ECON6300/7320/8300 Advanced Microeconometrics Introduction and Review

Christiern Rose

[1] University of Queensland

Lecture 1

Course Staffs and Meetings

- Christiern Rose (Coordinator & Lecturer)
  Colin Clark Building (39) Room 537
  Email: christiern.rose@uq.edu.au
  Consultation: Tuesdays 15:00-17:00

- Lectures: Fridays, 08:00-10:00
  Practicals: Fridays 10:00-12:00 or 12:00-14:00 (starting week 1)

- Lectures: mainly theoretical “which methods and why"

- Practicals: how to work with data and implement methods (STATA)

Course Information

- The course will focus on estimation and inference methods that are widely used in applied microeconomics.
- The course has a topics-based structure, and theory and applications are closely integrated.
- Whenever you learn a method theoretically (lecture) there will be a Stata exercise with data for the method (practical)
- The course will provide econometric skills that could be used in quantitative research at the Graduate level.
- The course has a strong practical focus, but we will use some basic mathematical/statistical concepts.

Learning Resources

- Required Textbook:
  Cameron, A.C. and P.K. Trivedi (2009) Microeconometrics Using Stata Revised Edition, Stata Press, College Station: TX. **CT-Stata**

- There are many useful texts, e.g., Greene, Wooldridge, etc. see ECP.
- As the semester progresses, lecture notes, slides, datasets, problem sets, etc. will be provided. It is however strongly encouraged that you read the relevant part of the textbook and other references

This week: math review

- **You are EXPECTED to have a basic grasp of mathematics/statistics. If most of the concepts we cover today are unfamiliar to you, this course may be unsuitable.**

- Random Variables, Expectation, Variance, Covariance
- Point Estimation, Hypothesis Testing, Interval Estimation
- Linear Algebra: vector and matrix and their operations

- Informally, a **Random Variable (RV)** takes on numerical values determined by an experiment.

- **Example:** Consider an experiment in which a fair coin is tossed. Then, the possible outcomes are Head and Tail, i.e., $\{H, T\}$. Then, we define a random variable X as follows;

$$X = \begin{cases} 0 & \text{if the outcome is } T \\ 1 & \text{if the outcome is } H \end{cases}$$

The X could be either 0 or 1 whenever the coin is tossed.
Each time, the realisation of X would be different.
The uncertainty can be summarised as $\Pr(X = 0) = 0.5$.

Math Review probability functions

- The uncertainty of a random variable, say X, is represented by its cumulative distribution function (CDF), defined as

$$F_X(x) := \Pr(X \le x)$$

- When $F_X(x)$ is a continuous function, X is said to be a continuous random variable. For a continuous random variable, the probability density function (PDF), denoted by $f_X(x)$, is a function that satisfies

$$\underbrace{\Pr(a < X < b)}_{=F_X(b)-F_X(a)} = \int_a^b f_X(t)dt$$

- If $F_X(x)$ is differentiable, we have

$$f_X(x) = \frac{d}{dx} F_X(x)$$

- If $F_X(x)$ is a step function, X is discrete, which we do not cover today.

Math Review expectation

- A function of a random variable is also a random variable. For example, $\exp(X)$, $\log(X)$, etc. More generally, $g(X)$ with some function $g(\cdot)$.

- Suppose X have the PDF $f_X(x)$. The expectation of $g(X)$ is

$$E[g(X)] := \int_{-\infty}^{\infty} g(t) \cdot f_X(t)dt$$

which represents the central tendency of the random variable, $g(X)$.

- As a special case with $g(X) = X$,

$$E[X] := \int_{-\infty}^{\infty} t \cdot f_X(t)dt$$

which is the expectation of X, or the expected value or the mean

- Consider another case with $g(X) = (X - E[X])^2$.

$$V(X) := E[(X - E[X])^2]$$

which is called the variance that represents the variability of X. Note that $\sqrt{V(X)}$ is the standard deviation of X.

Math Review expectation

- $E[X]$ and $V(X)$ are the population parameters, or theoretical moments.
  - You need to know the distribution to compute $E[X]$ and $V(X)$.

- $E[X]$ must be distinguished from the sample mean (average). For example, when you observe some numbers $x_1, \ldots, x_n$, its average is

$$\overline{x}_n := \frac{1}{n}\sum_{i=1}^{n} x_i$$

which is NOT the expectation of the distribution

- Similarly, you must distinguish $V(X)$ and the sample variance

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2$$

Math Review expectation

1. $E[c] = c$, for any constant c
2. $E[aX + b] = aE[X] + b$ for any constants a and b
3. $a_1, \ldots, a_n$ are constants, and $X_1, \ldots, X_n$ are RV's. Then,

$$E\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1} a_i E[X_i].$$

As a special case, we have

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1} E[X_i].$$

4. $V(c) = 0$, for any constant c.
5. $V(aX + b) = a^2 Var(X)$ for any constants a and c.

- Suppose X and Y are **jointly** distributed with the **joint PDF** $f_{XY}(x, y)$.

- The marginal PDF of X can be obtained by integrating Y out.

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)dy$$

- X and Y are **independent** if and only if for all x and y

$$f_{XY}(x, y) = f_X(x)f_Y(y)$$

- The expectation of $g(X, Y)$ for some function $g(X, Y)$ is

$$E[g(X, Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y)f_{XY}(x, y)dxdy$$

- If $g(X, Y) = (X - E[X])(Y - E[Y])$, then we have the covariance,

$$C(X, Y) = E[(X - E[X])(Y - E[Y])]$$

- If $C(X, Y) > 0$, the X and Y move in the same direction.
  If $C(X, Y) < 0$, they move in the opposite direction.
- $|C(X, Y)| \leq \sqrt{V(X)}\sqrt{V(Y)}$. So, the correlation coefficient

$$\rho_{XY} := \frac{C(X, Y)}{\sqrt{V(X)}\sqrt{V(Y)}}$$

  is always between -1 and 1.

- If X and Y are independent, $C(X, Y) = 0$. But, the converse is not generally true

Math Review conditional distribution

- Suppose X and Y are jointly distributed with the joint PDF $f_{XY}(x, y)$. Then, the conditional PDF of Y given $X = x$ is

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

which summarises the distribution of Y when X takes a value x.

- If X and Y are independent, $f_{Y|X}(y|x) = f_Y(y)$ and $f_{X|Y}(x|y) = f_X(x)$

Math Review conditional expectation

- The conditional expectation of Y given $X = x$ is

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

- The conditional variance of Y given $X = x$ is

$$V(Y|X = x) = E\left[(Y - E[Y|X = x])^2 \middle| X = x\right]$$

- $E[Y|X = x]$ and $V(Y|X = x)$ are functions of x.

- When x is not specified, $E[Y|X]$ and $V(Y|X)$ are random.

1. $E[c(X)|X] = c(X)$ for any function $c(\cdot)$
2. $E[a(X)Y + b(X)|X] = a(X)E[Y|X] + b(X)$ for any functions $a(\cdot), b(\cdot)$
3. If X and Y are independent, $E[Y|X] = E[Y]$
4. $E[E[Y|X]] = E[Y]$. This is called the **law of iterated expectations.**
5. If X and Y are independent, $V(Y|X) = V(Y)$

- Suppose a random variable Y follows a normal distribution with mean μ and variance σ^2. Or simply we can write

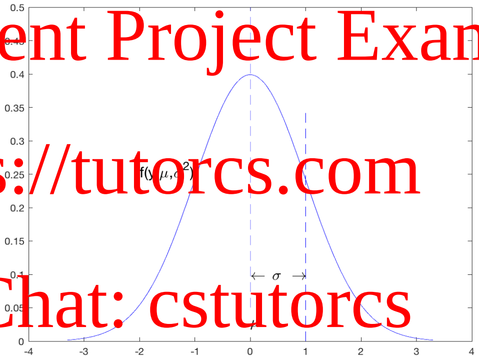
$$Y \sim \mathcal{N}(\mu, \sigma^2)$$

- The Y takes on a value from $\mathbb{R} = (-\infty, \infty)$.
Whenever it is drawn from $\mathcal{N}(\mu, \sigma^2)$, the realisation of Y will be different.
The uncertainty is summarised as its probability density function (PDF),

$$f_Y(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)$$

Math Review normal distribution

- The shape of $f_Y(y|\mu, \sigma^2)$ is given as follows when $\mu = 0$ and $\sigma^2 = 1$.



- It is symmetric about μ and the scale is determined by $\sigma := \sqrt{\sigma^2}$
- The whole shape (distribution) is completely determined by (μ, σ^2)

- $E[Y] = \mu$ and $V(Y) = \sigma^2$ for $Y \sim \mathcal{N}(\mu, \sigma^2)$,
- $\mathcal{N}(0, 1)$ is called the standard normal distribution (the graph above), often used for statistical inference

- I simulated three number from $\mathcal{N}(0, 1)$ using my computer and I obtained
$$\{0.5377, 1.8339, -2.2588\}$$
- These numbers are **realisations** of $\mathcal{N}(0, 1)$. Their average is NOT the expectation of $\mathcal{N}(0, 1)$. In fact the average is 0.0376 while the expectation is 0.

- But, the average seems quite close to the mean

Math Review point estimation for normal mean

- Suppose $Y_1, \ldots, Y_n$ are identically and independently distributed as $\mathcal{N}(\mu, \sigma^2)$, i.e.,

$$Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2),$$

  and also assume that μ is unknown but σ^2 is known. That is, if we knew the **parameter** μ, we would know all.

- We wish to learn μ from the **random sample** of size n, i.e., $Y_1, \ldots, Y_n$. Especially, we consider the **(point) estimator**

$$\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^{n} Y_i$$

- The estimator $\hat{\mu}_n$ is random because $Y_1, \ldots, Y_n$ are all random. (Its realisation is an **estimate** of $\hat{\mu}_n$.)

- **Result:** a linear transformation of normal variables is a normal.
- This implies that

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

- That is, we know that

$$\hat{\mu}_n \sim \mathcal{N}\left(E[\hat{\mu}_n], V(\hat{\mu}_n)\right)$$

- Then, what are the mean and the variance of the estimator $\hat{\mu}_n$?

Math Review sampling distribution of the estimator

- **Result 2:** $E[X + Z] = E[X] + E[Z]$ for any two RVs X and Z
- This implies that

$$E[\hat{\mu}_n] = \frac{1}{n}\sum_{i=1}^{n} E[Y_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}(n\mu) = \mu$$

which means

$$\hat{\mu}_n \sim \mathcal{N}(\mu, V(\hat{\mu}_n))$$

- The estimator $\hat{\mu}_n$ is **unbiased**: $E[\hat{\mu}_n] = \mu$

Math Review sampling distribution of the estimator

- **Result:** $V(aX) = a^2 V(X)$ when X is random and a is a constant

- **Result:** $V(X + Z) = V(X) + V(Z)$ if X and Z are independent
- Therefore,

$$V[\hat{\mu}_n] = V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

- Finally, we have

$$\hat{\mu}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

where μ is the unknown parameter and σ^2 is assumed to be known.

- Recall that

$$\hat{\mu}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

especially, $E[\hat{\mu}_n] = \mu$ and $V(\hat{\mu}_n) = \frac{\sigma^2}{n}$

- The estimator is distributed centred around μ and its variance gets smaller when n grows,
- This suggests that $\hat{\mu}_n$ gets 'very' close to μ as $n \to \infty$.
- This quality is referred to as the consistency of the estimator.

Math Review consistency of the estimator

- Suppose $X_1, X_2, \dots,$ are sequence of random variables such that

$$\lim_{n\to\infty} \Pr(|X_n - c| > \varepsilon) = 0,$$

for any $\varepsilon > 0$. Then X_n converges in probability to c, and we often write

$$X_n \xrightarrow{p} c$$

- $\hat{\mu}_n$ is **consistent** to μ if $\hat{\mu}_n \xrightarrow{p} \mu$.

- When

$$Y \sim \mathcal{N}(\mu, \sigma^2),$$

the properties of normal random variables implies

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

Exercise: verify this.

- Back to our estimator $\hat{\mu}_n$: since

$$\hat{\mu}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

we have

$$\frac{\hat{\mu}_n - \mu}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0, 1)$$

- Recall that when $Z \sim \mathcal{N}(0, 1)$

$$\Pr(|Z| \leq 1.96) = 0.95 \text{ or, equivalently, } \Pr(|Z| > 1.96) = 0.05$$

- That is, the event that $|Z| \leq 1.96$ happens with 95% probability. But, the event that $|Z| > 1.96$ happens with 5% probability.
- So, the event that $|Z| \leq 1.96$ is likely to happen. But, the event that $|Z| > 1.96$ is unlikely to happen.
- Conversely, when the event that $|Z| > 1.96$ happens, you might want to suspect whether Z really follows $\mathcal{N}(0, 1)$

- We know
$$\frac{\hat{\mu}_n - \mu}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0,1)$$
but we do not know the true value of μ.

- So let's assume that $\mu = \mu_0$ for some number μ_0 and study whether this value is reasonable with respect to our distributional knowledge of $\hat{\mu}_n$.

- We test the null hypothesis
$$H_0 : \mu = \mu_0$$

Math Review statistical inference

- Under the null hypothesis H_0, we know $\frac{\hat{\mu}_n - \mu_0}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0,1)$ and the event

$$\left| \frac{\hat{\mu}_n - \mu_0}{\sqrt{\sigma^2/n}} \right| > 1.96$$

happens with 5% of probability (low prob).

- Since μ_0 is a hypothesised value, we consider this unlikely event as a statistical evidence against H_0. So, we reject H_0.
- We admit that even if H_0 is correct, we could mistakenly reject H_0 with 5% of probability. This error rate is called the size (level) of the test.

- You could have a different size. For example, you could use

$$P(|Z| > 2.5758) = 0.01$$

- A confidence interval (CI) can be obtained by inverting the test

- For example, 95% CI is the set of all that values for μ_0 that would not be rejected by the test of 5% size. After some algebra, we construct

$$95\%\ \mathrm{CI} = \left[\hat{\mu}_n - 1.96\sqrt{\sigma^2/n}, \hat{\mu}_n + 1.96\sqrt{\sigma^2/n}\right]$$

or simply we may write $\hat{\mu}_n \pm 1.96\sqrt{\sigma^2/n}$

- CI = interval estimator

- We relax the normality assumption. That is, $Y_1, \ldots, Y_n$ are i.i.d from some distribution (may not be normal) with mean μ and variance σ^2

- Two useful asymptotic results:

$$\hat{\mu}_n \xrightarrow{p} \mu \qquad \text{Weak Law of Large Numbers (WLLN)}$$

$$\sqrt{n}(\hat{\mu}_n - \mu)/\sigma \xrightarrow{d} \mathcal{N}(0, 1) \qquad \text{Central Limit Theorem (CLT)}$$

- WLLN directly implies consistency of $\hat{\mu}_n$

- The CLT

$$\sqrt{n}(\hat{\mu}_n - \mu)/\sigma \xrightarrow{d} \mathcal{N}(0,1)$$

implies that when n is large

$$\sqrt{n}(\hat{\mu}_n - \mu)/\sigma \overset{a}{\sim} \mathcal{N}(0,1)$$

where $\overset{a}{\sim}$ is 'approximately distributed,' which can be rewritten as

$$\frac{\hat{\mu}_n - \mu}{\sqrt{\sigma^2/n}} \overset{a}{\sim} \mathcal{N}(0,1)$$

Math Review asymptotic inference

- So, the hypothesis testing and confidence interval are <u>asymptotically</u> valid even without the normality assumption.
- For example, we reject $H_0 : \mu = \mu_0$ at 5% level if

$$\left| \frac{\hat{\mu}_n - \mu_0}{\sqrt{\sigma^2/n}} \right| > 1.96$$

  and the asymptotic 95% confidence interval is

$$\hat{\mu}_n \pm 1.96\sqrt{\sigma^2/n}$$

- Under normality, $\frac{\hat{\mu}_n - \mu_0}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0,1)$ but
  without normality $\frac{\hat{\mu}_n - \mu_0}{\sqrt{\sigma^2/n}} \overset{a}{\sim} \mathcal{N}(0,1)$

Math Review asymptotic inference

- So far, we have assumed that σ^2 is known. Now, we relax this assumption.
- If σ^2 is unknown, we cannot compute the test statistic and the confidence interval.
- Suppose we have a consistent estimator for σ^2, i.e.,

$$\hat{\sigma}_n^2 \xrightarrow{p} \sigma^2,$$

- Exercise: show that the CLT and Slutsky lemma implies

$$\sqrt{n}(\hat{\mu}_n - \mu)/\hat{\sigma}_n \xrightarrow{d} \mathcal{N}(0,1)$$

- Then, this result

$$\sqrt{n}(\hat{\mu}_n - \mu)/\hat{\sigma}_n \xrightarrow{d} \mathcal{N}(0,1)$$

implies that when n is large

$$\sqrt{n}(\hat{\mu}_n - \mu)/\hat{\sigma}_n \overset{a}{\sim} \mathcal{N}(0,1)$$

where $\overset{a}{\sim}$ is 'approximately distributed,' which can be rewritten as

$$\frac{\hat{\mu}_n - \mu}{\sqrt{\hat{\sigma}_n^2/n}} \overset{a}{\sim} \mathcal{N}(0,1)$$

- Therefore, we can construct the asymptotic testing and confidence interval as before.
- Specifically, we reject $H_0 : \mu = \mu_0$ at 5% level if

$$\left| \frac{\hat{\mu}_n - \mu_0}{\sqrt{\hat{\sigma}_n^2 / n}} \right| > 1.96$$

and the asymptotic 95% confidence interval is

$$\hat{\mu}_n \pm 1.96 \sqrt{\hat{\sigma}_n^2 / n}$$

- Note that $\sqrt{\hat{\sigma}_n^2 / n}$ is called the standard error, $se(\hat{\mu}_n) = \sqrt{\hat{\sigma}_n^2 / n}$.

- Recall that $Z \sim \mathcal{N}(0,1)$. Then, $Z^2 \sim \chi^2(1)$.
- Also, $Z_1, \ldots, Z_K \overset{iid}{\sim} \mathcal{N}(0,1)$. Then, $\sum_{j=1}^{K} Z_j^2 \sim \chi^2(K)$
- Under $H_0 : \mu = \mu_0$, since

$$\frac{\hat{\mu}_n - \mu_0}{\sqrt{\hat{\sigma}_n^2/n}} \overset{a}{\sim} \mathcal{N}(0,1)$$

we have

$$W_n := \left(\frac{\hat{\mu}_n - \mu_0}{\sqrt{\hat{\sigma}_n^2/n}}\right)^2 \overset{a}{\sim} \chi^2(1)$$

- The LHS is the Wald statistic. At level 5%, we reject H_0 if $W_n > (1.96)^2$.

- We started from the result

$$\sqrt{n}(\hat{\mu}_n - \mu)/\sigma \xrightarrow{d} \mathcal{N}(0,1) \quad \text{or} \quad \sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{d} \mathcal{N}(0,\sigma^2)$$

and by squaring both sides and using $\hat{\sigma}_n^2 \xrightarrow{p} \sigma^2$, we obtain

$$W_n = n(\hat{\mu}_n - \mu)\left(\hat{\sigma}_n^2\right)^{-1}(\hat{\mu}_n - \mu) \overset{a}{\sim} \chi^2(1)$$

- More generally, for a parameter $\theta \in \mathbb{R}^k$ and and its estimator $\hat{\theta}_n$, if

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

we can construct the Wald statistic

$$n(\hat{\theta}_n - \theta)'\hat{V}^{-1}(\hat{\theta}_n - \theta) \overset{a}{\sim} \chi^2(k)$$

with a consistent estimator $\hat{V} \xrightarrow{p} V$, the asymptotic variance covariance matrix of $\hat{\theta}$. You will see this sandwich form often.

Math Review matrix algebra

- As mentioned briefly, we often use a vector and matrix notation in econometrics. Vectors and matrices are collections of numbers in a certain form.

- Let $\boldsymbol{a} := \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$ and $\boldsymbol{b} := \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$ be n dimensional column vectors.

- Then, $\boldsymbol{a}' = \begin{pmatrix} a_1 & \dots & a_n \end{pmatrix}$, the transpose of $\boldsymbol{a}$, is an n-dim row vector.

- Addition is defined when two vectors are in the same form.

$$\boldsymbol{a} + \boldsymbol{b} = \begin{pmatrix} a_1 + b_1 \\ \vdots \\ a_n + b_n \end{pmatrix}$$ and $\boldsymbol{a}' + \boldsymbol{b}' = (\boldsymbol{a} + \boldsymbol{b})'$. But $\boldsymbol{a}' + \boldsymbol{b}$ is not defined.

Math Review matrix algebra

- Inner product

$$\langle \boldsymbol{a}, \boldsymbol{b} \rangle := \boldsymbol{a}'\boldsymbol{b} = \begin{pmatrix} a_1 & \ldots & a_n \end{pmatrix} \cdot \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = \sum_{i=1}^{n} a_i b_i$$

- Consider an $n \times k$ dimensional matrix:

$$\boldsymbol{A} := \begin{pmatrix} a_{11} & \ldots & a_{1k} \\ \vdots & & \vdots \\ a_{n1} & \ldots & a_{nk} \end{pmatrix} = \begin{pmatrix} \boldsymbol{a}_1 & \ldots & \boldsymbol{a}_k \end{pmatrix}$$

where $\boldsymbol{a}_j = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{pmatrix}$ for $j = 1, \ldots, k$.

- a vector is a special case of a matrix (e.g., $k = 1$)

- Transpose:

$$A' := \begin{pmatrix} a_{11} & \dots & a_{n1} \\ \vdots & & \vdots \\ a_{1k} & \dots & a_{nk} \end{pmatrix} = \begin{pmatrix} \boldsymbol{a}_1' \\ \vdots \\ \boldsymbol{a}_k' \end{pmatrix}$$

- property: $(AB)' = B'A'$

- Addition is allowed between two matrices in the same dimension. Let

$$\boldsymbol{B} := \begin{pmatrix} b_{11} & \dots & b_{1k} \\ \vdots & & \vdots \\ b_{n1} & \dots & b_{nk} \end{pmatrix} = \begin{pmatrix} \boldsymbol{b}_1 & \dots & \boldsymbol{b}_k \end{pmatrix}$$

where $\boldsymbol{b}_i$ denotes the i^{th} column of $\boldsymbol{B}$. Then,

$$\boldsymbol{A} + \boldsymbol{B} = \begin{pmatrix} a_{11} + b_{11} & \dots & a_{1k} + b_{1k} \\ \vdots & & \vdots \\ a_{n1} + b_{n1} & \dots & a_{nk} + b_{nk} \end{pmatrix}$$

But, $\boldsymbol{A}' + \boldsymbol{B}$ is not defined unless $k = n$, i.e., $\boldsymbol{A}$ and $\boldsymbol{B}$ are square

- To explain multiplication, we define a $k \times m$ matrix

$$
\boldsymbol{C} := \begin{pmatrix} c_{11} & \dots & c_{1m} \\ \vdots & & \vdots \\ c_{k1} & \dots & c_{km} \end{pmatrix}
$$

Then,

$$
\begin{aligned}
\boldsymbol{A} \times \boldsymbol{C} &:= \begin{pmatrix} a_{11} & \dots & a_{1k} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nk} \end{pmatrix} \begin{pmatrix} c_{11} & \dots & c_{1m} \\ \vdots & & \vdots \\ c_{k1} & \dots & c_{km} \end{pmatrix} \\
&:= \begin{pmatrix} \sum_{\ell=1}^{k} a_{1\ell} c_{\ell 1} & \sum_{\ell=1}^{k} a_{1\ell} c_{\ell 2} & \dots & \sum_{\ell=1}^{k} a_{1\ell} c_{\ell m} \\ \sum_{\ell=1}^{k} a_{2\ell} c_{\ell 1} & \sum_{\ell=1}^{k} a_{2\ell} c_{\ell 2} & \dots & \sum_{\ell=1}^{k} a_{2\ell} c_{\ell m} \\ \vdots & & & \vdots \\ \sum_{\ell=1}^{k} a_{n\ell} c_{\ell 1} & \sum_{\ell=1}^{k} a_{n\ell} c_{\ell 2} & \dots & \sum_{\ell=1}^{k} a_{n\ell} c_{\ell m} \end{pmatrix}
\end{aligned}
$$

Math Review matrix algebra

- The $(i,j)^{\text{th}}$ entry of $\boldsymbol{A} \times \boldsymbol{C}$ is $\sum_{\ell=1}^{k} a_{i\ell} c_{\ell j}$

- Multiplication is defined only when the number of columns of the first matrix is equal to the number of rows of the second matrix.

- For example, $\underbrace{\boldsymbol{A}}_{n\times k} \times \underbrace{\boldsymbol{C}}_{k\times m}$ is well defined and the resulting matrix is $n \times m$ dimensional but $\underbrace{\boldsymbol{C}}_{k\times m} \times \underbrace{\boldsymbol{A}}_{n\times k}$ is not defined unless $m = n$.

- Suppose $\boldsymbol{A}$ is a square matrix, i.e., $n = k$.
  a_{ij} is called a diagonal element if $i = j$. Otherwise, it is off-diagonal.
  If off-diagonal elements are all zero, $\boldsymbol{A}$ is said to be a diagonal matrix.
  If $\boldsymbol{A}' = \boldsymbol{A}$, it is symmetric ($a_{ij} = a_{ji}$ for all (i,j))

Math Review matrix algebra

- If $\boldsymbol{A}$ is diagonal and all diagonal elements are ones, $\boldsymbol{A}$ is an n dimensional identity matrix, for which a reserved notation is $\boldsymbol{I}_n$. So,

$$\boldsymbol{I}_n := \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

- A multiplication with the identity does not change. That is, $\boldsymbol{A} \times \boldsymbol{I} = \boldsymbol{A}$ and $\boldsymbol{I} \times \boldsymbol{B} = \boldsymbol{B}$.

- For a square matrix $\boldsymbol{A}$, if there is another matrix $\boldsymbol{B}$ such that $\boldsymbol{I} = \boldsymbol{AB} = \boldsymbol{BA}$, then $\boldsymbol{B}$ is called the inverse of $\boldsymbol{A}$ and denoted by $\boldsymbol{A}^{-1}$, and $\boldsymbol{A}$ is said to be invertible or nonsingular.

- (i) $(\alpha A)^{-1} = \frac{1}{\alpha} A^{-1}$ for a constant α,
  (ii) $(AB)^{-1} = B^{-1}A^{-1}$,
  (iii) $(A')^{-1} = (A^{-1})'$

Math Review matrix algebra

- For any $n \times n$ matrix A, the trace of A, $tr(A)$, is the sum of diagonal elements

$$tr(A) = \sum_{i=1}^{n} a_{ii}$$

- (i) $tr(I_n) = n$, (ii) $tr(A') = tr(A)$, (iii) $tr(A + B) = tr(A) + tr(B)$,
  (iv) $tr(\alpha A) = \alpha tr(A)$ for a constant α
  (v) $tr(AB) = tr(BA)$ for any dimensionality conformable A and B

- Let $x_1, \ldots, x_n$ be a set of $n \times 1$ vectors. These are **linearly independent** if and only if

$$\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n = 0$$

implies $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$.

- When $x_1, \ldots, x_n$ are linearly independent, x_j cannot be represented as a linear combination of others.
- Let X be an $n \times k$ matrix ($n \geq k$). The rank of X, denoted by $rank(A)$, is the maximum number of linearly independent column vectors in X.

- If $rank(X) = k$, X has full column rank.

- Let A be an $n \times n$ symmetric matrix. The **quadratic form** with A is a function $f(x)$ defined for all $n \times 1$ vectors x

$$f(x) = x'Ax$$

  Then, A is **positive definite** (p.d.) if $x'Ax > 0$ for all non-zero x.
  Moreover, A is **positive semi-definite** (p.s.d.) if $x'Ax \geq 0$ for all x.

- (i) If A is p.d., then $a_{ii} > 0$ for all i.
  (ii) If A is p.s.d., then $a_{ii} \geq 0$ for all i.
  (iii) If A is p.d., then A is invertible, i.e., A^{-1} exists.
  (iv) If X is $n \times k$, then $X'X$ and XX' are p.s.d.
  (v) If X is $n \times k$ and $rank(X) = k$, then $X'X$ is p.d. (so invertible)

- For a given $n \times 1$ vector a, consider the linear function

$$f(x) = a'x$$

for all $n \times 1$ vectors x. Then,

$$\frac{\partial f(x)}{\partial x} = a$$

- For an $n \times n$ symmetric matrix A, define a quadratic function

$$g(x) = x'Ax$$

for all $n \times 1$ vectors x. Then,

$$\frac{\partial g(x)}{\partial x} = 2Ax$$

- Note $\frac{\partial f(x)}{\partial x'} = \left(\frac{\partial f(x)}{\partial x}\right)'$ and $\frac{\partial g(x)}{\partial x'} = \left(\frac{\partial g(x)}{\partial x}\right)'$

- Let y be an $n \times 1$ vector and X be $n \times k$. Consider

$$\min_{\beta} (y - X\beta)'(y - X\beta)$$

- $(y - X\beta)' = y' - (X\beta)' = y' - \beta' X'$. So, the objective function is

$$\begin{aligned}
&(y - X\beta)'(y - X\beta) \\
&= (y' - \beta' X')(y - X\beta) \\
&= y'y - y'X\beta - \beta' X' y + \beta' X' X \beta \\
&= y'y - 2y'X\beta + \beta' X' X \beta
\end{aligned}$$

-

$$\min_{\beta} (y - X\beta)'(y - X\beta) = \min_{\beta} (y'y - 2y'X\beta + \beta' X' X \beta)$$

- Then, the first order necessary condition (FOC) is

$$\frac{\partial}{\partial \beta}(y'y - 2y'X\beta + \beta'X'X\beta) = -2\frac{\partial}{\partial \beta}y'X\beta + \frac{\partial}{\partial \beta}\beta'X'X\beta$$

- Applying the rules above,

$$\frac{\partial}{\partial \beta}y'X\beta = X'y, \quad \frac{\partial}{\partial \beta}\beta'X'X\beta = 2X'X\beta$$

- The FOC is then

$$-2X'y + 2X'X\beta = 0 \iff X'y = X'X\beta$$

at the optimal solution for β, say $\hat{\beta}$. The solution exists if and only if $X'X$ is invertible;

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

$rank(X) = k \implies (X'X)$ is p.d. (therefore invertible)

Math Review

- Suppose we have n random variables, $Y_1, \ldots, Y_n$. Then, define a random vector

$$Y := \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \text{and then we have } E[Y] := \begin{pmatrix} E[Y_1] \\ \vdots \\ E[Y_n] \end{pmatrix}$$

- Also $V(Y) := E[(Y - E[Y])(Y - E[Y])']$.
- (i) The $(i,j)^{\text{th}}$ element of $V(Y)$ is $C(Y_i, Y_j)$.
  (ii) The i^{th} diagonal element of $V(Y)$ is $C(Y_i, Y_i) = V(Y_i)$.
  (iii) $V(Y)$ is symmetric because $C(Y_i, Y_j) = C(Y_j, Y_i)$,

Math Review

- Let X be another random variable. Then,

$$E[Y|X] := \begin{pmatrix} E[Y_1|X] \\ \vdots \\ E[Y_n|X] \end{pmatrix}$$

and

$$V(Y|X) := E\left[(Y - E[Y|X])(Y - E[Y|X])'|X\right]$$