Class 1: Course Intro & Review

MFE 402

Dan Yavorsky

Topics for Today

1. Introduction to course and instruction team
   - What's this course about? What makes this course different?
   - Should I be excited about econometrics (yes!)
   - Course logistics
2. Review fundamental concepts in probability and statistics
   - Probability theory is *foundational* for mathematical statistics
   - It's the mathematical language used to handle uncertainty
   - This course is all about fitting models that involve uncertainty
3. Review some linear algebra and vector calculus
   - This is a graduate-level course; we use matrices

Course Intro

What is Econometrics?

The unified study of economic models, mathematical statistics, and economic data.
–Hansen, 2022

The unification of statistics, economic theory, and mathematics.
–Frisch, 1933

Useful distinction between theoretical econometrics and applied econometrics.
–Greene, 2006

Or, if you prefer, by analogy:

- when Economists "do statistics", it's called econometrics
- when Psychologists do it, it's called psychometrics
- when Biologists do it, it's called biometrics
- when Political Scientists do it, it's not very good ($\leftarrow$ that's a joke)

Your Instructor and TAs

Prof. Dan Yavorsky

dan.yavorsky@anderson.ucla.edu
@dyavorsky
www.danyavorsky.com
D-419 or Zoom, by appointment

Teaching Assistants

Zechen Liu & Arpan Indora

zechen.liu.2024@anderson.ucla.edu
arpanindora96@ucla.edu

@zechen-liu
@arpanindora

Course Logistics

Course website is on Canvas ("BruinLearn")

- Syllabus & Slides
- Modules list the readings
- Problem sets and submissions links

Class sessions are Tuesdays in D-313

- Sec 2 from 8:30–11:20a
- Sec 1 from 1:00–3:50p

Important Dates

- **Nov 6** Midterm 1:00–2:30p (class follows 2:50–3:50p)
- No class Tues **Nov 26**; makeup Fri **Nov 22**
- Final exam Tues **Dec 10** 11:30a–2:30p

Course Design Principles

- **Attendance:** I don't take it, but I think it's a big part of what you're paying for.
- **Participation:** Do it. Why wouldn't you? It might help your grade.
- **Collaboration:** Permitted on homeworks.
- **ChatGPT:** Use and cite it, just like Wikipedia and Stack Overflow.
- **Contacting us:** One email to both Prof and TAs.
- **Late Submissions:** Generally not accepted.

Outline for the course

- **Week 1:** Prob/Stat/Matrix review
- **Weeks 2–5:** Linear Conditional Mean Models
- **Weeks 6:** Midterm and Computational Topics
- **Weeks 7:** Causal Estimation
- **Weeks 8–9:** Maximum Likelihood
- **Weeks 10:** Bayesian Statistics and/or Non-parametrics

Textbook

Required

- **Hansen, Bruce *Econometrics* (BHE)**
- **Hansen, Bruce *Probability & Statistics for Economists* (BHP)**

Highly Recommended

- Davidson, Russell & James MacKinnon *Econometric Theory and Methods*
- Goldberger, Arthur *A Course in Econometrics*
- Kennedy, Peter *A Guide to Econometrics* (KEN)
- Efron, Bradley & Trevor Hastie *Computer Age Statistical Inference* (CSI)
- Angrist, Joshua & Steffen Pischke *Mostly Harmless Econometrics*

Encyclopedic and/or Classic

- Greene, William *Econometric Analysis*
- Wooldridge, Jeffrey *Econometric Analysis of Cross Section and Panel Data*
- Cameron, Colin & Pravin Trivedi *Microeconometrics: Methods and Applications*
- Davidson, Russell & James MacKinnon *Estimation and Inference in Econometrics*

Assessment

5 Problem Sets (50%)

- PSet 1 due Oct 7 (6%)
- PSet 2 due Oct 21 (10%)
- PSet 3 due Nov 4 (10%)
- PSet 4 due Nov 18 (6%)
- PSet 5 due Dec 2 (12%)
- PSet 6 due Dec 7 (6%)

Midterm (20%)

- Wednesday November 6 1:00p–2:30p (class follows)

Final Exam (30%)

- Tuesday December 10 11:30a–2:30p

How to Succeed

You should go over the material **3 times** (or more):

1. Read the textbook sections
2. Attend class well-rested and ready to engage
3. Return to textbook and slides as you work on assignments or prepare for exams
4. Discuss with your classmates

A comment on grades

- Scores are curved onto a grade distribution
- Top 1–5 students get an A+
- About 40–50% earn A-, A, or A+
- C or F grades must be "earned"

Probability & Statistics

Sets

Given an "experiment," denote an outcome s, where $s \in S$
(S is the collection of all possible outcomes)

Consider a set of subsets of S, denote this collection Γ, we require:

- $S \in \Gamma$,
- if $A \in \Gamma$, then $\overline{A} \in \Gamma$, and
- if $A_i \in \Gamma$ for i countable, then $\bigcup_i A_i \in \Gamma$

For example, flip a coin. Then:

- $s = \{\mathrm{H}\}$ or $s = \{\mathrm{T}\}$
- $S = \{\mathrm{H}, \mathrm{T}\}$
- $\Gamma = \Big\{\{\emptyset\}, \{\mathrm{H}\}, \{\mathrm{T}\}, \{\mathrm{H}, \mathrm{T}\}\Big\}$

Probability

Let $\mathbb{P}$ be a real-valued function on Γ (i.e., $\mathbb{P} : \Gamma \to \mathbb{R}$) then $\mathbb{P}$ is a probability set function if:

- $\mathbb{P}(A) \geq 0 \quad \forall\ A \in \Gamma$,
- $\mathbb{P}(S) = 1$, and
- $\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i) \qquad \text{for } A_i \cap A_j = \emptyset,\ i \neq j$

Direct Consequences: $\forall\ A, B \in \Gamma$

- $0 \leq \mathbb{P}(A) \leq 1$, notably $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(S) = 1$
- if $A \subset B$, $\mathbb{P}(A) \leq \mathbb{P}(B)$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

Notice:

- $\mathbb{P}$ is a function from the space of events to the real line.
- Probability follows certain rules, but the definition does not describe the *meaning* of probability.

Conditional Probability

Conditional Probability

- $\mathbb{P}(A|B) = \mathbb{P}(A \cap B) \ / \ P(B)$

Law of Total Probability

- $\mathbb{P}(B) = \sum_i \mathbb{P}(B|C_i) \cdot \mathbb{P}(C_i)$

Bayes' Rule

- $\mathbb{P}(A|B) = \mathbb{P}(B|A) \cdot \mathbb{P}(A) \ / \ \mathbb{P}(B)$

Independence

- $A \perp B$ iff $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$
- Bayes' Rule implies $\mathbb{P}(A|B) = \mathbb{P}(A)$ iff $A \perp B$
- Note: equally-probable events do not have to be independent

Random Variables

Definition:

- A *function* $X : S \to \mathbb{R}$, which assigns to each element $s \in S$ one (and only one) number $X(s) = x \in \mathbb{R}$

Implications:

- The space or range of X becomes the sample space of interest: $S_X = \{X(s) \in \mathbb{R},\ s \in S\}$
- We can define Γ_X such that $\forall\ B \in \Gamma_X$, $X^{-1}(B) = \{s \in S : X(s) \in B\} \in \Gamma$
- And we can define an induced probability measure $\mathbb{P}_X$ as $\mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B)) \quad \forall\ B \in \Gamma_X$
- Then we've gone from $(S, \Gamma, \mathbb{P})$ to $(S_X, \Gamma_X, \mathbb{P}_X)$

Distribution

For a random variable X the cumulative distribution function (cdf) is:

$$\begin{aligned} F_X(x) &= P_X(X \leq x) \\ &= P_X(\ (\text{-}\infty, x]\ ) \\ &= P(\ \{s \in S : X(s) \leq x\}\ ) \end{aligned}$$

Note that we're assuming $(\text{-}\infty, x] \in \Gamma_X$ so that $X^{-1}((\text{-}\infty, x]) \in \Gamma$

Properties:

- $F_X(x) \in [0, 1] \quad \forall\ x \in \mathbb{R}$
- $x_1 \leq x_2 \Rightarrow F_X(x_1) \leq F_X(x_2)$
- $\lim_{x \to \text{-}\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$
- $F_X(x)$ is right-continuous

Mass and Density

Discrete r.v.

- $p_X(x) = P_X(X = x)$

Continuous r.v.

- $P_X(a \leq X \leq b) = \int_a^b f_X(x)dx$

By the Fundamental Theorem of Calculus:

- $F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$
- $\Rightarrow f_X(x) = dF_X(x)/dx = F'_X(x)$

Expected Value and (Co)Variance

Let X be a random variable with pdf $f_X(x)$ and let g be a real-valued function then:[1]

$$\begin{aligned}\mathbb{E}[g(X)] &= \int_{-\infty}^{\infty} g(x) f_X(x)\, dx \quad \text{for } X \text{ continuous} \\ &= \sum_i g(x_i) p_X(x_i) \quad \text{for } X \text{ discrete}\end{aligned}$$

Variance and covariance are defined to be:

$$\begin{aligned}\operatorname{var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2\end{aligned}$$

$$\begin{aligned}\operatorname{cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

[1] The integral must exist for the expectation to be defined

Linearity of Expectations

If $Y = a + bX_1 + cX_2$, then:

$$\mathbb{E}[Y] = a + b\mathbb{E}[X_1] + c\mathbb{E}[X_2]$$
$$\mathrm{var}[Y] = b^2\mathrm{var}[X_1] + c^2\mathrm{var}[X_2] + 2bc\,\mathrm{cov}(X_1, X_2)$$

Special property for independent variables:

$$\mathbb{E}\Big[\prod_i X_i\Big] = \prod_i \mathbb{E}[X_i]$$

The Normal Family of Distributions

Let $X \sim \mathcal{N}(\mu, \sigma^2)$, then $F_X(x)$ has no closed-form solution, but

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ - \frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2 \right\}$$

Why is this distribution so popular?

- Symmetric, unimodal, with thin tails
- Mathematical convenience
- Central Limit Theorem

See 3Blue1Brown Youtube Series for a great explanation of the Normal distribution and CLT.

Key Notation

One goal of statistics is to learn about features of the **data-generating process** (ie, the DGP or "population"). These features are often called **parameters** and typically denoted with Greek letters such as α, β, or γ.

A **statistic** is a function $f(X_1, \ldots, X_n)$ of the sample $\{X_i : \ i = 1, \ldots, n\}$.

An **estimator** $\hat{\theta}$ for parameter θ is a statistic intended to be a guess about θ.

There is an important distinction between a statistic/estimator as a function of a random sample (and thus is a random variable itself) and a statistic/estimate as a function of the realized sample (and thus is a realized value itself).

Random Variables Z and $\bar{X}$

Let $X \sim \mathcal{N}(\mu, \sigma^2)$, then:

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

Let $X_1, \ldots, X_n$ be *iid* $\mathcal{N}(\mu, \sigma^2)$, then:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim \mathcal{N}(\mu, \sigma^2/n)$$

Student's t-Distribution

Let $X \sim \mathcal{N}(\mu, \sigma^2)$, then:

$$V = Z^2 = (X - \mu)^2/\sigma^2 \sim \chi^2(1)$$

If $V_1, \ldots, V_n$ are independent, each distributed $\chi^2(r_i)$, then:

$$W = \sum_{i=1}^{n} V_i \sim \chi^2(\textstyle\sum_{i=1}^{n} r_i)$$

If Z and W are independent, then:

$$T = \frac{Z}{\sqrt{W/r}} \sim t(r)$$

Student's t "in the wild"

Let $X_1, \ldots, X_n$ be *iid* $\mathcal{N}(\mu, \sigma^2)$. Define:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Then: $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$,

$(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$, and

$$T = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{(n-1)S^2/(\sigma^2(n-1))}} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

Confidence Interval for $\bar{X}$ (X's normal)

Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ with $\bar{X}$ and S^2 as the sample mean and variance defined before.

Define t^* (sometimes denoted $t_{\alpha/2,n\text{-}1}$) to be the upper $\alpha/2$ critical point of a t-distribution with n - 1 degrees of freedom, i.e., $\mathbb{P}(T > t^*) = \alpha/2$. Then:

$$\begin{aligned} 1 - \alpha &= \mathbb{P}\Big(- t^* < \ T \ < t^* \Big) \\ &= \mathbb{P}\Big(- t^* < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t^* \Big) \\ &= \mathbb{P}\Big(\bar{X} - t^* \frac{S}{\sqrt{n}} < \ \mu \ < \bar{X} + t^* \frac{S}{\sqrt{n}} \Big) \end{aligned}$$

The intervals' bounds are random and depend on the sample; μ is a fixed (unknown) value.

In repeated hypothetical samples, the random interval $\left(\bar{X} - t^* \frac{S}{\sqrt{n}}, \bar{X} + t^* \frac{S}{\sqrt{n}} \right)$ will contain μ a specific percent of the time: $(1 - \alpha) \times 100$ percent

t-Test of the Mean

Let $X \sim \mathcal{N}(\mu, \sigma^2)$ and consider the hypothesis:

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0$$

Suppose $X_1, \ldots, X_n$ drawn *iid* from $\mathcal{N}(\mu, \sigma^2)$ distribution.

Under H_0, the statistic $T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$ is distributed $t(n-1)$.

So the t-test is to reject H_0 in favor of H_1 if $|T| \geq t^*$

Intuitively, the test says reject H_0 if $\bar{X}$ is very different than μ_0, where "very different" depends on S and n. This is a statement about precision. Suppose $\bar{X} - \mu_0 = 0.72$. Whether 0.72 is "small" or "large" depends on the variability of the data (S) and how much information you have (n).

Type I and Type II Errors

There are 3 potential outcomes when hypothesis testing:

- A correct decision is made
- H_0 is rejected when it is true – a Type I error
- H_0 is accepted[2] when H_1 is true – Type II error

Decisions are based on critical regions. The goal is to select a critical region from all possible critical regions which minimizes these errors.

In general, this is not possible, as lowering the probability of a Type I error increases the probability of a Type II error and vice versa. Most researchers pick an arbitrary value for α (often 0.05) and then try to minimize the probability of a Type II error given *alpha*.

[2] "Accepted" is just a synonym for "failed to reject" and does not mean H_0 is true – we'll never know the true μ!

LLN & CLT

What if $X_i \overset{iid}{\sim} F$ and F is not the normal distribution?

We can calculate $\hat{\mu} = \bar{X} = \frac{1}{n}\sum X_i$, an estimator of the mean. $\bar{X}$ is a random variable, so it has a distribution with features like a mean and a variance:

$$\mathbb{E}[\bar{X}] = \mu \qquad \text{and} \qquad \text{var}\left(\bar{X}\right) = \frac{\sigma^2}{n}$$

LLN: Let $X_1, \ldots, X_n$ denote observations of a random sample, then $\bar{X} = \frac{1}{n}\sum X_i$ converges to $\mathbb{E}[X]$ as n gets large.

- The mean is the limit of the average.

CLT: Let $X_1, \ldots, X_n$ denote observations of a random sample from a distribution that has mean μ and positive variance σ^2. Then the r.v. $Y = \sqrt{n}(\bar{X} - \mu)/\sigma$ converges to a r.v. with $\mathcal{N}(0, 1)$ distribution.

- $\bar{X}$ is asymptotically normally distributed (even if the X_i's are not normally distributed).

Confidence Interval for $\bar{X}$ (X's not normal)

Let $X_1, \ldots, X_n \overset{iid}{\sim} F$ with $\bar{X}$ and S^2 as the sample mean and variance defined before.

Define z^* (sometimes denoted $z_{\alpha/2}$) to be the upper $\alpha/2$ critical point of a standard normal distribution, i.e., $\mathbb{P}(Z > z^*) = \alpha/2$. Then a $(1 - \alpha) \times 100$ percent **confidence interval** for μ is:

$$
\begin{aligned}
1 - \alpha &= \mathbb{P}\Big(- z^* < \ Z \ < z^* \Big) \\
&= \mathbb{P}\Big(- z^* < \ \frac{\bar{X} - \mu}{S/\sqrt{n}} \ < z^* \Big) \\
&= \mathbb{P}\Big(\bar{X} - z^* \frac{S}{\sqrt{n}} < \ \mu \ < \bar{X} + z^* \frac{S}{\sqrt{n}} \Big)
\end{aligned}
$$

In practice, since $t_{\alpha/2, n-1} \geq z_{\alpha/2}$, the former is often chosen to be "conservative"

Sampling Distribution - Importance

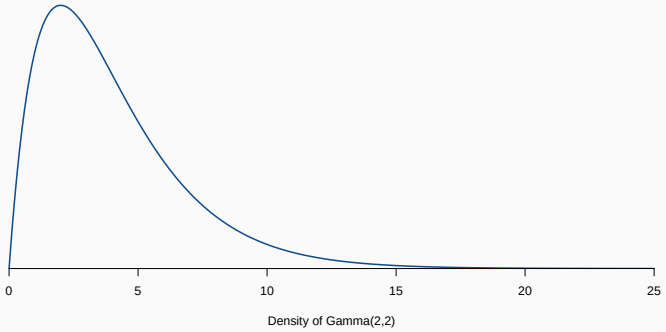
- Statistics are functions of the random sample and are therefore random variables themselves.
- Thus, statistics have probability distributions of their own.
- We often call the distribution of a statistic the **sampling distribution** because it is the distribution induced by sampling.

Let the statistic $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ be a function of an iid sample.

- The sampling distribution of $\hat{\theta}$ is often unknown because the distribution F_X is unknown.
- The goal of an estimator like $\hat{\theta}$ is to learn about the parameter θ.
- To make accurate inferences and to measure the accuracy of our measurements, we need to know something about $\hat{\theta}$'s sampling distribution.
- When $\hat{\theta}$ is a sum or average of random variables, the CLT indicates the sampling distribution of $\hat{\theta}$ is asymptotically normal.

Sampling Distribution - Example

```
par(mar=c(4.5,1,1.2,2))
plot(x = seq(0, 25, 0.1), y=dgamma(seq(0, 25, 0.1), shape=2, scale=2),
     type="l", lwd=2, col="dodgerblue4", yaxs="i", yaxt="n", bty="n",
     main="", ylab="", xlab="Density of Gamma(2,2)", ylim=c(0, 0.2))
```

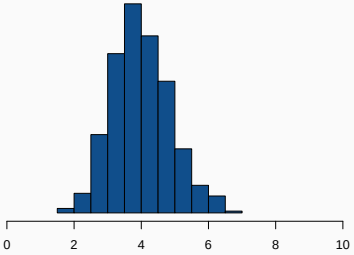


Sampling Distribution - Example

10,000 repetitions of mean

each mean has n=10

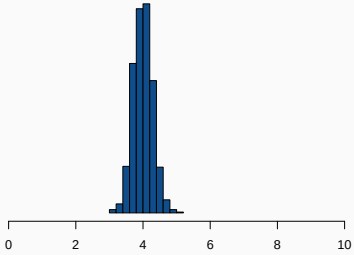
```
r <- 1000
n <- 10
mat <- matrix(rgamma(n*r, shape=2, scale=2),
              nrow=r, ncol=n)
xbar <- apply(mat, 1, mean)
hist(xbar, col="dodgerblue4", yaxt="n", ylab="",
     main="", xlab="", xlim=c(0,10))
```



10,000 repetitions of mean

each mean has n=100

```
r <- 1000
n <- 100
mat <- matrix(rgamma(n*r, shape=2, scale=2),
              nrow=r, ncol=n)
xbar <- apply(mat, 1, mean)
hist(xbar, col="dodgerblue4", yaxt="n", ylab="",
     main="", xlab="", xlim=c(0,10))
```



Linear Algebra

Vector

A vector is an ordered set of numbers arranged in either a row or a column

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

The default will be to denote vectors as column vectors

Matrix

A matrix of dimension $N \times K$ is a rectangular array of numbers

$$\mathbf{A} = [a_{ik}] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1K} \\ a_{21} & a_{22} & \cdots & a_{2K} \\ \vdots & & \ddots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NK} \end{bmatrix}$$

Can think of this as stacked row vectors or aligned column vectors

Special square matrices:

- Symmetric: $a_{ik} = a_{ki}$ for all i and k
- Diagonal: only non-zero elements are on the diagonal
- Scalar: diagonal matrix with same diagonal element values
- Identity: a scalar matrix of ones, denoted I or I_N

Transpose

The transpose of a matrix $\mathbf{A}$, denoted $\mathbf{A}'$ is obtained by creating the matrix whose kth row is the kth column of the original matrix.

Alternatively, consider the transpose operation as switching the indices: $\mathbf{B} = \mathbf{A}' \iff b_{ik} = a_{ki}$

It follows that

- $(\mathbf{A}')' = \mathbf{A}$
- $\mathbf{A} = \mathbf{A}'$ if an only if $\mathbf{A}$ is symmetric

Note that row vectors can be written as $\mathbf{b}' = \begin{bmatrix} b_1 & b_2 \cdots b_k \end{bmatrix}$

Matrix Addition

Matrix addition is element-wise addition:

$$\mathbf{A} + \mathbf{B} = [a_{ik} + b_{ik}]$$

Matrices can only be added if they have the same dimensions

Properties:

- Identity: $\mathbf{A} + \mathbf{0} = \mathbf{A}$ for $\mathbf{0}$ an $N \times K$ matrix of zeros
- Commutative: $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$
- Associative: $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$
- Transpose rule: $(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$

A column vector can be considered a $n \times 1$ matrix, and these rules then apply to vectors as well.

Vector Multiplication

Vector multiplication is called the inner product

$$\mathbf{a}'\mathbf{b} = a_1 b_1 + a_2 b_2 + \ldots a_n b_n$$

Many notations: $\mathbf{a}'\mathbf{b}$ or $\mathbf{a} \cdot \mathbf{b}$ or $< \mathbf{a}, \mathbf{b} >$

Note that

- $\mathbf{a}'\mathbf{b} = \mathbf{b}'\mathbf{a}$
- $\mathbf{a}'\mathbf{a} = \sum_{i=1}^{n} a_i^2$

$\mathbf{i} = [1, 1, \ldots, 1]'$

Let $\mathbf{i}$ be a vector of ones, a a constant, and $\mathbf{x}$ a vector:

- Then $\mathbf{i}'\mathbf{x} = 1x_1 + 1x_2 + \ldots + 1x_n = \sum_{i=1}^{n} x_i$
- If all $x_i = a$, then $\mathbf{x} = a\mathbf{i}$ and $\mathbf{i}'(a\mathbf{i}) = \sum_{i=1}^{n} x_i = na$
- For any a and $\mathbf{x}$:
  - $\sum_{i=1}^{n} ax_i = a\sum_{i=1}^{n} x_i = a(\mathbf{i}'\mathbf{x})$
  - If $a = \frac{1}{n}$, then $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\mathbf{i}'\mathbf{x}$
  - So we can write $\sum_{i=1}^{n} x_i = \mathbf{i}'\mathbf{x} = n\bar{x}$

Matrix Multiplication

NOT the same as scalar multiplication!

For $N \times K$ matrix $\mathbf{A}$ and $K \times M$ matrix $\mathbf{B}$, the product matrix $\mathbf{C} = \mathbf{AB}$ is an $N \times M$ matrix whose ikth element is the inner product of row i of $\mathbf{A}$ and column k of $\mathbf{B}$.

$$\mathbf{C} = \mathbf{AB} \Rightarrow c_{ik} = \mathbf{a}_i' \mathbf{b}_k$$

The one exception is $c\mathbf{A} = [ca_{ik}]$ i.e., just multiple each element of the matrix $\mathbf{A}$ by the number c

Matrix Multiplication Example

$$\underset{3\times 2}{\begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}} \underset{2\times 2}{\begin{bmatrix} 7 & 9 \\ 8 & 0 \end{bmatrix}} = \underset{3\times 2}{\begin{bmatrix} 1(7)+4(8) & 1(9)+4(0) \\ 2(7)+5(8) & 2(9)+5(0) \\ 3(7)+6(8) & 3(9)+6(0) \end{bmatrix}} = \begin{bmatrix} 39 & 9 \\ 54 & 18 \\ 69 & 27 \end{bmatrix}$$

Notice that:

- Inner dimensions must match
- Outer dimensions indicate product matrix dimensions

Matrix Multiplication Properties

Some general rules for matrix multiplication:

- Identity: $\mathbf{AI} = \mathbf{A}$
- Zeroing: $\mathbf{A0} = \mathbf{0}$
- Associative: $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$
- Distributive: $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$
- Transpose: $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$
- Extended transpose: $(\mathbf{ABC})' = \mathbf{C}'\mathbf{B}'\mathbf{A}'$

In general, and unless by coincidence or in certain special cases:

- $\mathbf{AB} \neq \mathbf{BA}$
- $\mathbf{A}'\mathbf{A} \neq \mathbf{AA}'$

And for vectors $\mathbf{a}'\mathbf{a} \neq \mathbf{aa}'$ unless $\mathbf{a}$ is the scalar a

Systems of Equations & Linear Combinations (1)

Let $\mathbf{c} = \mathbf{Ab}$ defined as

$$\underbrace{\begin{bmatrix} 5 \\ 4 \\ 1 \end{bmatrix}}_{\mathbf{c}} = \underbrace{\begin{bmatrix} 4 & 2 & 1 \\ 3 & 6 & 1 \\ 1 & 1 & 0 \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} x \\ y \\ z \end{bmatrix}}_{\mathbf{b}}$$

This can be interpreted as a compact way of writing 3 equations

$$5 = 4x + 2y + 1z$$
$$4 = 3x + 6y + 1z$$
$$1 = 1x + 1y + 0z$$

Or as a linear combinations of columns $\mathbf{a}_k$ of $\mathbf{A}$:

$$\begin{bmatrix} 5 \\ 4 \\ 1 \end{bmatrix} = x \begin{bmatrix} 4 \\ 3 \\ 1 \end{bmatrix} + y \begin{bmatrix} 2 \\ 6 \\ 1 \end{bmatrix} + z \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$$

Systems of Equations & Linear Combinations (2)

For matrix $\mathbf{B}$ instead of vector $\mathbf{b}$ we have

$$\mathbf{C} = \mathbf{AB} \Leftrightarrow \mathbf{c}_k = \mathbf{Ab}_k$$

For example,

$$\begin{bmatrix} 4 & 2 & 1 \\ 3 & 6 & 1 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ x_3 & y_3 \end{bmatrix} = \left[\begin{bmatrix} 4 & 2 & 1 \\ 3 & 6 & 1 \\ 1 & 1 & 0 \end{bmatrix} \mathbf{x} \;\middle|\; \begin{bmatrix} 4 & 2 & 1 \\ 3 & 6 & 1 \\ 1 & 1 & 0 \end{bmatrix} \mathbf{y} \right]$$

$$= \left[\begin{bmatrix} 4 \\ 3 \\ 1 \end{bmatrix} x_1 + \begin{bmatrix} 2 \\ 6 \\ 1 \end{bmatrix} x_2 + \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} x_3 \;\middle|\; \dots \right]$$

Vector Spaces, briefly

- A **vector space** is any set of vectors that is closed under scalar multiplication and addition (e.g., the set of real numbers $\mathbb{R}$)
- A set of vectors in a vectors space is a **basis** for that vector space if they are linearly independent and any vector in the vector space can be written as a linear combination of that set of vectors (e.g. $[1, 0]$ and $[0, 1]$ in $\mathbb{R}^2$)
- A set of $k \geq 2$ vectors is **linearly dependent** if at least one of the vectors in the set can be written as a linear combination of the others (e.g. $[0, 4]$, $[2, 0]$, and $[8, 8]$)
- A set of vectors $\{\mathbf{x}\}$ is **linearly independent** if and only if the solution to $a_1\mathbf{x}_1 + \ldots + a_k\mathbf{x}_k = 0$ is $a_i = 0$ for all i
- The set of all linear combinations of a set of vectors is the vectors space that is **spanned** by those vectors (e.g. $\mathbb{R}^2$ is spanned by the vectors $[1, 0, 0]$ and $[0, 1, 0]$ in $\mathbb{R}^3$)

From Vector Spaces to Matrices

- The **column space** of a matrix is the vector space that is spanned by its column vectors, and the **column rank** of a matrix is the dimension of its column space. Similarly for row space and row rank
- The column space and row space (although not necessarily the same spaces) have the same dimension
- Thus the column rank is equal to the row rank
- For $\mathbf{C} = \mathbf{AB}$, every column of $\mathbf{C}$ is a linear combination of the columns of $\mathbf{A}$, so each column of $\mathbf{C}$ is in the column space of $\mathbf{A}$, thus $\text{rank}(\mathbf{AB}) \leq \min\{\text{rank}(\mathbf{A}),\text{rank}(\mathbf{B})\}$
- If $\mathbf{A}_{m\times n}$ and $\mathbf{B}_{n\times n}$ then $\text{rank}(\mathbf{AB}) = \text{rank}(\mathbf{A})$
- Then, $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}'\mathbf{A}) = \text{rank}(\mathbf{A}\mathbf{A}')$

The Determinant of a Matrix

Denote the determinant of square matrix $\mathbf{A}$ as $|\mathbf{A}|$, defined as

$$|\mathbf{A}| = \sum_{i=1}^{k} a_{ik}(\text{-}1)^{i+k}|\mathbf{A}_{ik}|$$

for $k = 1, \ldots, K$ and where $\mathbf{A}_{ik}$ is the matrix obtained from $\mathbf{A}$ by deleting row i and column k.

For 2×2 and 3×3 matrices, it is:

$$\begin{vmatrix} a & c \\ b & d \end{vmatrix} = ad \text{ - } bc \quad \text{and} \quad \begin{vmatrix} a & d & g \\ b & e & h \\ c & f & i \end{vmatrix} = aei + dhc + gbf \text{ - } gec \text{ - } dbi \text{ - } ahf$$

In general, $|\mathbf{A}|$ is complicated and it involves all elements of $\mathbf{A}$ as it is the K-dim generalization of volume.

Determinant Properties

The determinant is only defined for square matrices.

For an $n \times n$ diagonal matrix $\mathbf{D}$ the determinant is simply

$$|\mathbf{D}| = \prod_{i=1}^{n} d_{ii}$$

For scalar c and $n \times n$ matrix $\mathbf{A}$,

$$|c\mathbf{A}| = c^n |\mathbf{A}|$$

For $n \times n$ matrices $\mathbf{A}$ an $\mathbf{B}$,

$$|\mathbf{AB}| = |\mathbf{A}| \times |\mathbf{B}|$$

One additional result that follows from cofactor expansion,

$$|\mathbf{A}'| = |\mathbf{A}|$$

Matrix Inverse

This is the matrix version of division. $\mathbf{A}$ must be square.

Let $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{b}$ a non-zero vector. To solve for $\mathbf{x}$ we need to find a matrix $\mathbf{B}$ such that $\mathbf{AB} = \mathbf{I}$. If such matrix exists, we say $\mathbf{B} = \mathbf{A}^{-1}$ and it is unique. Also, naturally, $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$.

Let's calculate it. $\mathbf{AB} = \mathbf{I}$ implies

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

with solution

$$\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix} = \frac{1}{|\mathbf{A}|} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$$

We call a matrix **nonsingular** if and only if its inverse exists

Some Computation Results Involving Inverses

If $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$ exist, then

- $|\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|}$
- $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
- $(\mathbf{A}^{-1})' = (\mathbf{A}')^{-1}$
- If $\mathbf{A}$ is symmetric, then $\mathbf{A}^{-1}$ is symmetric
- $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$

Non-singular Equivalences

A list of non-singular equivalences:

- The only solution to $\mathbf{Ax} = \mathbf{0}$ is $\mathbf{x} = \mathbf{0}$
- For any $\mathbf{b} \neq \mathbf{0}$ the linear system $\mathbf{Ax} = \mathbf{b}$ has a unique solution
- $\mathbf{A}$ has "full' ' rank n
- $|\mathbf{A}| \neq 0$
- The column vectors are linearly independent
- The row vectors are linearly independent
- 0 is not an eigenvalue of $\mathbf{A}$ (not covered here)
- $\mathbf{A}$ is row-equivalent to $\mathbf{I}$ (not covered here)
- $\mathbf{A}$ has n non-zero pivots (not covered here)

Geometric Least Squares (1)

Given a vector $\mathbf{y}$ and a matrix $\mathbf{X}$, suppose we are interested in expressing $\mathbf{y}$ as a linear combination of the columns of $\mathbf{X}$. Two possibilities exist:

1. $\mathbf{y}$ lies in the column space of $\mathbf{X}$ and so we can find some vector $\mathbf{b}$ such that $\mathbf{y} = \mathbf{Xb}$
2. $\mathbf{y}$ is not in the column space of $\mathbf{X}$ so $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ where $\mathbf{e}$ is defined as the difference between $\mathbf{Xb}$ and $\mathbf{y}$

In the second case, we are then interested in finding $\mathbf{b}$ such that $\mathbf{y}$ is as close as possible to $\mathbf{Xb}$ in the sense that $\mathbf{e}$ is as short as possible

We'll measure "short" using the l_2 norm, defined as

$$\|\mathbf{e}\| = \sqrt{\mathbf{e}'\mathbf{e}}$$

Geometric Least Squares (2)

The problem is to find the $\mathbf{b}$ for which $\|\mathbf{e}\| = \|\mathbf{y} - \mathbf{Xb}\|$ is as small as possible.

The solution is that $\mathbf{b}$ that makes $\mathbf{e}$ perpendicular (or *orthogonal*) to $\mathbf{Xb}$.

In general, nonzero vectors $\mathbf{a}$ and $\mathbf{b}$ are orthogonal (written $\mathbf{a} \perp \mathbf{b}$) if and only if $\mathbf{a}'\mathbf{b} = \mathbf{b}'\mathbf{a} = 0$.

Returning to our problem of $\mathbf{e} \perp \mathbf{Xb}$:

$$\begin{aligned}(\mathbf{Xb})'\mathbf{e} &= (\mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) \\ &= \mathbf{b}'\mathbf{X}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{Xb} \\ &= \mathbf{b}'(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{Xb}) = 0\end{aligned}$$

This implies $\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{Xb}$ or $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ as long as $\mathbf{X}'\mathbf{X}$ is non-singular or, equivalently, as long as $\mathbf{X}$ has full column rank

Eigenvalues and Eigenvectors

A useful set of results for analyzing a square matrix $\mathbf{A}$ arises from the solutions to the set of equations

$$\mathbf{Ac} = \lambda \mathbf{c}$$

The pairs of solutions are the eigenvectors $\mathbf{c}$ and eigenvalues λ.[3]

To solve, notice $\mathbf{Ac} = \lambda \mathbf{Ic}$ or that $(\mathbf{A} - \lambda \mathbf{I})\mathbf{c} = 0$.

That is, an eigenvalue of $\mathbf{A}$ is a scalar λ which, when subtracted from each of the diagonal entries of $\mathbf{A}$, converts $\mathbf{A}$ into a singular matrix.

[3] If $\mathbf{c}$ is any nonzero solution vector, then $k\mathbf{c}$ is also for any value of k. To remove the indeterminancy, $\mathbf{c}$ is normalized so that $\mathbf{c}'\mathbf{c} = 1$.

Characteristic Equation

If the square matrix $(\mathbf{A} - \lambda\mathbf{I})\mathbf{c} = 0$ is singular, it has a zero determinant. Therefore, if λ is a solution to $\mathbf{Ac} = \lambda\mathbf{c}$ then,

$$|\mathbf{A} - \lambda\mathbf{I}| = 0$$

This polynomial in λ is called the *characteristic polynomial* of $\mathbf{A}$. When we set that polynomial to zero, we call that equation the *characteristic equation* of $\mathbf{A}$.

Finding the eigenvalues of $\mathbf{A}$ is thus the same as finding the roots of the characteristic polynomial of $\mathbf{A}$.

Eigenvalue Example

Define $\mathbf{A}$ to be

$$\mathbf{A} = \begin{bmatrix} 5 & 1 \\ 2 & 4 \end{bmatrix}$$

Then the characteristic equation is

$$|\mathbf{A} - \lambda\mathbf{I}| = \begin{vmatrix} 5 - \lambda & 1 \\ 2 & 4 - \lambda \end{vmatrix} = (5 - \lambda)(4 - \lambda) - 2(1) = \lambda^2 - 9\lambda + 18$$

With roots $\lambda = 6$ and $\lambda = 3$ and the eigenvectors can be solved from $(\mathbf{A} - \lambda\mathbf{I})\mathbf{c} = 0$ and $\mathbf{c}'\mathbf{c} = 1$

Spectral (or Eigenvalue) Decomposition

A $K \times K$ symmetric matrix has K eigenvalues and K distinct orthogonal eigenvectors. Let's collect them into matrices $\mathbf{C}$ and $\mathbf{\Lambda}$:

$$\mathbf{C} = \begin{bmatrix} \mathbf{c}_1 & \mathbf{c}_2 & \cdots & \mathbf{c}_K \end{bmatrix} \quad \text{and} \quad \mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_K \end{bmatrix}$$

Because of orthogonality (i.e., $\mathbf{c}_i'\mathbf{c}_j = 0$ for $i \neq j$), we have

$$\mathbf{C}'\mathbf{C} = \mathbf{I} \Longrightarrow \mathbf{C}' = \mathbf{C}^{-1} \text{ and } \mathbf{C}'\mathbf{A}\mathbf{C} = \mathbf{\Lambda}$$

Putting it all together then,

$$\mathbf{A} = \mathbf{C}\mathbf{\Lambda}\mathbf{C}' = \sum_{i=1}^{K} \lambda_k \mathbf{c}_k \mathbf{c}_k'$$

So we've written $\mathbf{A}_{K \times K}$ as a sum of K rank one matrices.

Proof: From Matrices to Sums[4]

$$\mathbf{X}'\mathbb{E}\mathbf{X} = \begin{bmatrix} x_1 & x_2 & x_3 \\ z_1 & z_2 & z_3 \end{bmatrix} \begin{bmatrix} e_1 & 0 & 0 \\ 0 & e_2 & 0 \\ 0 & 0 & e_3 \end{bmatrix} \begin{bmatrix} x_1 & z_1 \\ x_2 & z_2 \\ x_3 & z_3 \end{bmatrix}$$

$$= \begin{bmatrix} x_1 & x_2 & x_3 \\ z_1 & z_2 & z_3 \end{bmatrix} \begin{bmatrix} x_1 e_1 & z_1 e_1 \\ x_2 e_2 & z_2 e_2 \\ x_3 e_3 & z_3 e_3 \end{bmatrix}$$

$$= \begin{bmatrix} x_1^2 e_1 + x_2^2 e_2 + x_3^2 e_3 & x_1 z_1 e_1 + x_2 z_2 e_2 + x_3 z_3 e_3 \\ z_1 x_1 e_1 + z_2 x_2 e_2 + z_3 x_3 e_3 & z_1^2 e_1 + z_2^2 e_2 + z_3^2 e_3 \end{bmatrix}$$

$$= \begin{bmatrix} x_1^2 & x_1 z_1 \\ z_1 x_1 & z_1^2 \end{bmatrix} e_1 + \begin{bmatrix} x_2^2 & x_2 z_2 \\ z_2 x_2 & z_2^2 \end{bmatrix} e_2 + \begin{bmatrix} x_3^2 & x_3 z_3 \\ z_3 x_3 & z_3^2 \end{bmatrix} e_3$$

$$= \sum \mathbf{x}_i \mathbf{x}_i' e_i$$

[4] Change in notation designed to minimize subscripting

What's the Point?

The diagonalization enables us to obtain the rank of a matrix very easily. We need one more fact though[5]

For any matrix $\mathbf{A}$ and nonsingular matrices $\mathbf{B}$ and $\mathbf{C}$

$$\text{rank}(\mathbf{BAC}) = \text{rank}(\mathbf{A})$$

Then we have:

- $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{\Lambda})$
- Because $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}'\mathbf{A})$, this is also the number of nonzero characteristic roots in $\mathbf{A}'\mathbf{A}$
- The nonzero characteristics roots of $\mathbf{AA}'$ are the same as those of $\mathbf{A}'\mathbf{A}$

[5]Proof repeatedly uses: if $\mathbf{A}_{m\times n}$ and $\mathbf{B}_{n\times n}$ then $\text{rank}(\mathbf{AB}) = \text{rank}(\mathbf{A})$

More Results Follow

Determinants: $\mathbf{C}'\mathbf{A}\mathbf{C} = \mathbf{\Lambda} \Longrightarrow |\mathbf{C}'\mathbf{A}\mathbf{C}| = |\mathbf{\Lambda}| = \prod_{i=1}^{K} \lambda_k$

Inverse: Each λ for $\mathbf{A}$ corresponds to $1/\lambda = \lambda^{-1}$ for $\mathbf{A}^{-1}$

- Proof: $\mathbf{A}^{-1} = (\mathbf{C}\mathbf{\Lambda}\mathbf{C}')^{-1} = (\mathbf{C}')^{-1}\mathbf{\Lambda}^{-1}\mathbf{C}^{-1} = \mathbf{C}\mathbf{\Lambda}^{-1}\mathbf{C}'$

Powers: $\mathbf{A}^2 = \mathbf{A}\mathbf{A} = (\mathbf{C}\mathbf{\Lambda}\mathbf{C}')(\mathbf{C}\mathbf{\Lambda}\mathbf{C}') = \mathbf{C}\mathbf{\Lambda}\mathbf{I}\mathbf{\Lambda}\mathbf{C} = \mathbf{C}\mathbf{\Lambda}^2\mathbf{C}'$

Matrix Square Root: $\mathbf{A}^{1/2} = \mathbf{C}\mathbf{\Lambda}^{1/2}\mathbf{C}'$

- Proof: $\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{C}\mathbf{\Lambda}^{1/2}\mathbf{C}'\mathbf{C}\mathbf{\Lambda}^{1/2}\mathbf{C} = \mathbf{C}\mathbf{\Lambda}\mathbf{C}' = \mathbf{A}$

Factoring: To find $\mathbf{P}$ such that $\mathbf{P}\mathbf{P}' = \mathbf{A}^{-1}$, choose $\mathbf{P} = \mathbf{\Lambda}^{-1/2}\mathbf{C}'$

- Proof: $\mathbf{P}\mathbf{P}' = (\mathbf{C}')'(\mathbf{\Lambda}^{-1/2})'\mathbf{\Lambda}^{-1/2}\mathbf{C}' = \mathbf{C}\mathbf{\Lambda}^{-1} = \mathbf{A}^{-1}$

Cholesky Factorization

Four main methods ro find a matrix square root:

- Spectral decomposition for square matrices (last slide)
- Singular Value Decomposition (SVD) for any matrix, which is the same as spectral decomposition if the matrix is square
- QR Decomposition where $\mathbf{X} = \mathbf{QR}$ for orthogonal matrix $\mathbf{Q}$ and upper-triangular matrix $\mathbf{R}$
- Cholesky Decomposition
  - For a symmetric positive definite matrix $\mathbf{A}$
  - $\mathbf{A} = \mathbf{L}\mathbf{L}' = \mathbf{L}\mathbf{U}$ for $\mathbf{L}' = \mathbf{U}$
  - Denote diagonal elements of $\mathbf{L}$ as d_i, arrange in $\mathbf{D}$
  - Write $\mathbf{A} = \mathbf{L}\mathbf{D}^{-1}\mathbf{D}^2\mathbf{D}^{-1}\mathbf{U} = \mathbf{L}^*\mathbf{D}^2\mathbf{U}^*$, like SVD
  - Also $\mathbf{A}^{-1} = \mathbf{U}^{-1}\mathbf{L}^{-1}$ is simple and numerically stable

Definiteness

Many optimization problems (including OLS) involve double sums of the form:

$$q = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$$

This quadratic form can be written as $q = \mathbf{x}'\mathbf{A}\mathbf{x}$ where $\mathbf{A}$ is symmetric

- If $\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0$ then $\mathbf{A}$ is positive semi-definite
- If $\mathbf{x}'\mathbf{A}\mathbf{x} > 0$ then $\mathbf{A}$ is positive definite
- For $<$ and $\leq$ then $\mathbf{A}$ negative (semi)definite

Proof: $\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{x}'\mathbf{C}\mathbf{\Lambda}\mathbf{C}'\mathbf{x} = \mathbf{y}'\mathbf{\Lambda}\mathbf{y} = \sum_{i=1}^{n} \lambda_i y_i^2 > 0$
(for $\mathbf{y} = \mathbf{C}'\mathbf{x}$)

Results from Definiteness

Denote positive semidefinite (psd) and positive definite (pos. def.)

- If $\mathbf{A}$ is psd, then $|\mathbf{A}| \geq 0$
  - Proof: $|\mathbf{A}| = |\mathbf{\Lambda}| = \prod \lambda_i$ and each $\lambda_i \geq 0$
- If $\mathbf{A}$ is psd, then so is $\mathbf{A}^{-1}$
  - Proof: $\lambda_i \geq 0$ so $1/\lambda_i \geq 0$
- $\mathbf{I}$ is psd
  - Proof: $\mathbf{x}'\mathbf{I}'x = \mathbf{x}'\mathbf{x} = \sum x_i > 0$
- If $\mathbf{A}$ is $N \times K$ with full column rank and $N > K$, then $\mathbf{A}'\mathbf{A}$ is pos. def. and $\mathbf{A}\mathbf{A}'$ is psd
  - Proof: $\mathbf{A}\mathbf{x} \neq \mathbf{0}$, so $\mathbf{x}'\mathbf{A}'\mathbf{A}\mathbf{x} = (\mathbf{A}\mathbf{x})'(\mathbf{A}\mathbf{x}) = \mathbf{y}'\mathbf{y} = \sum y_i > 0$
- If $\mathbf{A}$ is pos. def. and $\mathbf{B}$ is nonsingular, then $\mathbf{B}'\mathbf{A}\mathbf{B}$ is pos. def.
  - Proof: $\mathbf{x}\mathbf{B}'\mathbf{A}\mathbf{B}\mathbf{x} = \mathbf{y}'\mathbf{A}\mathbf{y} > 0$

PSD and Linear Regression

Least Squares minimizes the sum of squared residuals $\mathbf{e}'\mathbf{e}$ to yield

$$\hat{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

These coefficient estimates can only be identified if $(\mathbf{X}'\mathbf{X})$ is nonsingular. But if $N \times K$ matrix $\mathbf{X}$ has full column rank K, then

- $K \times K$ matrix $(\mathbf{X}'\mathbf{X})$ also has rank K
- thus $(\mathbf{X}'\mathbf{X})$ is nonsingular
- and $\hat{\beta}_{OLS}$ is identified

Also, the second order condition yields $\partial^2 \mathbf{e}'\mathbf{e}/\partial\mathbf{b}\partial\mathbf{b}' = 2\mathbf{X}'\mathbf{X}$ which is psd because $\mathbf{c}\mathbf{X}'\mathbf{X}\mathbf{c} = \mathbf{v}'\mathbf{v} = \sum v_i^2 > 0$, demonstrating that $\hat{\beta}_{OLS}$ is a minimum

Trace Operator

The trace of a square $K \times K$ matrix is the sum of its diagonal elements

$$\text{tr}(\mathbf{A}) = \sum_{k=1}^{K} a_{kk}$$

Some results:

- $\text{tr}(c\mathbf{A}) = c \times \text{tr}(\mathbf{A})$
- $\text{tr}(\mathbf{A}') = \text{tr}(\mathbf{A})$
- $\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B})$
- $\text{tr}(\mathbf{I}_K) = K$
- $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$
- $\mathbf{a}'\mathbf{a} = tr(\mathbf{a}'\mathbf{a}) = \text{tr}(\mathbf{a}\mathbf{a}')$
- $\text{tr}(\mathbf{A}'\mathbf{A}) = \sum_{k=1}^{K} \mathbf{a}_k'\mathbf{a}_k = \sum_{i=1}^{K}\sum_{k=1}^{K} a_{ik}^2$
- Cyclic trace: $\text{tr}(\mathbf{ABC}) = \text{tr}(\mathbf{BCA}) = \text{tr}(\mathbf{CAB}) \neq \text{tr}(\mathbf{BAC})$
- $\text{tr}(\mathbf{\Lambda}) = \text{tr}(\mathbf{C}'\mathbf{AC}) = \text{tr}(\mathbf{ACC}') = \text{tr}(\mathbf{AI}) = \text{tr}(\mathbf{A})$

Matrix Calculus

Matrix Calculus: Gradient and Hessian

Let $y = f(\mathbf{x}) = f(x_1, \ldots, x_n)$ be a scalar-valued function.

The vector of partial derivatives, or gradient vector, is

$$\mathbf{g} = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \partial y/\partial x_1 \\ \partial y/\partial x_2 \\ \vdots \\ \partial y/\partial x_n \end{bmatrix}$$

A second derivative matrix, or Hessian is computed as

$$\mathbf{H} = \frac{\partial^2 y}{\partial \mathbf{x} \partial \mathbf{x}'} = \begin{bmatrix} \partial^2 y/\partial x_1 \partial x_1 & \partial^2 y/\partial x_1 \partial x_1 & \cdots & \partial^2 y/\partial x_1 \partial x_n \\ \partial^2 y/\partial x_2 \partial x_1 & \partial^2 y/\partial x_2 \partial x_2 & \cdots & \partial^2 y/\partial x_2 \partial x_n \\ \vdots & \vdots & \ddots & \vdots \\ \partial^2 y/\partial x_n \partial x_1 & \partial^2 y/\partial x_n \partial x_2 & \cdots & \partial^2 y/\partial x_n \partial x_n \end{bmatrix}$$

Matrix Calculus: Linear Functions

A linear function can be written

$$y = \mathbf{a}'\mathbf{x} = \mathbf{x}'\mathbf{a} = \sum a_i x_i$$

So its derivative is

$$\frac{\partial(\mathbf{a}'\mathbf{x})}{\partial \mathbf{x}} = \mathbf{a} \qquad \text{and not } \mathbf{a}'$$

Then in a set of linear functions $\mathbf{y} = \mathbf{A}\mathbf{x}$, each element y_i of $\mathbf{y}$ is $y_i = \mathbf{a}_i'\mathbf{x}$ where $\mathbf{a}_i'$ is the ith row of $\mathbf{A}$. Therefore,

$$\frac{\partial y_i}{\partial \mathbf{x}'} = \mathbf{a}_i' \quad \text{and so} \quad \frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}'} = \begin{bmatrix} \partial y_1/\partial \mathbf{x} \\ \partial y_2/\partial \mathbf{x} \\ \vdots \\ \partial y_n/\partial \mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \\ \vdots \\ \mathbf{a}_n' \end{bmatrix} = \mathbf{A}$$

More commonly, $\partial \mathbf{A}\mathbf{x}/\partial \mathbf{x} = \mathbf{A}'$.

Matrix Calculus: Quadratic Forms

A quadratic form is written

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j a_{ij}$$

With derivative

$$\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$$

For example, let

$$\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix}$$

Then, $\mathbf{x}'\mathbf{A}\mathbf{x} = 1x_1^2 + 6x_1x_2 + 4x_2^2$ and

$$\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \begin{bmatrix} 2x_1 + 6x_2 \\ 6x_1 + 8x_2 \end{bmatrix} = \begin{bmatrix} 2 & 6 \\ 6 & 8 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2\mathbf{A}\mathbf{x}$$

Next Time

- Conditional Expectation Function
- General CEF Model
- Linear CEF Model (ie, linear regression)
- Three Reasons in Support of a Linear CEF Model
- Least Squares Estimator