TA Session of Econometrics 2 (2020-2021)

Hiroki Kato Pang Kan

Contents

1 About TA Session

- Class schedule: Friday pm 13:30-15:00 via zoom.
  - You can access the meeting ID and its pascode via CLE.
- Instructor (If you have any question, please contact us via e-mail)
  1. Hiroki Kato (D2, vge008kh@student.econ.osaka-u.ac.jp)
  2. Pang Kan (D1, member_1363710747@yahoo.co.jp)
- Purpose: We will review the contents of the main class "Econometics II." using R which is a free software environment for statistical computing.
  - We strongly recommend that you download R (https://www.r-project.org/) and its IDE called R studio (https://rstudio.com/products/rstudio/download/), and try to reproduct by yourself.

2 Reviews of Matrix Algebra and Probability

2.1 Matrix Algebra

2.1.1 Addition and Subtraction

Consider $A = (a_{ij}) \in \mathbb{R}^{m \times n}$ and $B = (b_{ij}) \in \mathbb{R}^{p \times q}$. Addition and subtraction require that the dimentions are same, that is, $m = p$ and $n = q$. Then, the sum of two matricies is

$$A + B = (a_{ij} + b_{ij}).$$

The difference of two matricies is

$$A - B = (a_{ij} - b_{ij}).$$

2.1.2 Multiplication

The standard matrix multiplication requires that the number of columns of the first matrix is equal to the number of rows of the second matrix ($n = p = l$). The product of two matricies is

$$AB = (\sum_{k=1}^{l} a_{ik} b_{kj}) \in \mathbb{R}^{m \times q}.$$

We should remark following important points about multiplication:

- it holds non-commutativity: $XY \neq YX$;
- it holds associative law: $(XY)Z = X(YZ)$;
- it holds distributive law: $X(Y + Z) = XY + XZ$;
- when $B = A$, we obtain the second power of a matrix A, that is, $A^2 = AA$. Especially, if a martix A holds $AA = A$, then the matrix is called an **idempotent matrix (べき等行列)**.

We introduce the another key product of matrix, called the **Kronecker product (クロネッカー積)**. This is defined by

$$A \otimes B = (a_{ij} B) \in \mathbb{R}^{mp \times nq}.$$

The Kronecker product has a following property:

- $X_1 X_2 \otimes Y_1 Y_2 = (X_1 \otimes Y_1)(X_2 \otimes Y_2)$

2.1.3 Transposed Matrix, Diagonal Matrix, and Inverse Matrix

Consider $X = (x_{ij}) \in \mathbb{R}^{m \times n}$ throughout this subsection.

2.1.3.1 Transposed Matrix The **transposed matrix (転置行列)** of X, denoted by X' is a $n \times m$ matrix whose element x'_{ij} holds

$$x'_{ij} = x_{ji}.$$

That is, i-th row and j-th column element of transposed matrix is j-th row and i-th column element of original matrix. We remark following important points:

- it holds $(XY)' = Y'X'$;
- it holds $(XYZ)' = Z'Y'X'$;
- $(X \otimes Y)' = X' \otimes Y'$;
- let $x_i = (x_{i1}, ... x_{ij})$ be a row vector of matrix X. Then, we have $X'X = \sum_{n=1}^{i} x'_n x_n$;
- if a matrix X holds $X' = X$, then the matrix is called a **symmetric matrix (対称行列)**.

2.1.3.2 Diagonal Matrix and Trace Suppose a matrix X is a **square matrix (正方行列)**, that is, $n = m$. The matrix X is called a **diagonal matrix (対角行列)** whose diagonal elements (i-th row and i-th column elements) consist of $(x_{11}, ... x_{nn})$, and other elements are zero. That is,

$$X = diag(x_{11}, x_{22}, ..., x_{ii}) = \begin{pmatrix} x_{11} & 0 & 0 & \cdots & 0 \\ 0 & x_{22} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & x_{nn} \end{pmatrix}.$$

Especially, a matrix $I = diag(1, 1, ..., 1)$ is called an **identity matrix (単位行列)**.

There is one important concept, called **trace (トレース)**, related with diagonal elements of matrix. The trace of matrix is derived by the sum of diagonal elements, that is,

$$tr(X) = \sum_{n=1}^{i} x_{nn}.$$

The trace has following properties:

- $tr(cX) = c \cdot tr(X)$ where c is scalar;
- $tr(X') = tr(X)$;
- $tr(X + Y) = tr(X) + tr(Y)$;
- $tr(XY) = tr(YX)$;
- $xx' = tr(x'x) = tr(xx')$ if x is a $1 \times j$ vector.

2.1.3.3 Inverse Matrix the matrix X is **regular matrix (正則行列)** if there exists a matrix Y such that

$$XY = I,$$

where I is an identity matrix. In this case, the matrix Y is called an **inverse matrix (逆行列)**, which is denoted by X^{-1}. The inverse matrix has following important properties:

- $(X^{-1})' = (X')^{-1}$;
- $(X \otimes Y)^{-1} = X^{-1} \otimes Y^{-1}$ if the inverse exists;
- $(X \otimes Y)(X^{-1} \otimes Y^{-1}) = I$

2.1.4 Quadratic Forms

Consider a symmetric and square matrix $A \in \mathbb{R}^{n \times n}$ and a vector $x \in \mathbb{R}^{n \times 1}$. Then, the quadratic form is written as

$$Q = x'Ax.$$

For example, consider $x = (x, y)'$ and A is a 2×2 matrix whose elements is one. Then, the quadratic form is

$$Q = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x+y & x+y \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = x^2 + 2xy + y^2 = (x+y)^2.$$

In this case, for any non-zero x and y, Q takes non-negative value. Then, we call the matrix A *positive semidefinite*. The definiteness of matrix is defined as follows:

- If $x'Ax > 0$ for all nonzero x, then A is **positive definite (正値定符号)**.
- If $x'Ax < 0$ for all nonzero x, then A is **negative definite (負値定符号)**.
- If $x'Ax \geq 0$ for all nonzero x, then A is **positive semidefinite (半正値定符号)**.
- If $x'Ax \leq 0$ for all nonzero x, then A is **negative semidefinite(半負値定符号)**.

2.1.4.1 Characteristic Roots and Characteristic Vectors Before describing useful theorm to check definiteness easily, we have to introduce two concepts: **characteristic roots (固有根)** and **characteristic vectors (固有ベクトル)**.

> If a scalar λ and a vector $c \in \mathbb{R}^{k \times 1}$, which is normalized as $c'c = 1$, satisfy the following equation, then they are called as the **characteristic root** and the **characteristic vector**, respectively;
>
> $$Ac = \lambda c \Leftrightarrow (A - \lambda I)c = 0,$$
>
> where I is an identity matrix.

These $(\lambda_1, \lambda_2, ..., \lambda_k)$ correspond to characteristic vectors $(c_1, c_2, ..., c_k)$. There is the following useful theorm that states the relationship between characteristic roots and definiteness:

> Let A be a symmetric matrix.
>
> 1. If all the characteristic roots of A are positive (negative), then A is positive definite (negative definite).
> 2. If some of roots are zero, then A is positive (negative) semidefinite if the reminder are positive (negative).
> 3. If A has both negative and positive roots, then A is indefinite.

2.1.4.2 Determinants Alternative way to check definiteness of matrix is to using **determinants (行列式)**, which is a scalar quantity defined by a square matrix $A = (a_{ij}) \in \mathbb{R}^{n \times n}$. The determinant of 2×2 matrix, i.e., $n = 2$, is obtained by

$$det(A) = a_{11}a_{22} - a_{12}a_{21}.$$

Using the determinant of a matrix with $n = 2$, we can calculate the determinant of 3×3 matrix. Let $det(A_{ij})$ be the determinant of the 2×2 submatrix obtained when i-th row and j-th column are removed from the original matrix, which is called **minor (小行列式)**. Furthermore, we define $C_{ij} = (-1)^{i+j} det(A_{ij})$, which is called **cofactor (余因子)**. We call a matrix in which each element a_{ij} is replaced by the corresponding cofactor C_{ij} **cofactor matrix (余因子行列)**. Then, the determinant of 3×3 matrix is

$$\begin{aligned} det(A) &= a_{11}C_{11} + a_{12}C_{12} + a_{13}C_{13} \\ &= a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12}\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13}\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}, \end{aligned}$$

or

$$det(A) = a_{11}C_{11} + a_{21}C_{21} + a_{31}C_{31}.$$

How about matrices with $n \geq 4$? Essentially, we can calculate the determinant, using cofactors. That is, for any i,

$$det(A) = \sum_{k=1}^{n} a_{ik}C_{ik},$$

or, for any j

$$det(A) = \sum_{k=1}^{n} a_{kj}C_{kj},$$

Before describing the important theorem to check definiteness, we introduce the *Cramer's rule*, which provides inverse matrices.

> Let $A \in \mathbb{R}^{n \times n}$ be a square matrix with $det(A) \neq 0$. Then, the inverse matrix of A is equal to the transposed cofactor matrix multiplied by $det(A)^{-1}$. That is,
>
> $$A^{-1} = \frac{1}{det(A)} \begin{pmatrix} C_{11} & \cdots & C_{n1} \\ \vdots & \vdots & \vdots \\ C_{1n} & \cdots & C_{nn} \end{pmatrix}.$$

Next, we introduce the useful theorem to check definiteness of matricies. The determinant has the follwoing relation with the definiteness of matricies.

Let $A \in \mathbb{R}^{n \times n}$ be a square matrix. Let

$$det(A_i) = \begin{vmatrix} a_{11} & \cdots & a_{1i} \\ \vdots & \vdots & \vdots \\ a_{i1} & \cdots & a_{ii} \end{vmatrix}.$$

A necessary and sufficient condition for a matrix A to be positive definite is that $det(A_i) > 0$ for all $i \in \{1, ..., n\}$. Moreover, a necessary sufficient condition for a matrix A to be negative definite is that $det(A_i) < 0$ for odd i and $det(A_i) > 0$ for even i.

As an illustration, consider the following matrix:

$$A = \begin{pmatrix} 6 & 4 \\ 4 & 5 \end{pmatrix}.$$

Then, we have $det(A_1) = 6 > 0$ and $det(A_2) = 30 - 16 > 0$. Thus, this matrix is positive definite.

To show another example, we use the following diagonal matrix:

$$A = \begin{pmatrix} -3 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

Since the determinant of a diagonal matrix can be computed as the product of diagonal elements, we have $det(A_1) = -3$, $det(A_2) = (-3)(-2) = 6 > 0$, and $det(A_3) = (-3)(-2)(-1) = -6 < 0$. Thus, this diagonal matrix is negative definite. Moreover, the inverse matrix of this diagonal matrix is given by

$$A^{-1} = -\frac{1}{6} \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 6 \end{pmatrix} = \begin{pmatrix} -1/3 & 0 & 0 \\ 0 & -1/2 & \\ 0 & 0 & -1 \end{pmatrix}$$

Cleary, we have $AA^{-1} = I$.

2.1.5 Differentiation

Consider two vectors: $a \in \mathbb{R}^{n \times 1}$ and $x \in \mathbb{R}^{n \times 1}$. We obtain the product of transposed vector of a and x, that is, $a'x = a_1x_1 + \cdots + a_nx_n$. Then, the differentiation of this scalar with respect to x is defined by

$$\frac{\partial a'x}{\partial x} = \begin{pmatrix} \frac{\partial a'x}{\partial x_1} \\ \vdots \\ \frac{\partial a'x}{\partial x_n} \end{pmatrix} = a.$$

Now, we expand to a symmetric and square matrix $A \in \mathbb{R}^{n \times n}$. Then, the differentiation of the quadratic form $x'Ax$ with respect to x is defined by

$$\frac{\partial x'Ax}{\partial x} = (A + A')x.$$

2.1.5.1 Optimization Consider function $y = g(x)$ where $x \in \mathbb{R}^n$, denoted as $g : \mathbb{R}^n \to \mathbb{R}$. We can obtain x^0 such that maximizing (minimizing) the function g, using the following theorem:

> If a function $g : \mathbb{R}^n \to \mathbb{R}$ is maximized (minimized) at the point $x^0 = (x_1^0, ..., x_n^0)$, then the following equation holds:
>
> $$\frac{\partial g(x)}{\partial x}\Big|_{x=x^0} = \begin{pmatrix} \frac{\partial g(x^0)}{\partial x_1} \\ \vdots \\ \frac{\partial g(x^0)}{\partial x_n} \end{pmatrix} = 0.$$
>
> x^0 is maximum (minimum) point if the following **Hessian matrix (ヘッセ行列)** is negative (positive) definite:
>
> $$H = \frac{\partial g(x)}{\partial x x'} = \begin{pmatrix} \frac{\partial^2 g(x)}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 g(x)}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 g(x)}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 g(x)}{\partial x_n \partial x_n} \end{pmatrix}.$$

2.2 Probability

This section refers to Wasserman (2013). Let Ω be a **(sample) space** which is the set of possible outcomes of an experiment. Let ω be **sample outcomes**, **reaizations**, or **elements**. Let A be **events** which are the subsets of Ω. Then, we can define the **probability** and **random variable** as follows:

> A function $\mathbb{P}$ that assigns a real number $\mathbb{P}(A)$ to each event A is a **probability** if it satisfies the following three axioms:
>
> 1. $\mathbb{P}(A) \geq 0$ for all A;
> 2. $\mathbb{P}(\Omega) = 1$;
> 3. If $A_1, A_2, ...$ are disjoint, then $\mathbb{P}(\bigcup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} \mathbb{P}(A_i)$.
>
> A **random variable** is a mapping $X : \Omega \to \mathbb{R}$ that assigns a real number $X(\omega)$ to each realization ω.

For illustlation, consider the situation where you try to flip a fair coin twice. Then, the sample space $\Omega = \{TT, HH, TH, HT\}$. The probability of each outcome is 1/4, that is, $P(\omega) = 1/4$ for all ω. Let the random variable X be the number of heads. Then, $X(TT) = 0$, $X(HH) = 2$, $X(TH) = 1$, and $X(HT) = 1$.

2.2.1 Distribution Functions

Given a random variable X, we define **probability mass function**, **probability density function** and **cumulative distribution function** as follows:

> Suppose that a random variable X is *discrete* taking countably many values $\{x_1, ...\}$. Then, the **probability mass function** for X is defined by $f_X(x) = \mathbb{P}(X = x)$.

Suppose that a random variable X is *continuous*. Then, there exists a **probability density function** f_X such that (i) $f_X(x) \geq 0$ for all x, (ii) $\int_{-\infty}^{+\infty} f_X(x)dx = 1$ and (iii) for every $a \leq b$, $\mathbb{P}(a < X < b) = \int_a^b f_X(x)dx$.

The **cumulative distribution function (CDF)** is the function $F_X : \mathbb{R} \to [0, 1]$ defined by $F_X(x) = \mathbb{P}(X \leq x)$.

We summarize the relationship among three distribution functions as follows:

$$F_X(x) = \begin{cases} \sum_{x_i \leq x} f_X(x_i) & X \text{ is discrete} \\ \int_{-\infty}^{x} f_X(t)dt & X \text{ is continuous} \end{cases}$$

From this, we obtain the property of CDF:

- F is non-decreasing: $x_1 < x_2$ implies $F(x_1) \leq F(x_2)$;
- F is normalized: $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$;
- F is right-contious: $F(x) = F(x^+)$ for all x, where $F(x^+) = \lim_{y \downarrow x} F(y)$;
- $\mathbb{P}(X = x) = F(x) - F(x^-)$ where $F(x^-) = \lim_{y \uparrow x} F(y)$;
- $\mathbb{P}(x < X \leq y) = F(y) - F(x)$;
- $\mathbb{P}(X > x) = 1 - F(x)$.

2.2.1.1 Bivariate Distributions When there are two random varibales, you can define bivariate distributions as follows:

Given discrete random variables X and Y, we define the **joint mass function** by $f_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$.

In the continuous case, we call a function $f(x, y)$ a **joint probability density function** for the random variables (X, Y) if (i) $f(x, y) \geq 0$, $\forall (x, y)$, (ii) $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y)dxdy = 1$, and, (iii) for any $A \subset \mathbb{R} \times \mathbb{R}$, $\mathbb{P}[(X, Y) \in A] = \int \int_A f(x, y)dxdy$.

In both cases, we define the **joint cumulative distribution function** as $F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y)$.

When you are interested to the probability of a single event occurring, there are two distributions called a **marginal distribution (周辺分布)** and a **conditional distribution (条件付き分布)**. The formar distribution is the probability of $X = x$ independent of Y. On the other hand, the latter distribution is the probability that $X = x$ occurs given the event $Y = y$ has already occured. Formally,

The **marginal distribution** is defined by

$$f_X(x) = \mathbb{P}(X = x) = \begin{cases} \sum_y \mathbb{P}(X = x, Y = y) & X \text{ is discrete} \\ \int f(x, y)dy & X \text{ is continuous} \end{cases}.$$

The **conditional distribution** is defined by

$$f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \begin{cases} \sum_y \frac{\mathbb{P}(X=x,Y=y)}{\mathbb{P}(Y=y)} & X \text{ is discrete} \\ \int \frac{f_{X,Y}(x,y)}{f_Y(y)} dy & X \text{ is continuous} \end{cases},$$

To define the conditional distribution function, we assume $\mathbb{P}(Y = y) > 0$ for discrete random variables and $f_Y(y) > 0$ for continuous random variables. In the case of continuous random variables, we must integrate to get a probability, that is,

$$\mathbb{P}(X \in A|Y = y) = \int_A f_{X|Y}(x|y)dx.$$

Finally, we introduce very important concept of probability, called **independence**, and define it as follows:

> Two random variables X and Y are **independent** if, for every event A and B, $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$.

In principle, to check independence, we need to check whether this relationship for all subsets A and B. But, there is useful theorem to check independence.

> Let X and Y have joint PDF $f_{X,Y}$. Then, X and Y are independent if and only if $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ for all values x and y.

Note that if two random variables are independent, then the conditional probability $\mathbb{P}(X = x|Y = y)$ reduces to $\mathbb{P}(X = x)$.

2.2.2 Expectations, Variance, and Covariance

Roughly speaking, the expectation of a random variable X is the average value of X. The variance of a random variable X measures the "spread" of a distribution. Formally,

> The **expected value (mean, first moment)** of X is defined to be
>
> $$E(X) = \mu_X = \int x dF(x) = \begin{cases} \sum_x x f(x) & X \text{ is discrete} \\ \int x f(x) dx & X \text{ is continuous} \end{cases}$$
>
> The **variance** of X is defined by $V(X) = \sigma^2 = E(X-\mu_X)^2 = \int (x-\mu_X)^2 dF(x)$.
> The **standard deviation** is $sd(X) = \sqrt{V(X)}$.

The mean of random variable, $E(X)$, exists if $\int_x |x| dF_X(x) < \infty$. Expectation and variance has some useful properties:

- Let $Y = r(X)$. Then, $E(Y) = E(r(X)) = \int r(x)dF(x)$;
- Suppose that $X_1, \ldots, X_n$ are random variables and $a_1, \ldots, a_n$ are constants. $E(\sum_i a_i X_i) = \sum_i a_i E(X_i)$;
- Suppose that $X_1, \ldots, X_n$ are independent random variables. Then, $E(\prod_{i=1}^n X_i) = \prod_i E(X_i)$;

- $V(X) = E(X^2) - \mu^2$;
- $V(aX + b) = a^2V(X)$ where a and b are constants;
- Suppose that $X_1, \dots, X_n$ are independent random variables and $a_1, \dots, a_n$ are constants. $V(\sum_i a_i X_i) = \sum_i a_i^2 V(X_i)$.
- If a is a vector and X is a random vector with mean μ and variance Σ, then $E(a'X) = a'\mu$ and $V(a'X) = a'\Sigma a$. If A is a matrix, then $E(AX) = A\mu$ and $V(AX) = A\Sigma A'$.

We introduce some theorems about probability inequalities which is used in the theory of convergence.

- Markov's inequality: Let X be a non-negative random variable and suppose that $E(X)$ exists. For any $t > 0$, $\mathbb{P}(X > t) \leq E(X)/t$.
- Chebyshev's inequality: Let $\mu = E(X)$ and $\sigma^2 = V(X)$. Then, $\mathbb{P}(|X-\mu| \geq t) \leq \sigma^2/t^2$ and $\mathbb{P}(|Z| \geq k) \leq 1/k^2$ where $Z = (X - \mu)/\sigma$.
- Jensen's inequality: If g is convex, then $E[g(X)] \geq g(E(X))$. If g is concave, then $E[g(X)] \leq g(E(X))$.

Next, we will introduce the definition of **covariance**. These measure how strong the linear relationship is between X and Y.

> Let X and Y be random variables with means μ_X and μ_Y, respectively. Then, the **covariance** between X and Y is defined by $\text{Cov}(X, Y) = E[(X-\mu_X)(Y-\mu_Y)]$.

Covariance has following properties:

- $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$;
- If X and Y are independent, $\text{Cov}(X, Y) = 0$. **The converse is not true.**
- $V(X + Y) = V(X) + V(Y) + 2\text{Cov}(X, Y)$
- $V(X - Y) = V(X) + V(Y) - 2\text{Cov}(X, Y)$
- $V(\sum_i a_i X_i) = \sum_i a_i^2 V(X_i) + 2\sum_i \sum_{i<j} a_i a_j \text{Cov}(X_i, X_j)$.

2.2.2.1 Conditional Expectation and Variance

The conditional expectation of X given $Y = y$ is

$$E(X|Y = y) = \begin{cases} \sum_x x f_{X|Y}(x|y) & X \text{ is discrete} \\ \int x f_{X|Y}(x|y)dx & X \text{ is continuous} \end{cases}$$

The conditional variance is defined as

$$V(X|Y = y) = \int (x - \mu(y))^2 f(x|y)dx.$$

where $\mu(y) = E(X|Y = y)$.

Even if $r(x, y)$ is a function of x and y, we can define the conditional expectation. For the continuous random variable, $E[r(X, Y)|Y = y] = \int r(x, y) f_{X|Y}(x|y)dx$. For the discrete random variable, $E[r(X, Y)|Y = y] = \sum_x r(x, y) f_{X|Y}(x|y)$.

We have following important properties:

- For random variables X and Y, assuming the expectations exist, we have that $E_X[E(Y|X)] = E(Y)$. More generally, for any function $r(x, y)$, we have $E_X[E(r(X, Y)|X)] = E(r(X, Y))$.
- For random variables X and Y, $V(Y) = E_X[V(Y|X)] + V_X[E(Y|X)]$.

2.2.2.2 Moment The expectation of a random variable X to the k-th power, $E(X^k)$ is called k-th **moment** of X. If the k-th moment exists and if $j < k$, then j-th moment exists. Now, we define the **moment generating function** which is used for finding moments.

> The **moment generating function** of X is defined by
>
> $$\psi_X(t) = E(e^{tX}) = \int e^{tX} dF(x),$$
>
> where t varies over the real numbers.

Now, we assume that the moment generating function is well defined for all t in some open interval around $t = 0$. Then, we can interchange the operations of differentiation and taking expectation. Thus, we obtain

$$\psi'(0) = \frac{d}{dt} E(e^{tX})|_{t=0} = E\left(\frac{d}{dt} e^{tX}\right)|_{t=0} = E(Xe^{tX})|_{t=0} = E(X).$$

This implies that the mean of random variable is derived by taking first-order derivatives at $t = 0$. Thus, we can conclude that $\psi^{(k)}(0) = E(X^k)$. We should remark properties of the moment generating function.

- If $Y = aX + b$, then $\psi_Y(t) = e^{bt}\psi_X(at)$.
- If $X_1, ..., X_n$ are independent and $Y = \sum_i X_i$, then $\psi_Y(t) = \prod_i \psi_i(t)$ where ψ_i is the moment generating function of X_i.
- Let X and Y be random variables. If $\psi_X(t) = \psi_Y(t)$ for all t in an open interval around 0, then X and Y have the same distribution function.

2.2.3 Convergence

When we are interested in what happens as we gather more and more data, we need to concern the limiting behavior of a sequence of random variables. This part of probability is called **large sample theory** or **asymptotic theory (漸近理論)**. First, we will define two types of convergence as follows:

> Let $X_1, X_2, ...$ be a sequence of random variables and let X be another random variable. Let F_n denote the cumulative distribution function (CDF) of X_n and let F denote the CDF of X.
>
> 1. X_n **converges to X in probability**, written $X_n \xrightarrow{p} X$, if for every $\epsilon > 0$, $\mathbb{P}(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$.
> 2. X_n **converges to X in distribution**, written $X_n \xrightarrow{d} X$, if $\lim_{n\to\infty} F_n(t) = F(t)$ for all t for which F is continuous.

We should remark the relationship bewteen two types of convergence and properties of each type of convergence. Especially, the property 4 and 6 are called the **Slutsky theorem**, and the property 7 and 8 are called the **continuous mapping theorem**.

1. $X_n \xrightarrow{p} X$ implies that $X_n \xrightarrow{d} X$
2. If $X_n \xrightarrow{d} X$ and $\mathbb{P}(X = c) = 1$ for some real number c, then $X_n \xrightarrow{p} X$
3. If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, then $X_n + Y_n \xrightarrow{p} X + Y$
4. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$, then $X_n + Y_n \xrightarrow{d} X + c$
5. If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, then $X_n Y_n \xrightarrow{p} XY$
6. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$, then $X_n Y_n \xrightarrow{d} cX$
7. If $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$
8. If $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$

2.2.3.1 Law of Large Numbers The first important theorem of asymptotic theory is the **(weak) law of large numbers**. This theorem says that the mean of a large sample is close to the mean of distribution. Now, we will state more precisely.

> Let $X_1, X_2, ...$ be an IID sample. Suppose that $\mu = E(X_i)$ for all i and $\sigma^2 = V(X_i)$ for all i. The sample mean is defined by $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$. Then, $\bar{X}_n \xrightarrow{p} \mu$.

As an illustlation, consider a situation where you flip a fair coin toss n times. The space is $\Omega = \{H, T\}$. The random variable X_i is the number of heads, that is, $X_i(H) = 1$ and $X_i(T) = 0$ for $i = 1, ..., n$, which is binomially distrbuted with one trial and probability 0.5, $B(1, 0.5)$ (Bernoulli distribution). The sample mean of this random variable represents the proportion of heads. WLLN says that the sample mean is close to 0.5 as n gets large.

We will simulate using R. First, the random variable of Bernoulli distribution is generated by `rbinom(n, size = 1, prob)` where `n` is the number of trials, `prob` is the probability of success (head). When you specify `size` is greater than one, this random variable indicated the number of sucess when you flip coin `size` times. We calculate the proportion of heads when $n = 1, ..., 20000$, and show line plot with logged number of trial on x-axis and the proportion of heads on y-axis.

```
set.seed(120504)

data <- data.frame(
  trial = 1:20000,
  success = rbinom(n = 20000, size = 1, prob = .5)
)
data$sum_success <- cumsum(data$success)
data$prob <- data$sum_success/data$trial

plot(
```

```
  log(data$trial), data$prob, type = "l", col = "blue",
  ylim = c(0.3, 0.7), xlab = "logged trials", ylab = "Pr(head)")
lines(c(0, 10), c(0.5, 0.5), lwd = 1, col = "red")
```

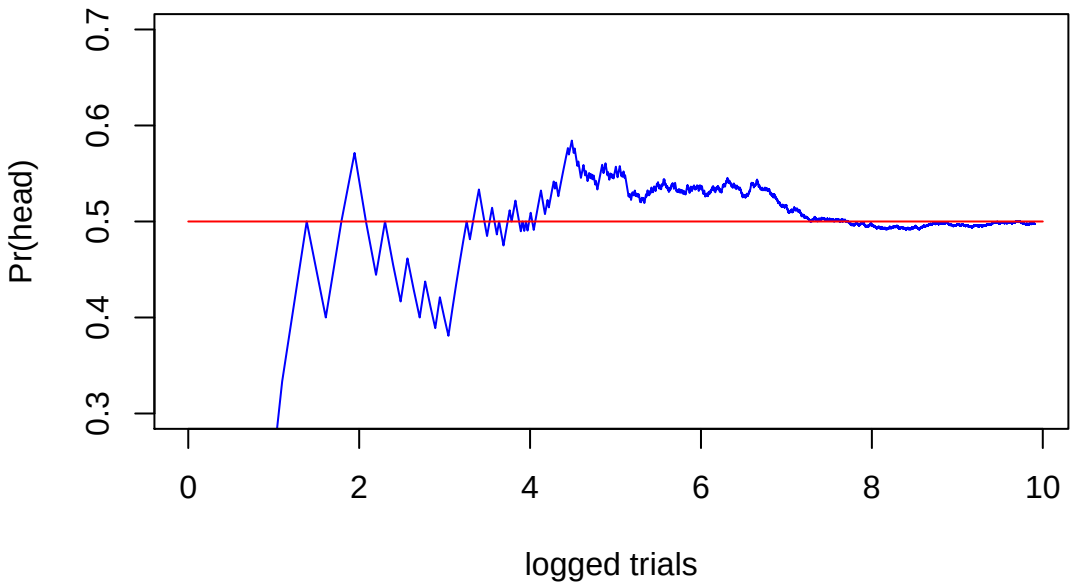

Figure 1: Simulation Result of WLLN

2.2.3.2 Central Limit Theorem The second important theorm is the **central limit theorem**. Suppose that $X_i, \ldots, X_n$ are IID sample with mean μ and variance σ^2. This theorem says that the sample mean $\bar{X}_n$ has a distribution which is approximately normal with mean μ and variance σ^2/n. This theorem does not assume the distribution of X_i, except the existence of the mean and variance. Formally,

> Let $X_1, \ldots, X_n$ be IID with mean μ and variance σ^2. Let $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$. Then,
>
> $$Z_n \equiv \frac{\bar{X}_n - \mu}{\sqrt{V(\bar{X}_n)}} = \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \xrightarrow{d} Z,$$
>
> where $Z \sim N(0, 1)$. In other words, $\bar{X}_n \xrightarrow{d} N(\mu, \sigma^2/n)$.

As an illustlation, consider a fair coin toss. The random variable is the number of heads. This random variable has the Bernoulli distribution with mean $\mu = 0.5$ and variance $\sigma^2 = 0.5(1-0.5) = 0.25$. Since we know μ and σ^2, we can calculate Z_n, using the sample mean $\bar{X}_n$.

We work this and plot its distribution, using R programming. We generate 10,000 sample means $\bar{X}_n$ for $n = 3, 5, 100, 1000$, and transform sample means to Z_n. To calculate Z_n, we use command `sqrt()`, which returns the saquare root value. Sometimes, this procedure is called Monte-Carlo simulation.

```
set.seed(120504)
m <- 10000; n <- c(3, 100, 1000); p <- 0.5
a <- seq(-4, 4, .01); b <- dnorm(a)

dt <- list("n = 3"=numeric(m), "n = 100"=numeric(m), "n = 1000"=numeric(m))
for (i in 1:3) {
  dt[[i]] <- rbinom(n = m, size = n[i], prob = p)
  dt[[i]] <- sqrt(n[i])*(dt[[i]]/n[i] - p)/sqrt(p*(1-p))
}

par(mfrow=c(2,2), mai = c(0.5, 0.5, 0.35, 0.35))
for (i in 1:3) {
  hist(dt[[i]], col = "grey", freq = FALSE,
    xlab = "", main = names(dt)[i], xlim = c(-4, 4))
  par(new = TRUE)
  plot(a, b, type = "l", col = "red", axes = FALSE,
    xlab = "", ylab = "", main = "")
}
```

3 Reviews of Ordinary Least Squares and Maximum Likelihood Estimation

This section refers to Johnston (1984) and Angrist and Pischke (2008). Consider the k-variables lienar regression model:

$$y_i = x_i\beta + u_i,$$

where $\beta = (\beta_0, \beta_1, \dots, \beta_k)'$ is a $k \times 1$ vector of regression coefficients, $x_i = (1, x_{i1}, \dots, x_{ik})$ is a $1 \times k$ vector of stochastic covariates, and u_i is the error term which is idependent and identically distributed (i.i.d.). Our parameter of interest is β.

3.1 Ordinary Least Squares Estimator (OLSE)

The **OLS estimators** are the value β such that minimizing the residual sums of squares, that is,

The **OLS estimators** $\hat{\beta}$ is defined by

$$\hat{\beta} \in \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i\beta)^2,$$

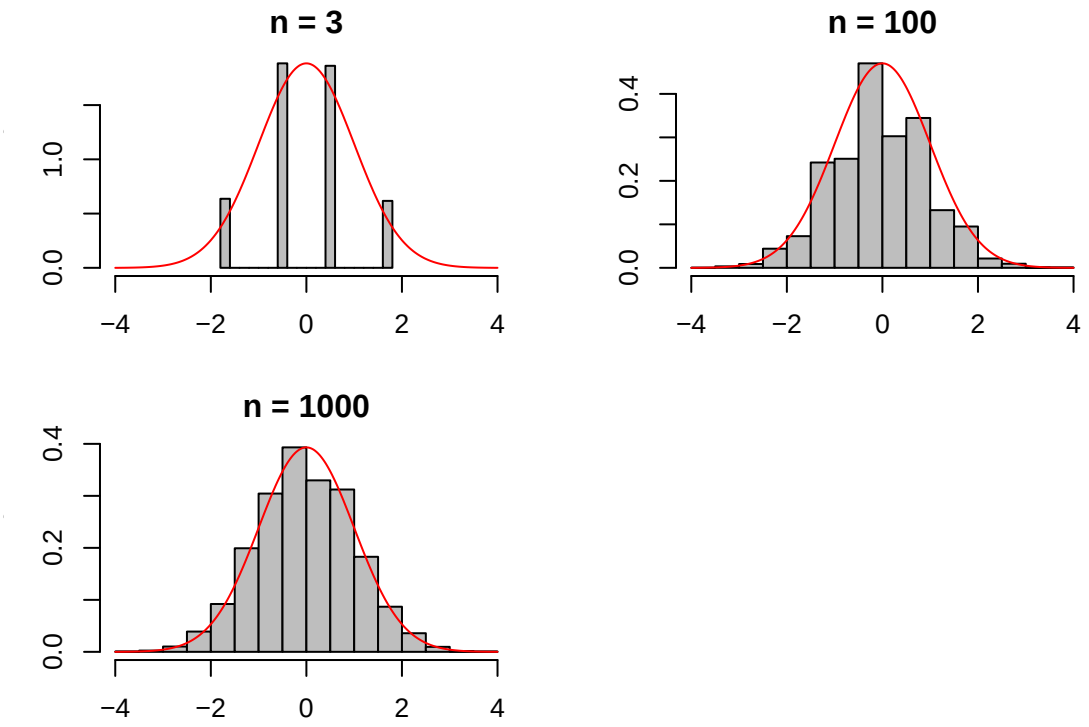

Figure 2: Simulation Result of CLT

or,

$$\hat{\beta} \in \arg\min_{\beta}(Y - X\beta)'(Y - X\beta),$$

where $Y = (y_1, \ldots, y_n)'$ is a $n \times 1$ vector, and $X = (x_1, \ldots, x_n)'$ is a $n \times k$.

Following this definition, the OLSE is given by

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

To exist the inverse matrix, we assume that the matrix $(X'X)$ is the regular matrix (i.e., there is no perfect correlation between any two covariates).

3.1.1 Best Linear Unbiased Estimator (BLUE)

We impose assumptions about the disturbance vector u: (i) $E(u|X) = 0$ (exogenity assumption or mean-idependence), and (ii) $V(u|X) = \sigma^2 I$ (homoscedasticity and pairwise uncorrelation). Under this condition, the OLS estimator is a linear unbiased estimator, that is, $E(\hat{\beta}) = \beta$ since

$$E(\hat{\beta}|X) = E[\beta + (X'X)^{-1}(X'u)|X] = \beta + (X'X)^{-1}X'E(u|X) = \beta.$$

Furthermore, the variance-covariance matrix of OLSE is

$$\begin{aligned} V(\hat{\beta}|X) &= E[(X'X)^{-1}X'uu'X(X'X)^{-1}|X] \\ &= (X'X)^{-1}X'\sigma^2 IX(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}. \end{aligned}$$

Note that $V(\hat{\beta}) = \sigma^2 E[(X'X)^{-1}]$. The most important result is that no other linear unbiased estimator can have smaller variances than those of OLSE. In other words, the OLSE has minimum variance within the class of linear unbiased estimators. Thus, the OLSE is a best linear unbiased estimator (**BLUE**). This result is known as the *Gauss-Markov theorem* (We omit proof).

3.1.2 Asymptotic Properties

First, the OLSE is a consistent estimator, that is,

$$\operatorname{plim}\hat{\beta} = \beta + \operatorname{plim}\left(\frac{1}{n}(X'X)\right)^{-1}\operatorname{plim}\left(\frac{1}{n}X'u\right) = \beta.$$

This is bacause $\operatorname{plim} n^{-1}(X'X) = \operatorname{plim} n^{-1}\sum_i x_i'x_i = E[x_i'x_i] = \Sigma$ and $\operatorname{plim} n^{-1}(X'u) = \operatorname{plim} n^{-1}\sum_i x_i'u_i = E[x_i'u_i] = 0$ by mean-independence assumption.

Second, the OLSE is asymptotically normally distributed. To show it, we derive the asymptotic distribution of $\sqrt{n}(\hat{\beta}-\beta)$ where

$$\sqrt{n}(\hat{\beta}-\beta) = \left(\frac{1}{n}\sum_i x_i'x_i\right)^{-1}\sqrt{n}\left(\frac{1}{n}\sum_i x_i'u_i\right).$$

By the central theorem, we have

$$\sqrt{n}\left(\frac{1}{n}\sum_i x_i'u_i\right) \xrightarrow{d} N(0,\sigma^2\Sigma).$$

Recall that $n^{-1}\sum_i x_i'x_i \xrightarrow{p} \Sigma$. By the Slutsky theorem (the 6th property of convergence), we get

$$\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{d} N(0,\sigma^2\Sigma^{-1}),$$

or,

$$\hat{\beta} \xrightarrow{d} N\left(\beta, \frac{1}{n}\sigma^2\Sigma^{-1}\right).$$

In a practical application, the unknown Σ is replaced by the sample estimate $n^{-1}X'X$, and the unknown σ^2 is estimated by $\hat{\sigma}^2 = \hat{u}'\hat{u}/(n-k)$ where $\hat{u} = Y - X\hat{\beta} = (I_n - X(X'X)^{-1}X')u = M_X u$ and M_X is a symmetric and idempotent matrix. Note that $\hat{\sigma}^2$ is an unbiased estimator of σ^2 since

$$E[\hat{\sigma}^2] = \frac{1}{n-k}E[tr(M_X uu')] = \frac{\sigma^2}{n-k}tr(M_X I_n) = \sigma^2.$$

3.1.3 Finite-sample Distribution and Inference

Now, we add the assumption with respect to the error term, $\epsilon_i|x_i \overset{iid}{\sim} N(0, \sigma^2)$. Then, we immediately obtain

$$\hat{\beta}|X \sim N(\beta, \sigma^2(X'X)^{-1}).$$

Consider the set of linear null hypothesis embodied in $R\beta = r$ where R is a arbitrary $q \times k$ matrix and r is a known q-element vector. To develop a test procedure, we derive the exact distribution of $R\hat{\beta}$. Cleary, we see $E(R\hat{\beta}) = R\beta$ and $V(R\hat{\beta}) = \sigma^2 R(X'X)^{-1}R'$. This leads to

$$R(\hat{\beta} - \beta) \sim N(0, \sigma^2 R(X'X)^{-1}R').$$

If the null hypothesis is true, then

$$R\hat{\beta} - r \sim N(0, \sigma^2 R(X'X)^{-1}R').$$

Using it and $\hat{u}'\hat{u} = u'M_X u$, we have follwing two distributions

$$\begin{aligned}
&(R\hat{\beta} - r)'[\sigma^2 R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r) \sim \chi^2(q), \\
&\frac{\hat{u}'\hat{u}}{\sigma^2} \sim \chi^2(n-k)
\end{aligned}$$

To derive these distributions, we use the following two properties about chi-squared distribution:

- If $x \sim N(0, \Sigma)$, then $x'\Sigma x \sim \chi^2(n)$ where x is n-element vector.
- If $x \sim N(0, \sigma^2 I)$ and A is idempotent matrix, then $(\sigma^2)^{-1}x'Ax \sim \chi^2(tr(A))$.

Finally, since $X_1 \sim \chi^2(d_1)$ and $X_2 \sim \chi^2(d_2)$ lead to $\frac{X_1}{d_1}/\frac{X_2}{d_2} \sim F(d_1, d_2)$, we have the distribution of test statistic, called the F-distribution,

$$\frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)/q}{\hat{u}'\hat{u}/(n-k)} \sim F(q, n-k).$$

The test procedure is to reject the null hypothesis $R\beta = r$ if the computed F-value exceeds a preselected cricial value.

Especially, when we test a single coefficient, we can use the t-value as an alternative test statistic. Suppose that $R = (0, 1, 0, ..., 0)$ and $r = 0$. The null hypothesis is $\hat{\beta}_2 = 0$. The matrix $R(X'X)^{-1}R'$ picks up the second diagonal element of $(X'X)^{-1}$ denoted by $(X'X)^{-1}_{22}$. Then, we have

$$\frac{\hat{\beta}_2^2}{\hat{\sigma}^2(X'X)^{-1}_{22}} \sim F(1, n-k).$$

Since $t \sim t(n)$ is equivalent to $t^2 \sim F(1,n)$ for any n, we finally obtain the test statistic following the Student's t-distribution

$$\frac{\hat{\beta}_2}{\hat{\sigma}\sqrt{(X'X)^{-1}_{22}}} \sim t(n-k).$$

When you use t-test of a single coefficient, you should *two-sided* t-test. If the computed t-statistic $\hat{t}$ holds $|\hat{t}| > t_{1-\alpha/2}(n-k)$ where $t_q(n-k)$ is the q-percentile t-value, then we can reject the null hyporhesis $\hat{\beta}_2 = 0$

3.2 Maximum Likelihood Estimator (MLE)

When we assume that the error term is normally distributed, we have $y_i|x_i \overset{iid}{\sim} N(x_i\beta, \sigma^2)$. Under this assumption, the estimator $\tilde{\beta}$ maximizing the log-likelihood function, called **maximum likelihood estimator**, is equivalent to the OLSE. The likelihood function is

$$\prod_{i=1}^n f(y_i, x_i) = \prod_{i=1}^n f_{Y|X}(y_i|x_i) \prod_{i=1}^n f_X(x_i) = \sum_{i=1}^n \log f_{Y|X}(y_i|x_i) + \sum_{i=1}^n \log f_X(x_i).$$

Since $f_X(x_i)$ does not involve the parameter vector β, the *conditional* MLE $\tilde{\beta}$ maximizes the conditional log-likelihood function $\sum_{i=1}^n \log f_i(y_i|x_i)$, that is,

$$\begin{aligned}
\log L(\theta) &= \sum_{i=1}^n \log f_{Y|X}(y_i|x_i) \\
&= \sum_{i=1}^n \log \left((2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(y_i - x_i\beta)^2}{2\sigma^2} \right) \right) \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta),
\end{aligned}$$

where $\theta = (\beta', \sigma^2)'$ is a $(k+1)\times 1$ vector of unknown parameters. The first-order derivatives of this function, sometimes called **score**, is given by

$$\frac{\partial \log L(\theta)}{\partial \theta} = \begin{pmatrix} -\frac{1}{2\sigma^2}(-2X'Y + 2X'X\beta) \\ -\frac{1}{2\sigma^2}\left(n - \frac{1}{\sigma^2}(Y - X\beta)'(Y - X\beta) \right) \end{pmatrix}.$$

The necessary condition of MLE is $\frac{\partial}{\partial\theta}\log L(\theta) = 0$. This leads to the following MLE:

$$\tilde{\beta} = (X'X)^{-1}(X'Y), \quad \tilde{\sigma}^2 = \frac{\hat{u}'\hat{u}}{n}.$$

The sufficient condition of MLE is the following Hessian matrix is negative definite.

$$H(\theta) = \begin{pmatrix} -\frac{1}{\sigma^2}X'X & \frac{1}{2\sigma^4}(-X'Y + X'X\beta) \\ \frac{1}{2\sigma^4}(-X'Y + X'X\beta) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}(Y - X\beta)'(Y - X\beta) \end{pmatrix}.$$

3.3 Properties of MLE

First, we provide the *Cramer-Rao theorem* that states that ML methods gives the lower bound of variance of unbiased estimators (proof is omitted).

> Let $\tilde{\theta}$ denote an unbiased estimator of θ. Then, $V(\tilde{\theta}) - I^{-1}(\theta)$ is a positive definite where $I(\theta)$ is a **Fisher information matrix**, which is defined by

$$I(\theta) = -E(H(\theta)) = -E\begin{pmatrix} \frac{\partial^2 \log L(\theta)}{\partial \theta_1^2} & \frac{\partial^2 \log L(\theta)}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \log L(\theta)}{\partial \theta_1 \partial \theta_k} \\ \frac{\partial^2 \log L(\theta)}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \log L(\theta)}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \log L(\theta)}{\partial \theta_2 \partial \theta_k} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 \log L(\theta)}{\partial \theta_k \partial \theta_1} & \frac{\partial^2 \log L(\theta)}{\partial \theta_k \partial \theta_2} & \cdots & \frac{\partial^2 \log L(\theta)}{\partial \theta_k^2} \end{pmatrix}.$$

Note that the Fisher information matrix conditional on some random variables also provides the Cramer-Rao lower bound. In the case of linear regression, the Cramer-Rao lower bound condtional on X gives

$$I^{-1}\begin{pmatrix} \beta \\ \sigma^2 \end{pmatrix} = \begin{pmatrix} \sigma^2 X'X^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}.$$

Although the ML estimator of β attains the Cramer-Rao lower bound, the ML estimator of σ^2 deviates.

Second, we summarize asymptotic properties of MLE as follows (proof is omitted):

> Under certain regularity conditions, (i) The ML estimator is consistent, i.e., $\tilde{\theta} \xrightarrow{p} \theta$, and (ii) The ML estimator is asymptotically normally distributed, i.e., $\tilde{\theta} \xrightarrow{d} N(\theta, I^{-1}(\theta))$

4 Reference

Angrist, Joshua D, and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton university press.

Johnston, J. 1984. *Econometric Methods 3rd*. McGraw-Hill book co.

Wasserman, Larry. 2013. *All of Statistics: A Concise Course in Statistical Inference*. Springer Science & Business Media.