correlation

- $\text{Corr}(X, Y)$ measures the strength of a linear relationship. X and Y are uncorrelated if $\text{Corr}(X, Y) = 0$.
- In general:

$$-1 \leq \text{Corr}(X, Y) \leq 1.$$

The correlation is 1 or -1 iff $X = a + bY$ for some constants a, b .

Proof.

By the Cauchy Schwarz inequality:

$$|\mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]|^2 \leq \mathbb{E}[(X - \mathbb{E}(X))^2] \mathbb{E}[(Y - \mathbb{E}(Y))^2].$$

or

$$|\text{Cov}(X, Y)| \leq \sqrt{\text{Var}(X) \text{Var}(Y)}.$$

with equality iff $X - \mathbb{E}(X) = b(Y - \mathbb{E}(Y))$ for some constant b , which holds iff $X = a + bY$ for some constants a, b . □

Chebyshev's Inequality

- Bounds the probability that X is large using expectations.
- Suppose X^r is a non-negative integrable random variable for some $r > 0$. Then for any $\delta > 0$:

$$P(X \geq \delta) \leq \frac{E(X^r)}{\delta^r}.$$

Proof.

Write $X^r \geq X^r \mathbf{1}_{X < \delta} + \delta^r \mathbf{1}_{X \geq \delta} \geq \delta^r \mathbf{1}_{X \geq \delta}$. Taking expectations on both sides of the inequality yields

$$E(X^r) \geq \delta^r P(X \geq \delta).$$



Existence of Moments

- Suppose $E(|X|^k) < \infty$ for some $k > 0$. Then for $0 < r < k$, $E(|X|^r) < \infty$.

Proof.

First note

$$|X|^r \leq 1 \cdot \mathbf{1}_{|X| < 1} + |X|^k \mathbf{1}_{|X| \geq 1}.$$

Recall that if $X \leq Y$, $E(X) \leq E(Y)$. So, if RHS of the inequality has finite expectation, $E(|X|^r) < \infty$ also. We have:

$$\begin{aligned} E\left(1 \cdot \mathbf{1}_{|X| < 1} + |X|^k \mathbf{1}_{|X| \geq 1}\right) &= P(|X| < 1) + E\left(|X|^k \mathbf{1}_{|X| \geq 1}\right) \\ &\leq P(|X| < 1) + E(|X|^k) \\ &< \infty. \end{aligned}$$



Matrices

- We will collect information about individuals together in a single matrix.
- e.g. Information about education levels and years of experience can be collected in a matrix X :

Expectation of Random Matrices

- If $X \in \mathbb{R}^{N \times M}$ is a random matrix, its expectation is defined as

$$E(X) := \begin{pmatrix} E(X_{11}) & E(X_{12}) & \cdots & E(X_{1M}) \\ E(X_{21}) & E(X_{22}) & \cdots & E(X_{2M}) \\ \vdots & \vdots & \ddots & \vdots \\ E(X_{N1}) & E(X_{N2}) & \cdots & E(X_{NM}) \end{pmatrix}.$$

- $M = 1$ corresponds to a random vector.
- The variance of an $(N \times 1)$ random vector is the matrix

$$Var(X) = E \left[(X - E(X))(X - E(X))' \right],$$

with (i, j) element equal to $Cov(X_i, X_j)$. Note $Cov(X_i, X_i) = Var(X_i)$.

(Co)variance of Random Vectors

- Let X be a random vector such that $\text{Var}(X)$ exists. Let A be a matrix of constants and b a vector of constants. Then

$$\text{Var}(AX + b) = A\text{Var}(X)A'$$

- If $X \in \mathbb{R}^N$ and $Y \in \mathbb{R}^M$, the covariance of X and Y is the $N \times M$ matrix

$$\text{Cov}(X, Y) := \text{E} \left[(X - \text{E}(X))(Y - \text{E}(Y))' \right].$$

Independence

- The elements of X are independent if

$$F_X(x_1, \dots, x_n) = \prod_{j=1}^n F_{X_j}(x_j)$$

for all (x_1, \dots, x_n) , where F_{X_j} is the distribution of X_j .

- A countable collection of random variables $\{X_i\}_{i \geq 1}$ is independently and identically distributed (iid) if:
 - $F_{X_j} = F_{X_1}$ for all $j \geq 1$. (Identical distribution)
 - Any finite combination of the X_i forms a vector of independent random variables.
- e.g. A stationary time series features identically distributed but not necessarily independent observations.

Marginal Distributions

- If $X = (X_1, \dots, X_n)$ has joint distribution P_X , define marginal distribution of X_1 as:

$$P_{X_1}(X_1 \in B) = P_X(X_1 \in B, X_2 \in \mathbb{R}, \dots, X_n \in \mathbb{R}).$$

- It also holds that

$$F_{X_1}(x_1) = \lim_{x_2, \dots, x_n \rightarrow +\infty} F_X(x_1, \dots, x_n).$$

Marginal Distributions

- If (X, Y) is continuously distributed, then for all x ,

$$\begin{aligned} F_X(x) &= \lim_{y \rightarrow +\infty} F_{XY}(x, y) \\ &= \lim_{y \rightarrow +\infty} \int_{s=-\infty}^x \int_{t=-\infty}^y f_{XY}(s, t) \, ds \, dt \\ &= \int_{s=-\infty}^x \int_{t=-\infty}^{\infty} f_{XY}(s, t) \, dt \, ds \end{aligned}$$

This shows that

$$f_X(s) = \int_{t=-\infty}^{\infty} f_{XY}(s, t) \, dt$$

since densities are unique (up to a set of lebesgue measure 0).

Questions?

Conditional Distributions

- Let (X, Y) have joint density f_{XY} . The conditional density of Y given X is

$$f_{Y|X}(y|x) := \frac{f_{XY}(x, y)}{f_X(x)}$$

and it follows that

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X}(y|x) f_X(x) dx$$

- View this as continuous version of LOTP.
- Discrete case follows as in definition of conditional probability to give conditional pmf $p_{Y|X}$.

Conditional Expectation

- Compute $E(Y|X)$ as

$$E(Y|X = x) = \begin{cases} \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy & (X, Y) \text{ continuous,} \\ \sum_y y p_{Y|X}(y|x) & (X, Y) \text{ discrete.} \end{cases}$$

- $E(Y|X)$ is itself a function of X , so it is a random variable.
- Properties:
 - (Linearity) $E(aY + bZ|X) = aE(Y|X) + bE(Z|X)$;
 - (Law of Iterated Expectation) $E(Y) = E(E(Y|X))$;
 - (Taking out what is known)
 $E(f(X) + g(X) Y|X) = f(X) + g(X) E(Y|X)$.

Conditional Variance

- Define

$$\begin{aligned}\text{Var}(Y|X) &:= \mathbb{E} \left([Y - \mathbb{E}(Y|X)]^2 | X \right) \\ &= \mathbb{E}(Y^2|X) - \mathbb{E}(Y|X)^2.\end{aligned}$$

- Same as variance, but with all expectations now conditioned on X .
- Can show:

$$\text{Var}(Y) = \mathbb{E}(\text{Var}(Y|X)) + \text{Var}(\mathbb{E}(Y|X)).$$

Conditional Mean Independence

- Y is mean independent of X if $E(Y|X) = E(Y)$.
- The following implications hold:

$$\begin{aligned} X \text{ independent of } Y &\implies Y \text{ mean independent of } X \\ &\implies \text{Cov}(X, Y) = 0 \end{aligned}$$

- Note that independence implies

$$P_{Y|X}(Y \in B|X \in A) = P_Y(Y \in B)$$

i.e. the conditional distribution of Y does not depend on X ,
so any feature of it (mean, variance) doesn't either.

Conditional Mean Independence

- Second implication follows from LIE:

$$\begin{aligned}\text{Cov}(X, Y) &= E(XY) - E(X)E(Y) \\ &= E(XE(Y|X)) - E(X)E(Y) \\ &= E(XE(Y)) - E(X)E(Y) \\ &= E(X)E(Y) - E(X)E(Y) \\ &= 0.\end{aligned}$$

- Reverse implications don't generally hold, except when (X, Y) are jointly normal.

Linear Regression

- Suppose the vector (Y, X_1, \dots, X_k) has distribution P_{YX} , where $X = (X_1, \dots, X_k)$.
- Suppose $E(Y^2) < \infty$ and $E(X_j^2) < \infty$ for each $j = 1, \dots, k$.
- A feature of this joint distribution is the function g^* which minimizes

$$E(Y - g(X))^2.$$

That is

$$g^* \in \arg \min_{g \in L^2(X)} E(Y - g(X))^2,$$

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

- Why learn this feature? g^* is the best predictor of Y under square loss.

Linear Regression

- Now show that $g^*(X)$ is given by $E(Y|X)$: Write

$$\begin{aligned} E(Y - g(X))^2 &= E(Y - E(Y|X) + E(Y|X) - g(X))^2 \\ &= E(Y - E(Y|X))^2 + E(E(Y|X) - g(X))^2 \\ &\quad + 2E[(Y - E(Y|X))(E(Y|X) - g(X))]. \end{aligned}$$

Use the LIE to show last term is 0! What remains is

$$\begin{aligned} E(Y - g(X))^2 &= E(Y - E(Y|X))^2 + E(E(Y|X) - g(X))^2 \\ &\geq E(Y - E(Y|X))^2. \end{aligned}$$

so $g(X) = E(Y|X)$ is the best predictor of Y under square loss.

Linear Regression

- We can restrict $L^2(X)$ to a smaller subset

$$H(X) = \{f : f(X) = X'a \text{ for some } a \in \mathbb{R}^N\}$$

- Then, the best linear predictor of Y given X is found by solving

$$\min_{b \in \mathbb{R}^N} E(Y - X'b)^2$$

- Expanding the square gives

$$\min_{b \in \mathbb{R}^N} E(Y^2) - 2b'E(XY) + b'E(XX')b.$$

Linear Regression

- Differentiate wrt. b to yield FOC:

$$2E(XX')b^* - 2E(XY) = 0.$$

- Provided X_j are not linearly dependent random variables, $E(XX')$ is full rank, so

$$b^* = E(XX')^{-1} E(XY).$$

- Define the prediction error $U = Y - X'b^*$. Then, by construction,

$$Y = X'b^* + U; \quad E(XU) = 0,$$

since

$$E(XU) = E(XY) - E(XX')b^* = 0$$

is the first order condition.

Questions?