Notes for Statistics

Julique

October 26, 2023

Contents

Chapter 1

Quick Review

1.1 Calculus

Omitted.

1.2 Matrix Algebra

1.2.1 Traces

Theorem 1.1. Properties of trace.

$$\begin{aligned}
\operatorname{tr}(c\boldsymbol{A}) &= c\operatorname{tr}(\boldsymbol{A}) \\
\operatorname{tr}\left(\boldsymbol{A}^{\mathrm{T}}\right) &= \operatorname{tr}(\boldsymbol{A}) \\
\operatorname{tr}(\boldsymbol{A}+\boldsymbol{B}) &= \operatorname{tr}(\boldsymbol{A}) + \operatorname{tr}(\boldsymbol{B}) \\
\operatorname{tr}(\boldsymbol{I}_k) &= k \\
\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) &= \operatorname{tr}(\boldsymbol{B}\boldsymbol{A})
\end{aligned}$$

1.2.2 Determinants

Theorem 1.2. If $\boldsymbol{A}$ is $k \times k$, then

$$\begin{aligned}
\det(\boldsymbol{A}) &= \det\left(\boldsymbol{A}^{\mathrm{T}}\right) \\
\det(c\boldsymbol{A}) &= c^k \det(\boldsymbol{A}) \\
\det(\boldsymbol{A}\boldsymbol{B}) &= \det(\boldsymbol{A})\det(\boldsymbol{B})
\end{aligned}$$

Moreover,

- If $\det \boldsymbol{A} \neq 0$, then $\det\left(\boldsymbol{A}^{-1}\right) = \det(\boldsymbol{A})^{-1}$.
- If $\det \boldsymbol{D} \neq 0$, then $\det \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} = \det(\boldsymbol{D}), \det\left(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\right)$

1.2.3 Inverse

Theorem 1.3. Properties of inverse. For nonsingular $\boldsymbol{A}$ and $\boldsymbol{B}$,

$$\begin{aligned}
\boldsymbol{A}^{-1} &= \frac{\operatorname{adj}(\boldsymbol{A})}{\det(\boldsymbol{A})} \\
\boldsymbol{A}\boldsymbol{A}^{-1} &= \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_k \\
\left(\boldsymbol{A}^{-1}\right)^{\mathrm{T}} &= \left(\boldsymbol{A}^{\mathrm{T}}\right)^{-1} \\
(\boldsymbol{A}\boldsymbol{B})^{-1} &= \boldsymbol{B}^{-1}\boldsymbol{A}^{-1} \\
(\boldsymbol{A}+\boldsymbol{B})^{-1} &= \boldsymbol{A}^{-1}\left(\boldsymbol{A}^{-1}+\boldsymbol{B}^{-1}\right)^{-1}\boldsymbol{B}^{-1} \\
\boldsymbol{A}^{-1}-(\boldsymbol{A}+\boldsymbol{B})^{-1} &= \boldsymbol{A}^{-1}\left(\boldsymbol{A}^{-1}+\boldsymbol{B}^{-1}\right)^{-1}\boldsymbol{A}^{-1}
\end{aligned}$$

Also, if $\boldsymbol{A}$ is an orthogonal matrix, then

$$\boldsymbol{A}^{-1} = \boldsymbol{A}^{\mathrm{T}}$$

Theorem 1.4. (**Woodbury matrix identity**) For a nonsingular $\boldsymbol{A}$,

$$(\boldsymbol{A}+\boldsymbol{B}\boldsymbol{C}\boldsymbol{D})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{C}\left(\boldsymbol{C}+\boldsymbol{C}\boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{C}\right)^{-1}\boldsymbol{C}\boldsymbol{D}\boldsymbol{A}^{-1}$$

In particular, for $\boldsymbol{C} = -1, \boldsymbol{B} = \boldsymbol{b}$ and $\boldsymbol{D} = \boldsymbol{b}^{\mathrm{T}}$ for vector $\boldsymbol{b}$ we find what is known as the **Sherman-Morrison formula**:

$$\left(\boldsymbol{A}-\boldsymbol{b}\boldsymbol{b}^{\mathrm{T}}\right)^{-1} = \boldsymbol{A}^{-1} + \left(1-\boldsymbol{b}^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{b}\right)^{-1}\boldsymbol{A}^{-1}\boldsymbol{b}\boldsymbol{b}^{\mathrm{T}}\boldsymbol{A}^{-1}$$

1.2.4 Eigenvalues

Definition 1.5. The **characteristic equation** of a $k \times k$ square matrix $\boldsymbol{A}$ is

$$\det(\boldsymbol{A} - \lambda\boldsymbol{I}_k) = 0$$

The left side is a polynomial of degree k in λ so it has exactly k roots, which are not necessarily distinct and may be real or complex. They are called the **latent roots** or **characteristic roots** or **eigenvalues** of $\boldsymbol{A}$.

If λ_i is an eigenvalue of $\boldsymbol{A}$, then $\boldsymbol{A} - \lambda_i\boldsymbol{I}_k$ is singular so there exists a non-zero vector h_i such that

$$(\boldsymbol{A} - \lambda_i\boldsymbol{I}_k)\,\boldsymbol{h}_i = \boldsymbol{0}$$

The vector $\boldsymbol{h}_i$ is called a **latent vector** or **characteristic vector** or **eigenvector** of $\boldsymbol{A}$ corresponding to λ_i.

Theorem 1.6. Let λ_i and $\boldsymbol{h}_i, i = 1, \ldots, k$ denote the k eigenvalues and eigenvectors of a square matrix $\boldsymbol{A}$. Let $\boldsymbol{\Lambda}$ be a diagonal matrix with the characteristic roots in the diagonal, and let $\boldsymbol{H} = [\boldsymbol{h}_1 \cdots \boldsymbol{h}_k]$.

- $\det(\boldsymbol{A}) = \prod_{i=1}^{k} \lambda_i$

- $\text{tr}(\boldsymbol{A}) = \sum_{i=1}^{k} \lambda_i$
- $\boldsymbol{A}$ is non-singular if and only if all its characteristic roots are non-zero.
- If $\boldsymbol{A}$ has distinct characteristic roots, there exists a nonsingular matrix $\boldsymbol{P}$ such that $\boldsymbol{A} = \boldsymbol{P}^{-1}\boldsymbol{\Lambda}\boldsymbol{P}$ and $\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1} = \boldsymbol{\Lambda}$.
- If $\boldsymbol{A}$ is symmetric, then $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ and $\boldsymbol{H}'\boldsymbol{A}\boldsymbol{H} = \boldsymbol{\Lambda}$, and the characteristic roots are all real. $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ is called the **spectral decomposition** of a matrix.

1.2.5 Positive Definiteness

Definition 1.7. A quadratic form on $\mathbb{R}^n$ is a real-valued function of the form

$$Q(x_1, x_2, \cdots, x_n) = \sum_{j=1}^{n} \sum_{i=1}^{j} a_{ij} x_i x_j$$

in which each term is a monomial of degree 2.

Remark 1.8. Each quadratic form Q can be represented by a symmetric matrix A, so that

$$Q(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} = (x_1, x_2, \cdots, x_n) \begin{pmatrix} a_{11} & \frac{1}{2}a_{12} & \cdots & \frac{1}{2}a_{1n} \\ \frac{1}{2}a_{12} & a_{22} & \cdots & \frac{1}{2}a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2}a_{1n} & \frac{1}{2}a_{2n} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Remark 1.9. At the point $\boldsymbol{x} = \boldsymbol{0}$, $Q(\boldsymbol{x}) = 0$. We may focus on the question of whether $\boldsymbol{x} = \boldsymbol{0}$ is a max, a min or neither of the quadratic forms, i.e., definiteness.

Definition 1.10. We say a square matrix $\boldsymbol{A}$ is

- **positive definite**: for any nonzero $\boldsymbol{x}$, $Q(\boldsymbol{x}) > 0$.

  Example: $Q(x) = x^2$.
- **negative definite**: for any nonzero $\boldsymbol{x}$, $Q(\boldsymbol{x}) < 0$.

  Example: $Q(x_1, x_2) = -x_1^2 - 3x_2^2$.
- **indefinite**: $\exists \boldsymbol{x}_1, \boldsymbol{x}_2, Q(\boldsymbol{x}_1) > 0, Q(\boldsymbol{x}_2) < 0$.

  Example: $Q(x_1, x_2) = -x_1^2 + x_2^2$.
- **positive semidefinite**: for any nonzero $\boldsymbol{x}$, $Q(\boldsymbol{x}) \geqslant 0$.

  Example: $Q(x_1, x_2) = (x_1 + x_2)^2 = x_1^2 + 2x_1x_2 + x_2^2$.

  Example: Covariance matrix is positive semidefinite.

Proof. Let $\forall \boldsymbol{y} = (y_1, y_2, \cdots, y_n)^{\mathrm{T}}$, then

$$\begin{aligned}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{y} &= \sum_{i=1}^{n}\sum_{j=1}^{n} y_i \mathrm{Cov}\left(X_i, X_j\right) y_j = \sum_{i=1}^{n}\sum_{j=1}^{n} y_i \mathbb{E}\left[\left(X_i - \mu_i\right)\left(X_j - \mu_j\right)\right] y_j \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}\left[y_i\left(X_i - \mu_i\right)\left(X_j - \mu_j\right) y_j\right]\end{aligned}$$

because y_i are real numbers. Moreover,

$$\begin{aligned}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{y} &= \mathbb{E}\left[\sum_{i=1}^{n} y_i\left(X_i - \mu_i\right)\sum_{j=1}^{n}\left(X_j - \mu_j\right) y_j\right] = \mathbb{E}\left[\sum_{i=1}^{n} y_i\left(X_i - \mu_i\right)\right]^2 \\ &= \mathrm{Var}\left[\sum_{i=1}^{n} y_i\left(X_i - \mu_i\right)\right] \geqslant 0\end{aligned}$$

Let $X_1 = X, X_2 = Y, X_3 = X + Y, y_1 = -1, y_2 = -1, y_3 = 1$, then $\exists \boldsymbol{y} \neq \boldsymbol{0}$, such that $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{y} = 0$. □

- **negative semidefinite**: for any nonzero $\boldsymbol{x}, Q\left(\boldsymbol{x}\right) \leqslant 0$.

  Example: $Q\left(x_1, x_2\right) = -\left(x_1 + x_2\right)^2$.

Remark 1.11. A matrix that is positive (negative) definite is automatically positive (negative) semidefinite.

Definition 1.12. Let $\boldsymbol{A}$ be a $n \times n$ matrix.

1. A $k \times k$ submatrix of $\boldsymbol{A}_{n\times n}$ formed by deleting $n - k$ columns and the same $n - k$ rows from $\boldsymbol{A}$ is called **principal submatrix** of $\boldsymbol{A}$.

2. The determinant of a $k \times k$ principal submatrix is called a **kth other principal minor**of $\boldsymbol{A}$.

   Example. $\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$, the second order principal minors of $\boldsymbol{A}$ are

   $$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix}, \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix}$$

3. The kth order principal submatrix of $\boldsymbol{A}_{n\times n}$ obtained by deleting the last $n - k$ rows and the last $n - k$ columns from $\boldsymbol{A}$ is called the **kth order leading principal submatrix** of $\boldsymbol{A}$, denoted by $\boldsymbol{A}_k$.

4. The determinant of the kth leading principal submatrix is called **kth order leading principal minor** of $\boldsymbol{A}$, denoted by $|\boldsymbol{A}_k|$.

Example. $\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$, all of the leading principal minors are

$$\boldsymbol{A}_1 = |a_{11}|, \boldsymbol{A}_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \boldsymbol{A}_3 = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

Theorem 1.13. Let $\boldsymbol{A}_{n\times n}$ be a symmetric matrix, then

- $\boldsymbol{A}$ is positive definite, if and only if $|\boldsymbol{A}_k| > 0, k = 1, 2, \cdots, n$.
- $\boldsymbol{A}$ is negative definite, if and only if

$$|\boldsymbol{A}_1| < 0, |\boldsymbol{A}_2| > 0, |\boldsymbol{A}_3| < 0, |\boldsymbol{A}_4| > 0 \cdots$$

- $\boldsymbol{A}$ is indefinite, if one of the 2 following cases happens, (LPMs are nonzero)
  1. $\exists k$ is even, $|\boldsymbol{A}_k| < 0$;
  2. $\exists k, j, k \neq j$ are odd, $|\boldsymbol{A}_k| < 0, |\boldsymbol{A}_j| > 0$.

Lemma 1.14. If $\boldsymbol{A}$ is positive definite, then $\boldsymbol{A}$ is nonsingular.

Proof. Suppose $\boldsymbol{A}$ is singular, then $\boldsymbol{Ax} = \boldsymbol{0}$ has a nonzero solution $\boldsymbol{x}$, then

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Ax} = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{0} = 0$$

a contradiction to positive definite.

□

Lemma 1.15. Suppose that $\boldsymbol{A}$ is a symmetric matrix and that $\boldsymbol{Q}$ is a nonsingular matrix. Then $\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{AQ}$ is a symmetric matrix and $\boldsymbol{A}$ is positive (negative) definite if and only if $\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{AQ}$ is positive (negative) definite.

Proof. Firstly, $\left(\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{AQ}\right)^{\mathrm{T}} = \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{Q} = \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{AQ}$, so $\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{Q}$ is symmetric. Secondly,

(**If**) $\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{AQ}$ is positive definite, then $\forall \boldsymbol{x} \neq \boldsymbol{0}$, since $\boldsymbol{Q}$ is nonsingular, then $\boldsymbol{Qx} = \boldsymbol{y}$ is nonzero

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{AQx} = (\boldsymbol{Qx})^{\mathrm{T}}\boldsymbol{AQx} = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{Ay} > 0$$

Therefore, $\boldsymbol{A}$ is positive definite.

(**Only if**) $\boldsymbol{A}$ is positive definite, then $\forall \boldsymbol{x} \neq \boldsymbol{0}$, $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{Ay} > 0$, let

$$\boldsymbol{y}^{\mathrm{T}}\boldsymbol{Ay} = (\boldsymbol{Qx})^{\mathrm{T}}\boldsymbol{AQx} = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{AQx} > 0$$

Therefore, $\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{AQ}$ is positive definite. □

Theorem 1.16. Let $\boldsymbol{A}_{n\times n}$ be a symmetric matrix, then

1. $\boldsymbol{A}$ is positive semidefinite if and only if every principal minor of $\boldsymbol{A}$ is nonnegative.

2. $\boldsymbol{A}$ is negative semidefinite if and only if every principal minor of odd orders is nonpositive and every principal minor of even orders is nonnegative.

Proof. (**Mathematical Induction**) For $n = 1$, if $\boldsymbol{A} = (a_{11})$ is positive definite$\Longleftrightarrow a_{11}x^2 > 0, x \neq 0 \Longleftrightarrow a_{11} > 0\Longleftrightarrow$all LPMs of $\boldsymbol{A}$ are positive. For $n = 2$, if $\boldsymbol{A} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$,

$$Q = ax_1^2 + 2bx_1x_2 + cx_2^2 = a\left(x_1 + \frac{b}{a}x_2\right)^2 + \frac{(ac - b^2)}{a}x_2^2 = |\boldsymbol{A}_1|\left(x_1 + \frac{b}{a}x_2\right)^2 + \frac{|\boldsymbol{A}_2|}{a}x_2^2$$

Then $Q > 0, \boldsymbol{x} \neq \boldsymbol{0} \Longleftrightarrow |\boldsymbol{A}_1| > 0, |\boldsymbol{A}_2| > 0$. Suppose that for n the theorem holds, then for $n + 1$ case, we have

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_n & \boldsymbol{a} \\ \boldsymbol{a}^{\mathrm{T}} & a_{n+1,n+1} \end{pmatrix}$$

And $\boldsymbol{A}_n$ is inversible. Let $d = a_{n+1,n+1} - \boldsymbol{a}^{\mathrm{T}}\left(\boldsymbol{A}_n\right)^{-1}\boldsymbol{a}$, then we can decompose $\boldsymbol{A}$ as

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{I}_n & \boldsymbol{0}_n \\ \left(\boldsymbol{A}_n^{-1}\boldsymbol{a}\right)^{\mathrm{T}} & 1 \end{pmatrix}\begin{pmatrix} \boldsymbol{A}_n & \boldsymbol{0}_n \\ \boldsymbol{0}_n^{\mathrm{T}} & d \end{pmatrix}\begin{pmatrix} \boldsymbol{I}_n & \left(\boldsymbol{A}_n^{-1}\boldsymbol{a}\right)^{\mathrm{T}} \\ \boldsymbol{0}_n & 1 \end{pmatrix} = \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{Q}$$

Then $\det \boldsymbol{Q} = \det \boldsymbol{Q}^{\mathrm{T}} = 1$, $\det \boldsymbol{A} = d \cdot \det \boldsymbol{A}_n$.

(**If**) If $\det \boldsymbol{A} > 0$, since $\det \boldsymbol{A}_n > 0$, we have $d > 0$, then $a_{n+1,n+1} > \boldsymbol{a}^{\mathrm{T}}\left(\boldsymbol{A}_n\right)^{-1}\boldsymbol{a}$. Let

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_n \\ x_{n+1} \end{pmatrix}$$

be an arbitrary $n + 1$- vector.

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_n^{\mathrm{T}} & x_{n+1} \end{pmatrix}\begin{pmatrix} \boldsymbol{A}_n & \boldsymbol{0}_n \\ \boldsymbol{0}_n^{\mathrm{T}} & d \end{pmatrix}\begin{pmatrix} \boldsymbol{x}_n \\ x_{n+1} \end{pmatrix} = \boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{A}_n\boldsymbol{x}_n + dx_{n+1}^2 > 0$$

Then $\boldsymbol{B}$ is positive definite, and so is $\boldsymbol{A}$.

(**Only if**) If $\boldsymbol{A}$ is positive definite, then $\boldsymbol{B}$ is also positive definite, we choose, $\boldsymbol{x} = \begin{pmatrix} \boldsymbol{0} \\ 1 \end{pmatrix}$,

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{x} = \begin{pmatrix} \boldsymbol{0}^{\mathrm{T}} & 1 \end{pmatrix}\begin{pmatrix} \boldsymbol{A}_n & \boldsymbol{0}_n \\ \boldsymbol{0}_n^{\mathrm{T}} & d \end{pmatrix}\begin{pmatrix} \boldsymbol{0} \\ 1 \end{pmatrix} = d > 0$$

Therefore, $\det \boldsymbol{A} > 0$. And we know all LPMs of $\boldsymbol{A}_n$ are positive, then we know all LPMs of $\boldsymbol{A}$ are positive.

□

Theorem 1.17. Properties of positive semi-definite matrices.

- For a positive semi-definite matrix $\boldsymbol{B}$, if $\exists \boldsymbol{G}$ such that $\boldsymbol{A} = \boldsymbol{G}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{G}$, then $\boldsymbol{A}$ is positive semi-definite.

  If $\boldsymbol{G}$ has full column rank and $\boldsymbol{B}$ is positive definite, then $\boldsymbol{A}$ is also positive definite.

- If $\boldsymbol{A}$ is positive definite, then $\boldsymbol{A}^{-1}$ exists, and $\boldsymbol{A}^{-1}$ is positive definite.
- $\boldsymbol{A}$ is also positive definite if and only if it is symmetric and all its eigenvalues are positive.
- The rank of $\boldsymbol{A}$ equals the number of strictly positive eigenvalues.

1.2.6 Matrix Calculus

Definition 1.18. Let $\boldsymbol{x}_{k\times 1} \in \mathbb{R}^k, y = f(\boldsymbol{x}) : \mathbb{R}^k \to \mathbb{R}$, then define the **derivative** of y w.r.t. $\boldsymbol{x}$ as

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) := \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\boldsymbol{x})}{\partial x_k} \end{bmatrix} \in \mathbb{R}^k$$

and

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}^{\mathrm{T}}} = \left[\frac{\partial f(\boldsymbol{x})}{\partial x_1}, \cdots, \frac{\partial f(\boldsymbol{x})}{\partial x_k}\right] \in \mathbb{R}^k$$

Let $\boldsymbol{x} \in \mathbb{R}^k, \boldsymbol{y} = f(\boldsymbol{x}) = (f_1, f_2, \cdots, f_p) : \mathbb{R}^k \to \mathbb{R}^p$, then define the **derivative** of $\boldsymbol{y}$ w.r.t. $\boldsymbol{x}$ as

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_2(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_q(\boldsymbol{x})}{\partial x_1} \\ \frac{\partial f_1(\boldsymbol{x})}{\partial x_2} & \frac{\partial f_2(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f_q(\boldsymbol{x})}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1(\boldsymbol{x})}{\partial x_k} & \frac{\partial f_2(\boldsymbol{x})}{\partial x_k} & \cdots & \frac{\partial f_q(\boldsymbol{x})}{\partial x_k} \end{bmatrix} \in \mathbb{R}^{k\times q}$$

Theorem 1.19. Let $\boldsymbol{y} = f(\boldsymbol{x}), \boldsymbol{z} = g(\boldsymbol{x})$, then

- Plus Rule:
$$\frac{\partial(\boldsymbol{y}+\boldsymbol{z})}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} + \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}}$$
- Product Rule:
$$\frac{\partial \boldsymbol{y}^{\mathrm{T}}\boldsymbol{z}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\boldsymbol{z} + \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}}\boldsymbol{y}$$
- Chain Rule:
$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} \cdot \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{y}}$$

Moreover, if $\boldsymbol{z} = f(\boldsymbol{y}), \boldsymbol{y} = g(\boldsymbol{X})$ where $\boldsymbol{X}$ is a matrix, then

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{X}_{ij}} = \operatorname{tr}\left(\left(\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{y}}\right)^{\mathrm{T}} \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{X}_{ij}}\right)$$

Theorem 1.20. Let $\boldsymbol{x}$ and $\boldsymbol{a}$ be k vectors, $\boldsymbol{A}$ be a $k \times k$ matrix.

$$\begin{aligned}
\frac{\partial \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}}{\partial \boldsymbol{x}} &= \frac{\partial \boldsymbol{x}^{\mathrm{T}}\boldsymbol{a}}{\partial \boldsymbol{x}} = \boldsymbol{a} \\
\frac{\partial \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} &= \boldsymbol{A}^{\mathrm{T}} \\
\frac{\partial \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}}{\partial \boldsymbol{x}} &= \frac{\partial \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}^{\mathrm{T}}} = \boldsymbol{A} \\
\frac{\partial \left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}\right)}{\partial \boldsymbol{x}} &= \left(\boldsymbol{A}^{\mathrm{T}} + \boldsymbol{A}\right) \cdot \boldsymbol{x} \\
\frac{\partial^2 \left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}\right)}{\partial \boldsymbol{x}\partial \boldsymbol{x}^{\mathrm{T}}} &= \boldsymbol{A} + \boldsymbol{A}^{\mathrm{T}}
\end{aligned}$$

If $\boldsymbol{A}$ is symmetric, then

$$\frac{\partial \left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}\right)}{\partial \boldsymbol{x}} = 2\boldsymbol{A}\boldsymbol{x}$$

Moreover,

$$\frac{\partial}{\partial \boldsymbol{A}} \mathrm{tr}\left(\boldsymbol{B}\boldsymbol{A}\right) = \boldsymbol{B}'$$

Proof. Note that

$$\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{a} = a_1 x_1 + \cdots + a_k x_k$$

then

$$\frac{\partial \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{x}^{\mathrm{T}}\boldsymbol{a}}{\partial \boldsymbol{x}} = \begin{pmatrix} a_1 \\ \vdots \\ a_k \end{pmatrix} = \boldsymbol{a}$$

Write

$$\boldsymbol{A} = (\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_k) \Longrightarrow \boldsymbol{A}\boldsymbol{x} = \boldsymbol{a}_1 x_1 + \cdots + \boldsymbol{a}_k x_k$$

$$\frac{\partial \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{A}^{\mathrm{T}}$$

then

$$\frac{\partial \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \begin{pmatrix} \boldsymbol{a}_1 \\ \vdots \\ \boldsymbol{a}_k \end{pmatrix} = \boldsymbol{A}^{\mathrm{T}}$$

Likewise, we can show that $\frac{\partial \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}^{\mathrm{T}}} = \boldsymbol{A}$.

For $\frac{\partial \left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}\right)}{\partial \boldsymbol{x}}$, note that $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}$, by the product rule,

$$\frac{\partial \left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}\right)}{\partial \boldsymbol{x}} = \frac{\partial \left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{I}_k\right)}{\partial \boldsymbol{x}} \cdot \boldsymbol{A}\boldsymbol{x} + \frac{\partial \left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\right)}{\partial \boldsymbol{x}} \cdot \boldsymbol{x} = \boldsymbol{I}_k\boldsymbol{A}\boldsymbol{x} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{x} = \left(\boldsymbol{A} + \boldsymbol{A}^{\mathrm{T}}\right)\boldsymbol{x}$$

And therefore,

$$\frac{\partial \left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}\right)}{\partial \boldsymbol{x}\partial \boldsymbol{x}^{\mathrm{T}}} = \frac{\partial}{\partial \boldsymbol{x}} \frac{\partial \left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}\right)}{\partial \boldsymbol{x}^{\mathrm{T}}} = \frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{x}^{\mathrm{T}}\left(\boldsymbol{A} + \boldsymbol{A}^{\mathrm{T}}\right) = \boldsymbol{A} + \boldsymbol{A}^{\mathrm{T}}$$

Last,

$$\mathrm{tr}\left(\boldsymbol{B}\boldsymbol{A}\right) = \sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}b_{ji} \Longrightarrow \frac{\partial}{\partial a_{ij}} \mathrm{tr}\left(\boldsymbol{B}\boldsymbol{A}\right) = b_{ji} \Longrightarrow \frac{\partial}{\partial \boldsymbol{A}} \mathrm{tr}\left(\boldsymbol{B}\boldsymbol{A}\right) = \boldsymbol{B}'$$

□

1.2.7 Kronecker Products and the Vec Operators

Definition 1.21. Let $\boldsymbol{A} = (a_{ij})_{m\times n}$ be an $m \times n$ matrix and let $\boldsymbol{B}$ be any matrix. The **Kronecker product of $\boldsymbol{A}$ and $\boldsymbol{B}$**, denoted $\boldsymbol{A} \otimes \boldsymbol{B}$, is the matrix

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \cdots & a_{1n}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots & a_{2n}\boldsymbol{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}\boldsymbol{B} & a_{m2}\boldsymbol{B} & \cdots & a_{mn}\boldsymbol{B} \end{bmatrix}$$

Definition 1.22. Let $\boldsymbol{A} = (a_{ij})_{m\times n} = (\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_n)_{m\times n}$. The **vec of $\boldsymbol{A}$** , denoted $\operatorname{vec}(\boldsymbol{A})$ is

$$\operatorname{vec}(\boldsymbol{A}) = \begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \vdots \\ \boldsymbol{a}_n \end{pmatrix}_{mn\times 1}$$

Theorem 1.23. Properties of the Kronecker product

- $(\boldsymbol{A} + \boldsymbol{B}) \otimes \boldsymbol{C} = \boldsymbol{A} \otimes \boldsymbol{C} + \boldsymbol{B} \otimes \boldsymbol{C}$
- $(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = \boldsymbol{AC} \otimes \boldsymbol{BD}$
- $\boldsymbol{A} \otimes (\boldsymbol{B} \otimes \boldsymbol{C}) = (\boldsymbol{A} \otimes \boldsymbol{B}) \otimes \boldsymbol{C}$
- $(\boldsymbol{A} \otimes \boldsymbol{B})^{\mathrm{T}} = (\boldsymbol{B} \otimes \boldsymbol{A})^{\mathrm{T}}$
- $\operatorname{tr}(\boldsymbol{A} \otimes \boldsymbol{B}) = \operatorname{tr}(\boldsymbol{A}) \operatorname{tr}(\boldsymbol{B})$
- Let $\boldsymbol{A}_{n\times n}$ and $\boldsymbol{B}_{m\times m}$, then $\det(\boldsymbol{A} \otimes \boldsymbol{B}) = (\det \boldsymbol{A})^n (\det \boldsymbol{B})^m$
- $(\boldsymbol{A} \otimes \boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$
- If $\boldsymbol{A}$ and $\boldsymbol{B}$ are both positive definite, then $\boldsymbol{A} \otimes \boldsymbol{B}$ is also positive definite.
- $\operatorname{vec}(\boldsymbol{ABC}) = \left(\boldsymbol{C}^{\mathrm{T}} \otimes \boldsymbol{A}\right) \operatorname{vec}(\boldsymbol{B})$
- $\operatorname{tr}(\boldsymbol{ABCD}) = \operatorname{vec}\left(\boldsymbol{D}^{\mathrm{T}}\right)^{\mathrm{T}} \left(\boldsymbol{C}^{\mathrm{T}} \otimes \boldsymbol{A}\right) \operatorname{vec}(\boldsymbol{B})$

Part I

Probability Theory

Chapter 2

Probability Space

2.1 Sets

Definition 2.1. Some basic definitions on probablity space.

- **Random experiment**: An experiment, in general sense, which has more than 1 outcome, and the outcome is uncertain.
- **Sample Space** Ω is a set consisting all the possible outcome.
- The elements $\omega \in \Omega$ is called **outcome/ realization/ state**.
- An **event** is subset of the universal set Ω.

Definition 2.2. Some basic knowledge for sets.

- Complement A^c
- Union $A \cup B$
- Intersection $A \cap B$
- Minus $A \setminus B$
- Containment $A \subset B$
- $A \cap B \subset A \setminus B \subset A \cup B$, $A = (A \cap B) \cup (A \cap B^c)$
- $\bigcup_{n=1}^{\infty} A_n = \{\omega \in \Omega : \exists i \in \mathbb{N}, \omega \in A_i\}$ $\bigcap_{n=1}^{\infty} A_n = \{\omega \in \Omega : \forall i \in \mathbb{N}, \omega \in A_i\}$
- De Morgan's Law: $(\bigcup_{n=1}^{\infty} A_n)^c = \bigcap_{n=1}^{\infty} A_n^c$, $(\bigcap_{n=1}^{\infty} A_n)^c = \bigcup_{n=1}^{\infty} A_n^c$
- Countability:
  1. Countable (finite or countably infinite 1-1 mapping with $\mathbb{N}$).
  2. Uncountable.

- Cardinality: $\#\Omega$ or $|\Omega|$. e.g.,

$$\Omega_1 = \{H, T\}, |\Omega| = 2$$

$$\Omega_2 = \{HH, HT, TH, TT\}, |\Omega| = 4\vdots$$

$$\Omega_\infty = \{\omega_1\omega_2, \cdots : \omega_i \in \{H, T\}\}, |\Omega| = 2^\infty$$

- We say $\{A_n\}_{n=1}^\infty$ is a partition of Ω, if $A_i \cap A_j = \varnothing$ (disjoint or mutually exclusive) and $\bigcup_{n=1}^\infty A_n = \Omega$.

Definition 2.3. Definitions on set limits.

- $\liminf\limits_{n\to\infty} A_n = \bigcup_{n=1}^\infty \bigcap_{k=n}^\infty A_k = \left\{\omega \in \Omega : \lim_{n\to\infty} \inf\limits_{k\geqslant n} \mathbb{I}_{A_k}(\omega) = 1\right\}$
- $\limsup\limits_{n\to\infty} A_n = \bigcap_{n=1}^\infty \bigcup_{k=n}^\infty A_k = \left\{\omega \in \Omega : \lim_{n\to\infty} \sup\limits_{k\geqslant n} \mathbb{I}_{A_k}(\omega) = 1\right\}$
- $\liminf\limits_{n\to\infty} A_n = \limsup\limits_{n\to\infty} A_n \iff \lim_{n\to\infty} A_n$ exists.

Theorem 2.4. $\liminf\limits_{n\to\infty} A_n \subset \limsup\limits_{n\to\infty} A_n$.

Remark 2.5. $\omega \in \liminf\limits_{n\to\infty} A_n \Longrightarrow \omega \notin A_n$, for finitely many $n \Longrightarrow \omega \in A_n$, for infinitely many n or $\omega \in A_n$ i.o.(infinitely often)

Theorem 2.6. $\lim_{n\to\infty} A_n$ exists $\iff \forall\omega \in \Omega, \lim_{n\to\infty} \mathbb{I}_{A_n}(\omega) = 1$.

Theorem 2.7. If $\{A_n\}_{n=1}^\infty$ is monotone, then $\lim_{n\to\infty} A_n$ exists.

Proof.

- Show that $A_n \subset A_{n+1}$.

$$\liminf_{n\to\infty} A_n = \bigcup_{n=1}^\infty \bigcap_{k=n}^\infty A_k = \bigcup_{n=1}^\infty A_n$$

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^\infty \bigcup_{k=n}^\infty A_k = \bigcap_{n=1}^\infty \bigcup_{k=1}^\infty A_k = \bigcup_{k=1}^\infty A_k$$

- Show that $A_n \supset A_{n+1}$.

$$\liminf_{n\to\infty} A_n = \bigcup_{n=1}^\infty \bigcap_{k=n}^\infty A_k = \bigcup_{n=1}^\infty \bigcap_{k=1}^\infty A_k = \bigcap_{k=1}^\infty A_k$$

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^\infty \bigcup_{k=n}^\infty A_k = \bigcap_{n=1}^\infty A_n$$

□

Example 2.8. Examples for set limits.

1. $A_n = \left[0, \frac{1}{n}\right)$, $\lim_{n\to\infty} A_n = \{0\}$.

   $\lim_{n\to\infty} \frac{1}{n} = 0, \frac{1}{n} > 0 \Longrightarrow \forall n \geqslant 1, 0 \in A_n \Longrightarrow 0 \in \bigcap_{n=1}^\infty A_n$.

2. $A_n = \left(0, \frac{1}{n}\right), \lim_{n\to\infty} A_n = \varnothing$.

3. $A_n = \left(-\frac{1}{n}, 1 - \frac{1}{n}\right]$.

$$\liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} \left(-\frac{1}{k}, 1 - \frac{1}{k}\right] = \bigcup_{n=1}^{\infty} \left[0, 1 - \frac{1}{n}\right] = [0, 1)$$

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} \left(-\frac{1}{k}, 1 - \frac{1}{k}\right] = \bigcap_{n=1}^{\infty} \left(-\frac{1}{n}, 1\right) = [0, 1)$$

4. $A_n = \left(\frac{(-1)^n}{n}, 1 - \frac{(-1)^n}{n}\right]$.

$$\liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} \left(\frac{(-1)^k}{k}, 1 - \frac{(-1)^k}{k}\right] = \bigcup_{\substack{n=1 \\ n \text{ is even}}}^{\infty} \left(\frac{1}{n}, 1 - \frac{1}{n}\right] = (0, 1)$$

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} \left(\frac{(-1)^k}{k}, 1 - \frac{(-1)^k}{k}\right] = \bigcap_{\substack{n=1 \\ n \text{ is odd}}}^{\infty} \left(-\frac{1}{n}, 1 + \frac{1}{n}\right] = [0, 1]$$

2.2 Probability Measure

Definition 2.9. For a set Ω, and its power set $\mathcal{P}(\Omega)$, a set $\mathcal{F} \subset \mathcal{P}(\Omega)$ is called **σ-algebra**, if

1. $\varnothing \in \mathcal{F}$.
2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.
3. If $\{A_i : i = 1, 2, \cdots\} \subset \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \subset \mathcal{F}$.

Definition 2.10. **Borel set** is a set in which elements can be obtained by relative complement countable union, intersection of open sets on $\mathbb{R}$ (Borel set is the smallest σ-algebra containing all open sets on $\mathbb{R}$, $\mathcal{B} \subset \sigma(\mathbb{R})$).

Definition 2.11. **Measure** $\mu : \Sigma \to \mathbb{R}$.

- (Nonnegative) $\forall A \in \Sigma, \mu(A) \geqslant 0$.
- (Null-empty set) $\mu(\varnothing) = 0$.
- (Countably additivity or σ-additivity) for$\{A_n\}_{n=1}^{\infty}$ disjoint,

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n)$$

Remark 2.12. Remarks on measures.

- If μ can take negative values, then μ is a signed measure.
- If $\exists A \in \Sigma, \mu(A) < \infty$, then $\mu(A \cup \varnothing) = \mu(A) - \mu(\varnothing) \Longrightarrow \mu(\varnothing) = 0$.
- For probability measure, $\mathbb{P}(\Omega) = 1$.
- Probability measure is not unique.
- Probability is a measure, and measure is a function.

Theorem 2.13. (**Continuity of Probabilities**) If $\lim_{n\to\infty} A_n = A$, then $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A)$.

Proof. Three steps to prove.

- **Step 1**: If $\{A_n\}_{n=1}^{\infty}$ is increasing, we have $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A)$;
- **Step 2**: If $\{A_n\}_{n=1}^{\infty}$ is decreasing, we have $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A)$;
- **Step 3**: Since $\bigcap_{k=n}^{\infty} A_k \subset A_n \subset \bigcup_{k=n}^{\infty} A_k$ Then $\mathbb{P}(\bigcap_{k=n}^{\infty} A_k) \leqslant \mathbb{P}(A_n) \leqslant \mathbb{P}(\bigcup_{k=n}^{\infty} A_k)$. And $\bigcap_{k=n}^{\infty} A_k$ is increasing, $\bigcup_{k=n}^{\infty} A_k$ is decreasing, so

$$\begin{aligned}\mathbb{P}\left(\liminf_{n\to\infty} A_n\right) &= \mathbb{P}\left(\bigcup_{n=1}^{\infty}\bigcap_{k=n}^{\infty} A_k\right) = \lim_{n\to\infty} \mathbb{P}\left(\bigcap_{k=n}^{\infty} A_k\right) \leqslant \lim_{n\to\infty} \mathbb{P}(A_n) \\ &\leqslant \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{k=n}^{\infty} A_k\right) = \mathbb{P}\left(\lim_{n\to\infty}\bigcup_{k=n}^{\infty} A_k\right) = \mathbb{P}\left(\bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} A_k\right) = \mathbb{P}\left(\limsup_{n\to\infty} A_n\right)\end{aligned}$$

Note that $\lim_{n\to\infty} A_n = A$ exists, and thus, $\liminf_{n\to\infty} A_n = \limsup_{n\to\infty} A_n = A$, then

$$\mathbb{P}(A) \leqslant \lim_{n\to\infty} \mathbb{P}(A_n) \leqslant \mathbb{P}(A)$$

which leads to $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A)$.

□

2.3 Conditional Probabilities and Independence

Definition 2.14. If $\mathbb{P}(B) > 0, \mathbb{P}(A|B) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)}$ $\forall A \in \Sigma$ is the **conditional probability** of A given B.

Remark 2.15. Conditional probability is a probability measure.

Definition 2.16. Given $A, B \in \Sigma$, if $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$, then A, B are called **independent**, noted by $A \perp B$.

Remark 2.17. For independence,

- Sometimes we assume events are independence (Naturally);
- Sometimes we need to derive independence (By definition).

Theorem 2.18. (**Law of Total Probability**) $\{A_k\}_{k=1}^n$ is a partition of Ω, then

$$\mathbb{P}(B) = \sum_{k=1}^{n} \mathbb{P}(B|\, A_k)\, \mathbb{P}(A_k)$$

Theorem 2.19. (Bayes' Theorem) $\mathbb{P}(B) > 0, \{A_k\}_{k=1}^n$ disjoint.

$$\mathbb{P}(A_k|\, B) = \frac{\mathbb{P}(B|\, A_k)\, \mathbb{P}(A_k)}{\sum_{i=1}^n \mathbb{P}(B|\, A_k)\, \mathbb{P}(A_k)}$$

Remark 2.20. From Bayesian point of view,

- A- parameter
- B- data
- $\mathbb{P}(A_i|\, B)$- posterior
- $\mathbb{P}(B|\, A_k)$- model
- $\mathbb{P}(A_k)$- prior

Theorem 2.21. Some inequalities.

- Bonferroni's Inequality: $\mathbb{P}(A \cap B) \geqslant \mathbb{P}(A) + \mathbb{P}(B) - 1$.
- Boole's Inequality: $\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) \leqslant \sum_{n=1}^{\infty} \mathbb{P}(A_n)$.
- Therefore, we have $\mathbb{P}(\bigcap_{i=1}^{n} A_i) \leqslant \sum_{i=1}^{n} \mathbb{P}(A_i) - n - 1$.

Proof.

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i^c\right) \leqslant \sum_{i=1}^{n} \mathbb{P}(A_i^c) \Longrightarrow 1 - \mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right) \leqslant n - \sum_{i=1}^{n} \mathbb{P}(A_i^c)$$

□

Remark 2.22. Bonferroni and beyond.

For sets $A_1, A_2, \cdots, A_n$, we create a new set of nested intersections as follows. Let

$$\begin{aligned}
P_1 &= \sum_{i=1}^{n} \mathbb{P}(A_i) \\
P_2 &= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{P}(A_i \cap A_j) \\
P_2 &= \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^{n} \mathbb{P}(A_i \cap A_j \cap A_k) \\
&\vdots \\
P_n &= \mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n)
\end{aligned}$$

Then the inclusion-exclusion identity says that

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = P_1 - P_2 + P_3 - P_4 + \cdots \pm P_n$$

Moreover, $\forall i \leqslant j, P_i \geqslant P_j$, then we have the upper and lower bounds

$$\begin{aligned} P_1 &\geqslant \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \geqslant P_1 - P_2 \\ P_1 - P_2 + P_3 &\geqslant \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \geqslant P_1 - P_2 + P_3 - P_4 \\ &\vdots \end{aligned}$$

Remark 2.23. Counting theory: n numbers, a lottery needs to choose k numbers.

1. Ordered, without replacement. $n \times (n-1) \times \cdots \times (n-k+1) = \frac{n!}{(n-k)!}$
2. Ordered, with replacement. n^k
3. Unordered, without replacement. $\begin{pmatrix} n \\ k \end{pmatrix}$
4. Unordered, with replacement. $\begin{pmatrix} n+k-1 \\ k \end{pmatrix}$

Chapter 3

Random Variables, Distributions and Expectations

3.1 Random Variables

Definition 3.1. A function $f : \Omega \to \mathbb{R}$ is **measurable** if

$$\forall B \in \mathcal{B}, f^{-1}(B) = \{\omega \in \Omega : f(\omega) \in B\} \in \sigma(\Omega)$$

Definition 3.2. Given $(\Omega, \Sigma, \mathbb{P})$, a **random variable** is a measurable function that $X : \Omega \to \mathbb{R}$.

Remark 3.3. Notes.

- $\mathbb{P}(X \in B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\})$;
- $\mathbb{P}(X = x) = \mathbb{P}(X^{-1}(x)) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\})$.

Definition 3.4. The **cumulative distribution function** (CDF) of a random variable X is

$$F(x) = F_X(x) = \mathbb{P}(X \leqslant x)$$

Theorem 3.5. A function $F(x)$ is a CDF if and only if the following three conditions hold:

1. Normalized. $\lim_{x \to +\infty} F(x) = 1, \lim_{x \to -\infty} F(x) = 0$.
2. Non-decreasing. $F(x)$ is a nondecreasing function of x.
3. Right-continuous. $F(x)$ is right-continuous; that is, for every number x_0,

$$\lim_{x \to x_0^+} F(x) = F(x_0)$$

Note: (different notation) $F(x_0+) = \lim_{x \to x_0^+} F(x) = \lim_{x \downarrow x_0} F(x)$

Definition 3.6. The **inverse CDF** or **quantile function** of $X \sim F(x)$is defined by

$$F^{-1}(q) = \inf\{x : F(x) > q\}$$

for $q \in [0,1]$. If F is strictly increasing and continuous then $F^{-1}(q)$ is the unique real number x such that $F(x) = q$.

1. $F^{-1}(0.25)$- first quantile.
2. $F^{-1}(0.5)$- median or second quantile.
3. $F^{-1}(0.75)$- third quantile.

Definition 3.7. The random variables X and Y are **identically distributed** if,

$$\forall A \in \mathcal{B}, \mathbb{P}(Y \in A) = \mathbb{P}(X \in A)$$

Noted by $X \overset{d}{=} Y$.

Remark 3.8. Notes.

- $X \overset{d}{=} Y \iff F_X(x) = F_Y(x), \forall x \in \mathbb{R} \iff \mathbb{P}(Y \in A) = \mathbb{P}(X \in A), \forall A \in \mathcal{B}$
- $X \overset{d}{=} Y \not\Leftrightarrow X = Y$.

Definition 3.9. Continuous and discrete r.v..

- A r.v. X is **continuous** if $F_X(x)$ is a continuous function.
- A r.v. X is **discrete** if $F_X(x)$ is step function.

Definition 3.10. The **probability mass function** for a discrete r.v. X

$$f_X(x) = \mathbb{P}(X = x)$$

Note: For discrete r.v., $f_X(x) > 0$ for countable numbers of x. Otherwise, $f_X(x) = 0$.

Definition 3.11. The **probability density function** for a continuous r.v. with CDF $F(x)$ is defined by

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t, \forall x \in \mathbb{R}$$

Note:

- "X has a distribution", we can say $X \sim f(x)$ or $X \sim F(x)$.
- $0 = \mathbb{P}(X = x) \neq f_X(x)$.

Remark 3.12. An stronger definition is that $X \sim F$ is continuous if there exists a function f, such that $F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$.

By the stronger definition, continuous random variables have density functions, i.e., we can represent their distribution function as $\int_{-\infty}^{x} f(u)\,\mathrm{d}u$, so the distribution function must be continuous by this representation. Therefore, any continuous random variable has a continuous CDF. However, the continuity of integral F does not imply the continuity of integrand f.

- Since f is nonnegative and integrable, then f is bounded on $\mathbb{R}$. Suppose $|f| < M$, then $\forall \varepsilon > 0$, $\forall x \in \mathbb{R}$

$$F(x+\varepsilon) - F(x) = \int_x^{x+\varepsilon} f(u)\,\mathrm{d}u \leqslant \varepsilon M$$

  Since ε is arbitrary, we know that F is continuous.

- Does any continuous random variable has a continuous density? No.

Example 3.13. We construct the Cantor Set by:

- Step 1 divide $[0,1]$ into 3 subintervals with equal length, drop the interval $A_1^{(1)} = \left(\frac{1}{3}, \frac{2}{3}\right)$, we get 2 subintervals $\left[0, \frac{1}{3}\right], \left[\frac{2}{3}, 1\right]$.

- Step 2 divide the 2 subintervals into 3 subintervals with the same lenghth respectively, and drop $A_1^{(2)} = \left(\frac{1}{3^2}, \frac{2}{3^2}\right)$,$A_2^{(2)} = \left(\frac{7}{3^2}, \frac{8}{3^2}\right)$, we have $2^2 = 4$ closed subintervals

$$\left[0, \frac{1}{9}\right], \left[\frac{2}{9}, \frac{1}{3}\right], \left[\frac{2}{3}, \frac{7}{9}\right], \left[\frac{8}{9}, 1\right]$$

- $\cdots$

- Step n: drop 2^{n-1} open intervals.

$$A_1^{(n)} = \left(\frac{1}{3^n}, \frac{2}{3^n}\right), A_2^{(n)} = \left(\frac{7}{3^n}, \frac{8}{3^n}\right), A_3^{(n)} = \left(\frac{19}{3^n}, \frac{20}{3^n}\right) \cdots, A_{2^{n-1}}^{(n)} = \left(\frac{3^n-2}{3^n}, \frac{3^n-1}{3^n}\right)$$

  Then we have 2^n closed intervals with length $\frac{1}{3^n}$, they are disjoint with each other.

- Continue in this fashion, in the end, we drop countable disjoint open intervals subset to $[0,1]$, denote their union as G, we call $P = [0,1] \setminus G$ the Cantor Set.

Figure 3.1. The Cantor Set.

Let $B = \bigcup_{k=1}^{\infty} \bigcup_{n=1}^{2^k-1} A_n, C = [0,1] \setminus B$, then

$$\mu(C) = \mu([0,1]) - \mu(B) = 1 - \sum_{n=1}^{\infty} \frac{2^{n-1}}{3} = 0$$

On the open set B, we define $f(x) = \frac{2k-1}{2^n}, x \in A_k^{(n)}, k = 1, 2, \cdots, 2^{n-1}, n = 1, 2, \cdots$. Trivially, $f(x) \in (0,1)$, and $f(x)$ is non-decreasing on B.

Define on $\mathbb{R}$:

$$F(x)=\begin{cases} f(x), & x\in B \\ \sup\limits_{t>x,t\in B} f(t), & x\in C \\ 0, & x\leqslant 0 \\ 1, & x\geqslant 1 \end{cases}$$

Note that $F(x)$ is continuous, but if $X\sim F$, X is not continuous (under the stronger definition), since there is no density function such that $F(x)=\int_{-\infty}^{x} f(t)\,\mathrm{d}t$.

Proof. If $\exists f:\mathbb{R}\to\mathbb{R}$, such that that $F(x)=\int_{-\infty}^{x} f(t)\,\mathrm{d}t$, then $F'(x)=f(x)$, a.e. .

Since $\mu(C)=0$, and on $\mathbb{R}\setminus C$, $F'(x)=0$, then $f(x)=0$ a.e. . Therefore, $\int_{\mathbb{R}} f\mathrm{d}x=0$, contradicting to the property of density function that $\int_{\mathbb{R}} f\mathrm{d}x$. □

Remark 3.14. Notes.

- All r.v. have CDFs.
- Discrete r.v. have probability mass function (PMF).
- Continuous r.v. have probability distribution function (PDF).
- There are also other kinds of r.v. that are neither discrete or continuous.

Theorem 3.15. If X is a discrete random variable, Y is a continuous random variable, X,Y are independent, then $X+Y$ is a continuous r.v..

Proof. Y is a continuous, then

$$\mathbb{P}(Y\leqslant y)=\int_{-\infty}^{y} f_Y(t)\,\mathrm{d}t$$

$$\begin{aligned}
\mathbb{P}(X+Y\leqslant z) &= \int_{\{X+Y\leqslant z\}} \mathrm{d}F_{X,Y}(x,y)=\int_{\{Y\leqslant z-X\}} \mathrm{d}F_X(x)\,\mathrm{d}F_Y(x) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{z-x} \mathrm{d}F_Y(x)\,\mathrm{d}F_X(x)=\int_{-\infty}^{\infty}\int_{-\infty}^{z-x} f_Y(t)\,\mathrm{d}t\mathrm{d}F_X(x) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{z} f_Y(s-x)\,\mathrm{d}s\mathrm{d}F_X(x)\ \ (\mathrm{Let} s=t+x) \\
&= \int_{-\infty}^{z}\int_{-\infty}^{\infty} f_Y(s-x)\mathrm{d}F_X(x)\,\mathrm{d}s \\
&= \int_{-\infty}^{z} f_{X+Y}(s)\,\mathrm{d}s
\end{aligned}$$

□

Theorem 3.16. A function $f(x)$ is a PDF for some r.v. X, if

1. $f(x)\geqslant 0,\forall x\in\mathbb{R}$.
2. $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x=1$.
3. $\mathbb{P}(a<X<b)=\int_a^b f(x)\,\mathrm{d}x$

3.2 Expectations

Definition 3.17. Let X be a r.v., and g be a Borel-measurable function, then $g(X)$ is also a random variable. The **expectation** of $g(x)$ is defined as

$$\mathbb{E}[g(X)] = \int_{\Omega} g(X)\,\mathrm{d}\mathbb{P} = \int_{-\infty}^{\infty} g(x)\,\mathrm{d}F_X(x)$$

Remark 3.18. Notations and other definitions.

- To ensure that $\mathbb{E}X$ is well defined, we say that $\mathbb{E}|X| < \infty$, rather than $\mathbb{E}X < \infty$.
- For discrete cases, $\mathbb{E}[g(X)] = \sum_{x\in X(\Omega)} g(x)\,\mathbb{P}(X = x)$.
- For continuous cases, $\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)\,f_X(x)\,\mathrm{d}x$.
- Mean/ first moment: $\mu = \mathbb{E}X$
- Variance: $\sigma^2 = \mathbb{E}\left[(X - \mathbb{E}X)^2\right]$
- Covariance: $\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$
- Correlation: $\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \in [-1, 1]$
- Cauchy inequality: $[\mathbb{E}(XY)]^2 \leqslant \mathbb{E}X^2 \cdot \mathbb{E}Y^2$

$$\rho^2 = \frac{[\mathrm{Cov}(X,Y)]^2}{\mathrm{Var}(X)\,\mathrm{Var}(Y)} = \frac{\mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]}{\mathbb{E}(X - \mathbb{E}X)^2 \cdot \mathbb{E}(Y - \mathbb{E}Y)^2} \leqslant 1$$

Definition 3.19. X, Y are **independent**, if $\forall A \in \sigma(X), B \in \sigma(Y)$, such that

$$\mathbb{P}(AB) = \mathbb{P}(A)\,\mathbb{P}(B)$$

or $\forall A, B \in \mathcal{B}(\mathbb{R})$, $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\,\mathbb{P}(Y \in B)$.

Theorem 3.20. Some trivial results.

- $\mathbb{E}[aX + bY] = a\mathbb{E}X + b\mathbb{E}Y$.
- If $X_1, X_2, \cdots, X_n$ are independent variables, then $\mathbb{E}\left(\prod_{i=1}^n X_i\right) = \prod_{i=1}^n \mathbb{E}(X_i)$.
- $\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2$.
- $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$.
- $\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}X\mathbb{E}Y$.
- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$
- $\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\mathrm{Cov}(X, Y)$

- Generally,

$$\begin{aligned}\text{Var}\left[\sum_{i=1}^{n} a_i X_i\right] &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \text{Cov}\left(X_i, X_j\right) \\ &= \sum_{i=1}^{n} a_i^2 \text{Var}\left(X_i\right) + \sum_{i=1}^{n}\sum_{j=i}^{n} 2a_i a_j \text{Cov}\left(X_i, X_j\right)\end{aligned}$$

3.3 Examples of Distribution

Example 3.21. Some Discrete Random Variables and distribution.

1. Point mass distribution. $X \sim \delta_a, \mathbb{P}\left(X = a\right) = 1$.

2. Bernoulli distribution. $X \sim \text{Bernoulli}\left(p\right)$.

$$\mathbb{P}\left(X = 1\right) = p, \mathbb{P}\left(X = 0\right) = 1 - p$$

   - $\mathbb{E}X = p$
   - $\mathbb{E}X^2 = p^2 \Longrightarrow \text{Var}\left(X\right) = p\left(1 - p\right)$

3. Binomial distribution. $Y \sim \text{Binomial}\left(n, p\right)$,

$$Y = \sum_{i=1}^{n} X_i, \text{ where } X_i \overset{\text{i.i.d}}{\sim} \text{Bernoulli}\left(p\right)$$

$$\mathbb{P}\left(X = k\right) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k \left(1 - p\right)^{n-k}, k = 0, 1, \cdots, n$$

   - If $Y_1 \sim \text{Binomial}\left(n_1, p\right), Y_2 \sim \text{Binomial}\left(n_2, p\right)$, independent, then

$$Y_1 + Y_2 \sim \text{Binomial}\left(n_1 + n_2, p\right)$$

   - Let $Z \sim \text{Binomial}\left(n - 1, p\right)$

$$\begin{aligned}\mathbb{E}\left[Y^m\right] &= \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) p^k \left(1 - p\right)^{n-k} k^m = \sum_{k=1}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) p^k \left(1 - p\right)^{n-k} k^m \\ &= \sum_{k=1}^{n} \left(\begin{array}{c} n - 1 \\ k - 1 \end{array} \right) p^{k-1} \left(1 - p\right)^{n-k} k^{m-1} \cdot np \\ &= np \cdot \sum_{k=0}^{n-1} \left(\begin{array}{c} n - 1 \\ k \end{array} \right) p^k \left(1 - p\right)^{n-1-k} \cdot \left(k + 1\right)^{m-1} \\ &= np \cdot \mathbb{E}\left[\left(Z + 1\right)^{m-1}\right]\end{aligned}$$

   - $m = 1$,$\mathbb{E}\left(Y\right) = np$.
   - $m = 2, \mathbb{E}\left[Y^2\right] = np\left(\mathbb{E}\left[Z\right] + 1\right) \Longrightarrow \text{Var}\left(Y\right) = np\left(1 - p\right)$

- $Y = \sum_{i=1}^{n} X_i, X_i \overset{\text{i.i.d}}{\sim} \text{Bernoulli}\,(p)$.

$$\mathbb{E}\,(Y) = n \cdot \mathbb{E}X_1 = np$$
$$\text{Var}\,(Y) = n \cdot \text{Var}\,(X_1) = np\,(1-p)$$

4. Poisson Distribution.

$$\mathbb{P}\,(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, k = 0, 1, \cdots$$

- $\mathbb{E}X = \lambda$.
- $\text{Var}\,(X) = \lambda$.
- Let $p = \frac{\lambda}{n}$, then

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}\,(Y = k) &= \lim_{n\to\infty} \left(\begin{array}{c} n \\ k \end{array} \right) p^k\,(1-p)^{n-k} \\
&= \lim_{n\to\infty} \frac{n!}{k!\,(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \lim_{n\to\infty} \left[1 \cdot \frac{n-1}{n} \cdots \frac{n-k+1}{n}\right] \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \\
&= \frac{\lambda^k}{k!} e^{-\lambda}
\end{aligned}$$

- If $X_1 \sim \text{Poisson}\,(\lambda_1)\,, X_2 \sim \text{Poisson}\,(\lambda_2)$, and X_1,X_2 are independent, then X_1+$X_2 \sim$ $\text{Poisson}\,(\lambda_1 + \lambda_2)$.

5. Geometric Distribution.

$$\mathbb{P}\,(X = n) = (1-p)^{n-1}\,p, n = 1, 2, \cdots$$

- $\mathbb{E}X = \frac{1}{p}$.
- $\text{Var}\,(X) = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}$.

Example 3.22. Some Continuous Random Variables and distribution.

1. Uniform distribution. $X \sim \text{Uniform}\,(a, b)$

$$f_X\,(x) = \begin{cases} \frac{1}{b-a}, & x \in (a, b) \\ 0, & \text{o.w.} \end{cases}$$

- $\mathbb{E}X = \frac{a-b}{2}$
- $\text{Var}\,(X) = \frac{b^3-a^3}{3(b-a)} - \frac{(a-b)^2}{4} = \frac{(a-b)^2}{12}$

2. Gaussian distribution. $X \sim \mathcal{N}\,(\mu, \sigma^2)$

$$f\,(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- $\mathbb{E}\left(X-\mu\right)^k = \begin{cases} 0, & k \text{ isodd} \\ (k-1)!!\sigma^2, & k \text{ iseven} \end{cases}$.
- $X \sim \mathcal{N}\left(\mu, \sigma^2\right) \Longrightarrow Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}\left(0,1\right)$.
- $X_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$ independent $\Longrightarrow \sum_{i=1}^n X_i \sim \mathcal{N}\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right)$.

3. Exponential distribution. $X \sim \text{Exp}\left(\lambda\right)$

$$f_X\left(x\right) = \begin{cases} \lambda e^{-\lambda x}, & x \geqslant 0 \\ 0, & \text{o.w.} \end{cases}$$

- $\mathbb{E}X = \frac{1}{\lambda}$. (Integration by parts)
- $\text{Var}\left(X\right) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$.

4. Cauchy distribution.

$$f\left(x\right) = \frac{1}{\pi}\frac{1}{1+x^2}, x \in \mathbb{R}$$

- $\int_{-\infty}^{\infty} f\left(x\right)\mathrm{d}x = \int_{-\infty}^{\infty} \frac{1}{\pi}\frac{1}{1+x^2}\mathrm{d}x = \frac{1}{\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{1}{1+\tan^2 t}\mathrm{d}\tan t = 1$
- $\int_0^{\infty} \frac{1}{\pi}\frac{x}{1+x^2}\mathrm{d}x = \frac{1}{2\pi}\ln\left(1+x^2\right)\Big|_0^{\infty} = \infty$, then $\mathbb{E}X = \int_{-\infty}^{\infty} \frac{1}{\pi}\frac{x}{1+x^2}\mathrm{d}x = \infty - \infty$ is not defined.

$$\mathbb{E}\left|X\right| = \int_{-\infty}^{\infty} \left|x\right| f\left(x\right)\mathrm{d}x = 2\int_0^{\infty} \frac{1}{\pi}\frac{x}{1+x^2}\mathrm{d}x = \frac{1}{\pi}\ln\left(1+x^2\right)\Big|_0^{\infty} = \infty$$

Distribution	Mean	Variance
Point mass at a	a	0
Bernoulli(p)	p	$p(1-p)$
Binomial(n,p)	np	$np(1-p)$
Geometric(p)	$1/p$	$(1-p)/p^2$
Poisson(λ)	λ	λ
Uniform(a,b)	$(a+b)/2$	$(b-a)^2/12$
Normal(μ,σ^2)	μ	σ^2
Exponential(β)	β	β^2
Gamma(α,β)	$\alpha\beta$	$\alpha\beta^2$
Beta(α,β)	$\alpha/(\alpha+\beta)$	$\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$
t_ν	0 (if $\nu > 1$)	$\nu/(\nu-2)$ (if $\nu > 2$)
χ^2_p	p	$2p$
Multinomial(n,p)	np	see below
Multivariate Normal(μ,Σ)	μ	Σ

Figure 3.2. Examples of Distribution.

3.4 Bivariable Random Variables

Definition 3.23. Given a pair of random variables X, Y **joint PDF** is defined as

$$F_{X,Y}(x,y) = \mathbb{P}(X \leqslant x, Y \leqslant y)$$

For discrete X, Y, joint PMF is defined as

$$f_{X,Y}(x,y) = \mathbb{P}(X = x, Y = \mathbb{P}(X \leqslant x, Y \leqslant y) = \mathbb{P}(X \leqslant x, Y \leqslant y)\, y)$$

For continuous X, Y, joint PDF is defined as

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y} = \frac{\partial^2 F_{XY}(x,y)}{\partial y \partial x}$$

Theorem 3.24. In the continuous case, a function $f_{X,Y}(x,y)$ is the joint PDF for some pair of r.v.s X, Y, if it satisfies

1. $f_{X,Y}(x,y) \geqslant 0$,
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y = 1$,
3. $\forall A \in \mathcal{B}(\mathbb{R}^2), \mathbb{P}(A) = \iint_A f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y$.

Definition 3.25. **Marginal distribution**:

- Marginal PMF: $f_X(x) = \sum_i f_{X,Y}(x, y_i)$
- Marginal PDF: $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y$

Theorem 3.26. Let $X_1, X_2, \cdots, X_n$ be i.i.d. variables, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$,

$$S_n = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$$

Then, $\mathbb{E}\left(\bar{X}_n\right) = \mu, \mathrm{Var}\left(\bar{X}_n\right) = \frac{\sigma^2}{n}, \mathbb{E}\left(S_n^2\right) = \sigma^2$.

3.5 Conditional Expectations

Definition 3.27. **Conditional distribution** of X given Y is defined as

- Discrete
$$\mathbb{P}(X = x|\, Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{f_{X,Y}(x,y)}{\sum_i f_{X,Y}(x_i, y)}$$
- Continuous
$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x} = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

Note: Conditional probability is a probability.

$$\mathbb{P}(X \in A|\, Y) = \int_A f_{X|Y}(x|\, y)\,\mathrm{d}x$$

Definition 3.28. **Conditional Expectation**.

- Discrete

$$\mathbb{E}[X|Y] = \sum_i x_i f_{X|Y}(x_i|y)$$
$$\mathbb{E}[r(X,Y)|Y] = \sum_i r(x_i,y) f_{X|Y}(x_i|y)$$

- Continuous

$$\mathbb{E}[X|Y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\,\mathrm{d}x$$
$$\mathbb{E}[r(X,Y)|Y] = \int_{-\infty}^{\infty} r(x,y) f_{X|Y}(x|y)\,\mathrm{d}y$$

Theorem 3.29. Law of iterated expectation.

$$\mathbb{E}(\mathbb{E}[X|Y]) = \mathbb{E}X$$
$$\mathbb{E}(\mathbb{E}[r(X,Y)|Y]) = \mathbb{E}[r(X,Y)]$$

Proof for Continuous Case.

$$\begin{aligned}\mathbb{E}(\mathbb{E}[X|Y]) &= \int_{-\infty}^{\infty} \mathbb{E}[X|Y] f_Y(y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x \cdot f_{X|Y}(x|y) f_Y(y)\,\mathrm{d}x\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x \cdot \frac{f_{X,Y}(x,y)}{f_Y(y)} f_Y(y)\,\mathrm{d}x\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x \cdot f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y = \mathbb{E}X\end{aligned}$$

□

Definition 3.30. **Conditional variance** of Y given X is defined as

$$\mathrm{Var}(Y|X=x) = \mathbb{E}\left([Y - \mathbb{E}(Y|X=x)]^2 \middle| X=x\right) = \int_{-\infty}^{\infty} (y-\mu(x))^2 f_{Y|X}(y|x)\,\mathrm{d}y$$

where $\mu(x) = \mathbb{E}[Y|X=x]$.

Theorem 3.31. Law of total variance.

$$\mathrm{Var}(Y) = \mathbb{E}(\mathrm{Var}[Y|X]) + \mathrm{Var}(\mathbb{E}(Y|X))$$

Proof.

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

Let $Z = \mathbb{E}[Y|X]$, since $\mathrm{Var}Z = \mathbb{E}Z^2 - (\mathbb{E}Z)^2$, we get

$$\mathbb{E}\left[\left[\mathbb{E}(Y|X)\right]^2\right] - \left[\mathbb{E}\left(\mathbb{E}[Y|X]\right)\right]^2 = \mathrm{Var}\left(\mathbb{E}(Y|X)\right)$$

□

Definition 3.32. X, Y are **independent**, $\forall A, B \in \sigma(\mathbb{R})$,

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\,\mathbb{P}(Y \in B)$$

Theorem 3.33. Continuous r.v. X, Y are independent $\Longleftrightarrow$

$$F_{X,Y}(x, y) = F_X(x)\, F_Y(y)$$

Theorem 3.34. X, Y are independent, $\Longleftarrow \forall x, y \in \mathbb{R}$,

$$f_{X,Y}(x, y) = f_X(x)\, f_Y(y)$$

(NOT necessary).

Example 3.35. Let X, Y has joint PDF:

$$f(x, y) = \begin{cases} 1, & 0 < x < 1, 0 < y < 1 \\ 1, & x = 2, y = 2 \\ 0, & \text{o.w.} \end{cases}$$

Then

$$f_X(x) = \begin{cases} 1, & 0 < x < 1 \\ 0, & \text{o.w.} \end{cases}, f_Y(y) = \begin{cases} 1, & 0 < y < 1 \\ 0, & \text{o.w.} \end{cases}$$

At $(2, 2)$, we have

$$f(2, 2) = 1 \neq f_X(2)\, f_Y(2) = 0$$

However, if we examine the CDF,

$$F_X(x) = \begin{cases} 0, & x \leqslant 0 \\ x, & 0 < x \leqslant 1 \\ 0, & x > 0 \end{cases}, F_Y(y) = \begin{cases} 0, & y \leqslant 0 \\ y, & 0 < y \leqslant 1 \\ 0, & y > 0 \end{cases}$$

$$F(x, y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(x, y)\,\mathrm{d}x\mathrm{d}y = \begin{cases} 0, & x \leqslant 0 \text{ or } y \leqslant 0 \\ xy, & 0 < x \leqslant 1, 0 < y \leqslant 1 \\ x, & 0 < x \leqslant 1, y > 0 \\ y, & x > 0, 0 < y \leqslant 1 \\ 1, & x > 0, y > 0 \end{cases} = F_X(x) \cdot F_Y(y)$$

By definition, X, Y are independent.

Theorem 3.36. If $X(\Omega) \times Y(\Omega)$ is a rectangle, and

$$f_{XY}(xy) = g(x)\, h(y)$$

then X, Y are independent.

Proof.

$$f_X(x) = \int_{Y(\Omega)} f_{XY}(xy)\,\mathrm{d}y = g(x)\int_{Y(\Omega)} h(y)\,\mathrm{d}y$$
$$f_Y(y) = \int_{X(\Omega)} f_{XY}(xy)\,\mathrm{d}x = h(y)\int_{X(\Omega)} g(x)\,\mathrm{d}x$$

Then

$$\begin{aligned} f_X(x)\cdot f_Y(y) &= g(x)\int_{Y(\Omega)} h(y)\,\mathrm{d}y \cdot h(y)\int_{X(\Omega)} g(x)\,\mathrm{d}x \\ &= f_{X,Y}(x,y)\cdot \iint_{X(\Omega)\times Y(\Omega)} g(x)\,h(y)\,\mathrm{d}x\mathrm{d}y \\ &= f_{X,Y}(x,y)\cdot \iint_{X(\Omega)\times Y(\Omega)} f_{XY}(xy)\,\mathrm{d}x\mathrm{d}y \\ &= f_{X,Y}(x,y) \end{aligned}$$

□

Theorem 3.37. Independence $\Longrightarrow$ uncorrelated.

Proof.

$$\mathrm{Cov}(X,Y) = \mathbb{E}([X-\mathbb{E}X][Y-\mathbb{E}Y]) = \mathbb{E}[X-\mathbb{E}X]\cdot\mathbb{E}[Y-\mathbb{E}Y] = 0$$

Here is a counter example for $\not\Leftarrow$. Suppose $X\sim N(0,1), Y=X^2$,

$$\mathrm{Cov}(X,Y) = \mathbb{E}[X(Y-\mathbb{E}Y)] = \mathbb{E}\left[X^3 - X\right] = 0$$

But X and X^2 are evidently not independent, or we have the example:

$$\begin{aligned} \mathbb{P}(X\leqslant 1, Y\leqslant 1) &= \mathbb{P}(X\leqslant 1, -1\leqslant X\leqslant 1) \\ &= \mathbb{P}(-1\leqslant X\leqslant 1) = \Phi(1)-\Phi(-1) \\ &\neq \mathbb{P}(X\leqslant 1)\,\mathbb{P}(Y\leqslant 1) = \Phi(1)[\Phi(1)-\Phi(-1)] \end{aligned}$$

□

3.6 Moment Generating Function

Definition 3.38. If $\exists h>0$, such that $t\in(-h,h)$, $\mathbb{E}\left(e^{tX}\right)<\infty$, then we call $M_X(t)=\mathbb{E}\left(e^{tX}\right)$ is the **moment generating function** or **Laplace transformation** of X.

If not, we say MGF of X doesn't exist.

$$M_X(t) = \mathbb{E}\left(e^{tX}\right) = \int_{-\infty}^{\infty} e^{tx}\mathrm{d}F(x)$$

Theorem 3.39. Some conclusions on MGF.

- $M_X^{(n)}(0) = \mathbb{E}(X^n)$. (Can limit and integral interchange? See Statistical Inference 2.4 Page.94)

- $Y = aX + b \Longrightarrow M_Y(t) = \mathbb{E}\left(e^{taX+tb}\right) = e^{tb} M_X(ta)$.
- If $X_1, X_2, \cdots, X_n$ are independent, $Y = \sum_{i=1}^n X_i \Longrightarrow M_Y(t) = \prod_{i=1}^n M_{X_i}(t)$.
- $\exists h > 0, \forall t \in (-h, h)$, $M_X(t) = M_Y(t) \Longleftrightarrow X \overset{d}{=} Y$.

Moment Generating Functions for Some Common Distributions

Distribution	MGF $\psi(t)$
Bernoulli(p)	$pe^t + (1-p)$
Binomial(n,p)	$(pe^t + (1-p))^n$
Poisson(λ)	$e^{\lambda(e^t-1)}$
Normal(μ,σ)	$\exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\}$
Gamma(α,β)	$\left(\frac{1}{1-\beta t}\right)^\alpha$ for $t < 1/\beta$

Figure 3.3. Some moment generating functions.

Example 3.40.

$$X_1 \sim f_1(x) = \frac{1}{\sqrt{2\pi}x} e^{-\frac{(\log x)^2}{2}}, 0 \leqslant x < \infty$$
$$X_2 \sim f_2(x) = f_1(x)\left[1 + \sin(2\pi \log x)\right], 0 \leqslant x < \infty$$

Since $\log X_1 \sim N(0,1)$, $\mathbb{E}\left[X_1^k\right] = \mathbb{E}\left[e^{k \log X_1}\right] = e^{\frac{1}{2}k^2}$.

$$\begin{aligned}\mathbb{E}X_2^k &= \int_0^\infty x^k f_1(x)\left[1 + \sin(2\pi \log x)\right] \mathrm{d}x \\ &= \mathbb{E}X_1^k + \int_0^\infty x^k f_1(x) \sin(2\pi \log x)\, \mathrm{d}x\end{aligned}$$

Let $y = \log x - k \Longrightarrow x = e^{k+y}$,

$$\begin{aligned}\mathbb{E}X_2^k &= \mathbb{E}X_1^k + \int_\infty^\infty e^{(y+k)^k} f_1\left(e^{y+k}\right) \sin(2\pi y + 2k\pi)\, \mathrm{d}\left(e^{y+k}\right) \\ &= \mathbb{E}X_1^k + \int_\infty^\infty e^{(y+k)^k} \frac{1}{\sqrt{2\pi}e^{y+k}} e^{-\frac{y^2+k^2+2ky}{2}} \sin(2\pi y) \cdot e^{y+k} \mathrm{d}y \\ &= \mathbb{E}X_1^k + \int_\infty^\infty \frac{1}{\sqrt{2\pi}} e^{\frac{-y^2+k^2}{2}} \sin(2\pi y) \cdot e^{y+k} \mathrm{d}y \ \text{ (odd function)} \\ &= \mathbb{E}X_1^k\end{aligned}$$

But X_1 and X_2 are different random variables.

Note: The problem of uniqueness of moments does not occur if the random variables have bounded support. In the next section, we will introduce the support set.

Theorem 3.41. Some conclusions.

1. If X, Y have bounded support, then $F_X(u) = F_Y(u), \forall u \Longleftrightarrow \mathbb{E}X^k = \mathbb{E}Y^k, \forall k = 1, 2, \cdots$.

2. If the MGFs exist and $M_X(t), M_Y(t)$ for all t in some neighborhood of 0, then $F_X(u) = F_Y(u), \forall u$, or $X \overset{d}{=} Y$.

3. For a sequence $X_1, X_2, \cdots, X_n$, if $\exists h > 0$,

$$\lim_{n\to\infty} M_{X_n}(t) = M_X(t), \forall t \in (-h, h)$$

Then there is a unique CDF F_X determined by M_X, and

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

For $t \in (-h, h)$, MGFs to an MGF implies convergence of CDFs.

Example 3.42. $X \sim \text{Poisson}(\lambda)$, $Y \sim \text{Binomial}(n, p)$.

$$M_X(t) = e^{\lambda\left(e^t - 1\right)} M_Y(t) = \left[pe^t + (1-p)\right]^n$$

Let $p = \frac{\lambda}{n}$, then as $n \to \infty$,

$$M_Y(t) = \left[\frac{\lambda}{n}e^t + \left(1 - \frac{\lambda}{n}\right)\right]^n = \left[\frac{\lambda\left(e^t - 1\right)}{n} + 1\right]^n \to e^{\lambda\left(e^t - 1\right)}$$

3.7 Transformation of Distributions

Remark 3.43. For r.v.s $X : \Omega \to \mathbb{R}$, $Y = g(X) : \Omega \to \mathbb{R}$, define the **support set** as

$$\mathcal{X} = \{x \in \mathbb{R} : f_X(x) > 0\}, \mathcal{Y} = \{y \in \mathbb{R} : y = g(x), x \in \mathcal{X}\}$$

Then $X : \Omega \to \mathcal{X}$, $Y = g(X) : \Omega \to \mathcal{Y}$.

Theorem 3.44. $X \sim F_X(x), Y = g(x)$, and $\mathcal{X}, \mathcal{Y}$ are support set of X, Y.

1. If g is an increasing function on $\mathcal{X}$,

$$F_Y(y) = F_X\left(g^{-1}(y)\right), y \in \mathcal{Y}$$

2. If g is a decreasing function on $\mathcal{X}$, and X is a continuous r.v.,

$$F_Y(y) = 1 - F_X\left(g^{-1}(y)\right), y \in \mathcal{Y}$$

3. If g is monotone, and X is a continuous r.v., and $g^{-1}(y)$ has continuous derivative, then

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)\right|, y \in \mathcal{Y}$$

Proof. For $x \in \mathcal{X}$. If g is increasing,

$$F_Y(y) = \mathbb{P}(Y \leqslant y) = \mathbb{P}(g(X) \leqslant y) = \mathbb{P}\left(X \leqslant g^{-1}(y)\right) = F_X\left(g^{-1}(y)\right)$$
$$f_Y(y) = f_X\left(g^{-1}(y)\right) \cdot \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y)$$

If g is decreasing,

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leqslant y) = \mathbb{P}(g(X) \leqslant y) = \mathbb{P}\left(X \geqslant g^{-1}(y)\right) \\ &= 1 - \mathbb{P}\left(X \leqslant g^{-1}(y)\right) + \mathbb{P}\left(X = g^{-1}(y)\right) \\ &= 1 - F_X\left(g^{-1}(y)\right) \\ f_Y(y) &= -f_X\left(g^{-1}(y)\right) \cdot \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) \end{aligned}$$

Then

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y)\right|, y \in \mathcal{Y}$$

□

Theorem 3.45. Let X have PDF $f_X(x)$, let $Y = g(X)$ and define the sample space as $\mathcal{X} = \{x : f_X(x) > 0\}$. Suppose there exists a partition, $A_0, A_1, \cdots, A_k$ of $\mathcal{X}$ such that $\mathbb{P}(X \in A_0) = 0$, and $f_X(x)$ is continuous on each A_i.

Suppose function $g_1, g_2, \cdots, g_k$ defined on $A_0, A_1, \cdots, A_k$, respectively, satisfying

1. $g(x) = g_i(x), \forall x \in A_i$,
2. $g_i(x)$ is monotone on A_i,
3. $\mathcal{Y} = \{y : y = g_i(x) \text{ forsome} x \in A_i\}$ is the same for each A_i,
4. $g_i^{-1}(y)$ has a continuous derivative on $\mathcal{Y}$, for each i.

Then

$$f_Y(y) = \begin{cases} \sum_{i=1}^{k} f_X\left(g_i^{-1}(y)\right)\left|\frac{\mathrm{d}}{\mathrm{d}y} g_i^{-1}(y)\right|, & y \in \mathcal{Y} \\ 0, & \text{o.w.} \end{cases}$$

Example 3.46. Let $X \sim f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, and $Y = X^2$.

The function $g(x)$ is monotone on $(-\infty, 0)$ and $(0, \infty)$. Then

$$f_Y(y) = f_X(-\sqrt{y})\left|-\frac{1}{2\sqrt{y}}\right| + f_X(\sqrt{y})\left|\frac{1}{2\sqrt{y}}\right| = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} e^{-\frac{y}{2}}, y > 0$$

Theorem 3.47. (Probability Integral Transformation) Let continuous r.v., $X \sim F_X(x)$, and $Y = F_X(X)$, then $Y \sim U(0, 1)$.

Proof. For $x \in \mathcal{X}$, $F_X(X)$ is strictly increasing, and thus has inverse function.

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leqslant y) = \mathbb{P}(F_X(X) \leqslant y) = \mathbb{P}\left(X \leqslant F_X^{-1}(y)\right) \\ &= F_X\left(F_X^{-1}(y)\right) = y \end{aligned}$$

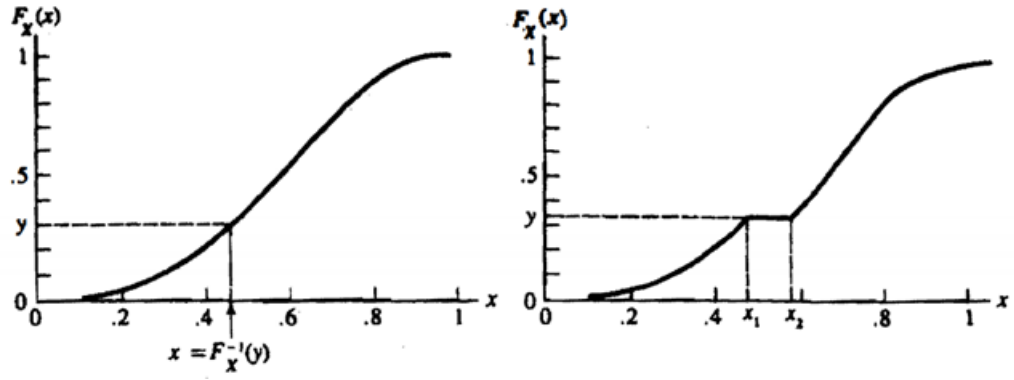

Figure 3.4. $F_X^{-1}(y)$.

Remark: If we have a sample from $Y \sim U(0,1)$, then we can generate an arbitrary r.v. with CDF $F_X(x)$ by letting $X = F_X^{-1}(Y)$. Note that $F_X^{-1}(y) = \inf\{x : F_X(x) \geqslant y\}$ $F_X^{-1}(1) = \infty, F_X^{-1}(0) = -\infty$. The equality

$$\mathbb{P}\left(F_X^{-1}(F_X(X)) \leqslant F_X^{-1}(y)\right) = \mathbb{P}\left(X \leqslant F_X^{-1}(y)\right)$$

still holds, even for F_X is flat. See the right figure, $x \in [x_1, x_2]$,

$$F_X^{-1}(F_X(x)) = x_1$$

may not hold, but, $\forall x \in [x_1, x_2]$, $F_X(x) = F_X(x_1) \iff \mathbb{P}(X \leqslant x) = \mathbb{P}(X \leqslant x_1)$, therefore, $\mathbb{P}(x_1 < X \leqslant x) = 0$. □

Example 3.48. We want a sample from $\mathcal{N}(0,1)$, but only for $X \in [c_1, c_2]$.

$$F_X(x) = \frac{\Phi(x) - \Phi(c_1)}{\Phi(c_2) - \Phi(c_1)}$$

Remark 3.49. Transformations of several random variables. Suppose X, Y are given r.v.s, let $Z = r(X, Y)$.

- For each z, find $A_z = \{(x, y) : r(x, y) \leqslant z\}$
- $F_Z(z) = \mathbb{P}(Z \leqslant z) = \iint_{A_z} f_{X,Y}(x, y)\,\mathrm{d}x\mathrm{d}y$.
- $f_Z(z) = \frac{\partial}{\partial z} F_Z(z)$.

Chapter 4

Asymptotic Theory

4.1 Inequalities

Theorem 4.1. (Markov Inequality) X is a nonnegative r.v., and $\mathbb{E}|X| < \infty$.

$$\mathbb{P}(X \geqslant t) \leqslant \frac{\mathbb{E}X}{t}$$

Proof.

$$\mathbb{P}(X \geqslant t) = \int_{\{x \geqslant t\}} \mathrm{d}F(x) \leqslant \int_{\{x \geqslant t\}} \frac{x}{t} \mathrm{d}F(x) \leqslant \int_{[0,\infty)} \frac{x}{t} \mathrm{d}F(x) = \frac{\mathbb{E}X}{t}$$

□

Theorem 4.2. (Chebyshev's Inequality) $\mu = \mathbb{E}X, \sigma^2 = \mathrm{Var}(X)$, then

$$\mathbb{P}(|X - \mu| \geqslant t) \leqslant \frac{\sigma^2}{t}$$

and

$$\mathbb{P}(|Z| \geqslant k) \leqslant \frac{1}{k^2}, Z = \frac{X - \mu}{\sigma^2}$$

Proof. Note that $(X - \mu)^2, Z^2$ are nonnegative r.v.s, then by Markov Inequality,

$$\mathbb{P}(|X - \mu| \geqslant t) = \mathbb{P}\left((X - \mu)^2 \geqslant t^2\right) \leqslant \frac{\sigma^2}{t}$$

and

$$\mathbb{P}(|Z| \geqslant k) = \mathbb{P}\left(|Z|^2 \geqslant k^2\right) \leqslant \frac{1}{k^2}$$

□

Theorem 4.3. (Jensen's Inequality)

1. g is convex, $g[\mathbb{E}X] \leqslant \mathbb{E}[g(X)]$;
2. g is concave, $g[\mathbb{E}X] \geqslant \mathbb{E}[g(X)]$.

Theorem 4.4. (Cauchy-Schwarz Inequality)

$$[\mathbb{E}(XY)]^2 \leqslant \mathbb{E}\left(X^2\right)\mathbb{E}\left(Y^2\right) \text{ or } \operatorname{Cov}(X,Y) \leqslant \sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}$$

Note: Let $\tilde{X} = X - \mathbb{E}X, \tilde{Y} = Y - \mathbb{E}Y$.

Definition 4.5. An **inner product** on V is a map, $\langle\cdot,\cdot\rangle : V \times V \to \mathbb{R}$ or $\mathbb{C}$, such that $\forall \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in V$,

1. Linearity in the first slot. $\langle \boldsymbol{u}+\boldsymbol{v}, \boldsymbol{w}\rangle = \langle \boldsymbol{u}, \boldsymbol{w}\rangle + \langle \boldsymbol{v}, \boldsymbol{w}\rangle$.
2. Nonnegativity. $\langle \boldsymbol{u}, \boldsymbol{u}\rangle \geqslant \mathbf{0}$.
3. Positive definite. $\langle \boldsymbol{u}, \boldsymbol{u}\rangle = 0 \Longleftrightarrow \boldsymbol{u} = \mathbf{0}$.
4. Conjugate symmetric. $\langle \boldsymbol{u}, \boldsymbol{v}\rangle = \langle \boldsymbol{v}, \boldsymbol{u}\rangle$.

Theorem 4.6.

$$\langle \boldsymbol{u}, \boldsymbol{v}\rangle^2 \leqslant \langle \boldsymbol{u}, \boldsymbol{u}\rangle \cdot \langle \boldsymbol{v}, \boldsymbol{v}\rangle$$

Proof. Let

$$g(t) = \langle t\boldsymbol{u}+\boldsymbol{v}, t\boldsymbol{u}+\boldsymbol{v}\rangle^2 = \langle \boldsymbol{u}, \boldsymbol{u}\rangle t^2 + 2\langle \boldsymbol{u}, \boldsymbol{v}\rangle t + \langle \boldsymbol{v}, \boldsymbol{v}\rangle \geqslant 0$$

So, $\Delta = 4\langle \boldsymbol{u}, \boldsymbol{v}\rangle^2 - 4\langle \boldsymbol{u}, \boldsymbol{u}\rangle \cdot \langle \boldsymbol{v}, \boldsymbol{v}\rangle \leqslant 0$. □

Remark 4.7. Let V be a space of r.v.s for which $\mathbb{E}|X| < \infty$, and for $\forall X, Y \in V$, define the inner product as $\mathbb{E}[XY]$, we get Cauchy-Schwarz Inequality for r.v.s.

Theorem 4.8. (Hölder's Inequality)

$$\mathbb{E}|XY| \leqslant [\mathbb{E}|X|^p]^{\frac{1}{p}} \cdot [\mathbb{E}|Y|^q]^{\frac{1}{q}}$$

where $\frac{1}{q} + \frac{1}{p} = 1$.

If $p = 2, q = 2$, we have $\mathbb{E}[XY] \leqslant \mathbb{E}|XY| \leqslant \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}$, which gives us the Cauchy-Schwarz Inequality.

Theorem 4.9. (Minkowski's Inequality) $\forall X, Y \in \mathcal{L}^p$,

$$\|X+Y\|_p \leqslant \|X\|_p + \|Y\|_p$$

Proof.

$$\begin{aligned}
&\mathbb{E}|X+Y|^p \\
=&\mathbb{E}\left[|X+Y|^{p-1}\cdot|X+Y|\right] \\
\leqslant&\mathbb{E}\left[|X+Y|^{p-1}(|X|+|Y|)\right] \\
=&\mathbb{E}\left[|X+Y|^{p-1}\cdot|X|\right] + \mathbb{E}\left[|X+Y|^{p-1}\cdot|Y|\right] \\
=&\mathbb{E}\left[|X+Y|^{p\cdot\frac{p-1}{p}}\cdot|X|^{p\cdot\frac{1}{p}}\right] + \mathbb{E}\left[|X+Y|^{p\cdot\frac{p-1}{p}}\cdot|Y|^{p\cdot\frac{1}{p}}\right] \\
=&\mathbb{E}\left[|X+Y|^{p\cdot\frac{1}{q}}\cdot|X|^{p\cdot\frac{1}{p}}\right] + \mathbb{E}\left[|X+Y|^{p\cdot\frac{1}{q}}\cdot|Y|^{p\cdot\frac{1}{p}}\right] \quad \left(\text{Let } q = \left(1-\frac{1}{p}\right)^{-1} = \frac{p}{p-1}\right) \\
\leqslant&\left(\mathbb{E}\left[|X+Y|^p\right]\right)^{\frac{1}{q}}\cdot\left(\mathbb{E}|X|^p\right)^{\frac{1}{p}} + \left(\mathbb{E}\left[|X+Y|^{p-1}\right]\right)^{\frac{1}{q}}\left(\mathbb{E}|Y|^p\right)^{\frac{1}{p}} \text{ (By Hölder's)} \\
=&\left(\mathbb{E}\left[|X+Y|^p\right]\right)^{1-\frac{1}{p}}\left[\left(\mathbb{E}|X|^p\right)^{\frac{1}{p}} + \left(\mathbb{E}|Y|^p\right)^{\frac{1}{p}}\right]
\end{aligned}$$

Then

$$\left[\mathbb{E}\left|X+Y\right|^{p}\right]^{\frac{1}{p}} \leqslant \left[\left(\mathbb{E}\left|X\right|^{p}\right)^{\frac{1}{p}}+\left(\mathbb{E}\left|Y\right|^{p}\right)^{\frac{1}{p}}\right]$$
$$\left\|X+Y\right\|_{p} \leqslant \left\|X\right\|_{p}+\left\|Y\right\|_{p}$$

□

Theorem 4.10. If $q \geqslant p \geqslant 1$, $\mathcal{L}^{q} \subset \mathcal{L}^{p}$, i.e., $\mathbb{E}\left|X\right|^{q} < \infty \Longrightarrow \mathbb{E}\left|X\right|^{p} < \infty$.

Proof. **Method 1**. By Jense's inequality.

$$\mathbb{E}\left|X\right|^{p} = \mathbb{E}\left[\left|X\right|^{q\cdot\frac{p}{q}}\right] \leqslant \mathbb{E}\left[\left|X\right|^{q}\right]^{\frac{p}{q}} < \infty$$

Method 2. By Hölder's Inequality,

$$\begin{aligned}
\mathbb{E}\left|X\right|^{p} &= \mathbb{E}\left[\left|X\right|^{p}\cdot 1\right] \\
&= \mathbb{E}\left[\left|X\right|^{q\cdot\frac{p}{q}}\cdot 1^{\left(1-\frac{q}{p}\right)}\right] \quad \left(\text{Let } \tilde{p}=\frac{p}{q}, \text{and } \tilde{q}=\left(1-\frac{p}{q}\right)^{-1}=\frac{q}{q-p}\right) \\
&= \mathbb{E}\left[\left|X\right|^{q\cdot\frac{1}{\tilde{p}}}\cdot 1^{\frac{1}{\tilde{q}}}\right] \\
&\leqslant \left(\mathbb{E}\left|X\right|^{q}\right)^{\frac{1}{\tilde{p}}}\cdot 1^{\frac{1}{\tilde{q}}} < \infty
\end{aligned}$$

Method 3.

$$\begin{aligned}
\mathbb{E}\left|X\right|^{p} &= \int_{\mathbb{R}}\left|x\right|^{p}\mathrm{d}F\left(x\right) \\
&= \int_{\{|x|\leqslant 1\}}\left|x\right|^{p}\mathrm{d}F\left(x\right)+\int_{\{|x|>1\}}\left|x\right|^{p}\mathrm{d}F\left(x\right) \\
&\leqslant \int_{\{|x|\leqslant 1\}}\mathrm{d}F\left(x\right)+\int_{\{|x|>1\}}\left|x\right|^{q}\mathrm{d}F\left(x\right) \\
&\leqslant 1+\mathbb{E}\left|x\right|^{q} < \infty
\end{aligned}$$

□

4.2 Convergence

Definition 4.11. We say that X_n **converges to** X **in probability**, and write $X_n \xrightarrow{\mathbb{P}} X$, if

$$\forall\varepsilon > 0, \mathbb{P}\left(\left|X_n - X\right| > \varepsilon\right) \to 0, \text{ as } n \to \infty$$

Note: For any given ε, $\left\{\mathbb{P}\left(\left|X_n - X\right| > \varepsilon\right)\right\}_{n=1}^{\infty} \subset \mathbb{R}$. Therefore, we convert the definition of the convergence for r.v.s to the definition of the convergence for real sequences.

Definition 4.12. We say that X_n **converges to** X **in distribution**, and write $X_n \xrightarrow{D} X$ if

$$\forall t \text{ at which } F \text{ is continuous }, \lim_{n\to\infty} F_n\left(t\right) = F\left(t\right)$$

Note: For any given t, $\left\{F_n\left(t\right)\right\}_{n}^{\infty} \subset \mathbb{R}$.

Definition 4.13. We say that X_n **converges to** X **almost surely**, and write $X_n \xrightarrow{\text{a.s.}} X$

$$\mathbb{P}\left(\{\omega : X_n(\omega) \to X(\omega)\}\right) = 1$$

Note:

1. It's a pointwise converge.
2. For any given ω, $\{X_n(\omega)\}_n^\infty \subset \mathbb{R}$.

Definition 4.14. We say that X_n **converges to** X **in** $\mathcal{L}^p$, and write $X_n \xrightarrow{\mathcal{L}_p} X$

$$\mathbb{E}\left|X_n - X\right|^p \to 0$$

1. $X_n \xrightarrow{\mathcal{L}_1} X$, then $\mathbb{E}\left|X_n - X\right| \to 0$
2. $X_n \xrightarrow{\mathcal{L}_2} X$ or $X_n \xrightarrow{\text{qm}} X$, then $\mathbb{E}\left|X_n - X\right|^2 \to 0$, we say X_n converges to X in quadratic mean.
3. $\mathcal{L}^p$ norm of X: $\|X\|_p = \left(\mathbb{E}\left|X^p\right|\right)^{\frac{1}{p}}$.

Theorem 4.15. Another form of Hölder's Inequality.

$$\|XY\|_1 \leqslant \|X\|_p \cdot \|Y\|_q, \text{ where } \frac{1}{p} + \frac{1}{q} = 1$$

Remark 4.16. If the limiting r.v. X is a point mass, for simplicity,

1. $X_n \xrightarrow{\mathbb{P}} c$ means $X_n \xrightarrow{\mathbb{P}} X, \mathbb{P}(X = c) = 1$.
2. $X_n \xrightarrow{d} c$ means $X_n \xrightarrow{d} X, \mathbb{P}(X = c) = 1$.

Remark 4.17. Overview: Relationships between convergences.

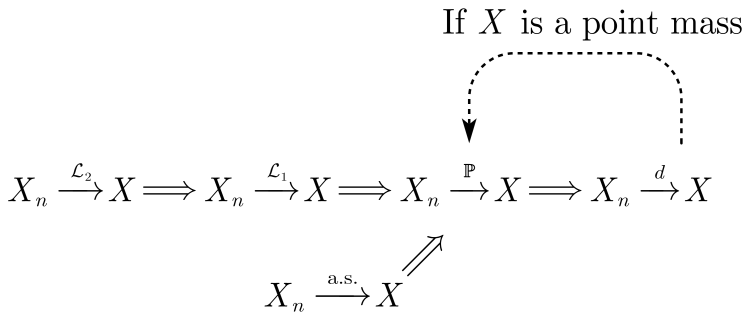

Figure 4.1. Relationships between convergences

Theorem 4.18. $\forall q, p,$ with $q \geqslant p$,

$$X_n \xrightarrow{\mathcal{L}^q} X \Longrightarrow X_n \xrightarrow{\mathcal{L}^p} X$$

Proof. Since $g(x) = x^{\frac{q}{p}}$ is convex, by Jensen's Inequality,

$$\mathbb{E}|X_n - X|^q = \mathbb{E}\left[|X_n - X|^{p\cdot\frac{q}{p}}\right] \geqslant \left[\mathbb{E}|X_n - X|^p\right]^{\frac{q}{p}}$$

Then

$$\left[\mathbb{E}|X_n - X|^q\right]^{\frac{1}{q}} \geqslant \left[\mathbb{E}|X_n - X|^p\right]^{\frac{1}{p}}$$

□

Theorem 4.19.

$$X_n \xrightarrow{\mathcal{L}_1} X \Longrightarrow X_n \xrightarrow{\mathbb{P}} X$$

But not vice versa.

Proof. For any fixed $\varepsilon > 0$,

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(|X_n - X|^p > \varepsilon^p) \leqslant \frac{\mathbb{E}|X_n - X|^p}{\varepsilon^p} \to 0$$

Now we consider a counter example for $\nLeftarrow$, Construct a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ that satisfies

$$\Omega = [0,1], \mathcal{F} = \mathcal{B}([0,1]), \mathbb{P}([a,b]) = b - a, 0 \leqslant a \leqslant b \leqslant 1$$

Define a scaled indicator function,

$$X_n(\omega) = 2^n \cdot \mathbb{I}_{\left[0,\frac{1}{n}\right]}(\omega) = \begin{cases} 2^n, & \omega \in \left[0, \frac{1}{n}\right] \\ 0, & \text{o.w.} \end{cases}$$

We may think that $X = 0$ is the limit of X_n, then we prove it:

$$\forall \varepsilon > 0, \mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(|X_n| > \varepsilon) < \frac{1}{n} \to 0$$

Therefore, $X_n \xrightarrow{\mathbb{P}} X$. But for $\mathbb{E}|X_n - X| = \mathbb{E}|X_n| = \frac{2^n}{n} \nrightarrow 0$, therefore,

$$X_n \xrightarrow{\mathcal{L}_1} X \nLeftarrow X_n \xrightarrow{\mathbb{P}} X$$

□

Theorem 4.20.

$$X_n \xrightarrow{\mathbb{P}} X \Longrightarrow X_n \xrightarrow{d} X$$

But not vice versa.

Proof. Note that $\{X_n \leqslant x - \varepsilon, X > x\} \subset \{|X_n - X| < \varepsilon\}$

$$\begin{aligned} F_n(x) &= \mathbb{P}(X_n \leqslant x) \\ &= \mathbb{P}(X_n \leqslant x, X \leqslant x+\varepsilon) + \mathbb{P}(X_n \leqslant x, X > x+\varepsilon) \\ &\leqslant \mathbb{P}(X \leqslant x+\varepsilon) + \mathbb{P}(|X_n - X| < \varepsilon) \\ &= F(x+\varepsilon) + \mathbb{P}(|X_n - X| < \varepsilon) \end{aligned}$$

$$\begin{aligned} F(x-\varepsilon) &= \mathbb{P}(X \leqslant x-\varepsilon) \\ &= \mathbb{P}(X \leqslant x-\varepsilon, X_n \leqslant x) + \mathbb{P}(X \leqslant x-\varepsilon, X_n > x) \\ &\leqslant \mathbb{P}(X_n \leqslant x) + \mathbb{P}(|X_n - X| < \varepsilon) \end{aligned}$$

Then

$$F(x+\varepsilon) + \mathbb{P}(|X_n - X| < \varepsilon) \geqslant F_n(x) \geqslant F(x-\varepsilon) + \mathbb{P}(|X_n - X| < \varepsilon)$$

Since $\forall \varepsilon > 0, \mathbb{P}(|X_n - X| > \varepsilon) \to 0$, let $n \to \infty$, we get

$$F(x+\varepsilon) \geqslant \limsup_{n\to\infty} F_n(x) \geqslant \liminf_{n\to\infty} F_n(x) \geqslant F(x-\varepsilon)$$

If F is continuous at t, then

$$F(t+0) \geqslant \lim_{n\to\infty} F_n(t) \geqslant F(t-0) \Longrightarrow \lim_{n\to\infty} F_n(t) = F(t)$$

Now, consider a counter example for $\nLeftarrow$, $X \sim N(0,1)$, $X_n = -X$, $F_n(x) = \Phi(x) \to \Phi(x)$, but

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(|-X - X| > \varepsilon) = \mathbb{P}\left(|X| > \frac{\varepsilon}{2}\right) = 2\Phi\left(-\frac{\varepsilon}{2}\right) > 0$$

□

Theorem 4.21. If X is a point mass,

$$X_n \xrightarrow{\mathbb{P}} c \Longleftarrow X_n \xrightarrow{d} c$$

Proof. $\forall \varepsilon > 0$,

$$\begin{aligned} \mathbb{P}(|X_n - c| > \varepsilon) &= \mathbb{P}(X_n < c-\varepsilon) + \mathbb{P}(X_n > c+\varepsilon) \\ &\leqslant \mathbb{P}(X_n \leqslant c-\varepsilon) + \mathbb{P}(X_n > c+\varepsilon) \\ &= F_n(c-\varepsilon) + 1 - F_n(c+\varepsilon) \end{aligned}$$

For a point mass, we have $F(\varepsilon + c) = 1, F(\varepsilon - c) = 0$, then

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P}(|X_n - c| > \varepsilon) &\leqslant \lim_{n\to\infty} F_n(c-\varepsilon) + 1 - \lim_{n\to\infty} F_n(c+\varepsilon) \\ &= F(c-\varepsilon) + 1 - F(c+\varepsilon) = 0 \end{aligned}$$

□

Theorem 4.22. (Slutzky's theorem) Let $\{X_n\}_{n=1}^{\infty}, \{Y_n\}_{n=1}^{\infty}$ be two sequences of random variables, and g is a continuous function, c is a constant, then

1. $X_n \xrightarrow{\mathbb{P}} X, Y_n \xrightarrow{\mathbb{P}} Y \Longrightarrow X_n + Y_n \xrightarrow{\mathbb{P}} X + Y, X_n Y_n \xrightarrow{\mathbb{P}} XY$
2. $X_n \xrightarrow{\mathcal{L}^2} X, Y_n \xrightarrow{\mathcal{L}^2} Y \Longrightarrow X_n + Y_n \xrightarrow{\mathcal{L}^2} X + Y$
3. $X_n \xrightarrow{d} X, Y_n \xrightarrow{d} c \Longrightarrow X_n + Y_n \xrightarrow{d} X + c, X_n Y_n \xrightarrow{d} cX$
4. $X_n \xrightarrow{\mathbb{P}} X \Longrightarrow g(X_n) \xrightarrow{\mathbb{P}} g(X)$
5. $X_n \xrightarrow{d} X \Longrightarrow g(X_n) \xrightarrow{d} g(X)$

Note: $X_n \xrightarrow{d} X, Y_n \xrightarrow{d} Y \nRightarrow X_n + Y_n \xrightarrow{d} X + Y$.

4.3 Law of Large Numbers and Central Limit Theorem

Theorem 4.23. (Weak Law of Large Number) Suppose $X_1, X_2, \cdots, X_n$ are i.i.d., $\mathbb{E}[X_i] = \mu$, $\mathrm{Var}(X_i) = \sigma^2 < \infty$, then $\bar{X}_n \xrightarrow{\mathbb{P}} \mu$.

Note: $\mathrm{Var}(X_i) = \sigma^2 < \infty$ is **not necessary**, without the condition, the WLLN still holds.

Proof. By Markov' inequality, for $\forall \varepsilon_1 > 0$, and any given $\varepsilon > 0$, we can choose $N = \left(\frac{\sigma^2}{\varepsilon^2} + 1\right)\frac{1}{\varepsilon_1}$, for all $n \geqslant N$

$$\mathbb{P}\left(\left|\bar{X}_n - \mu\right| < \varepsilon\right) = \mathbb{P}\left(\left|\bar{X}_n - \mu\right|^2 < \varepsilon^2\right) \leqslant \frac{\mathbb{E}\left|\bar{X}_n - \mu\right|^2}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} < \varepsilon_1$$

□

Theorem 4.24. (Strong Law of Large Number) Suppose $X_1, X_2, \cdots, X_n$ are i.i.d., $\mathbb{E}[X_i] = \mu$, $\mathrm{Var}(X_i) = \sigma^2 < \infty$, then $\bar{X}_n \xrightarrow{\text{a.s.}} \mu$.

Theorem 4.25. (Central Limit Theorem) Suppose $X_1, X_2, \cdots, X_n$ are i.i.d., $\mathbb{E}[X_i] = \mu$, $\mathrm{Var}(X_i) = \sigma^2 < \infty$, then

$$\frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{d} N(0,1)$$

Note:

1. Prove by MGF and Taylor's expansion.
2. $\frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu}{\sqrt{\frac{S_n}{n}}} \xrightarrow{d} N(0,1)$, where $S_n = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$.
3. Actually, the theorem still holds if $S_n = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$.

Theorem 4.26. (CLT for independent r.v.s) Let $X_1, X_2, \cdots$ be independent r.v.s each having finite mean μ_i and finite variance σ_i, then

$$\frac{\sum_{i=1}^{n}(X_i - \mu_i)}{\sqrt{\sum_{i=1}^{n}\sigma_i^2}} \xrightarrow{d} N(0,1)$$

Definition 4.27. $\boldsymbol{X} = (X_1, X_2, \cdots, X_k)$ is **multivariate normal distributed**, denoted by $\boldsymbol{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if $\boldsymbol{X}$ has the density,

$$f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{k/2} \cdot \sqrt{\det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu})\right)$$

where $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}]$, $\boldsymbol{\Sigma} = (\sigma_{ij})_{k\times k}$ is the covariance matrix.

Theorem 4.28. (Multivariate CLT) Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n, \cdots$, be i.i.d. random vectors, with mean $\boldsymbol{\mu}_{k\times 1}$, covariance $\boldsymbol{\Sigma}_{k\times k}$, then

$$\sqrt{n}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \xrightarrow{d} N_k(0, \boldsymbol{\Sigma})$$

where $\bar{\boldsymbol{X}} = (\bar{\boldsymbol{X}}_1, \bar{\boldsymbol{X}}_2, \cdots, \bar{\boldsymbol{X}}_k)^{\mathrm{T}}$.

Note that $\boldsymbol{\Sigma}$ is symmetric and positive definiteness, so is $\boldsymbol{\Sigma}^{-1}$, by decomposition,

$$\boldsymbol{P}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{P} = \boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_k \end{pmatrix} = \operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_k)$$

where $\lambda_1, \lambda_2, \cdots, \lambda_k > 0$, and $\boldsymbol{\Sigma}^{-1} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{-1}$. Let

$$\boldsymbol{\Lambda}_1 = \operatorname{diag}\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \cdots, \sqrt{\lambda_k}\right)$$

and $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}_1\boldsymbol{P}^{-1}$, then $\boldsymbol{\Lambda}_1^2 = \operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_k)$

$$\boldsymbol{A}^2 = \boldsymbol{P}\boldsymbol{\Lambda}_1\boldsymbol{P}^{-1}\boldsymbol{P}\boldsymbol{\Lambda}_1\boldsymbol{P}^{-1} = \boldsymbol{P}\boldsymbol{\Lambda}_1^2\boldsymbol{P}^{-1} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{-1} = \boldsymbol{\Sigma}^{-1}$$

Therefore, we may guess that

$$\sqrt{n}\left(\boldsymbol{P}\boldsymbol{\Lambda}_1\boldsymbol{P}^{-1}\right)(\bar{\boldsymbol{X}} - \boldsymbol{\mu})_{k\times 1} \xrightarrow{d} N_k(0, \boldsymbol{I}_n)$$

Theorem 4.29. (The Delta Method) if $\frac{\sqrt{n}(Y_n - \mu)}{\sigma} \xrightarrow{d} N(0,1)$, and g is differentiable function such that $g'(\mu) \neq 0$. Then

$$\frac{\sqrt{n}(g(Y_n) - g(\mu))}{|g'(\mu)|\,\sigma} \xrightarrow{d} N(0,1)$$

i.e.,

$$Y_n \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{n}\right) \Longrightarrow g(Y_n) \xrightarrow{d} N\left(g(\mu), [g'(\mu)]^2 \frac{\sigma^2}{n}\right)$$

Idea of Proof.

$$g(Y_n) \approx g(\mu) + g'(\mu)(Y_n - \mu)$$

where $g'(\mu)$ is a constant, and $Y_n - \mu \xrightarrow{d} N\left(0, \frac{\sigma^2}{n}\right)$

□

Theorem 4.30. (The Multivariate Delta Method) if $\sqrt{n}(\boldsymbol{Y}_n - \mu) \xrightarrow{d} N(0, \boldsymbol{\Sigma})$, where$\boldsymbol{Y}_n = (Y_{n1}, Y_{n2}, \cdots, Y_{nk})$, and $g: \mathbb{R}^k \to \mathbb{R}$ is differentiable function such that $\nabla g(\boldsymbol{\mu}) \neq \boldsymbol{0}$. Then

$$\sqrt{n}(g(\boldsymbol{Y}_n) - g(\boldsymbol{\mu})) \xrightarrow{d} N\left(0, [\nabla g(\boldsymbol{\mu})]^{\mathrm{T}} \boldsymbol{\Sigma} \nabla g(\boldsymbol{\mu})\right)$$

Note: Have a guess, if $g: \mathbb{R}^k \to \mathbb{R}^m$, then $\nabla \boldsymbol{g}(\boldsymbol{\mu})_{k\times m}$ is a matrix instead of vectors. $\sqrt{n}(\boldsymbol{g}(\boldsymbol{Y}_n) - \boldsymbol{g}(\boldsymbol{\mu})) \xrightarrow{d} N\left(0, [\nabla \boldsymbol{g}(\boldsymbol{\mu})]^{\mathrm{T}} \boldsymbol{\Sigma}_{k\times k} \nabla \boldsymbol{g}(\boldsymbol{\mu})\right)$

Part II

Estimation Theory

Chapter 5

Introduction to Statistical Inference

Remark 5.1. Notations.

- Statistical model: $\mathfrak{F}$ is a set of distributions or densities.
- Parametric model: $\mathfrak{F}$ can be parameterized by finite number of parameters.
- Parameter space: Θ.
- Parameter: θ.
- **Parameters of interest** α, and **nuisance parameters** β
- $\mathfrak{F} = \{f(x;\theta) : \theta \in \Theta\}$.
- $\mathbb{P}_\theta(X \in A) = \int_A f(x;\theta)\,\mathrm{d}x$.
- $\mathbb{E}_\theta(r(X)) = \int_\Omega r(x) f(x;\theta)\,\mathrm{d}x$.

Remark 5.2. Parametric and nonparametric models.

1. Parametric models.
   - The complexity of the model is bounded.
   - In other words, the number of parameters of the model is finite.
   - If we get more observations, $\dim(\Theta)$ won't become larger.
     If we only are concerned about μ, then σ is the nuisance parameter.
2. Nonparametric models.
   - Parameter is infinite.
   - If we get more observations, $\dim(\Theta)$ can become larger.
     The amount of information that θ can capture about data can grow as the amount of data grows.
   - E.g., stochastic processes.

Example 5.3. Parametric and nonparametric models.

1. (Parametric) $X_1, X_2, \cdots, X_n \sim F$, and PDF $f \in \mathfrak{F}$

$$\mathfrak{F} = \left\{ f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \mu \in \mathbb{R}, \sigma > 0 \right\}$$

   Then $\theta = (\mu, \sigma) \in \Theta = \mathbb{R} \times (0, \infty) \subset \mathbb{R}^2$.

2. (Nonparametric) Let $X_1, \cdots, X_n$ be independent observations from a CDF F. The problem to estimate F by assuming only $F \in \mathfrak{F}_{all} = \{\text{all CDFs}\}$ is a nonparametric model.

   In this case, the mean value $\mu = T(F) = \int x \mathrm{d}F(x)$, and the median $T(F) = F^{-1}(1/2)$ are functions of F, we call that F is a **statistical functional**.

Example 5.4. Let $X_1, \cdots, X_n$ be independent observations from a CDFF, and let $f = F\prime$ be the PDF. We want to estimate f. We can assume that

$$f \in \mathfrak{F} = \mathfrak{F}_{dens} \cap \mathfrak{F}_{sob}$$

where $\mathfrak{F}_{dens}$ is a set of all PDFs, and

$$\mathfrak{F}_{sob} = \left\{ f : \int_{-\infty}^{\infty} (f''(x))^2 \mathrm{d}x < \infty \right\}$$

is called **Sobolev space**, it is the set of functions that are not too wiggly.

Example 5.5. Suppose $\theta \in \Theta \subset \mathbb{R}^d$, sample: $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$, X can be a vector.

1. X: regressors/ predictors/ independent variables.
2. Y: outcomes/ response variables/ dependent variables.
3. Regression function or **conditional expectation function** (CEF):

$$m(x) = \mathbb{E}[Y | X = x] : \mathbb{R}^k \to \mathbb{R}$$

4. Define $e = y - m(x)$.
   - $Y = m(X) + e$.
   - $\mathbb{E}[e | X] = \mathbb{E}[Y | X] - \mathbb{E}[m(X) | X] = \mathbb{E}[Y | X] - m(X) = 0$.
5. Given $\mathfrak{F} = \{f(x; \theta) : \theta \in \Theta\}$, if $m \in \mathfrak{F}$, we call it a **parametric model** e.g., $m(X) = X^3\theta_1 + X^2\theta_2 + \theta_3$, in this case, particularly, we cay m is linear in $\theta = (\theta_1, \theta_2, \theta_3)$.
6. Prediction model: predict Y_{n+1} based on X_{n+1}.
7. If Y is discrete, we call **classification** instead of regression.
8. Curve estimate: estimate m in function form.

Chapter 6

Point Estimate

6.1 Frequentists' Approach (θ is Fixed, but unknown)

6.1.1 MM and MLE

Definition 6.1. Point Estimation.

$\hat{\theta}_n$ or $\hat{\theta}$ is a function of data, i.e., if we have i.i.d. $(X_1, X_2, \cdots, X_n)$, $\hat{\theta}_n = g(X_1, X_2, \cdots, X_n)$ is a **point estimation** of θ.

Definition 6.2. Method of Moments (Karl Pearson 1800s).

Let $X_1, X_2, \cdots, X_n$ be a sample from $f(x;\theta)$, $\theta = (\theta_1, \theta_2, \cdots, \theta_k)$. For $1 \leqslant j \leqslant k$,

$$\alpha_j := \mathbb{E}\left[X^j\right] = \int_{\mathbb{R}} x^j f(x;\theta)\,\mathrm{d}x$$

and define j-th sample moment

$$\hat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$$

The method of moments estimator $\hat{\theta}_n$ is defined to be the value of $\hat{\theta}$, such that

$$\begin{aligned}
\alpha_1\left(\hat{\theta}_n\right) &= \hat{\alpha}_1 \\
\alpha_2\left(\hat{\theta}_n\right) &= \hat{\alpha}_2 \\
&\vdots \\
\alpha_n\left(\hat{\theta}_n\right) &= \hat{\alpha}_n
\end{aligned}$$

Definition 6.3. Likelihood Function.

Let $X_1, X_2, \cdots, X_n \sim f(x;\theta)$ and be i.i.d., the **likelihood function** is defined as the joint density of data

$$\begin{aligned}
\mathcal{L}_n(\theta) = \mathcal{L}(\theta|\, X) &= \mathcal{L}(\theta_1, \theta_2, \cdots, \theta_k|\, X_1, X_2, \cdots, X_n) \\
&:= \prod_{i=1}^{n} f(x_i;\theta_1, \theta_2, \cdots, \theta_k) = \prod_{i=1}^{n} f(x_i;\theta)
\end{aligned}$$

Definition 6.4. 1. Maximum Likelihood Estimate. The **maximum likelihood estimator** (MLE) is the value of θ that maximize $\mathcal{L}_n(\theta)$, i.e.,

$$\hat{\theta}_{MLE} = \arg\max_{\theta\in\Theta} \mathcal{L}_n(\theta)$$

Note: $\ell_n(\theta) := \log(\mathcal{L}_n(\theta)) = \sum_{i=1}^n \log f(x_i;\theta)$, then

$$\hat{\theta}_{MLE} = \arg\max_{\theta\in\Theta} \mathcal{L}_n(\theta) = \arg\max_{\theta\in\Theta} \ell_n(\theta)$$

Moreover, if $\ell_n(\theta) = Ag(\theta) + B$ where $A > 0, B$ are constants, then

$$\hat{\theta}_{MLE} = \arg\max_{\theta\in\Theta} \mathcal{L}_n(\theta) = \arg\max_{\theta\in\Theta} \ell_n(\theta) = \arg\max_{\theta\in\Theta} g(\theta)$$

Since

$$\forall\theta\in\Theta, \mathcal{L}_n\left(\hat{\theta}_{MLE}\right) \geqslant \mathcal{L}_n(\theta) \Longleftrightarrow \ell_n\left(\hat{\theta}_{MLE}\right) \geqslant \ell_n(\theta) \Longleftrightarrow g\left(\hat{\theta}_{MLE}\right) \geqslant g(\theta)$$

The MLE is the parameter point for which the observed sample is most likely.

Definition 6.5. Some concepts.

1. The distribution of $\hat{\theta}_n$ is called **sampling distribution**.
2. $\mathsf{bias}\left(\hat{\theta}_n\right) = \mathbb{E}_\theta\left(\hat{\theta}_n\right) - \theta$.

   Note:

$$\begin{aligned}\mathbb{E}_\theta\left(\hat{\theta}_n\right) &= \mathbb{E}_{f(x;\theta)}\left(\hat{\theta}_n\right) = \mathbb{E}_{f(x;\theta)}\left(g(X_1, X_2, \cdots, X_n)\right) \\ &= \int\cdots\int_{\mathbb{R}^n} g(x_1, x_2, \cdots, x_n) \prod_{i=1}^n f(x_i;\theta)\mathrm{d}x_1\cdots\mathrm{d}x_n\end{aligned}$$

3. The standard deviation of $\hat{\theta}_n$ is called **standard error**, denoted by $\mathsf{se} = \mathsf{se}\left(\hat{\theta}_n\right) = \sqrt{\mathrm{Var}\left(\hat{\theta}_n\right)}$, estimated standard error is denoted by $\widehat{\mathsf{se}}$.

   Note: Similarly,

$$\begin{aligned}\mathrm{Var}_\theta\left(\hat{\theta}_n\right) &= \mathbb{E}_\theta\left[\hat{\theta}_n - \mathbb{E}_\theta\left(\hat{\theta}_n\right)\right]^2 = \mathbb{E}_{f(x;\theta)}\left[\hat{\theta}_n - \mathbb{E}_\theta\left(\hat{\theta}_n\right)\right]^2 \\ &= \mathbb{E}_\theta\left[\hat{\theta}_n^2\right] - \left[\mathbb{E}_\theta\left(\hat{\theta}_n\right)\right]^2 \\ &= \int\cdots\int_{\mathbb{R}^n} g^2(x_1, x_2, \cdots, x_n) \prod_{i=1}^n f(x_i;\theta)\mathrm{d}x_1\cdots\mathrm{d}x_n - \left[\mathbb{E}_\theta\left(\hat{\theta}_n\right)\right]^2\end{aligned}$$

4. We say that $\hat{\theta}$ is **unbiased** if $\mathbb{E}\left(\hat{\theta}_n\right) = \theta$.
5. We say that $\hat{\theta}_n$ is **consistent**, if $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$.
6. We say that $\hat{\theta}_n$ is **asymptotically normal,** if $\frac{\hat{\theta}_n - \theta}{\mathsf{se}\left(\hat{\theta}_n\right)} \xrightarrow{d} N(0,1)$.

7. Mean squared error: $\text{MSE} = \mathbb{E}\left(\hat{\theta}_n - \theta\right)^2$.

$$\begin{aligned}
\text{MSE} &= \mathbb{E}\left(\hat{\theta}_n - \theta\right)^2 \\
&= \mathbb{E}\hat{\theta}_n^2 + \theta^2 - 2\theta\mathbb{E}\hat{\theta}_n \\
&= \left[\mathbb{E}\hat{\theta}_n\right]^2 + \theta^2 - 2\theta\mathbb{E}\hat{\theta}_n + \mathbb{E}\hat{\theta}_n^2 - \left[\mathbb{E}\hat{\theta}_n\right]^2 \\
&= \left(\mathbb{E}\hat{\theta}_n - \theta\right)^2 + \text{Var}\left(\hat{\theta}_n\right) \\
&= \mathsf{bias}^2\left(\hat{\theta}_n\right) + \text{Var}\left(\hat{\theta}_n\right)
\end{aligned}$$

Note: If $n \to \infty, \mathsf{bias}^2\left(\hat{\theta}_n\right) \to 0, \text{Var}\left(\hat{\theta}_n\right) \to 0$, then $\hat{\theta}_n \xrightarrow{\mathcal{L}^2} \theta$, then $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$.

Example 6.6. Let $X_1, X_2, \cdots, X_n \sim N(\mu, \sigma^2)$ and be i.i.d., from MM and MLE we both get $\hat{\sigma} = \frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}\right)^2$, but it is biased, since $\mathbb{E}\left[\hat{\sigma}^2\right] = \frac{n-1}{n}\sigma^2$ But if $n \to \infty$, $\mathbb{E}\left[\hat{\sigma}^2\right] \to \sigma^2$.

Definition 6.7. Best Unbiased Estimator (BUE)

We say that $\hat{\theta}_n$ is the **Best Unbiased Estimator**, if $\mathbb{E}_\theta\left[\hat{\theta}_n\right] = \theta$ and for all $\tilde{\theta}_n$ with $\mathbb{E}_\theta\left[\tilde{\theta}_n\right] = \theta$,

$$\text{Var}_\theta\left[\hat{\theta}_n\right] \leqslant \text{Var}_\theta\left[\tilde{\theta}_n\right]$$

Theorem 6.8. For all estimator of μ, if $\mathbb{E}_\mu\left[\tilde{\mu}_n\right] = \mu$, then

$$\text{Var}_\mu\left(\tilde{\mu}_n\right) \geqslant \frac{\sigma^2}{n}$$

Therefore, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$ is the best unbiased estimator.

Proof. Skipped. We need to know the C-R lower bound first. □

Example 6.9. Show that OLS estimator $\hat{\beta}$ is the BUE.

Moreover, OLS estimator is Best Linear Unbiased Estimator, which is a subset of BUE because of the assumption of Linearity.

Proof. See Woodridge Page.127. □

Theorem 6.10. Properties of MM: For the method of moments estimator $\hat{\theta}_n$,

1. $\hat{\theta}_n$ exists with probability 1.
2. $\hat{\theta}_n$ is consistent.
3. $\hat{\theta}_n$ is asymptotical normal.

Theorem 6.11. Properties of MLE: For the MLE $\hat{\theta}_n$,

1. $\hat{\theta}_n$ is consistent.
2. $\hat{\theta}_n$ is equivariant, or invariant.

3. $\hat{\theta}_n$ is asymptotical normal.
4. $\hat{\theta}_n$ is asymptotical optimal.
5. $\hat{\theta}_n$ is asymptotically equals to Bayesian Estimator.

6.1.2 Properties of MLE

Definition 6.12. **K-L distance** (Kullback-Leibler Divergence).

For f, g are PDFs, and they share the same support $\mathcal{X}$, i.e.,

$$\int_{\mathcal{X}} f = \int_{\mathcal{X}} g = 1$$

Define

$$D(f, g) = \int_{\mathcal{X}} f(x) \ln \frac{f(x)}{g(x)} \mathrm{d}x$$

Remark 6.13. Properties of K-L distance.

1. For $f \neq g, D(f, g) \neq D(g, f)$.
2. $D(f, f) = 0$.
$$D(f, f) = \int_{\mathcal{X}} f(x) \log \frac{f(x)}{f(x)} \mathrm{d}x = \int_{\mathcal{X}} f(x) \log 1 \mathrm{d}x = 0$$
3. $D(f, g) \geqslant 0$. By Jensen's Inequality,
$$\begin{aligned} D(f, g) &= \int_{\mathcal{X}} f(x) \log \frac{f(x)}{g(x)} \mathrm{d}x = \mathbb{E}_f \left[\log \frac{f(x)}{g(x)} \right] = -\mathbb{E}_f \left[\log \frac{g(x)}{f(x)} \right] \\ &\geqslant -\log \mathbb{E}_f \left[\frac{g(x)}{f(x)} \right] = \log \int_{\mathcal{X}} g(x) \mathrm{d}x = 0 \end{aligned}$$

Definition 6.14. We call a model is **identifiable** if

$$\forall \theta_1 \neq \theta_2, D(\theta_1, \theta_2) := D(f(x; \theta_1), f(x; \theta_2)) > 0$$

Theorem 6.15. Asymptotic properties of MLE.

Let $X_1, X_2, \cdots, X_n \sim f(x; \theta)$ and be i.i.d.. Let θ^* denote the true value of θ. Define

$$M_n = \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(X_i; \theta)}{f(X_i; \theta^*)}$$

and $M(\theta) = -D(\theta^*, \theta)$. Suppose that

$$\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\mathbb{P}} 0$$

and that for $\forall \varepsilon > 0$,

$$\sup_{\theta : |\theta - \theta^*| \geqslant \varepsilon} M(\theta) < M(\theta^*)$$

Let $\hat{\theta}_n$ denote the MLE, then

$$\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta^*$$

Intuition. The log-likelihood function is

$$\ell_n(\theta) = \sum_{i=1}^{n} \log f(x;\theta)$$

then

$$\arg\max_{\theta} M_n(\theta) = \arg\max_{\theta} \ell_n(\theta)$$

By WLLN, as $n \to \infty$,

$$M_n(\theta) \xrightarrow{\mathbb{P}} \mathbb{E}_\theta\left[\log \frac{f(X;\theta)}{f(X;\theta^*)}\right]$$

and

$$\begin{aligned}
\mathbb{E}_\theta\left[\log \frac{f(X;\theta)}{f(X;\theta^*)}\right] &= \int_{\mathcal{X}} f(x;\theta^*) \log \frac{f(x;\theta)}{f(x;\theta^*)} \mathrm{d}x \\
&= -\int_{\mathcal{X}} f(x;\theta^*) \log \frac{f(x;\theta^*)}{f(x;\theta)} \mathrm{d}x \\
&= -D(\theta^*, \theta) \leqslant 0
\end{aligned}$$

Therefore,

$$M_n(\theta) \to -D(\theta^*,\theta) \Longrightarrow \arg\max_{\theta} M_n(\theta) \to \arg\max_{\theta} \{-D(\theta^*,\theta)\} = \theta \Longrightarrow \hat{\theta}_n \to \theta$$

□

Proof. Since $\hat{\theta}_n$ maximizes $M_n(\theta)$, then $M_n\left(\hat{\theta}_n\right) \geqslant M_n(\theta^*)$. Hence,

$$\begin{aligned}
M(\theta^*) - M\left(\hat{\theta}_n\right) &= M_n(\theta^*) - M\left(\hat{\theta}_n\right) + M(\theta^*) - M_n(\theta^*) \\
&\leqslant M_n\left(\hat{\theta}_n\right) - M\left(\hat{\theta}_n\right) + M(\theta^*) - M_n(\theta^*) \\
&\leqslant \sup_{\theta} \left| M_n\left(\hat{\theta}_n\right) - M\left(\hat{\theta}_n\right) \right| + M(\theta^*) - M_n(\theta^*) \\
&\xrightarrow{\mathbb{P}} 0
\end{aligned}$$

Then $\forall \delta > 0$,

$$\mathbb{P}\left(M\left(\hat{\theta}_n\right) < M(\theta^*) - \delta\right) \to 0$$

By the condition that $\forall \varepsilon > 0$, $\sup_{\theta:|\theta-\theta^*|\geqslant\varepsilon} M(\theta) < M(\theta^*)$, we can pick $\forall \varepsilon > 0$, then $\exists \delta > 0$, such that

$$|\theta - \theta^*| \geqslant \varepsilon \Longrightarrow M\left(\hat{\theta}_n\right) < M(\theta^*) - \delta$$

Therefore,

$$\mathbb{P}(|\theta - \theta^*| > \varepsilon) \leqslant \mathbb{P}\left(M\left(\hat{\theta}_n\right) < M(\theta^*) - \delta\right) \to 0$$

□

Theorem 6.16. Equivariant or Invariant Property of MLE.

Let g be a function of θ, whose inverse image is single valued. $\hat{\theta}_n$ denotes the MLE for θ, and $\hat{\tau}_n$ denotes the MLE of $\tau = g(\theta)$. Then

$$\hat{\tau}_n = g\left(\hat{\theta}_n\right)$$

Note: Parametrization won't change MLE.

Proof. For g, there exists inverse function g^{-1}, then $g^{-1}(\hat{\tau}_n) = \hat{\theta}_n$.

$$\begin{aligned}
&\hat{\theta}_n = \arg\max_{\theta} \mathcal{L}_1(\theta) \\
&\hat{\tau}_n = \arg\max_{\tau} \mathcal{L}_2(\tau) = \arg\max_{g^{-1}(\theta)} \mathcal{L}_2\left(g^{-1}(\theta)\right) = g\left(\arg\max_{\theta} \mathcal{L}_2\left(g^{-1}(\theta)\right)\right) \\
&\quad = g\left(\arg\max_{\theta} \mathcal{L}_1(\theta)\right) = g\left(\hat{\theta}_n\right)
\end{aligned}$$

□

Example 6.17. Dose MM estimator have equivariant property?Yes!

Remark 6.18. Asymptotic Normality of MLE

$$\frac{\hat{\theta}_n - \theta}{\mathsf{se}\left(\hat{\theta}_n\right)} \to N(0,1)$$

We don't know θ and $f(x;\theta)$, therefore cannot compute $\mathsf{se}\left(\hat{\theta}_n\right)$. Actually, we need more consideration.

Definition 6.19. Based on previous notation,

- **Score function** is defined as
$$s(X;\theta) = \frac{\partial \log f(X;\theta)}{\partial \theta}$$
- **Fisher information**, for i.i.d. sample, is defined as
$$I_n(\theta) = \mathrm{Var}_\theta\left[\sum_{i=1}^{n} s(X_i;\theta)\right] = \sum_{i=1}^{n} \mathrm{Var}_\theta\left[s(X_i;\theta)\right] = n \cdot I(\theta)$$
where $I(\theta) := \mathrm{Var}_\theta\left[s(X;\theta)\right], X \sim f(x;\theta)$.

Remark 6.20.

$$I(\theta) = \mathrm{Var}_\theta\left[s(X;\theta)\right] = \mathbb{E}_\theta\left[s(X;\theta) - \mathbb{E}_\theta s(X;\theta)\right]^2$$

and

$$\begin{aligned}
\mathbb{E}_\theta s(X;\theta) &= \int_{\mathcal{X}} \frac{\partial \log f(x;\theta)}{\partial \theta} f(x;\theta)\,\mathrm{d}x \\
&= \int_{\mathcal{X}} \frac{1}{f(x;\theta)} \frac{\partial f(x;\theta)}{\partial \theta} \cdot f(x;\theta)\,\mathrm{d}x \\
&= \int_{\mathcal{X}} \frac{\partial f(x;\theta)}{\partial \theta}\mathrm{d}x = \frac{\partial}{\partial \theta}\int_{\mathcal{X}} f(x;\theta)\,\mathrm{d}x \\
&= 0
\end{aligned}$$

then

$$I(\theta) = \mathbb{E}_\theta [s(X;\theta)]^2 = \int_{\mathcal{X}} f(x;\theta) \cdot \left[\frac{\partial \log f(x;\theta)}{\partial \theta}\right]^2 \mathrm{d}x$$

Since we have

$$\mathbb{E}_\theta s(X;\theta) = \int_{\mathcal{X}} \frac{\partial \log f(x;\theta)}{\partial \theta} f(x;\theta)\,\mathrm{d}x = 0$$

then

$$\frac{\partial}{\partial \theta}\mathbb{E}_\theta s(X;\theta) = \frac{\partial}{\partial \theta}\int_{\mathcal{X}} \frac{\partial \log f(x;\theta)}{\partial \theta} f(x;\theta)\,\mathrm{d}x = 0$$

Moreover,

$$\begin{aligned}
&\int_{\mathcal{X}} \left[\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} f(x;\theta) + \frac{\partial \log f(x;\theta)}{\partial \theta}\frac{\partial f(x;\theta)}{\partial \theta}\right]\mathrm{d}x \\
=&\mathbb{E}_\theta\left[\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2}\right] + \int_{\mathcal{X}} \frac{\partial \log f(x;\theta)}{\partial \theta}\frac{\partial f(x;\theta)}{\partial \theta}\frac{1}{f(x;\theta)} \cdot f(x;\theta)\,\mathrm{d}x \\
=&\mathbb{E}_\theta\left[\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2}\right] + \mathbb{E}_\theta\left[\frac{\partial \log f(x;\theta)}{\partial \theta}\right]^2 \\
=&0
\end{aligned}$$

Therefore,

$$I(\theta) = \mathbb{E}_\theta [s(X;\theta)]^2 = -\mathbb{E}_\theta\left[\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2}\right]$$

and

$$I_n(\theta) = n \cdot I(\theta) = -n \cdot \mathbb{E}_\theta\left[\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2}\right]$$

Note: The intuition for $I_n(\theta) = -n \cdot \mathbb{E}_\theta\left[\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2}\right]$.

If n increases, means we have more samples, then $|I_n(\theta)|$ becomes greater.

As for $\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2}$, there are 2 different log-likelihood function f and g.

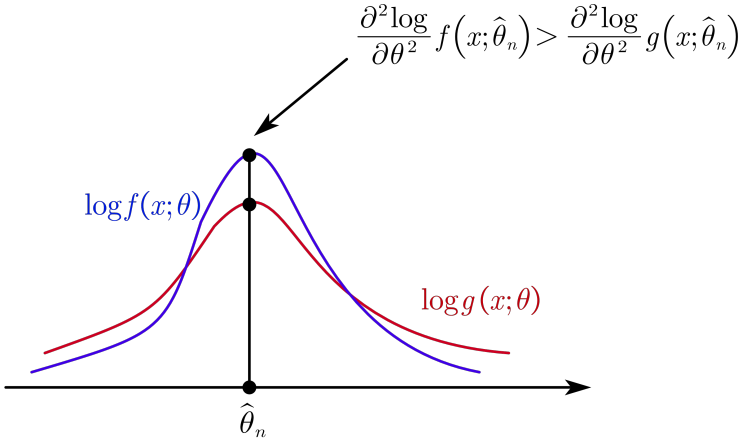

Figure 6.1. Example for comparing $\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2}$.

Theorem 6.21. Asymptotic Normality of MLE.

Let $\mathsf{se}\left(\hat{\theta}_n\right) = \sqrt{\mathrm{Var}_\theta\left(\hat{\theta}_\theta\right)}$.

1. (Theoretically) $\mathsf{se}\left(\hat{\theta}_n\right) \approx \sqrt{\frac{1}{I_n(\theta)}}$, and $\frac{\hat{\theta}_n - \theta}{\mathsf{se}\left(\hat{\theta}_n\right)} \to N(0,1)$.
2. (Computationally) $\widehat{\mathsf{se}}\left(\hat{\theta}_n\right) = \sqrt{\frac{1}{I_n\left(\hat{\theta}_n\right)}}$. Then $\frac{\hat{\theta}_n - \theta}{\widehat{\mathsf{se}}\left(\hat{\theta}_n\right)} \to N(0,1)$.

Proof. See All of Statistics Theorem 9.18. □

Theorem 6.22. Let

$$\hat{C}_n = \left(\hat{\theta}_n - z_{\alpha/2} \cdot \widehat{\mathsf{se}}, \hat{\theta}_n + z_{\alpha/2} \cdot \widehat{\mathsf{se}}\right)$$

Then, $\mathbb{P}_\theta\left(\theta \in C_n\right) \to 1 - \alpha$ as $n \to \infty$.

Proof.

$$\begin{aligned}
\mathbb{P}_\theta\left(\theta \in C_n\right) &= \mathbb{P}_\theta\left(\hat{\theta}_n - z_{\alpha/2} \cdot \widehat{\mathsf{se}} \leqslant \theta \leqslant \hat{\theta}_n + z_{\alpha/2} \cdot \widehat{\mathsf{se}}\right) \\
&= \mathbb{P}_\theta\left(-z_{\alpha/2} \leqslant \frac{\theta - \hat{\theta}_n}{\widehat{\mathsf{se}}} \leqslant z_{\alpha/2}\right) \\
&\to \mathbb{P}\left(-z_{\alpha/2} \leqslant Z \leqslant z_{\alpha/2}\right) \\
&= 1 - \alpha
\end{aligned}$$

□

Note: $z_{0.01} = 2.33, z_{0.025} = 1.96, z_{0.05} = 1.65$.

Theorem 6.23. Asymptotical Optimality for MLE or Asymptotical efficient.

If $\hat{\theta}_n$ is the MLE for θ, $\tilde{\theta}_n$ is another estimator for θ, and

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} N\left(0, u^2\right)$$

$$\sqrt{n}\left(\tilde{\theta}_n - \theta\right) \xrightarrow{d} N\left(0, v^2\right)$$

then $u^2 \leqslant v^2$.

And thus, MLE has the smallest asymptotical variance.

Example 6.24. Let $X_1, X_2, \cdots, X_n \sim N\left(\theta, \sigma^2\right)$ and be i.i.d., the MLE $\hat{\theta}_n = \bar{X}_n$, another reasonable estimator of θ is the sample median $\tilde{\theta}_n$. It can be shown that

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} N\left(0, \sigma^2\right)$$

$$\sqrt{n}\left(\tilde{\theta}_n - \theta\right) \xrightarrow{d} N\left(0, \sigma^2 \frac{\pi}{2}\right)$$

6.2 Bayesian Approach (θ is Random)

Remark 6.25. Introduction.

1. Before we observed data, θ can be characterized by some probability distribution $\pi(\theta)$, which is called **subjective distribution**.

2. After we observed $X_1, X_2, \cdots, X_n$, we can update the distribution.

3. $\pi(\theta)$ is called **prior**.

4. $\pi(\theta|\, X_1, X_2, \cdots, X_n)$- is called **posterior**.

$$\pi(\theta|\, x_1, x_2, \cdots, x_n) := \frac{\pi(\theta) \cdot f(x_1, x_2, \cdots, x_n|\, \theta)}{m(x_1, x_2, \cdots, x_n)}$$

where $m(x_1, x_2, \cdots, x_n) = \int_\Theta \pi(\theta) \cdot f(x_1, x_2, \cdots, x_n|\, \theta)\, \mathrm{d}\theta$ is called **marginal likelihood**, which is a constant, then we have

$$\int_\Theta \pi(\theta)\, \mathrm{d}\theta = 1 \Longrightarrow \int_\Theta \pi(\theta|\, x_1, x_2, \cdots, x_n)\, \mathrm{d}\theta = 1$$

5. Note:

$$m(x_1, x_2, \cdots, x_n) = \int_\Theta \pi(\theta) \cdot f(x_1, x_2, \cdots, x_n|\, \theta)\, \mathrm{d}\theta = \int_\Theta f(x_1, x_2, \cdots, x_n)\, \mathrm{d}\theta$$

since $\pi(\theta) = f_\theta(\theta)$ is the PDF or PMF of θ, we just change the notation.

Definition 6.26. Let $\mathfrak{F}$ be the class of PDFs or PMFs $f(x|\, \theta)$. A class Π is called a **conjugate prior** is the posterior is also in Π.

Example 6.27. Binomial-Beta. $X \sim \text{Binomial}(n, p)$, $p \sim \text{Beta}(\alpha, \beta)$.

Let $X_1, X_2, \cdots, X_N \overset{\text{i.i.d.}}{\sim} \text{Binomial}(n, p)$, the posterior

$$\begin{aligned}
\pi(p|\, x) &\propto \prod_{i=1}^{N} f(x_i|\, p)\, \pi(p) \\
&= \pi(p) \cdot p^{\sum_{i=1}^N x_i} (1-p)^{n\cdot N - \sum_{i=1}^N x_i} \prod_{i=1}^{N} \binom{n}{x_i} \\
&\propto \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \cdot p^{\sum_{i=1}^N x_i} (1-p)^{n\cdot N - \sum_{i=1}^N x_i} \\
&\propto p^{\alpha-1+\sum_{i=1}^N x_i} (1-p)^{\beta-1+n\cdot N-\sum_{i=1}^N x_i} \\
&:= p^{\tilde{\alpha}-1}(1-p)^{\tilde{\beta}-1}
\end{aligned}$$

Then

$$p|\, x \sim \text{Beta}\left(\tilde{\alpha}, \tilde{\beta}\right)$$

and the Bayes estimator is

$$\mathbb{E}_{\pi(p|x)}[p] = \frac{\tilde{\alpha}}{\tilde{\alpha}+\tilde{\beta}} = \frac{\sum_{i=1}^N x_i + \alpha}{\alpha+\beta+nN}$$

To calculate Bayes risk $r(\delta, \pi)$, which equals to

$$\begin{aligned}
r(\delta, \pi) &= \mathbb{E}_{\pi(p)}\left(\mathbb{E}_p (p-\delta)^2\right) \\
&= \mathbb{E}_{\pi(p)}(\mathsf{MSE}) \\
&= \mathbb{E}_{\pi(p)}\left(\mathsf{bias}^2(\delta) + \text{Var}(\delta)\right)
\end{aligned}$$

First compute

$$\mathbb{E}_p(\delta) = \mathbb{E}\left[\frac{\sum_{i=1}^N x_i + \alpha}{\alpha+\beta+nN}\right] = \frac{Np + \alpha}{\alpha+\beta+nN}$$

$$\mathrm{Var}_p(\delta) = \frac{\mathrm{Var}_p\left[\sum_{i=1}^N x_i\right]}{(\alpha+\beta+nN)^2} = \frac{Np(1-p)}{(\alpha+\beta+nN)^2}$$

then

$$\mathsf{bias}^2(\delta) = \left[\frac{Np+\alpha}{\alpha+\beta+nN} - p\right]^2 = \frac{[\alpha - p(\alpha+\beta)]^2}{(\alpha+\beta+nN)^2}$$

and thus, Bayes risk is

$$\begin{aligned}
r(\delta,\pi) =& \mathbb{E}_{\pi(p)}\left[\frac{\alpha^2 + (\alpha+\beta)^2 p^2 - 2\alpha(\alpha+\beta)p}{(\alpha+\beta+nN)^2} + \frac{Np(1-p)}{(\alpha+\beta+nN)^2}\right] \\
=& \frac{\alpha^2}{(\alpha+\beta+nN)^2} + \frac{(\alpha+\beta)^2}{(\alpha+\beta+nN)^2}\frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} \\
& - \frac{2\alpha(\alpha+\beta)}{(\alpha+\beta+nN)^2}\frac{\alpha}{\alpha+\beta} + \frac{Nn}{(\alpha+\beta+nN)^2}\frac{\alpha}{\alpha+\beta} - \frac{Nn}{(\alpha+\beta+nN)^2}\frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} \\
=& \frac{Nn\alpha(\alpha+\beta+1) - Nn\alpha(\alpha+1) - \alpha^2(\alpha+\beta)(\alpha+\beta+1) + (\alpha+\beta)^2\alpha(\alpha+1)}{(\alpha+\beta+nN)^2(\alpha+\beta)(\alpha+\beta+1)} \\
=& \frac{Nn\alpha\beta + \alpha(\alpha+\beta)(\alpha^2+\alpha+\alpha\beta+\beta-\alpha^2-\alpha\beta-\alpha)}{(\alpha+\beta+nN)^2(\alpha+\beta)(\alpha+\beta+1)} \\
=& \frac{Nn\alpha\beta + \alpha(\alpha+\beta)\beta}{(\alpha+\beta+nN)^2(\alpha+\beta)(\alpha+\beta+1)} \\
=& \frac{\alpha\beta}{(\alpha+\beta+nN)(\alpha+\beta)(\alpha+\beta+1)}
\end{aligned}$$

Example 6.28. Poisson-Gamma. $X \sim \mathrm{Poisson}(\lambda)$, $\lambda \sim \mathrm{Gamma}(\alpha,\beta)$.

Let $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Poisson}(\lambda)$, the posterior

$$\pi(\lambda|x) \propto \prod_{i=1}^n \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\frac{1}{\beta^\alpha\Gamma(\alpha)}e^{-\frac{\lambda}{\beta}}\lambda^\alpha \propto e^{-n\lambda-\frac{\lambda}{\beta}}\lambda^{\alpha+\sum_{i=1}^n x_i}$$

Then

$$\lambda|x \sim \mathrm{Gamma}\left(\alpha + \sum_{i=1}^n x_i, \frac{1}{n+\frac{1}{\beta}}\right)$$

The Bayes estimator is

$$\delta = \mathbb{E}_{\pi(\lambda|x)}[\lambda] = \frac{\alpha+\sum_{i=1}^n x_i}{n+\frac{1}{\beta}}$$

Likewise,

$$\mathbb{E}_\lambda[\delta] = \frac{n\lambda+\alpha}{n+\frac{1}{\beta}}$$

$$\mathrm{Var}_\lambda[\delta] = \frac{n\lambda}{\left(n+\frac{1}{\beta}\right)^2}$$

then

$$\mathsf{bias}^2(\delta) = \left[\frac{n\lambda+\alpha}{n+\frac{1}{\beta}} - \lambda\right]^2 = \frac{(\alpha\beta-\lambda)^2}{(n\beta+1)^2}$$

and thus, Bayes risk is

$$\begin{aligned} r(\delta,\pi) &= \mathbb{E}_{\pi(\lambda)}\left(\mathsf{bias}^2(\delta) + \mathrm{Var}_\lambda[\delta]\right) \\ &= \mathbb{E}_{\pi(\lambda)}\left[\frac{(\alpha\beta-\lambda)^2}{(n\beta+1)^2} + \frac{n\lambda\beta^2}{(n\beta+1)^2}\right] \\ &= \mathbb{E}_{\pi(\lambda)}\left[\frac{\alpha^2\beta^2+\lambda^2-2\alpha\beta\lambda}{(n\beta+1)^2} + \frac{n\lambda\beta^2}{(n\beta+1)^2}\right] \\ &= \frac{\alpha^2\beta^2+\alpha(\alpha+1)\beta^2-2\alpha\beta\cdot\alpha\beta+n\alpha\beta^3}{(n\beta+1)^2} \\ &= \frac{\alpha\beta^2+n\alpha\beta^3}{(n\beta+1)^2} \\ &= \frac{\alpha\beta^2}{n\beta+1} \end{aligned}$$

Example 6.29. Normal-Normal. $X \sim N(\theta,\sigma^2)$, where σ^2 is known and $\theta \sim N(a,b^2)$.

Let $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta,\sigma^2)$, the posterior

$$\begin{aligned} \pi(\theta|x) &\propto \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\theta)^2}{2\sigma^2}} \cdot \frac{1}{\sqrt{2\pi}b} e^{-\frac{(\theta-a)^2}{2b^2}} \\ &\propto \exp\left\{-\frac{\sum_{i=1}^n (x_i-\theta)^2}{2\sigma^2} - \frac{(\theta-a)^2}{2b^2}\right\} \\ &= \exp\left\{-\frac{\sum_{i=1}^n x_i^2 + n\theta^2 - 2\theta\sum_{i=1}^n x_i}{2\sigma^2} - \frac{\theta^2+a^2-2a\theta}{2b^2}\right\} \\ &\propto \exp\left\{-\frac{nb^2\theta^2 - 2b^2\theta\sum_{i=1}^n x_i + \sigma\theta^2 - 2a\sigma^2\theta}{2\sigma^2b^2}\right\} \\ &\propto \exp\left\{\frac{-1}{2\sigma^2b^2(nb^2+\sigma^2)}\left(\theta - \frac{b^2\sum_{i=1}^n x_i + a\sigma^2}{nb^2+\sigma^2}\right)^2\right\} \\ &:= \exp\left(-\frac{(x-\tilde{\mu})^2}{2\tilde{\sigma}^2}\right) \end{aligned}$$

Then

$$\theta|x \sim N\left(\tilde{\mu},\tilde{\sigma}^2\right)$$

The Bayes estimator is

$$\delta = \tilde{\mu} = \frac{b^2\sum_{i=1}^n x_i + a\sigma^2}{nb^2+\sigma^2}$$

$$\mathrm{Var}_\theta[d] = \frac{b^4 n\sigma^2}{(nb^2+\sigma^2)^2}$$

then

$$\mathsf{bias}^2(\delta) = \left[\frac{b^2 n\theta + a\sigma^2}{nb^2+\sigma^2} - \theta\right]^2 = \frac{(a-\theta)^2}{(nb^2+\sigma^2)^2}\sigma^4$$

and thus, Bayes risk is

$$\begin{aligned} r(\delta,\pi) &= \mathbb{E}_{\pi(\theta)}\left(\mathsf{bias}^2(\delta) + \mathrm{Var}_\lambda[\delta]\right) \\ &= \mathbb{E}_{\pi(\theta)}\left(\frac{(a-\theta)^2}{(nb^2+\sigma^2)^2}\sigma^4 + \frac{b^4 n\sigma^2}{(nb^2+\sigma^2)^2}\right) \\ &= \frac{b^2\sigma^4}{(nb^2+\sigma^2)^2} + \frac{b^4 n\sigma^2}{(nb^2+\sigma^2)^2} \\ &= \frac{\sigma^2 b^2}{nb^2+\sigma^2} \end{aligned}$$

Example 6.30. Normal-Inverse Gamma. $X \sim N(\mu,\sigma^2)$, where μ is known and $\sigma^2 \sim IG\left(\frac{a_0}{2},\frac{b_0}{2}\right)$.

Let $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu,\sigma^2)$, the posterior

$$\begin{aligned} \pi\left(\sigma^2 \middle| x\right) &\propto f\left(x_1,x_2,\cdots,x_n;\sigma^2\right)\pi\left(\sigma^2\right) \\ &= \prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right)\cdot\frac{\left(\frac{b_0}{2}\right)^{\frac{a_0}{2}}}{\Gamma\left(\frac{a_0}{2}\right)}\left(\frac{1}{\sigma^2}\right)^{\frac{a_0}{2}+1}e^{-\frac{b_0}{2\sigma^2}} \\ &\propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}}\left(\frac{1}{\sigma^2}\right)^{\frac{a_0}{2}+1}e^{-\frac{b_0}{2\sigma^2}} \\ &= \left(\frac{1}{\sigma^2}\right)^{\frac{n+a_0}{2}+1} e^{-\frac{\sum_{i=1}^n (x_i-\mu)^2+b_0}{2\sigma^2}} \\ &:= \left(\frac{1}{\sigma^2}\right)^{\frac{\tilde{a}}{2}+1} e^{-\frac{\tilde{b}}{2\sigma^2}} \end{aligned}$$

Thus,

$$\sigma^2\big| x \sim IG\left(\frac{\tilde{a}}{2},\frac{\tilde{b}}{2}\right)$$

6.3 Decision Theory

Definition 6.31. **Loss function** $L(\theta,d) : \Theta\times\Theta \to \mathbb{R}_+$ is a function that measures the discrepancy between θ and d.

E.g., squared error loss $L(\theta,d) = (\theta-d)^2$; absolute error loss $L(\theta,d) = |d-\theta|$.

Definition 6.32. **Frequentist's Risk Function**.

$$R(\theta,d) = \mathbb{E}_\theta[L(\theta,d)] = \int_{\mathcal{X}} L(\theta,d)\, f(x;\theta)\,\mathrm{d}x$$

In particular, if $L(\theta,d) = (\theta-d)^2$, then $R(\theta,d) = \mathbb{E}_\theta(\theta-d)^2 = \mathrm{MSE}$.

Example 6.33. If we want more penalty for $d > \theta$, then we can define

$$L(d,\theta) = \begin{cases} |d-\theta|, & d \leqslant \theta \\ 10\,|d-\theta|, & d > \theta \end{cases}$$

Remark 6.34. Make a decision. For frequentist, θ is fixed but unknown.

1. If $\forall \theta \in \Theta, R(\theta, d_1) \leqslant R(\theta, d_2)$, then d_1 if preferred.

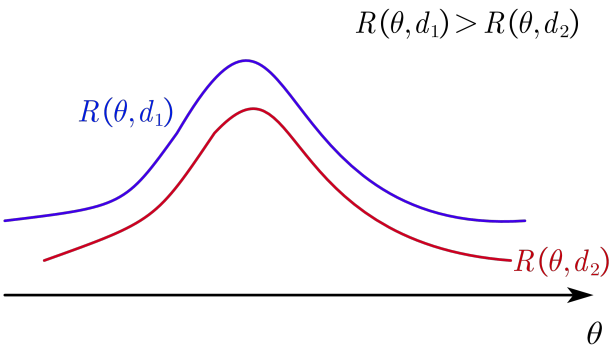

Figure 6.2. Case I.

2. If $R(\theta,d_1), R(\theta,d_2)$ as a function of θ, they cross.

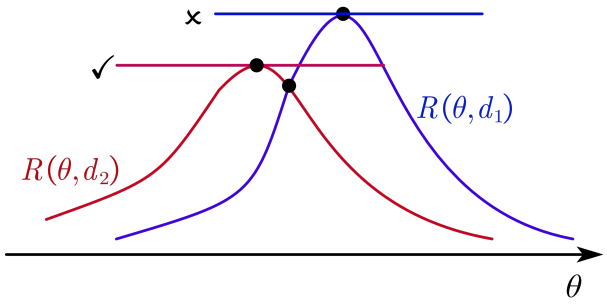

Figure 6.3. Case II.

A possible choice is Min-max Rule, we choose $\inf_d \sup_\theta R(\theta,d)$.

Definition 6.35. **Bayesian Risk Function**.

For Bayesian approach, θ is random.

$$r(d,\pi) = \mathbb{E}_{\pi(\theta)}[R(\theta,d)] = \int_\Theta R(\theta,d)\,\pi(\theta)\,\mathrm{d}\theta$$

where $\pi(\theta)$ is a prior for θ.

Note: It's just a weighted average of frequentists' $R(\theta,d)$.

Definition 6.36. Posterior expected loss.

$$r_p(d,\pi) = \mathbb{E}_{\pi(\theta|x)}[L(\theta,d)] = \int_\Theta L(\theta,d)\,\pi(\theta|x)\,\mathrm{d}\theta$$

Theorem 6.37.

$$\arg\min_d \; r(d,\pi) = \arg\min_d \; r_p(d,\pi)$$

Proof. First note that

$$\pi(\theta|\boldsymbol{x}) = \frac{\pi(\theta)\,f(\boldsymbol{x}|\theta)}{\int_{\mathcal{X}}\pi(\theta)\,f(\boldsymbol{x}|\theta)\,\mathrm{d}\boldsymbol{x}} = \frac{\pi(\theta)\,f(\boldsymbol{x}|\theta)}{\int_{\mathcal{X}} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}} = \frac{\pi(\theta)\,f(\boldsymbol{x}|\theta)}{m(\boldsymbol{x})}$$

Then

$$\begin{aligned}
r(d,\pi) = \mathbb{E}_{\pi(\theta)}[R(\theta,d)] &= \int_\Theta R(\theta,d)\,\pi(\theta)\,\mathrm{d}\theta \\
&= \int_\Theta \mathbb{E}_\theta[L(\theta,d)]\,\pi(\theta)\,\mathrm{d}\theta \\
&= \int_\Theta\int_{\mathcal{X}} L(\theta,d)\,f(\boldsymbol{x};\theta)\,\pi(\theta)\,\mathrm{d}\boldsymbol{x}\mathrm{d}\theta \\
&= \int_\Theta\int_{\mathcal{X}} L(\theta,d)\,f(\boldsymbol{x}|\theta)\,\pi(\theta)\,\mathrm{d}\boldsymbol{x}\mathrm{d}\theta \\
&= \int_\Theta\int_{\mathcal{X}} L(\theta,d)\,m(\boldsymbol{x})\,\pi(\theta|\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\mathrm{d}\theta \\
&= \int_{\mathcal{X}} m(x)\int_\Theta L(\theta,d)\,\pi(\theta|\boldsymbol{x})\,\mathrm{d}\theta\mathrm{d}\boldsymbol{x} \\
&= \int_{\mathcal{X}} m(x)\,r_p(d,\pi)\,\mathrm{d}\boldsymbol{x}
\end{aligned}$$

and $m(\boldsymbol{x}) > 0$ only depends on observed values $x_1, x_2, \cdots, x_n$, which shows the theorem. Note that d is the point estimation of θ, then $d = g(x_1, x_2, \cdots, x_n)$, we choose g to determine d and finally to have the best guess of θ. □

Theorem 6.38. Two useful examples.

1. Decision Rule 1: If $L(\theta,d) = (\theta-d)^2$, then

$$\arg\min_d r_p(d,\pi) = \mathbb{E}_{\pi(\theta|x)}(\theta)$$

2. Decision Rule 2: If $L(\theta,d) = |\theta-d|$, then

$$\arg\min_d r_p(d,\pi) = \text{median of } \; \theta|\,x$$

Proof. Firstly,

$$r_p(d,\pi) = \mathbb{E}_{\pi(\theta|x)}[L(\theta,d)] = \int_\Theta (\theta-d)^2\,\pi(\theta|x)\,\mathrm{d}\theta$$

then

$$\frac{\partial}{\partial d} r_p(d,\pi) = \int_\Theta -2(\theta-d)\,\pi(\theta|x)\,\mathrm{d}\theta = 0$$

or

$$\mathbb{E}_{\pi(\theta|x)}(\theta) = \int_\Theta \theta\pi(\theta|x)\,\mathrm{d}\theta = \int_\Theta d\cdot\pi(\theta|x)\,\mathrm{d}\theta = d\int_\Theta \pi(\theta|x)\,\mathrm{d}\theta = d$$

Secondly,

$$\begin{aligned} r_p(d,\pi) &= \mathbb{E}_{\pi(\theta|x)}[L(\theta,d)] = \int_\Theta |\theta-d|\,\pi(\theta|x)\,\mathrm{d}\theta \\ &= \int_{-\infty}^{d}(d-\theta)\,\pi(\theta|x)\,\mathrm{d}\theta + \int_d^{+\infty}(\theta-d)\,\pi(\theta|x)\,\mathrm{d}\theta \end{aligned}$$

then

$$\begin{aligned} \frac{\partial}{\partial d}r_p(d,\pi) &= \frac{\partial}{\partial d}\int_{-\infty}^{d}(d-\theta)\,\pi(\theta|x)\,\mathrm{d}\theta + \frac{\partial}{\partial d}\int_d^{+\infty}(\theta-d)\,\pi(\theta|x)\,\mathrm{d}\theta \\ &= \int_{-\infty}^{d}\pi(\theta|x)\,\mathrm{d}\theta - \int_d^{+\infty}\pi(\theta|x)\,\mathrm{d}\theta \\ &= \mathbb{P}_{\theta|x}(\theta\leqslant d) - \mathbb{P}_{\theta|x}(\theta>d) \\ &= 0 \end{aligned}$$

or

$$\mathbb{P}_{\theta|x}(\theta\leqslant d) = \mathbb{P}_{\theta|x}(\theta>d) = \frac{1}{2}$$

which leads to d = median of $\theta|x$. □

6.4 Hypothesis Testing

Definition 6.39. A **hypothesis** is a statement of a population parameter θ, $\theta\in\Theta$.

- Null Hypothesis: $H_0:\theta\in\Theta_0$.
- Alternative Hypothesis: $H_1:\theta\in\Theta_1=\Theta\setminus\Theta_0$.

Example 6.40. Two examples.

- Simple hypothesis: $H_0:\theta=\theta_0$.
- Composite hypothesis: $H_1:\theta\geqslant\theta_0$.

Definition 6.41. **Hypothesis testing** is about a decision rule:

- For which sample values, we cannot reject the null H_0.
- For which sample values, we can reject the null H_0.

$$\begin{aligned} &X=(X_1,\cdots,X_n),X_1,\cdots X_n\sim f(x;\theta),\ \text{i.i.d.} \\ &\text{Test statistics}: T(X) \\ &\text{Reject region}: R=\{X:T(X)>c\} \\ &\text{Critical value}: c \\ &X\in R,\text{can reject } H_0 \\ &X\notin R,\text{cannot reject } H_0 \end{aligned}$$

Definition 6.42. Some concepts.

1. **Type I error**: $\mathbb{P}\left(\text{reject}H_0 \middle| H_0 \text{ istrue}\right) = \mathbb{P}\left(X \in R \middle| \theta \in \Theta_0\right)$
2. **Type II error**: $\mathbb{P}\left(\text{cannotreject}H_0 \middle| H_0 \text{ isfalse}\right) = \mathbb{P}\left(X \notin R \middle| \theta \in \Theta_1\right)$
3. **Power function**: $\beta(\theta) = \mathbb{P}_\theta\left(\text{reject}H_0\right) = \mathbb{P}_\theta\left(X \in R\right)$
4. **Size**: $\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$
5. **Power**: $\beta = \inf_{\theta \in \Theta_1} \beta(\theta)$

Example 6.43. $X \in \text{Binomial}(5, \theta)$, consider testing $H_0 : \theta \leqslant \frac{1}{2}, H_1 : \theta > \frac{1}{2}$.

- Decision rule 1: reject H_0 only if all observations are "success", i.e., $X \in \{5\} = R$

$$\beta(\theta) = \mathbb{P}_\theta\left(X \in R\right) = \theta^5$$

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta) = \sup_{\theta \leqslant \frac{1}{2}} \theta^5 = \frac{1}{2^5} = 0.03125$$

Type I error:

$$\mathbb{P}\left(X \in R \middle| \theta \in \Theta_0\right) = \left[\theta^5 \middle| \theta \leqslant \frac{1}{2}\right] = \beta(\theta) \middle| \theta \in \Theta_0 \leqslant \alpha$$

Type II error:

$$\mathbb{P}\left(X \notin R \middle| \theta \in \Theta_1\right) = \left[1 - \theta^5 \middle| \theta > \frac{1}{2}\right] = 1 - \beta(\theta) \middle| \theta \in \Theta_1 \leqslant 1 - \frac{1}{2^5} = 1 - \beta$$

- Decision rule 2: reject H_0 if $X \in \{3, 4, 5\}$

$$\beta(\theta) = \binom{5}{3} \theta^3 (1-\theta)^2 + \binom{5}{4} \theta^4 (1-\theta) + \theta^5$$

- Compare, see Figure. 6.4.

Remark 6.44.

$$\beta(\theta) = \begin{cases} \mathbb{P}\left(\text{Type I error}\right), & \theta \in \Theta_0 \\ 1 - \mathbb{P}\left(\text{Type II error}\right), & \theta \in \Theta_1 \end{cases}$$

Definition 6.45. **Wald Test**.

Under the null $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$.

$$Z = \frac{\hat{\theta}_n - \theta}{\widehat{\mathsf{se}}\left(\hat{\theta}_n\right)} \xrightarrow{d} N(0, 1)$$

The size α test is to reject H_0 if $|Z| > z_{\alpha/2}$.

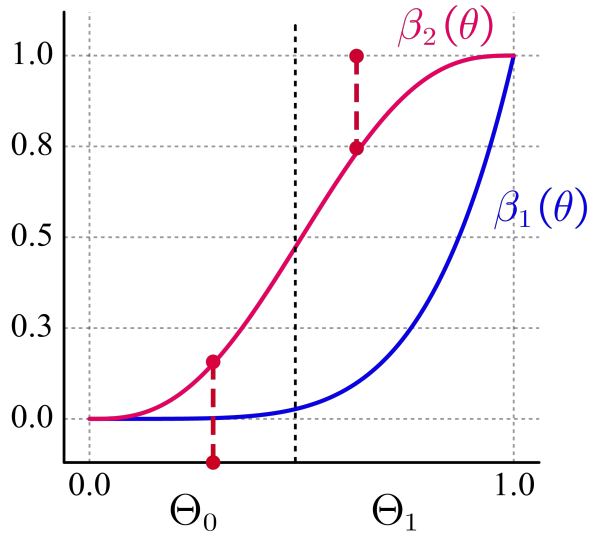

Figure 6.4. Example

Remark 6.46.

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta) = \beta(\theta_0) = \mathbb{P}_{\theta_0}(X \in R) = \mathbb{P}_{\theta_0}\left(\left|\frac{\hat{\theta}_n - \theta}{\widehat{\mathsf{se}}\left(\hat{\theta}_n\right)}\right| > z_{\alpha/2}\right)$$

Remark 6.47. If θ^* is the true value, but $\theta \neq \theta^* \in \Theta_1$, then $\frac{\hat{\theta}_n - \theta^*}{\widehat{\mathsf{se}}\left(\hat{\theta}_n\right)} \xrightarrow{d} N(0,1)$, and

$$\begin{aligned}
\beta(\theta^*) &= \mathbb{P}_{\theta^*}(X \in R) = \mathbb{P}_{\theta^*}\left(|Z| > z_{\alpha/2}\right) \\
&= \mathbb{P}_{\theta^*}\left(\left|\frac{\hat{\theta}_n - \theta}{\widehat{\mathsf{se}}\left(\hat{\theta}_n\right)}\right| > z_{\alpha/2}\right) \\
&= \mathbb{P}_{\theta^*}\left(\left|\frac{\hat{\theta}_n - \theta^*}{\widehat{\mathsf{se}}\left(\hat{\theta}_n\right)} - \frac{\theta - \theta^*}{\widehat{\mathsf{se}}\left(\hat{\theta}_n\right)}\right| > z_{\alpha/2}\right) \\
&= \mathbb{P}_{\theta^*}\left(\frac{\hat{\theta}_n - \theta^*}{\widehat{\mathsf{se}}\left(\hat{\theta}_n\right)} > z_{\alpha/2} + \frac{\theta - \theta^*}{\widehat{\mathsf{se}}\left(\hat{\theta}_n\right)}\right) + \mathbb{P}_{\theta^*}\left(\frac{\hat{\theta}_n - \theta^*}{\widehat{\mathsf{se}}\left(\hat{\theta}_n\right)} < -z_{\alpha/2} + \frac{\theta - \theta^*}{\widehat{\mathsf{se}}\left(\hat{\theta}_n\right)}\right) \\
&= 1 - \Phi\left(z_{\alpha/2} + \frac{\theta - \theta^*}{\widehat{\mathsf{se}}\left(\hat{\theta}_n\right)}\right) + \Phi\left(-z_{\alpha/2} + \frac{\theta - \theta^*}{\widehat{\mathsf{se}}\left(\hat{\theta}_n\right)}\right)
\end{aligned}$$

Definition 6.48. Confidence interval.

$$\hat{C}_n = \left(\hat{\theta}_n \pm z_{\alpha/2}\widehat{\mathsf{se}}\left(\hat{\theta}_n\right)\right)$$

The size α Z-test reject $H_0 : \theta = \theta_0 \Longleftrightarrow \theta \notin \hat{C}_n$.

Definition 6.49. Suppose that $\forall \alpha \in (0,1)$, we have a size α test with rejection region R_α, then ***p*-value** is defined as

$$p = \inf_{\alpha} \{\alpha : X \in R_\alpha\}$$

- $\alpha < p$, cannot reject.
- $\alpha > p$, reject.

Part III

Econometrics

Chapter 7

Linear Regression and Preparations

7.1 Ordinary Least Squares Regression Review

7.1.1 Assumptions and Estimations

Remark 7.1. Notation of linear regression.

$$\hat{\beta} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$$

where

- $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \sum_{i=1}^{n} x_i x_i^{\mathrm{T}}$.
- x_i is a $k \times 1$.
- $\boldsymbol{X} = [x_1, x_2, \cdots, x_n]$ is $K \times n$.

Some conventions:

- X: a random variable or vector X, sometime x_i denotes a random variable as well.
- x: a realization of the random variable or vector X.
- $\boldsymbol{X}$: a stacked data matrix of n-random variables or vectors, sometimes it is written X without bold for simplicity and without confusion.

Remark 7.2. Assumption of linear regression. Consider the model

$$Y = X^{\mathrm{T}}\beta + u$$

where Y is a scalar, $X_{k\times 1}$ is a random vector.

Usually, we assume that

1. Random sampling: Samples (x_i, y_i) are i.i.d. over i.
2. Linearity: The model is correctly specified as a linear model (linear in parameters).

   Note:

- Relax the assumption → e.g., nonlinear model, nonparametric model.

3. $\mathbb{E}\left[XX^{\mathrm{T}}\right]$ is invertible.

   Note:

   - For finite sample, we need $\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}$ exists to compute β.
   - Relax the assumption → e.g., high dimension linear model, Lasso.

4. $\mathbb{E}\left(\boldsymbol{u}|\,\boldsymbol{X}\right)=0$.

   Note:

   - If it is violated, then the estimator is biased.
   - $\mathbb{E}\left(u|\,X\right)=0\Longrightarrow\mathbb{E}\left[uX\right]=\mathbb{E}\left[\mathbb{E}\left(u|\,X\right)X\right]=0$.
   - Relax the assumption → e.g., Instrument variable (IV), fixed effect model in panel data.

5. $\mathbb{E}Y^2<\infty,\mathbb{E}\left(\|X\|^2\right)<\infty$.

   Note:

   - Relax the assumption → e.g., median or quantile estimation.

6. Homoskedasticity (optional): $\mathrm{Var}\left(\boldsymbol{u}|\,\boldsymbol{X}\right)=\sigma_u^2\boldsymbol{I}$.

   Note:

   - With assumptions 1 to 5, we can get $\hat{\beta}\xrightarrow{\mathbb{P}}\beta,\mathbb{E}\left[\hat{\beta}-\beta\right]=0$.
   - With assumption 6, $\mathrm{Var}\left[\hat{\beta}-\beta\right]$ becomes small (efficiency).
   - Relax the assumption → e.g., general least square (GLS), maximum likelihood estimation (MLE).

7. Normality (optional): $u|\,X\sim\mathcal{N}\left(0,\sigma_u^2\right)$.

   Note:

   - With this assumption, we can get more concrete results about normality.
   - Relax the assumption → e.g., asymptotic theory, M-estimator and GMM.

Remark 7.3. The OLS estimator.

1. The OLS estimator as an optimizer.

$$\hat{\beta}_{\mathrm{OLS}}=\arg\min_{\beta}\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\beta_i^{\mathrm{T}}x_i\right)^2$$

   The F.O.C. is

$$-\frac{2}{n}\sum_{i=1}^{n}\left(y_i-\hat{\beta}_n^{\mathrm{T}}x_i\right)\cdot x_i=0\Longleftrightarrow\sum_{i=1}^{n}y_ix_i=\left(\sum_{i=1}^{n}x_ix_i^{\mathrm{T}}\right)\hat{\beta}_n^{\mathrm{T}}$$

The second order condition of a minimize should demand

$$-\frac{2}{n}\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}$$

is positive definite (and thus, invertible).

Therefore,

$$\hat{\beta}_n = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right)$$

2. The OLS estimator as a MLE.

$$y_i = \beta_0^{\mathrm{T}} x_i + \varepsilon_i, \text{ where } x_i \perp \varepsilon_i \text{ and } \varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$$

The likelihood function is

$$f\left(y_i \middle| x_i; \beta\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(y_i - \beta^{\mathrm{T}} x_i\right)^2}{2\sigma^2}}$$

then

$$\ln f\left(y_i \middle| x_i; \beta\right) = -\frac{1}{2}\ln\left(2\pi\right) - \ln\sigma - \frac{\left(y_i - \beta^{\mathrm{T}} x_i\right)^2}{2\sigma^2}$$

The MLE is given by

$$\begin{aligned}\hat{\beta}_{\mathrm{MLE}} &= \arg\max_{\beta} \frac{1}{n}\sum_{i=1}^{n} \ln f\left(y_i \middle| x_i; \beta\right) = \arg\max_{\beta} \frac{1}{n}\sum_{i=1}^{n} -\frac{\left(y_i - \beta^{\mathrm{T}} x_i\right)^2}{2\sigma^2} \\ &= \arg\min_{\beta} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \beta^{\mathrm{T}} x_i\right)^2\end{aligned}$$

3. The OLS estimator as a MM estimator.

The key assumption is

$$\mathbb{E}\left[\varepsilon_i \middle| x_i\right] = 0$$

then

$$\mathbb{E}\left[\varepsilon_i x_i\right] = \mathbb{E}\left[\mathbb{E}\left[\varepsilon_i \middle| x_i\right] x_i\right] = 0 \Longrightarrow \mathbb{E}\left[x_i\left(y_i - \beta_0^{\mathrm{T}} x_i\right)\right] = 0 \Longrightarrow \mathbb{E}\left[x_i y_i\right] = \mathbb{E}\left[x_i x_i^{\mathrm{T}}\right]\beta_0$$

assume that

$$\mathbb{E}\left[x_i x_i^{\mathrm{T}}\right]$$

is positive definite or invertible, we have

$$\beta_0 = \mathbb{E}\left[x_i x_i^{\mathrm{T}}\right]^{-1}\mathbb{E}\left[x_i y_i\right]$$

By method of moment,

$$\hat{\beta}_n = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right)$$

7.1.2 Properties of OLS

Remark 7.4. Estimator in matrix form.

Define the fitted value as

$$\hat{y}_i := \hat{\beta}_n^{\mathrm{T}} x_i$$

Define the residual as

$$\hat{e}_i := y_i - \hat{y}_i = y_i - \hat{\beta}_n^{\mathrm{T}} x_i$$

In matrix form,

$$\hat{\boldsymbol{Y}}_{n\times 1} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix}, \hat{\boldsymbol{e}} = \begin{pmatrix} \hat{e}_1 \\ \hat{e}_2 \\ \vdots \\ \hat{e}_n \end{pmatrix}, \boldsymbol{X}_{n\times k} = \begin{pmatrix} x_1^{\mathrm{T}} \\ x_2^{\mathrm{T}} \\ \vdots \\ x_n^{\mathrm{T}} \end{pmatrix}, \boldsymbol{Y}_{n\times 1} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

Then

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\beta}_n = \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$$

and

$$\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} = \left[\boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\right]\boldsymbol{Y}$$

Two properties.

- $\frac{1}{n}\sum_{i=1}^{n} x_i\hat{e}_i = \frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{e}} = 0$.

  Proof.

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} x_i\hat{e}_i = \frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{e}} &= \frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\left[\boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\right]\boldsymbol{Y} \\ &= \frac{1}{n}\left[\boldsymbol{X}^{\mathrm{T}} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\right]\boldsymbol{Y} = \frac{1}{n}\left[\boldsymbol{X}^{\mathrm{T}} - \boldsymbol{X}^{\mathrm{T}}\right]\boldsymbol{Y} = 0\end{aligned}$$

  □

- $\frac{1}{n}\sum_{i=1}^{n} \hat{e}_i = 0$.

  Proof. Because the first element of x_i is 1. □

Definition 7.5. We can generate two important matrices:

- **Projection matrix**: $\boldsymbol{P}_X = \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}$.
- **Residual matrix**: $\boldsymbol{M}_X = \boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}} = \boldsymbol{I}_n - \boldsymbol{P}_X$.

Theorem 7.6. Properties of projection matrix and residual matrix.

1. $\boldsymbol{P}_X$ and $\boldsymbol{M}_X$ are both symmetric.
2. $\boldsymbol{P}_X$ and $\boldsymbol{M}_X$ are both idempotent.

3. $\boldsymbol{P}_X\boldsymbol{M}_X = \boldsymbol{M}_X\boldsymbol{P}_X = \mathbf{0}$.

4. The eigenvalues of $\boldsymbol{P}_X$ and $\boldsymbol{M}_X$ are either 0 or 1.

5. $\operatorname{rank}(\boldsymbol{P}_X) = \operatorname{tr}(\boldsymbol{P}_X) = k$,

6. $\operatorname{rank}(\boldsymbol{M}_X) = \operatorname{tr}(\boldsymbol{M}_X) = n - k$.

7. $\hat{\boldsymbol{Y}} = \boldsymbol{P}_X\boldsymbol{Y}, \hat{\boldsymbol{e}} = \boldsymbol{M}_X\boldsymbol{Y}$.

8. $\boldsymbol{M}_X\boldsymbol{X} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{M}_X = 0, \boldsymbol{P}_X\boldsymbol{X} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}_X = 0$.

Remark 7.7. Geometric interpretation.

Consider $k = 2$, denote

$$X_1 = \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix}, X_2 = \begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{pmatrix}$$

Then $\boldsymbol{X} = (X_1, X_2)$, thus,

$$\hat{\beta}_n = \arg\min_{\beta} \|Y - X\beta\|^2 = \arg\min_{\beta} \|Y - X_1\beta_1 - X_2\beta_2\|$$

Therefore, OLS projects $\boldsymbol{Y}$ onto the space spanned by X_1 and X_2, so $\boldsymbol{P}_X$ is called the projection Matrix and $\boldsymbol{M}_X$ is called the residual Matrix.

Definition 7.8. Define

$$\boldsymbol{P}_0 = \mathbf{1}_n \left(\mathbf{1}_n^{\mathrm{T}}\mathbf{1}_n\right)^{-1} \mathbf{1}_n = \frac{1}{n}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}$$

then

$$\boldsymbol{M}_0 = \boldsymbol{I}_n - \boldsymbol{P}_0 = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix} - \frac{1}{n}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} = \begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & 1-\frac{1}{n} \end{pmatrix}$$

Moreover, since

$$\boldsymbol{P}_0\hat{\boldsymbol{e}} = 0 \Longrightarrow (\boldsymbol{I}_n - \boldsymbol{P}_0)\hat{\boldsymbol{e}} = \hat{\boldsymbol{e}} \Longrightarrow \boldsymbol{M}_0\hat{\boldsymbol{e}} = \hat{\boldsymbol{e}}$$

then

$$\boldsymbol{M}_0\boldsymbol{Y} = \boldsymbol{M}_0\boldsymbol{X}\hat{\beta} + \boldsymbol{M}_0\hat{\boldsymbol{e}} = \boldsymbol{M}_0\boldsymbol{X}\hat{\beta} + \hat{\boldsymbol{e}}$$

Therefore,

$$\begin{aligned}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{M}_0\boldsymbol{Y} &= \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{M}_0^{\mathrm{T}}\boldsymbol{M}_0\boldsymbol{Y} \\ &= \left(\boldsymbol{M}_0\boldsymbol{X}\hat{\beta}+\hat{\boldsymbol{e}}\right)^{\mathrm{T}}\left(\boldsymbol{M}_0\boldsymbol{X}\hat{\beta}+\hat{\boldsymbol{e}}\right) \\ &= \left(\boldsymbol{M}_0\boldsymbol{X}\hat{\beta}\right)^{\mathrm{T}}\boldsymbol{M}_0\boldsymbol{X}\hat{\beta}+\hat{\boldsymbol{e}}^{\mathrm{T}}\hat{\boldsymbol{e}}+\left(\boldsymbol{M}_0\boldsymbol{X}\hat{\beta}\right)^{\mathrm{T}}\hat{\boldsymbol{e}}+\hat{\boldsymbol{e}}^{\mathrm{T}}\boldsymbol{M}_0\boldsymbol{X}\hat{\beta} \\ &= \left(\boldsymbol{M}_0\boldsymbol{X}\hat{\beta}\right)^{\mathrm{T}}\boldsymbol{M}_0\boldsymbol{X}\hat{\beta}+\hat{\boldsymbol{e}}^{\mathrm{T}}\hat{\boldsymbol{e}}\end{aligned}$$

Note that

$$\begin{aligned}\left(\boldsymbol{M}_0\boldsymbol{X}\hat{\beta}\right)^{\mathrm{T}}\hat{\boldsymbol{e}}+\hat{\boldsymbol{e}}^{\mathrm{T}}\boldsymbol{M}_0\boldsymbol{X}\hat{\beta} &= \hat{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{M}_0^{\mathrm{T}}\hat{\boldsymbol{e}}+\hat{\boldsymbol{e}}^{\mathrm{T}}\boldsymbol{M}_0\boldsymbol{X}\hat{\beta} \\ &= \hat{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{e}}+\hat{\boldsymbol{e}}^{\mathrm{T}}\boldsymbol{X}\hat{\beta} \\ &= \boldsymbol{0}\end{aligned}$$

Then we can define R^2 as

$$R^2 := 1-\frac{\hat{\boldsymbol{e}}^{\mathrm{T}}\hat{\boldsymbol{e}}}{\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{M}_0\boldsymbol{Y}}$$

and adjusted R^2 as

$$\bar{R}^2 := 1-\frac{\frac{\hat{\boldsymbol{e}}^{\mathrm{T}}\hat{\boldsymbol{e}}}{n-k}}{\frac{\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{M}_0\boldsymbol{Y}}{n-1}} = 1-\frac{n-1}{n-k}\frac{\hat{\boldsymbol{e}}^{\mathrm{T}}\hat{\boldsymbol{e}}}{\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{M}_0\boldsymbol{Y}}$$

Example 7.9. Consider 2 OLS regressions:

1. Regress $\boldsymbol{Y}$ on $\boldsymbol{X}$.
2. Regress $\boldsymbol{Y}$ on $\boldsymbol{X}$ and $\boldsymbol{Z}$.

where $\boldsymbol{X}$ is a $n\times k_1$ matrix of regressors, and $\boldsymbol{Z}$ is a $n\times k_2$ matrix of additional regressors. Show that the R^2 in regression 2 is larger than or equal to the R^2 in regression 1.

Proof 1. Consider

$$\begin{aligned}\boldsymbol{Y} &= \boldsymbol{X}\hat{\beta}_0+\hat{\boldsymbol{e}} \\ \boldsymbol{Y} &= \boldsymbol{X}\hat{\beta}_1+\boldsymbol{Z}\hat{\beta}_2+\hat{\boldsymbol{u}}\end{aligned}$$

And thus

$$\hat{\beta}_0 = \arg\min_{\beta_0}\|\boldsymbol{Y}-\boldsymbol{X}\beta_0\|^2 \tag{7.1}$$

$$\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix} = \arg\min_{\beta_1,\beta_2}\|\boldsymbol{Y}-\boldsymbol{X}\beta_1-\boldsymbol{Z}\beta_2\|^2 \tag{7.2}$$

and

$$\begin{aligned}\hat{\boldsymbol{e}}^{\mathrm{T}}\hat{\boldsymbol{e}} &= \left\|\boldsymbol{Y}-\boldsymbol{X}\hat{\beta}_0\right\|^2 \\ \hat{\boldsymbol{u}}^{\mathrm{T}}\hat{\boldsymbol{u}} &= \left\|\boldsymbol{Y}-\boldsymbol{X}\hat{\beta}_1-\boldsymbol{Z}\hat{\beta}_2\right\|^2\end{aligned}$$

Note that if we take $\beta_1 = \hat{\beta}_0, \beta_2 = 0$ in (7.2), then

$$\left\| \boldsymbol{Y} - \boldsymbol{X}\hat{\beta}_0 - \boldsymbol{Z} \cdot 0 \right\|^2 = \hat{\boldsymbol{e}}^{\mathrm{T}}\hat{\boldsymbol{e}}$$

which implies that the minimum value of function

$$\|\boldsymbol{Y} - \boldsymbol{X}\beta_1 - \boldsymbol{Z}\beta_2\|^2$$

is less than a special value $\hat{\boldsymbol{e}}^{\mathrm{T}}\hat{\boldsymbol{e}}$, i.e., $\hat{\boldsymbol{e}}^{\mathrm{T}}\hat{\boldsymbol{e}} \geqslant \hat{\boldsymbol{u}}^{\mathrm{T}}\hat{\boldsymbol{u}}$. Then $R^2_{YX} \leqslant R^2_{YXZ}$.

□

Proof 2. Consider again,

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{X}\hat{\beta}_0 + \hat{\boldsymbol{e}} \\ \boldsymbol{Y} &= \boldsymbol{X}\hat{\beta}_1 + \boldsymbol{Z}\hat{\beta}_2 + \hat{\boldsymbol{u}} \end{aligned}$$

And thus,

$$\boldsymbol{X}\hat{\beta}_1 + \boldsymbol{Z}\hat{\beta}_2 + \hat{\boldsymbol{u}} = \boldsymbol{X}\hat{\beta}_0 + \hat{\boldsymbol{e}} \tag{7.3}$$

Recall the property that

$$\hat{\boldsymbol{e}}^{\mathrm{T}}\boldsymbol{X}_{n\times k_1} = \boldsymbol{0}_{1\times k_1} \text{ and } \hat{\boldsymbol{u}}^{\mathrm{T}}(\boldsymbol{X}, \boldsymbol{Z}) = (\boldsymbol{0}_{1\times k_1}, \boldsymbol{0}_{1\times k_2})$$

Multiply both sides of (7.3) by $\hat{\boldsymbol{e}}^{\mathrm{T}}$, we have

$$\hat{\boldsymbol{e}}^{\mathrm{T}}\boldsymbol{Z}\hat{\beta}_2 + \hat{\boldsymbol{e}}^{\mathrm{T}}\hat{\boldsymbol{u}} = \hat{\boldsymbol{e}}^{\mathrm{T}}\hat{\boldsymbol{e}}$$

Similarly, multiply both sides of (7.3) by $\hat{\boldsymbol{u}}^{\mathrm{T}}$, we have

$$\hat{\boldsymbol{u}}^{\mathrm{T}}\hat{\boldsymbol{u}} = \hat{\boldsymbol{u}}^{\mathrm{T}}\hat{\boldsymbol{e}}$$

Note that $\hat{\boldsymbol{u}}^{\mathrm{T}}\hat{\boldsymbol{e}} = \hat{\boldsymbol{e}}^{\mathrm{T}}\hat{\boldsymbol{u}}$, then

$$\hat{\boldsymbol{e}}^{\mathrm{T}}\boldsymbol{Z}\hat{\beta}_2 + \hat{\boldsymbol{u}}^{\mathrm{T}}\hat{\boldsymbol{u}} = \hat{\boldsymbol{e}}^{\mathrm{T}}\hat{\boldsymbol{e}} \tag{7.4}$$

Finally,

$$\begin{aligned} \hat{\boldsymbol{u}}^{\mathrm{T}}\hat{\boldsymbol{u}} &= \left(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}_1 - \boldsymbol{Z}\hat{\beta}_2\right)^{\mathrm{T}}\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}_1 - \boldsymbol{Z}\hat{\beta}_2\right) \\ &= \left(\boldsymbol{X}\hat{\beta}_0 + \hat{\boldsymbol{e}} - \boldsymbol{X}\hat{\beta}_1 - \boldsymbol{Z}\hat{\beta}_2\right)^{\mathrm{T}}\left(\boldsymbol{X}\hat{\beta}_0 + \hat{\boldsymbol{e}} - \boldsymbol{X}\hat{\beta}_1 - \boldsymbol{Z}\hat{\beta}_2\right) \\ &= \left(\boldsymbol{X}\left(\hat{\beta}_0 - \hat{\beta}_1\right) - \boldsymbol{Z}\hat{\beta}_2 + \hat{\boldsymbol{e}}\right)^{\mathrm{T}}\left(\boldsymbol{X}\left(\hat{\beta}_0 - \hat{\beta}_1\right) - \boldsymbol{Z}\hat{\beta}_2 + \hat{\boldsymbol{e}}\right) \\ &= \left(\boldsymbol{X}\left(\hat{\beta}_0 - \hat{\beta}_1\right) - \boldsymbol{Z}\hat{\beta}_2\right)^{\mathrm{T}}\left(\boldsymbol{X}\left(\hat{\beta}_0 - \hat{\beta}_1\right) - \boldsymbol{Z}\hat{\beta}_2\right) + 2\hat{\boldsymbol{e}}^{\mathrm{T}}\left(\boldsymbol{X}\left(\hat{\beta}_0 - \hat{\beta}_1\right) - \boldsymbol{Z}\hat{\beta}_2\right) + \hat{\boldsymbol{e}}^{\mathrm{T}}\hat{\boldsymbol{e}} \\ &:= a - 2\hat{\boldsymbol{e}}^{\mathrm{T}}\boldsymbol{Z}\hat{\beta}_2 + \hat{\boldsymbol{e}}^{\mathrm{T}}\hat{\boldsymbol{e}} \\ &= a - 2\hat{\boldsymbol{e}}^{\mathrm{T}}\hat{\boldsymbol{e}} + 2\hat{\boldsymbol{u}}^{\mathrm{T}}\hat{\boldsymbol{u}} + \hat{\boldsymbol{e}}^{\mathrm{T}}\hat{\boldsymbol{e}} \\ &= a - \hat{\boldsymbol{e}}^{\mathrm{T}}\hat{\boldsymbol{e}} + 2\hat{\boldsymbol{u}}^{\mathrm{T}}\hat{\boldsymbol{u}} \end{aligned}$$

where $a = \left(\boldsymbol{X}\left(\hat{\beta}_0 - \hat{\beta}_1\right) - \boldsymbol{Z}\hat{\beta}_2\right)^{\mathrm{T}}\left(\boldsymbol{X}\left(\hat{\beta}_0 - \hat{\beta}_1\right) - \boldsymbol{Z}\hat{\beta}_2\right) \geqslant 0$, then

$$\hat{\boldsymbol{e}}^{\mathrm{T}}\hat{\boldsymbol{e}} = a + \hat{\boldsymbol{u}}^{\mathrm{T}}\hat{\boldsymbol{u}} \geqslant \hat{\boldsymbol{u}}^{\mathrm{T}}\hat{\boldsymbol{u}}$$

□

Example 7.10. Show that in an OLS regression:

1. $\bar{R}^2 < R^2$ when $k > 1$.
2. $\bar{R}^2$ does not necessarily increase as k increases.
3. $\bar{R}^2$ can be negative.

Proof. Based on the definitions.

1. $$R^2 - \bar{R}^2 = \left(\frac{n-1}{n-k} - 1\right) \frac{\hat{\boldsymbol{e}}^{\mathrm{T}}\hat{\boldsymbol{e}}}{\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{M}_0\boldsymbol{Y}} > 0 \Longleftrightarrow \frac{n-1}{n-k} > 1 \Longleftrightarrow k > 1$$
2. If $k \uparrow \Longrightarrow \hat{\boldsymbol{e}}^{\mathrm{T}}\hat{\boldsymbol{e}} \downarrow, (n-k) \uparrow$.

   For example, if we add a variable z_i into the model, where z_i is independent of $x_{1i}, x_{2i}, \cdots, x_{in}, y_i$, then $\hat{\boldsymbol{e}}^{\mathrm{T}}\hat{\boldsymbol{e}}$ remains the same, but $n-k$ decreases.
3. If k is close to n, then $\frac{n-1}{n-k}$ can be large, possibly leading to a negative $\bar{R}^2$.

   For example, consider the model $y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \varepsilon_i$ where x_{i1} and x_{i2} are independent with y_i, then $\hat{\boldsymbol{e}}^{\mathrm{T}}\hat{\boldsymbol{e}} = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{M}_0\boldsymbol{Y}$, and thus, $\bar{R} = 1 - \frac{n-1}{n-2} < 0$.

□

Theorem 7.11. (The Firsch-Waugh-Lovell Theorem) Consider the model:

$$\boldsymbol{Y}_{n\times 1} = \boldsymbol{X}_{n\times k}\beta_0 + \boldsymbol{Z}_{n\times p}\gamma_0 + \varepsilon_{n\times 1}$$

Let $\hat{\beta}_n$ and $\hat{\gamma}_n$ be the OLS estimator of β_0 and γ_0 when regressing $\boldsymbol{Y}$ on $\boldsymbol{X}$ and $\boldsymbol{Z}$.

Then we have

$$\begin{aligned}
&\boldsymbol{Y} = \boldsymbol{X}\beta_0 + \boldsymbol{Z}\gamma_0 + \hat{\boldsymbol{e}} \\
&\Longrightarrow \boldsymbol{M}_Z\boldsymbol{Y} = \boldsymbol{M}_Z\boldsymbol{X}\beta_0 + \hat{\boldsymbol{e}} \ \ (\text{Since } \boldsymbol{M}_Z\boldsymbol{Z} = \boldsymbol{0}, \boldsymbol{M}_Z\hat{\boldsymbol{e}} = \hat{\boldsymbol{e}}) \\
&\Longrightarrow \boldsymbol{X}^{\mathrm{T}}\boldsymbol{M}_Z\boldsymbol{Y} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{M}_Z\boldsymbol{X}\beta_0 \ \ \left(\text{Since } \boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{e}} = \boldsymbol{0}\right) \\
&\Longrightarrow \beta_0 = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{M}_Z\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{M}_Z\boldsymbol{Y}
\end{aligned}$$

Note that $\boldsymbol{M}_X$ is symmetric and idempotent, then

$$\beta_0 = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{M}_Z\boldsymbol{M}_Z^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{M}_Z\boldsymbol{M}_Z^{\mathrm{T}}\boldsymbol{Y} := \left(\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{Y}}$$

where

- $\tilde{\boldsymbol{X}} = \boldsymbol{M}_Z^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{M}_Z\boldsymbol{X}$ is the residuals of regressing $\boldsymbol{X}$ on $\boldsymbol{Z}$;
- $\tilde{\boldsymbol{Y}} = \boldsymbol{M}_Z^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{M}_Z\boldsymbol{Y}$ is the residuals of regressing $\boldsymbol{Y}$ on $\boldsymbol{Z}$.

Remark 7.12. Consistency.

- Any estimator is a function of the sample size, then it depends on the sample size n. For example, in OLS,

$$\hat{\beta}_n = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} = \left(\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}\right)^{-1}\sum_{i=1}^{n} x_i y_i$$

- Three confusing notions. (Let β_0 be the true value of our estimator.)

  1. Consistency: $\lim_{n\to\infty}\hat{\beta}_n = \beta_0$.
  2. Unbiasedness: $\mathbb{E}\left[\hat{\beta}_n\right] = \beta_0$.
  3. Asymptotic unbiasedness: $\lim_{n\to\infty}\mathbb{E}\left[\hat{\beta}_n\right] = \beta_0$.

Remark 7.13. Asymptotic theory for OLS.

Under assumptions mentioned in Remark 7.2, we have

1. Consistency: By Central Limit Theorem, $y_i = x_i^{\mathrm{T}}\beta_0 + u_i$, and $\mathbb{E}\left[x_i u_i\right] = 0$, we have

$$\begin{aligned}
\hat{\beta}_n &= \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}\right)^{-1}\frac{1}{n}\sum_{i=1}^{n} x_i y_i \xrightarrow{d} \left(\mathbb{E}\left[x_i x_i^{\mathrm{T}}\right]\right)^{-1}\mathbb{E}\left[x_i y_i\right] \\
&= \left(\mathbb{E}\left[x_i x_i^{\mathrm{T}}\right]\right)^{-1}\mathbb{E}\left[x_i\left(x_i^{\mathrm{T}}\beta_0 + u_i\right)\right] \\
&= \left(\mathbb{E}\left[x_i x_i^{\mathrm{T}}\right]\right)^{-1}\mathbb{E}\left[x_i\mathbb{E}\left[x_i^{\mathrm{T}}\beta_0 + u_i \middle| x_i\right]\right] \\
&= \left(\mathbb{E}\left[x_i x_i^{\mathrm{T}}\right]\right)^{-1}\mathbb{E}\left[x_i x_i^{\mathrm{T}}\right]\cdot\beta_0 \\
&= \beta_0
\end{aligned}$$

2. Unbiasedness: First consider

$$\begin{aligned}
\mathbb{E}\left[\hat{\beta}\middle|\boldsymbol{X}\right] &= \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}\right)^{-1}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\left(x_i^{\mathrm{T}}\beta_0 + u_i\right)\middle|\boldsymbol{X}\right] \\
&= \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}\right)^{-1}\left[\beta_0\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}\right) + \frac{1}{n}\sum_{i=1}^{n} x_i\mathbb{E}\left[u_i\middle|\boldsymbol{X}\right]\right] = \beta_0
\end{aligned}$$

   Then by the Law of iterated expectation, we get

$$\mathbb{E}\left[\hat{\beta}_n\right] = \mathbb{E}\left[\mathbb{E}\left[\hat{\beta}_n\middle|\boldsymbol{X}\right]\right] = \beta_0$$

3. Asymptotic normality:

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \to \mathcal{N}\left(0, \sigma_u\mathbb{E}\left[x_i x_i^{\mathrm{T}}\right]^{-1}\right)$$

   Proof. Firstly,

$$\hat{\beta}_n = \left(\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}\right)^{-1}\sum_{i=1}^{n} x_i\left(x_i^{\mathrm{T}}\beta_0 + u_i\right) = \beta_0 + \left(\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}\right)^{-1}\sum_{i=1}^{n} x_i u_i$$

Then

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}_n-\beta_0\right) &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^{\mathrm{T}}\right)^{-1}\frac{1}{n}\sum_{i=1}^n x_i u_i \\ &= \left(\frac{1}{n}\sum_{i=1}^n x_i x_i^{\mathrm{T}}\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n x_i u_i\end{aligned}$$

Define $\Sigma_{xx} := \mathbb{E}\left[x_i x_i^{\mathrm{T}}\right]$ is a symmetric matrix, $\sigma_u^2 := \mathbb{E}\left[u_i^2 \middle| x_i\right]$, and since

$$\begin{aligned}\mathrm{Var}\left[x_i u_i\right] &= \mathbb{E}\left[x_i u_i \left(x_i u_i\right)^{\mathrm{T}}\right] - \mathbb{E}\left[x_i u_i\right]\mathbb{E}\left[x_i u_i\right]^{\mathrm{T}} \\ &= \mathbb{E}\left[x_i x_i^{\mathrm{T}} u_i^2\right] - 0 = \mathbb{E}\left[x_i x_i^{\mathrm{T}}\mathbb{E}\left[u_i^2 \middle| x_i\right]\right] = \Sigma_{xx}\sigma_u^2\end{aligned}$$

Then by CLT,

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^n x_i u_i \xrightarrow{d} \mathcal{N}\left(\mathbb{E}\left[x_i u_i\right], \mathrm{Var}\left[x_i u_i\right]\right) = \mathcal{N}\left(0, \Sigma_{xx}\sigma_u^2\right)$$

By LLN,

$$\frac{1}{n}\sum_{i=1}^n x_i x_i^{\mathrm{T}} \xrightarrow{\mathbb{P}} \Sigma_{xx}$$

Therefore,

$$\sqrt{n}\left(\hat{\beta}_n-\beta_0\right) \to \mathcal{N}\left(0, \Sigma_{xx}^{-1}\Sigma_{xx}\sigma_u^2\left(\Sigma_{xx}^{-1}\right)^{\mathrm{T}}\right) = \mathcal{N}\left(0, \sigma_u^2\Sigma_{xx}^{-1}\right)$$

□

Remark 7.14. Best predictor and best linear predictor.

1. Review: If $\mathbb{E}\left[Y^2\right] < \infty$, then the conditional mean, $m(x) := \mathbb{E}(Y|X)$ is the best predictor in the sense that it is the solution of $\min_g \mathbb{E}\left[Y - g(X)\right]^2$, where g is all the measurable function of X.
2. The best predictor solves $\min_g \mathbb{E}\left[Y - g(X)\right]^2$.
3. The best linear predictor solves $\min_\beta \mathbb{E}\left[Y - X^{\mathrm{T}}\beta\right]^2$.
   - If we use the sample, the problem becomes $\min_\beta \frac{1}{n}\sum_{i=1}^n \left(Y_i - X_i^{\mathrm{T}}\beta\right)^2$.

7.2 Framework and Overview

Remark 7.15. Parametric or nonparametric model.

Consider the model

$$y = \beta^{\mathrm{T}}x + \varepsilon, \text{ where } \mathbb{E}\left[\varepsilon \middle| x\right] = 0$$

It is a parametric model.

$$y = g(x) + \varepsilon \text{ where } \varepsilon \perp x \text{ and } \mathbb{E}(\varepsilon) = 0$$

If g is a measurable function, it can be a nonparametric model. If we specify g, say a quadratic function, then it becomes a parametric model.

Remark 7.16. The goal of the course.

1. Estimation.
   - Estimation and inference of β in parametric and semi-parametric models.
   - Estimation of g in non-parametric models.
2. Some notations.
   - $\{z_i\}_{i=1}^n$ are observed. Sometimes, $z_i = (y_i, x_i)$ where y_i is a scalar called the dependent variable, and x_i is a vector called the regressors (or independent variables).
   - Let β_0 be the TRUE value that generates observed sample.
   - An estimation is a measurable function of the observed sample: $\hat{\beta}_n := \hat{\beta}_n(z_1, z_2, \cdots, z_n)$.
3. We hope to
   - Show that $\hat{\beta}_n$ is consistent: $\mathbb{P}\left(\left\|\hat{\beta}_n - \beta_0\right\| > \varepsilon\right) \to 0$ as $n \to \infty$ for any $\varepsilon > 0$.
   - Establish the asymptotic distribution of $\hat{\beta}_n$, in order to make inference about β_0.

Remark 7.17. General framework.

- Let $\beta_0 \in \mathbb{R}^k, B \subseteq \mathbb{R}^k$ be a parameter space such that $\beta_0 \in B$.
- Let $M_n(z_1, \cdots, z_n; \beta)$ be a real-valued function that depends on the sample size n. For simplicity, we ignore $z_1, \cdots, z_n$ in the notation.

  Note that $M_n(\cdot) : \mathbb{R}^k \to \mathbb{R}$ is a **random** function.

The extremum estimator of β_0 is defined as:

$$\hat{\beta} = \arg\max_{\beta \in B} M_n(\beta)$$

The key to define an estimator : the choice of $M_n(\cdot)$.

- $M_n(\cdot)$ usually takes the form $M_n(\beta) = \frac{1}{n}\sum_{i=1}^n m(z_i, \beta)$ so the choice of $m(\cdot, \cdot)$ is the key.
- Different $M_n(\cdot)$ defines a different estimator.

Example 7.18. Consider

$$y_i = \beta_0^{\mathrm{T}} x_i + \varepsilon_i, \text{ where } x_i \perp \varepsilon_i \text{ and } \varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$$

1. $M_n(\beta) = -\frac{1}{n}\sum_{i=1}^n \left(y_i - \beta^{\mathrm{T}} x_i\right)^2$ generates the ordinary least squares (OLS) estimator, associated with the mean.
2. $M_n(\beta) = -\frac{1}{n}\sum_{i=1}^n \left|y_i - \beta^{\mathrm{T}} x_i\right|$ generates the least absolute deviation (LAD) estimator, associated with the median.

3. Note that Both estimators are consistent but they have different asymptotic distributions.

Example 7.19. Consider[1]

$$y_i = \mathbb{I}_{\left\{\beta_0^{\mathrm{T}} x_i - \varepsilon_i > 0\right\}}, \text{ where } x_i \perp \varepsilon_i \text{ and } \varepsilon_i \sim \text{Logistic}\,(0,1)$$

1. $M_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left\{y_i \ln\left(\frac{e^{\beta^{\mathrm{T}} x_i}}{1+e^{\beta^{\mathrm{T}} x_i}}\right) + (1-y_i)\ln\left(\frac{1}{1+e^{\beta^{\mathrm{T}} x_i}}\right)\right\}$ generates the maximum likelihood estimator.

2. $M_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left\{y_i \mathbb{I}_{\{\beta^{\mathrm{T}} x_i > 0\}} + (1-y_i)\,\mathbb{I}_{\{\beta^{\mathrm{T}} x_i \leqslant 0\}}\right\}$ generates the maximum score estimator.

3. Again, both estimators are consistent but they have very different asymptotic distributions.

Example 7.20. Examples of Extremum Estimators.

$$y_i = \beta_0^{\mathrm{T}} x_i + \varepsilon_i, \mathbb{E}\left[\varepsilon_i | \, x_i\right] = 0$$

1. OLS estimator:
$$\hat{\beta}_{\text{OLS}} = \arg\min_{\beta \in B} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \beta^{\mathrm{T}} x_i\right)^2$$

2. Let $f(z;\beta_0)$ be the joint density of z_i at z, then the MLE is
$$\hat{\beta}_{\text{OLS}} = \arg\max_{\beta \in B} \frac{1}{n}\sum_{i=1}^{n} \ln f(z_i;\beta)$$

Example 7.21. Generalized Method of Moments.

Let $g(z_i,\beta)$ be an $\mathbb{R}^p$-valued function satisfying

$$\mathbb{E}\left[g(z_i,\beta_0)\right] = 0$$

Let $\beta \in \mathbb{R}^k$, where $p \geqslant k$, then the GMM estimator is

$$\hat{\beta}_{\text{GMM}} = \arg\min_{\beta \in B}\left(\frac{1}{n}\sum_{i=1}^{n} g(z_i,\beta)\right)^{\mathrm{T}} \hat{W}_{p\times p}\left(\frac{1}{n}\sum_{i=1}^{n} g(z_i,\beta)\right)$$

where $\hat{W}$ is a $p \times p$ positive definition symmetric matrix.

7.3 Consistency

Theorem 7.22. (A General Consistency Result)

Suppose β_0 is the real parameter, and $\hat{\beta}_n = \arg\max_{\beta \in B} M_n(\beta)$ is its estimator where n is the sample size. If there is a **non-random** function $M_0(\beta) : B \to \mathbb{R}$ such that

[1]The CDF for Logistic $(0,1)$ is $F(x) = \frac{1}{1+e^{-x}}$.

1. $M_0(\beta)$ is uniquely maximized at β_0;
2. B is compact;
3. $M_0(\beta)$ is continuous;
4. $M_n(\beta)$ converges uniformly in probability to $M_0(\beta)$, that is

$$\sup_{\beta \in B} |M_n(\beta) - M_0(\beta)| \xrightarrow{\mathbb{P}} 0$$

then

$$\left\| \hat{\beta}_n - \beta_0 \right\| \xrightarrow{\mathbb{P}} 0$$

Proof. By the definition of $\hat{\beta}_n$, we have

$$M_n\left(\hat{\beta}_n\right) \geqslant M_n(\beta_0), \forall n$$

Then

$$\begin{aligned} M_0\left(\hat{\beta}_n\right) - M_0(\beta_0) &\geqslant M_0\left(\hat{\beta}_n\right) - M_0(\beta_0) + M_n(\beta_0) - M_n\left(\hat{\beta}_n\right) \\ &= [M_n(\beta_0) - M_0(\beta_0)] - \left[M_n\left(\hat{\beta}_n\right) - M_0\left(\hat{\beta}_n\right)\right] \end{aligned}$$

By condition that $M_n(\beta)$ converges uniformly in probability to $M_0(\beta)$, we have $\forall \varepsilon > 0$,

$$|M_n(\beta_0) - M_0(\beta_0)| < \frac{\varepsilon}{2} \text{ and } \left|M_n\left(\hat{\beta}_n\right) - M_0\left(\hat{\beta}_n\right)\right| < \frac{\varepsilon}{2}$$

hold with probability approaching 1 (w.p.a.1). Then

$$0 \geqslant M_0\left(\hat{\beta}_n\right) - M_0(\beta_0) > -\varepsilon \text{ w.p.a.1} \tag{7.5}$$

Note that the LHS is from definition of $\beta_0 = \arg\max_{\beta \in B} M_0(\beta)$.

Let C be any open subset of B such that $\beta_0 \in C$. Then $B \cap C^c$ is a closed subset of a compact set B, and thus, $B \cap C^c$ is compact. Moreover, the continuity of M_0 ensures that[2]

$$\sup_{\beta \in B \cap C^c} M_0(\beta) < M_0(\beta_0)$$

[2]If the continuity of M_0 fails, then it could be that

$$\sup_{\beta \in B \cap C^c} M_0(\beta) = M_0(\beta_0)$$

instead of a strict inequality, then we cannot find a strictly positive ε afterwards.

An example of a bounded function defined on a compact set without a maximum nor minimum.

Let the domain be a compact set $[0,1]$, define f as

$$f = \begin{cases} (-1)^n \frac{n}{n+1}, & \text{if } x = \frac{m}{n},\ \gcd(m,n) = 1, n > 0 \\ 0, & \text{if } x \in [0,1] \setminus \mathbb{Q} \end{cases}$$

Pick

$$\varepsilon = M_0(\beta_0) - \sup_{\beta \in B \cap C^c} M_0(\beta) > 0$$

together with (7.5), we get

$$M_0\left(\hat{\beta}_n\right) > \sup_{\beta \in B \cap C^c} M_0(\beta) \text{ w.p.a.1}$$

Then $\hat{\beta}_n \in C$ w.p.a.1, for any open subset C of B such that $\beta_0 \in C$, then

$$\left\|\hat{\beta}_n - \beta_0\right\| \xrightarrow{\mathbb{P}} 0$$

□

Remark 7.23. Remark on Theorem 7.22.

1. $M_n\left(\hat{\beta}_n\right) \leqslant \sup_{\beta \in B} M_n(\beta) + o_p(1)$ is allowed.[3]

   This modification is useful to prove the consistency of the maximum score estimator and the simulated moment estimator.

2. The compact set B can be relaxed to a convex set (see Theorem 7.25).

3. Continuity of $M_0(\cdot)$ is a very weak condition, it can be true even when $M_n(\cdot)$ is not continuous.

4. How to verify uniform convergence? See Theorem 7.24.

Theorem 7.24. If

1. data are i.i.d.;

2. B is compact;

3. $\forall \beta \in B$, $a(z_i, \beta)$ is almost surely (a.s.) continuous;

4. $\exists d(z)$ with $\forall i, \mathbb{E}|d(z_i)| < \infty$ such that $\forall \beta \in B, |a(z, \beta)| \leqslant d(z)$,

then

1. $\mathbb{E}[a(z_i, \beta)]$ is continuous;

2. and
$$\sup_{\beta \in B} \left| \frac{1}{n} \sum_{i=1}^{n} a(z_i, \beta) - \mathbb{E}[a(z_i, \beta)] \right| \xrightarrow{\mathbb{P}} 0$$

[3] o_p refers to convergence in probability towards zero:

$$X_n = o_p(1)$$

is equivalent with

$$\lim_{n \to \infty} \mathbb{P}(|X_n| \geqslant \varepsilon) = 0, \forall \varepsilon > 0$$

Generally, we denote $X_n = o_p(a_n) \Longleftrightarrow \frac{X_n}{a_n} \xrightarrow{\mathbb{P}} 0$.

Theorem 7.25. If there is a function $M_0(\beta)$ such that

1. $M_0(\beta)$ is uniquely maximized at β_0;
2. β_0 is an interior point of a convex set B, and $M_n(\beta)$ is concave;
3. $\forall \beta \in B$, $M_n(\beta) \xrightarrow{\mathbb{P}} M_0(\beta)$,

then

1. $\hat{\beta}_n$ exists w.p.a.1.[4]
2. $\hat{\beta}_n \xrightarrow{\mathbb{P}} \beta$.

7.4 Asymptotic Distribution

With consistency, now we want to derive the asymptotic distribution of $\hat{\beta}_n$.

Theorem 7.26. Suppose that $\hat{\beta}_n$ satisfies:

1. $\hat{\beta}_n \xrightarrow{\mathbb{P}} \beta_0$;
2. $\beta_0 \in \operatorname{int}(B)$;
3. $M_n(\beta)$ is twice continuously differentiable in a neighborhood D of β_0;
4. $\sqrt{n}\nabla_\beta M_n(\beta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma)$;
5. $\exists H(\beta) \in \mathbb{R}^{k\times k}$ that is continuous at β_0 and
$$\sup_{\beta\in D} \left\| \nabla_{\beta\beta^{\mathrm{T}}} M_n(\beta) - H(\beta) \right\| \xrightarrow{\mathbb{P}} 0$$
6. $H := H(\beta_0)$ is nonsingular,

then
$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, H^{-1}\Sigma H^{-1}\right)$$

Intuitive Proof. Since $\hat{\beta}_n = \arg\max_{\beta\in B} M_n(\beta)$, by the F.O.C., we have
$$\nabla_\beta M_n\left(\hat{\beta}_n\right) = 0$$

Expanding $\nabla_\beta M_n\left(\hat{\beta}_n\right)$ around β_0 as (or by the mean-value theorem)
$$0 = \nabla_\beta M_n\left(\hat{\beta}_n\right) = \nabla_\beta M_n(\beta_0) + H\left(\bar{\beta}\right)\cdot\left(\hat{\beta}_n - \beta_0\right)$$

[4] As $n \to \infty$, for function $M_n(\beta)$ there exists a maximizer $\hat{\beta}_n$. For example, the Hessian matrix (or second derivative) can show that the solution of F.O.C.s is a maximizer.

where $H\left(\bar{\beta}\right) = \nabla^2_{\beta\beta^{\mathrm{T}}} M_n\left(\bar{\beta}\right)$, and $\bar{\beta}$ is between $\hat{\beta}_n$ and β_0.[5] By

$$\hat{\beta}_n \xrightarrow{\mathbb{P}} \beta_0$$

we have

$$H\left(\bar{\beta}\right) \xrightarrow{\mathbb{P}} H\left(\beta_0\right) = H$$

which is invertible. Moreover, by

$$\sqrt{n}\nabla_\beta M_n\left(\beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma\right)$$

then

$$-H^{-1} \cdot \sqrt{n}\nabla_\beta M_n\left(\beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, H^{-1}\Sigma\left(H^{-1}\right)^{\mathrm{T}}\right)$$

where H is symmetric, then $\left(H^{-1}\right)^{\mathrm{T}} = H^{-1}$.

Then by Slutzky Theorem, [6]

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, H^{-1}\Sigma H^{-1}\right)$$

□

Remark 7.27. Remarks on Theorem 7.26.

1. Condition 1 is prerequisite.
2. Condition 2 is essential.
3. Condition 3 can be relaxed (Discussed in Quantile Regression).
4. Condition 4 usually follows from first order condition $\mathbb{E}\left[\nabla_\beta M_n\left(\beta_0\right)\right] = 0$ and CLT.
5. Condition 5 can be verified using Theorem 7.24.
6. Condition 6 is usually related to rank conditions.

Example 7.28. Take OLS and an example. Consider

$$M_n\left(\beta\right) = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i^{\mathrm{T}}\beta\right)^2$$

Then

$$\nabla_\beta M_n\left(\beta_0\right) = \frac{1}{n}\sum_{i=1}^{n} 2\left(y_i - x_i^{\mathrm{T}}\beta_0\right)x_i = \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i x_i$$

and thus,

$$\nabla_{\beta\beta^{\mathrm{T}}} M_n\left(\beta_0\right) = -\frac{2}{n}\sum_{i=1}^{n} x_i x_i^{\mathrm{T}} \xrightarrow{\mathbb{P}} -2\mathbb{E}\left[x_i x_i^{\mathrm{T}}\right]$$

[5] If $\beta_0 > \hat{\beta}_n$, then $\bar{\beta} \in \left(\hat{\beta}_n, \beta_0\right)$; otherwise, $\bar{\beta} \in \left(\beta_0, \hat{\beta}_n\right)$.

[6] Here we use that $X_n \xrightarrow{d} X, Y_n \xrightarrow{d} c \Longrightarrow X_n + Y_n \xrightarrow{d} X + c, X_n Y_n \xrightarrow{d} cX$.

If we assume $H = -2\mathbb{E}\left[x_i x_i^{\mathrm{T}}\right]$ is invertible and impose that

$$\sqrt{n}\nabla_\beta M_n(\beta_0) = \frac{2}{n}\sqrt{n}\sum_{i=1}^{n}\varepsilon_i x_i \xrightarrow{d} \mathcal{N}(0,\Sigma)$$

then

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, H^{-1}\Sigma H^{-1}\right)$$

Example 7.29. Consider the GMM estimator

$$\hat{\beta}_n := \hat{\beta}_{\mathrm{GMM}} = \arg\min_{\beta\in B}\left(\frac{1}{n}\sum_{i=1}^{n} g(z_i,\beta)\right)^{\mathrm{T}} \hat{W}_{p\times p}\left(\frac{1}{n}\sum_{i=1}^{n} g(z_i,\beta)\right)$$

Suppose that $\hat{W} \xrightarrow{\mathbb{P}} W$, where W is positive definite, and $\hat{\beta}_n \xrightarrow{\mathbb{P}} \beta_0$, and that $\mathbb{E}\left[g(z_i,\beta_0)\right] = 0$.

Find the asymptotic distribution of $\hat{\beta}_n$.

Solution 1. Define

$$M_n(\beta) := \left(\frac{1}{n}\sum_{i=1}^{n} g(z_i,\beta)\right)^{\mathrm{T}} \hat{W}_{p\times p}\left(\frac{1}{n}\sum_{i=1}^{n} g(z_i,\beta)\right)$$

and

$$G(z_i,\beta)_{p\times k} = \frac{\partial g}{\partial \beta}(z_i,\beta)$$

Then the F.O.C. is (suppose $\hat{W}$ is symmetric)

$$\begin{aligned}
&\nabla_\beta M_n\left(\hat{\beta}_n\right)\\
=&\left(\frac{1}{n}\sum_{i=1}^{n} G\left(z_i,\hat{\beta}_n\right)\right)^{\mathrm{T}} \hat{W}\left(\frac{1}{n}\sum_{i=1}^{n} g\left(z_i,\hat{\beta}_n\right)\right) + \left(\frac{1}{n}\sum_{i=1}^{n} g\left(z_i,\hat{\beta}_n\right)\right)^{\mathrm{T}} \hat{W}_{p\times p}\left(\frac{1}{n}\sum_{i=1}^{n} G\left(z_i,\hat{\beta}_n\right)\right)\\
=&\left(\frac{1}{n}\sum_{i=1}^{n} G\left(z_i,\hat{\beta}_n\right)\right)^{\mathrm{T}} \left(\hat{W}+\hat{W}^{\mathrm{T}}\right)\left(\frac{1}{n}\sum_{i=1}^{n} g\left(z_i,\hat{\beta}_n\right)\right)\\
=&2\left(\frac{1}{n}\sum_{i=1}^{n} G\left(z_i,\hat{\beta}_n\right)\right)^{\mathrm{T}} \hat{W}\left(\frac{1}{n}\sum_{i=1}^{n} g\left(z_i,\hat{\beta}_n\right)\right)\\
=&0
\end{aligned}$$

Moreover,

$$\begin{aligned}
\nabla^2_{\beta\beta^{\mathrm{T}}} M_n(\beta) =& 2\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\beta} G(z_i,\beta)\right)^{\mathrm{T}} \hat{W}\left(\frac{1}{n}\sum_{i=1}^{n} g(z_i,\beta)\right)\\
&+ 2\left(\frac{1}{n}\sum_{i=1}^{n} G(z_i,\beta)\right)^{\mathrm{T}} \hat{W}\left(\frac{1}{n}\sum_{i=1}^{n} G(z_i,\beta)\right)
\end{aligned}$$

Note that $\nabla^2_{\beta\beta^{\mathrm{T}}} M_n\left(\hat{\beta}_n\right)$ should be positive definite to make $\hat{\beta}_n$ a maximizer.

By the mean-value theorem,

$$0 = \nabla_\beta M_n\left(\hat{\beta}_n\right) = \nabla_\beta M_n\left(\beta_0\right) + \nabla^2_{\beta\beta^{\mathrm{T}}} M_n\left(\bar{\beta}\right)\left(\hat{\beta}_n - \beta_0\right)$$

Therefore,

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) = -\left[\nabla^2_{\beta\beta^{\mathrm{T}}} M_n\left(\bar{\beta}\right)\right]^{-1}\sqrt{n}\nabla_\beta M_n\left(\beta_0\right)$$

By $\hat{\beta}_n \xrightarrow{\mathbb{P}} \beta_0$, we have $\bar{\beta} \xrightarrow{\mathbb{P}} \beta_0$ then

$$\frac{1}{n}\sum_{i=1}^{n} g\left(z_i, \bar{\beta}\right) \xrightarrow{\mathbb{P}} \mathbb{E}g\left(z_i, \beta_0\right) = 0$$

then

$$\nabla^2_{\beta\beta^{\mathrm{T}}} M_n\left(\bar{\beta}\right) \xrightarrow{\mathbb{P}} 2G_0^{\mathrm{T}}\hat{W}G_0$$

where $G_0 := \mathbb{E}\left[G\left(z_i, \hat{\beta}_n\right)\right]$.

Moreover, CLT implies

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} g\left(z_i, \hat{\beta}_n\right) \xrightarrow{d} \mathcal{N}\left(0, \mathrm{Var}\left[g\left(z_i, \beta_0\right)\right]\right) := \mathcal{N}\left(0, S_0\right)$$

Note that $\hat{W}_n \xrightarrow{\mathbb{P}} W_0$ then

$$\begin{aligned}
\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) &= \left[\nabla^2_{\beta\beta^{\mathrm{T}}} M_n\left(\bar{\beta}\right)\right]^{-1}\sqrt{n}\nabla_\beta M_n\left(\beta_0\right) \\
&\xrightarrow{\mathbb{P}} \left(2G_0^{\mathrm{T}}WG_0\right)^{-1} 2G_0^{\mathrm{T}}W\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} g\left(z_i, \hat{\beta}_n\right) \\
&= \left(G_0^{\mathrm{T}}WG_0\right)^{-1} G_0^{\mathrm{T}}W\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} g\left(z_i, \hat{\beta}_n\right) \\
&\xrightarrow{d} \mathcal{N}\left(0, \left(G_0^{\mathrm{T}}WG_0\right)^{-1} G_0^{\mathrm{T}}WS_0W^{\mathrm{T}}G_0\left(G_0^{\mathrm{T}}WG_0\right)^{-1}\right)
\end{aligned}$$

□

Solution 2. Define $g_n\left(\beta\right) := \frac{1}{n}\sum_{i=1}^{n} g\left(z_i, \beta\right)$, then

$$\hat{\beta}_n = \arg\min_{\beta\in B} g_n\left(\beta\right)^{\mathrm{T}}\hat{W}_{p\times p} g_n\left(\beta\right) \tag{7.6}$$

By the mean-value theorem, we have

$$g_n\left(\hat{\beta}_n\right) - g_n\left(\beta_0\right) = \left.\frac{\partial g_n}{\partial\beta}\right|_{\bar{\beta}}\left(\hat{\beta}_n - \beta_0\right)$$

where $\bar{\beta}$ is between $\hat{\beta}_n$ and β_0.

Then

$$\left(\left.\frac{\partial g_n}{\partial\beta}\right|_{\hat{\beta}_n}\right)^{\mathrm{T}}\hat{W}\left[g_n\left(\hat{\beta}_n\right) - g_n\left(\beta_0\right)\right] = \left(\left.\frac{\partial g_n}{\partial\beta}\right|_{\hat{\beta}_n}\right)^{\mathrm{T}}\hat{W}\left.\frac{\partial g_n}{\partial\beta}\right|_{\bar{\beta}}\left(\hat{\beta}_n - \beta_0\right) \tag{7.7}$$

Note that the F.O.C. of (7.6) is (Suppose $\hat{W}$ is symmetric) [7]

$$\left(\left.\frac{\partial g_n}{\partial \beta}\right|_{\hat{\beta}_n}\right)^{\mathrm{T}} \hat{W} g_n\left(\hat{\beta}_n\right) + g_n(\beta)^{\mathrm{T}} \hat{W} \left.\frac{\partial g_n}{\partial \beta}\right|_{\hat{\beta}_n} = 2\left(\left.\frac{\partial g_n}{\partial \beta}\right|_{\hat{\beta}_n}\right)^{\mathrm{T}} \hat{W} g_n\left(\hat{\beta}_n\right) = 0$$

then

$$\left(\left.\frac{\partial g_n}{\partial \beta}\right|_{\hat{\beta}_n}\right)^{\mathrm{T}} \hat{W} g_n\left(\hat{\beta}_n\right) = 0$$

Together with (7.7),

$$\hat{\beta}_n - \beta_0 = -\left[\left(\left.\frac{\partial g_n}{\partial \beta}\right|_{\hat{\beta}_n}\right)^{\mathrm{T}} \hat{W} \left.\frac{\partial g_n}{\partial \beta}\right|_{\bar{\beta}}\right]^{-1} \left(\left.\frac{\partial g_n}{\partial \beta}\right|_{\hat{\beta}_n}\right)^{\mathrm{T}} \hat{W} g_n(\beta_0)$$

To simplify the notation, denote that $d_n(\beta) := \left.\frac{\partial g_n}{\partial \beta}\right|_{\beta}$, then

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) = -\left[d\left(\hat{\beta}_n\right)^{\mathrm{T}} \hat{W} d\left(\bar{\beta}\right)\right]^{-1} d\left(\hat{\beta}_n\right)^{\mathrm{T}} \hat{W} \sqrt{n} g_n(\beta_0)$$

CLT implies that

$$\begin{aligned}\sqrt{n} g_n(\beta_0) &\xrightarrow{d} \mathcal{N}\left(\mathbb{E}\left[g(z_i, \beta_0)\right], n \cdot \frac{1}{n}\mathrm{Var}\left[g(z_i, \beta_0)\right]\right) \\ &= \mathcal{N}\left(0, \mathbb{E}\left[g(z_i, \beta_0) g(z_i, \beta_0)^{\mathrm{T}}\right]\right) := \mathcal{N}(0, S_0)\end{aligned}$$

Moreover, by $\hat{\beta}_n \xrightarrow{\mathbb{P}} \beta_0$, we have $\bar{\beta} \xrightarrow{\mathbb{P}} \beta_0$ then

$$d_n\left(\hat{\beta}_n\right) = \left.\frac{\partial g_n}{\partial \beta}\right|_{\hat{\beta}_n} \xrightarrow{\mathbb{P}} \left.\frac{\partial g_n}{\partial \beta}\right|_{\beta_0} := d_0 \text{ and } d_n\left(\bar{\beta}\right) = \left.\frac{\partial g_n}{\partial \beta}\right|_{\bar{\beta}} \xrightarrow{\mathbb{P}} d_0$$

Also recall that $\hat{W} \xrightarrow{\mathbb{P}} W$, then

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) &= -\left[d_n^{\mathrm{T}} \hat{W} d_n\right]^{-1} d_n^{\mathrm{T}} \hat{W} \sqrt{n} g_n(\beta_0) \\ &\xrightarrow{\mathbb{P}} -\left[d_0^{\mathrm{T}} W d_0\right]^{-1} d_0^{\mathrm{T}} W \sqrt{n} g_n(\beta_0) \\ &\xrightarrow{d} \mathcal{N}\left(0, \left[d_0^{\mathrm{T}} W d_0\right]^{-1} d_0^{\mathrm{T}} W S_0 W^{\mathrm{T}} d_0 \left[d_0^{\mathrm{T}} W d_0\right]^{-1}\right)\end{aligned}$$

Note that $d_0^{\mathrm{T}} W d_0 = \left(d_0^{\mathrm{T}} W d_0\right)^{\mathrm{T}}$, and since W is positive definite, $d_0^{\mathrm{T}} W d_0 > 0$.

□

7.5 Endogeneity

7.5.1 Endogeneity Problem

In a linear regression model:

$$y_i = \beta_0^{\mathrm{T}} x_i + \varepsilon_i$$

[7] You can take $p = 2$ as an example to verify.

Endogeneity means that [8]

$$\mathbb{E}\left[x_i\varepsilon_i\right] \neq 0$$

Here are some examples of the problem of endogeneity.

Example 7.30. Omitted variables.

The true model is that

$$y_i = \beta_0^{\mathrm{T}} x_i + \underbrace{\gamma_0^{\mathrm{T}} z_i + u_i}_{\varepsilon_i}$$

where

$$\mathbb{E}\left[u_i \middle| x_i, z_i\right] = 0$$

and x_i and z_i are correlated.

$$\mathbb{E}\left[x_i\varepsilon_i\right] = \mathbb{E}\left[x_i\left(\gamma_0^{\mathrm{T}} z_i + u_i\right)\right] = \mathbb{E}\left[\gamma_0^{\mathrm{T}} z_i x_i^{\mathrm{T}} + x_i u_i\right] = \gamma_0^{\mathrm{T}}\mathbb{E}\left[z_i x_i^{\mathrm{T}}\right] \neq 0$$

Example 7.31. Measurement error.

The true model is that

$$y_i = \beta_0^{\mathrm{T}} x_i^* + u_i, \mathbb{E}\left[u_i \middle| x_i^*\right] = 0$$

But x_i^* is measured with error:

$$x_i = x_i^* + v_i$$

Therefore,

$$y_i = \beta_0^{\mathrm{T}}\left(x_i - v_i\right) + u_i = \beta_0^{\mathrm{T}} x_i \underbrace{-\beta_0^{\mathrm{T}} v_i + u_i}_{\varepsilon_i}$$

If $\mathbb{E}\left[v_i x_i\right] \neq 0$, then

$$\mathbb{E}\left[x_i\varepsilon_i\right] \neq 0$$

Example 7.32. Demand and supply (Simultaneity).

The demand curve and supply curve are given by

$$q^d = \beta_0 P + \varepsilon_d, \beta_0 < 0$$
$$q^s = \gamma_0 P + \varepsilon_s, \gamma_0 > 0$$

In equilibrium,

$$q^d = q^s \Longrightarrow P^* = \frac{\varepsilon_s - \varepsilon_d}{\beta_0 - \gamma_0}$$

So,

$$q^* = \beta_0 P^* + \varepsilon_d$$
$$q^* = \gamma_0 P^* + \varepsilon_d$$

The equilibrium price P^* is correlated with ε_d and ε_s.

[8] Recall that $\mathbb{E}\left[x_i\varepsilon_i\right] \neq 0 \Longrightarrow \mathbb{E}\left[\varepsilon_i \middle| x_i\right] \neq 0$.

7.5.2 Instrument Variable Estimation

Remark 7.33. Afterwards, in particular, we let

$$x_i = (1, x_{i2}, \cdots, x_{ik})^{\mathrm{T}} \in \mathbb{R}^k$$

and assume that

$$\begin{aligned} \mathbb{E}[\varepsilon_i] &= 0, \forall i = 1, 2, \cdots, n \\ \mathbb{E}[x_{ij}\varepsilon_i] &= 0, \forall j \neq k \\ \mathbb{E}[x_{ik}\varepsilon_i] &\neq 0 \end{aligned}$$

That is $x_{i2}, \cdots, x_{i,k-1}$ are exogenous variables, and x_{ik} is an endogenous variable.

Let w_i be a random variable satisfying

$$\mathbb{E}[w_i\varepsilon_i] = 0$$

Define a new vector

$$z_i = (1, x_{i2}, \cdots, x_{i,k-1}, w_i)^{\mathrm{T}} \in \mathbb{R}^k$$

Then

$$\begin{aligned} &\mathbb{E}[z_i\varepsilon_i] = 0_{k\times 1} \\ \Longrightarrow\ &\mathbb{E}\left[z_i\left(y_i - \beta_0^{\mathrm{T}} x_i\right)\right] = 0_{k\times 1} \\ \Longrightarrow\ &\mathbb{E}[z_i y_i] = \mathbb{E}\left[z_i x_i^{\mathrm{T}}\right]\beta_0 \\ \Longrightarrow\ &\beta_0 = \left(\mathbb{E}\left[z_i x_i^{\mathrm{T}}\right]\right)^{-1}\mathbb{E}[z_i y_i] \end{aligned}$$

given that $\mathbb{E}\left[z_i x_i^{\mathrm{T}}\right]$ is positive definite and thus, invertible.

Therefore, by minimizing the square loss or method of moments or maximum likelihood, we can have

$$\hat{\beta}_{\mathrm{IV}} = \left(\frac{1}{n}\sum_{i=1}^n z_i x_i^{\mathrm{T}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n z_i y_i\right)$$

Theorem 7.34. (Unbiasedness)

$$\mathbb{E}\left[\hat{\beta}_{\mathrm{IV}}\right] = \beta_0$$

Proof. Write

$$\begin{aligned} \hat{\beta}_{\mathrm{IV}} &= \left(\frac{1}{n}\sum_{i=1}^n z_i x_i^{\mathrm{T}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n z_i y_i\right) \\ &= \left(\frac{1}{n}\sum_{i=1}^n z_i x_i^{\mathrm{T}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n z_i\left(x_i^{\mathrm{T}}\beta_0 + \varepsilon_i\right)\right) \\ &= \beta_0 + \left(\frac{1}{n}\sum_{i=1}^n z_i x_i^{\mathrm{T}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n z_i\varepsilon_i\right) \end{aligned}$$

Then

$$\mathbb{E}\left[\hat{\beta}_{\mathrm{IV}}\right] = \beta_0 + \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} z_i x_i^{\mathrm{T}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} z_i \mathbb{E}\left[\varepsilon_i \middle| z_i\right]\right)\right] = \beta_0$$

□

Theorem 7.35. (Consistency)

$$\hat{\beta}_{\mathrm{IV}} \xrightarrow{\mathbb{P}} \beta_0$$

Proof. Since

$$\hat{\beta}_{\mathrm{IV}} - \beta_0 = \left(\frac{1}{n}\sum_{i=1}^{n} z_i x_i^{\mathrm{T}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} z_i \varepsilon_i\right)$$

By LLN,

$$\frac{1}{n}\sum_{i=1}^{n} z_i x_i^{\mathrm{T}} \xrightarrow{\mathbb{P}} \mathbb{E}\left[z_i x_i^{\mathrm{T}}\right]$$
$$\frac{1}{n}\sum_{i=1}^{n} z_i \varepsilon_i \xrightarrow{\mathbb{P}} \mathbb{E}\left[z_i \varepsilon_i\right] = 0$$

Therefore,

$$\hat{\beta}_{\mathrm{IV}} - \beta_0 \xrightarrow{\mathbb{P}} 0$$

□

Remark 7.36. Endogeneity makes $\mathbb{E}\left[x_i \varepsilon_i\right] \neq 0 \Longrightarrow \mathbb{E}\left[x_i \middle| \varepsilon_i\right] \neq 0$, thus,

$$\hat{\beta}_{\mathrm{OLS}} = \beta_0 + \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i \varepsilon_i\right)$$

is not consistency and unbaised.

Theorem 7.37. (Asymptotic distribution)

$$\sqrt{n}\left(\hat{\beta}_{\mathrm{IV}} - \beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_{zx}^{-1} V \left(\Sigma_{zx}^{-1}\right)^{\mathrm{T}}\right)$$

where

$$V = \mathbb{E}\left[z_i z_i^{\mathrm{T}} \varepsilon_i^2\right], \Sigma_{zx} = \mathbb{E}\left[z_i x_i^{\mathrm{T}}\right]$$

Proof. Firstly,

$$\frac{1}{n}\sum_{i=1}^{n} z_i x_i^{\mathrm{T}} \xrightarrow{\mathbb{P}} \mathbb{E}\left[z_i x_i^{\mathrm{T}}\right] := \Sigma_{zx}$$

Second, by CLT,

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} z_i \varepsilon_i \xrightarrow{d} \mathcal{N}\left(0, \mathrm{Var}\left(z_i \varepsilon_i\right)\right)$$

where

$$\mathrm{Var}\left(z_i \varepsilon_i\right) = \mathbb{E}\left[z_i \varepsilon_i \left(z_i \varepsilon_i\right)^{\mathrm{T}}\right] = \mathbb{E}\left[z_i z_i^{\mathrm{T}} \varepsilon_i^2\right] := V$$

Note that ε_i is a scalar.

Therefore,

$$\sqrt{n}\left(\hat{\beta}_{\mathrm{IV}}-\beta_0\right)\xrightarrow{d}\mathcal{N}\left(0,\Sigma_{zx}^{-1}V\left(\Sigma_{zx}^{-1}\right)^{\mathrm{T}}\right)$$

□

Remark 7.38. Several remarks on IV.

1. The original model is
$$y_i=x_i^{\mathrm{T}}\beta_0+\varepsilon_i$$
or
$$y_i=\mathbb{E}\left[x_i^{\mathrm{T}}\middle|z_i\right]\beta_0+\underbrace{\left(x_i^{\mathrm{T}}-\mathbb{E}\left[x_i^{\mathrm{T}}\middle|z_i\right]\right)\beta_0+\varepsilon_i}_{\hat{\varepsilon}_i}=\mathbb{E}\left[x_i^{\mathrm{T}}\middle|z_i\right]\beta_0+\hat{\varepsilon}_i$$
2. OLS estimator is a special case of IV estimator (taking x_k as the IV).

7.5.3 Two-stage Least Square Estimation

Remark 7.39. Now suppose that x_{ik} is an endogenous variable and we have a vector of IV:

$$w_i=(w_{i1},\cdots,w_{ip})$$

where $p>1$.

Now, consider

$$z_i=(1,x_{i2},\cdots,x_{i,k-1},w_{i1},\cdots,w_{ip})_{(k+p-1)\times 1}^{\mathrm{T}}$$

as the independent variable. Note that

$$\mathbb{E}\left[z_i\varepsilon_i\right]=\mathbb{E}\left[z_i\left(y_i-\beta_0^{\mathrm{T}}x_i\right)\right]=0\Longrightarrow\mathbb{E}\left[z_iy_i\right]_{(k+p-1)\times 1}=\mathbb{E}\left[z_ix_i^{\mathrm{T}}\right]_{(k+p-1)\times k}\beta_0$$

Since $p>1$, then $\mathbb{E}\left[z_ix_i^{\mathrm{T}}\right]$ cannot be invertible.

To deal with it, we can use 2-step Least Squares (2SLS) estimator. The 2SLS Estimator is

$$\hat{\beta}_{\mathrm{2SLS}}=\left(\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i\hat{x}_i^{\mathrm{T}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{x}_iy_i\right)$$

where $\hat{x}_i=(1,x_{i2},\cdots,x_{i,k-1},\hat{x}_{ik})^{\mathrm{T}}$, $\hat{x}_{ik}=\hat{\gamma}^{\mathrm{T}}z_i$ and $\hat{\gamma}$ is the OLS estimator in the regression of x_{ik} on z_i.

In matrix form,

$$\hat{\beta}_{\mathrm{2SLS}}=\left(\hat{\boldsymbol{X}}^{\mathrm{T}}\hat{\boldsymbol{X}}\right)^{-1}\left(\hat{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{Y}\right)$$

where [9]

$$\hat{\boldsymbol{X}}=\boldsymbol{P}_Z\boldsymbol{X}\text{ where }\boldsymbol{P}_Z=\boldsymbol{Z}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^{\mathrm{T}}$$

[9]Denote $\hat{\boldsymbol{X}}=\left(\hat{\boldsymbol{X}}_1,\hat{\boldsymbol{X}}_2,\cdots,\hat{\boldsymbol{X}}_k\right)$, then by the properties of OLS, $\hat{\boldsymbol{X}}_k=\boldsymbol{P}_Z\boldsymbol{X}_k$. Moreover, note that

$$z_i=(1,x_{i2},\cdots,x_{i,k-1},w_{i1},\cdots,w_{ip})^{\mathrm{T}}$$
$$x_i=(1,x_{i2},\cdots,x_{i,k-1},x_{ik})$$

The first k entries of z_i and x_i meet, then $\boldsymbol{X}_m=\boldsymbol{P}_Z\boldsymbol{X}_m\ (1\leqslant m\leqslant k-1)$.

Note that $\boldsymbol{P}_Z$ is symmetric and idempotent, then

$$\hat{\beta}_{\text{2SLS}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}_Z\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}_Z\boldsymbol{Y}\right)$$

Example 7.40. Prove that $\hat{\beta}_{\text{2SLS}}$ is consistent.

Proof.

$$\begin{aligned}\hat{\beta}_{\text{2SLS}} &= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}_Z\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}_Z\boldsymbol{Y}\right)\\ &= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}_Z\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}_Z\left(\boldsymbol{X}\beta_0+\boldsymbol{\varepsilon}\right)\right)\\ &= \beta_0 + \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}_Z\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\varepsilon}\end{aligned}$$

Then since

$$\frac{1}{n}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\varepsilon} \xrightarrow{\mathbb{P}} 0$$

we have

$$\hat{\beta}_{\text{2SLS}} \xrightarrow{\mathbb{P}} \beta_0$$

□

Example 7.41. Provide assumptions that are necessary to derive the asymptotic distribution of $\hat{\beta}_{\text{2SLS}}$.

Solution.

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}_{\text{2SLS}}-\beta_0\right) &= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\right)^{-1}\sqrt{n}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\varepsilon}\\ &= \left[\left(\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z}\right)\left(\frac{1}{n}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\right)^{-1}\left(\frac{1}{n}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{X}\right)\right]^{-1}\left(\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z}\right)\left(\frac{1}{n}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\right)^{-1}\frac{\sqrt{n}}{n}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\varepsilon}\\ &\xrightarrow{\mathbb{P}} \left(\Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx}\right)^{-1}\Sigma_{xz}\Sigma_{zz}^{-1}\frac{\sqrt{n}}{n}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\varepsilon}\end{aligned}$$

where

$$\sqrt{n}\frac{1}{n}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\varepsilon} \xrightarrow{d} \mathcal{N}\left(0, \operatorname{Var}\left(z_i^{\mathrm{T}}\varepsilon_i\right)\right) = \mathcal{N}\left(0, \mathbb{E}\left(z_i z_i{}^{\mathrm{T}}\varepsilon_i^2\right)\right) := \mathcal{N}\left(0, V\right)$$

Then

$$\sqrt{n}\left(\hat{\beta}_{\text{2SLS}}-\beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \left(\Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx}\right)^{-1}\Sigma_{xz}\Sigma_{zz}^{-1}V\left[\left(\Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx}\right)^{-1}\Sigma_{xz}\Sigma_{zz}^{-1}\right]^{\mathrm{T}}\right)$$

□

Chapter 8

Nonlinear Regression

8.1 M-estimation

Remark 8.1. In statistics, M-estimators are a broad class of extremum estimators for which the objective function is a sample average. We consider three types of M-estimation:

- Maximum Likelihood Estimation (MLE): to estimate discrete choice model.
- Method of Moments (generalized to Generalized Method of Moments, GMM).
- Quantile Regression.

Remark 8.2. M-estimation is a class of parameter estimations by

- either minimizing a sample average of $m(z, \beta)$,

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} m\left(z_i, \beta\right)$$

  The true value is $\beta_0 = \arg\min_{\beta} \mathbb{E}\left[m\left(z_i, \beta\right)\right]$.

- or solving the equation of the type

$$\frac{1}{n} \sum_{i=1}^{n} g\left(z_i, \beta\right) = 0$$

  The true value β_0 satisfies $\mathbb{E}\left[g\left(z_i, \beta_0\right)\right] = 0$.

Example 8.3. OLS as an M-estimation, $z_i = (x_i, y_i)$.

- Minimize an objective. (we take $m(z_i, \beta) = \left(y_i - x_i^{\mathrm{T}}\beta\right)^2$)

$$\hat{\beta}_n = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - x_i^{\mathrm{T}}\beta\right)^2 \xrightarrow{\mathbb{P}} \beta_0 = \arg\min_{\beta} \mathbb{E}\left[\left(Y - X^{\mathrm{T}}\beta\right)\right]$$

- Solve equations. F.O.C. of OLS gives us (we take $g(z_i, \beta) = x_i\left(y_i - x_i^{\mathrm{T}}\beta\right)$)

$$\frac{1}{n}\sum_{i=1}^{n} x_i\left(y_i - x_i^{\mathrm{T}}\hat{\beta}_n\right) = \frac{1}{n}\sum_{i=1}^{n} x_i\hat{\varepsilon}_i = 0$$

  or in population,

$$\mathbb{E}\left[x_i\left(y_i - x_i^{\mathrm{T}}\beta_0\right)\right] = \mathbb{E}\left[x_i\varepsilon_i\right] = 0$$

Example 8.4. Estimate the mean.

We are interested in the mean of Z, $\beta_0 = \mathbb{E}Z$.

- Minimize an objective.

$$\hat{\beta}_n = \arg\min_{\beta} \frac{1}{n}\sum_{i=1}^{n}(z_i - \beta)^2 \xrightarrow{\mathbb{P}} \beta_0 = \arg\min_{\beta}\mathbb{E}(Z - \beta)^2$$

- Solve equations. F.O.C. implies

$$\frac{1}{n}\sum_{i=1}^{n}(z_i - \beta) = \frac{1}{n}\sum_{i=1}^{n} z_i - \beta = 0$$

  In population,

$$\beta_0 = \mathbb{E}z_i$$

Example 8.5. MLE as an M-estimation.

we are interested in estimating β_0 of a parametric density function $f(z, \beta)$.

- Minimize an objective.

$$\hat{\beta}_n = \arg\min_{\beta} \frac{1}{n}\sum_{i=1}^{n} -\log f(z_i, \beta) \xrightarrow{\mathbb{P}} \beta_0 = \arg\min_{\beta}\mathbb{E}\left[-\log f(z_i, \beta)\right]$$

- Solve equations. F.O.C. implies

$$\frac{1}{n}\sum_{i=1}^{n} -\frac{\partial \log f(z_i, \beta)}{\partial \beta} = 0$$

  In population,

$$\mathbb{E}\left[-\frac{\partial \log f(z_i, \beta)}{\partial \beta}\right] = 0$$

8.2 Maximum Likelihood Estimation

8.2.1 Introduction

Remark 8.6. why we are interested in estimating density function?

- Density function is a type of distribution function, and

- A distribution function contains almost everything of the data generating process

Any parameter β can be expressed as a functional T of distribution f.

$$\beta = T(f)$$

Example 8.7. OLS estimator as a functional of density function.

$$\beta_0 = \left[\mathbb{E}\left(XX^{\mathrm{T}}\right)\right]^{-1}\mathbb{E}\left(XY\right)$$

Let X be a scalar, then

$$\begin{aligned}\beta_0 &= \left[\mathbb{E}\left(X^2\right)\right]^{-1}\mathbb{E}\left(XY\right)\\ &= \left[\int x^2 f_X\left(x\right)\mathrm{d}x\right]^{-1}\int\int xy f_{X,Y}\left(x,y\right)\mathrm{d}x\mathrm{d}y\\ &:= T_1\left(f_X, f_{X,Y}\right)\\ &:= T\left(f_{X,Y}\right)\end{aligned}$$

where $f_X = \int f_{X,Y}\mathrm{d}y$.

Example 8.8. OLS model with normality assumption.

$$y_i = x_i^{\mathrm{T}}\beta + u_i,\, u_i|\, x_i \sim \mathcal{N}\left(0, \sigma_u^2\right)$$

Note that

$$\frac{1}{n}\sum_{i=1}^{n}\log f\left(y_i, x_i, \beta\right) = \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{1}{\sqrt{2\pi}\sigma_u}e^{-\frac{1}{2}\frac{\left(y_i - x_i^{\mathrm{T}}\beta\right)^2}{\sigma_u^2}}\right) = \log\frac{1}{\sqrt{2\pi}\sigma_u} - \frac{1}{2n\sigma_u^2}\sum_{i=1}^{n}\left(y_i - x_i^{\mathrm{T}}\beta\right)^2$$

so,

$$\max_{\beta}\frac{1}{n}\sum_{i=1}^{n}\log f\left(y_i, x_i, \beta\right) \Longleftrightarrow \min_{\beta}\frac{1}{2n\sigma_u^2}\sum_{i=1}^{n}\left(y_i - x_i^{\mathrm{T}}\beta\right)^2$$

Remark 8.9. Advantages and disadvantages of MLE.

- Pros:
  - Wide applicability (binary data, count data, panel data, etc.) MLE can handle complicated data and economic models
  - MLE is efficient under correct specification
- Cons:
  - Strong distributional assumption (trade-off) between efficiency and robustness
  - If the distribution is true, MLE is efficient. If not, MLE could be inconsistent.

8.2.2 Identification

Remark 8.10. MLE is

$$\hat{\beta} = \arg\max_{\beta} \frac{1}{n} \sum_{i=1}^{n} \ln f\left(z_i; \beta\right) = \arg\max_{\beta} M_n\left(\beta\right)$$

and we have

$$M_n\left(\beta\right) \xrightarrow{\mathbb{P}} M_0\left(\beta\right) = \mathbb{E}\left[\ln f\left(z_i; \beta\right)\right]$$

According to the inequality:

$$\ln x \leqslant 2\left(\sqrt{x} - 1\right)$$

we have

$$\begin{aligned}
M_0\left(\beta\right) - M_0\left(\beta_0\right) &= \mathbb{E}\left[\ln \frac{f\left(z_i; \beta\right)}{f\left(z_i; \beta_0\right)}\right] \\
&\leqslant 2\mathbb{E}\left[\sqrt{\frac{f\left(z_i; \beta\right)}{f\left(z_i; \beta_0\right)}} - 1\right] \\
&= 2\int \sqrt{\frac{f\left(z; \beta\right)}{f\left(z; \beta_0\right)}} \cdot f\left(z; \beta_0\right) \mathrm{d}z - 2 \\
&= 2\int \sqrt{f\left(z; \beta\right)} \cdot \sqrt{f\left(z; \beta_0\right)} \mathrm{d}z - 2 \\
&= 2\int \sqrt{f\left(z; \beta\right)} \cdot \sqrt{f\left(z; \beta_0\right)} \mathrm{d}z - \int f\left(z; \beta\right) \mathrm{d}z - \int f\left(z; \beta_0\right) \mathrm{d}z \\
&= -\int \left[\sqrt{f\left(z; \beta\right)} - \sqrt{f\left(z; \beta_0\right)}\right]^2 \mathrm{d}z \\
&\leqslant 0
\end{aligned}$$

If $\int \left[\sqrt{f\left(z; \beta\right)} - \sqrt{f\left(z; \beta_0\right)}\right]^2 \mathrm{d}z \neq 0$, then β_0 uniquely maximizes $M_0(\beta)$.

8.2.3 Efficiency

Definition 8.11. Let $Z = \left(z_1, z_2, \cdots, z_n\right)^{\mathrm{T}}$ with joint PDF $f\left(\cdot, \theta_0\right)$, We call

- $S = \frac{\partial \log f(Z,\theta)}{\partial \theta}$ the *score function.*
- $J = \mathbb{E}\left[\frac{\partial \log f(z,\theta)}{\partial \theta} \frac{\partial \log f(z,\theta)}{\partial \theta^{\mathrm{T}}}\right]$ the *Fisher information matrix.*

Lemma 8.12. Variance of MLE $\mathrm{Var}\left(\sqrt{n}\hat{\theta}_n\right)$ attains the lower bound J^{-1} in the limit.

Lemma 8.13. If $\tilde{\theta}$ is unbiased, then $n\mathrm{Var}\left(\tilde{\theta}\right) = n\mathbb{E}\left(\tilde{\theta} - \theta\right)^2 \geqslant J^{-1}$.

Proof. Since $\tilde{\theta}$ is unbiased, it holds that

$$\theta = \int \tilde{\theta}\left(z\right) f\left(z, \theta\right) \mathrm{d}z$$

Let I denote the unit matrix, then

$$I := \frac{\partial \theta}{\partial \theta^{\mathrm{T}}} = \int \tilde{\theta}(z) \frac{\partial f(z,\theta)}{\partial \theta^{\mathrm{T}}} \mathrm{d}z = \int \tilde{\theta}(z) \frac{\partial \log f(z,\theta)}{\partial \theta^{\mathrm{T}}} f(z,\theta)\, \mathrm{d}z$$

Evaluating at θ_0, we obtain

$$I = \mathbb{E}\left[\tilde{\theta}(z) \frac{\partial \log f(z,\theta_0)}{\partial \theta^{\mathrm{T}}}\right] = \mathbb{E}\left[\left[\tilde{\theta}(z) - \theta_0\right] \frac{\partial \log f(z,\theta)}{\partial \theta^{\mathrm{T}}}\right] = \mathbb{E}\left[\left[\tilde{\theta}(z) - \theta_0\right] S^{\mathrm{T}}\right]$$

The second equality comes from

$$\begin{aligned}\mathbb{E}\left[\frac{\partial \log f(z,\theta_0)}{\partial \theta^{\mathrm{T}}}\right] &= \int \frac{1}{f(z,\theta_0)} \frac{\partial f(z,\theta_0)}{\partial \theta^{\mathrm{T}}} f(z,\theta_0)\, \mathrm{d}z \\ &= \int \frac{\partial f(z,\theta_0)}{\partial \theta^{\mathrm{T}}} \mathrm{d}z = \frac{\partial}{\partial \theta^{\mathrm{T}}} \int f(z,\theta_0)\, \mathrm{d}z = 0\end{aligned}$$

By matrix Cauchy-Schwarz inequality [1], $I = \mathbb{E}\left[\left(\tilde{\theta} - \theta_0\right) S^{\mathrm{T}}\right]$, also note that we assume that i.i.d. sampling and $\mathbb{E}\left[\frac{\partial \log f(z_i,\theta_0)}{\partial \theta^{\mathrm{T}}}\right] = 0$, we have

$$\begin{aligned}\mathrm{Var}\left(\tilde{\theta}\right) &= \mathbb{E}\left[\left(\tilde{\theta} - \theta_0\right)\left(\tilde{\theta} - \theta_0\right)^{\mathrm{T}}\right] \\ &\geqslant \mathbb{E}\left[\left(\tilde{\theta} - \theta_0\right) S^{\mathrm{T}}\right] \left[\mathbb{E}\left(SS^{\mathrm{T}}\right)\right]^{-1} \mathbb{E}\left[S\left(\tilde{\theta} - \theta_0\right)^{\mathrm{T}}\right] \\ &= \left[\mathbb{E}\left(SS^{\mathrm{T}}\right)\right]^{-1} \\ &= \left(\mathbb{E}\left[\left(\sum_{i=1}^{n} \frac{\partial \log f(z_i,\theta_0)}{\partial \theta}\right)\left(\sum_{i=1}^{n} \frac{\partial \log f(z_i,\theta_0)}{\partial \theta}\right)^{\mathrm{T}}\right]\right)^{-1} \\ &= \left(n\mathbb{E}\left[\frac{\partial \log f(z_i,\theta_0)}{\partial \theta} \frac{\partial \log f(z_i,\theta_0)}{\partial \theta^{\mathrm{T}}}\right]\right)^{-1} \\ &= \frac{1}{n} J^{-1}\end{aligned}$$

or

$$n\mathrm{Var}\left(\tilde{\theta}\right) = n\mathbb{E}\left[\left(\tilde{\theta} - \theta_0\right)^2\right] \geqslant J^{-1}$$

□

Theorem 8.14. If an estimator is unbaised, then MLE is the most efficient one.

Example 8.15. For the case of sample mean of a normal distribution $\mathcal{N}(\beta_0, 1)$, show that MLE is the most efficient unbiased estimator.

[1] For any random vectors x, y, we have

$$\mathbb{E}\left[yy^{\mathrm{T}}\right] \geqslant \mathbb{E}\left[yx^{\mathrm{T}}\right] \left[\mathbb{E}\left(xx^{\mathrm{T}}\right)\right]^{-1} \mathbb{E}\left[xy^{\mathrm{T}}\right]$$

where $A \geqslant B$ means $A - B$ is positive semi-definite.

Proof. Firstly, the MLE estimator is

$$\hat{\beta}_n = \arg\max_{\beta} \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\beta)^2}{2}} = \arg\min_{\beta} \sum_{i=1}^{n} (x_i - \beta)^2 = \frac{1}{n} \sum_{i=1}^{n} x_i$$

then

$$n\mathrm{Var}\left(\hat{\beta}_n\right) = \frac{1}{n}\mathrm{Var}\left(\sum_{i=1}^{n} x_1\right) = 1$$

And for any other estimator $\tilde{\beta}_n$ with $\mathbb{E}\beta_n = \beta_0$, we have

$$\beta_0 = \mathbb{E}\tilde{\beta}_n = \int \tilde{\beta}_n(z) f(z, \beta_0) \, \mathrm{d}z$$

then

$$1 = \int \tilde{\beta}_n(z) \frac{\partial f(z, \beta_0)}{\partial \beta} \mathrm{d}z = \mathbb{E}\left[\tilde{\beta}_n(z) \frac{\partial \log f(z, \beta_0)}{\partial \beta}\right]$$

since

$$\mathbb{E}\left[\frac{\partial \log f(z, \beta_0)}{\partial \beta}\right] = \int \frac{1}{f(z, \beta_0)} \frac{\partial f(z, \beta_0)}{\partial \beta} f(z, \beta_0) \, \mathrm{d}z = \frac{\partial}{\partial \beta} \int f(z, \beta_0) \, \mathrm{d}z = 1$$

we have

$$1 = \mathbb{E}\left[\left(\tilde{\beta}_n - \beta_0\right) \frac{\partial \log f(z, \beta_0)}{\partial \beta}\right]$$

thus,

$$\begin{aligned}
\mathrm{Var}\left(\tilde{\beta}_n\right) &= \mathbb{E}\left[\left(\tilde{\beta}_n - \beta_0\right)^2\right] \\
&\geqslant \mathbb{E}\left[\left(\tilde{\beta}_n - \beta_0\right) \frac{\partial \log f(z, \beta_0)}{\partial \beta}\right] \left(\mathbb{E}\left[\frac{\partial \log f(z, \beta_0)}{\partial \beta}\right]^2\right)^{-1} \mathbb{E}\left[\frac{\partial \log f(z, \beta_0)}{\partial \beta}\left(\tilde{\beta}_n - \beta_0\right)\right] \\
&= \left(\mathbb{E}\left[\frac{\partial \log f(z, \beta_0)}{\partial \beta}\right]^2\right)^{-1} \\
&= \left(\mathbb{E}\left[\sum_{i=1}^{n} -(x_i - \beta_0)\right]^2\right)^{-1} \\
&= \left(n\mathbb{E}(x_i - \beta_0)^2\right)^{-1} \\
&= \frac{1}{n}
\end{aligned}$$

Therefore,

$$\mathrm{Var}\left(\tilde{\beta}_n\right) \geqslant \mathrm{Var}\left(\hat{\beta}_n\right)$$

□

8.2.4 Asymptotic Normality

Theorem 8.16. If the density function f is differentiable, let $\hat{\beta}_n$ the MLE, then

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \Omega^{-1}\right)$$

where

$$\Omega = \mathbb{E}\left[\frac{f_\beta(z_i; \beta_0) f_{\beta^{\mathrm{T}}}(z_i; \beta_0)}{f^2(z_i; \beta_0)}\right]$$

Proof. Firstly note that

$$\mathbb{E}\left[\frac{f_\beta\left(z_i;\beta_0\right)}{f\left(z_i;\beta_0\right)}\right]=\int\frac{\partial}{\partial\beta}f\left(z;\beta_0\right)\mathrm{d}z=0$$

then by CLT,

$$\sqrt{n}\nabla_\beta M_n\left(\beta_0\right)=\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\frac{f_\beta\left(z_i;\beta_0\right)}{f\left(z_i;\beta_0\right)}\xrightarrow{d}\mathcal{N}\left(\mathbb{E}\left[\frac{f_\beta\left(z_i;\beta_0\right)}{f\left(z_i;\beta_0\right)}\right],\mathrm{Var}\left[\frac{f_\beta\left(z_i;\beta_0\right)}{f\left(z_i;\beta_0\right)}\right]\right):=\mathcal{N}\left(0,\Omega\right)$$

where

$$\Omega=\mathbb{E}\left[\frac{f_\beta\left(z_i;\beta_0\right)f_{\beta^{\mathrm{T}}}\left(z_i;\beta_0\right)}{f^2\left(z_i;\beta_0\right)}\right]$$

Also note that

$$\begin{aligned}\nabla_{\beta\beta^{\mathrm{T}}}M_n\left(\beta\right)&=\frac{1}{n}\sum_{i=1}^{n}\frac{f_\beta\left(z_i;\beta_0\right)f_{\beta^{\mathrm{T}}}\left(z_i;\beta_0\right)-f_{\beta\beta^{\mathrm{T}}}\left(z_i;\beta_0\right)f\left(z_i;\beta_0\right)}{f^2\left(z_i;\beta_0\right)}\\&=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{f_\beta\left(z_i;\beta_0\right)f_{\beta^{\mathrm{T}}}\left(z_i;\beta_0\right)}{f^2\left(z_i;\beta_0\right)}-\frac{f_{\beta\beta^{\mathrm{T}}}\left(z_i;\beta_0\right)}{f\left(z_i;\beta_0\right)}\right]\end{aligned}$$

where we have

$$\mathbb{E}\left[\frac{f_{\beta\beta^{\mathrm{T}}}\left(z_i;\beta_0\right)}{f\left(z_i;\beta_0\right)}\right]=\int\frac{\partial^2}{\partial\beta\partial\beta^{\mathrm{T}}}f\left(z;\beta_0\right)\mathrm{d}z=0$$

So,

$$\mathbb{E}\left[\frac{f_\beta\left(z_i;\beta_0\right)f_{\beta^{\mathrm{T}}}\left(z_i;\beta_0\right)}{f^2\left(z_i;\beta_0\right)}-\frac{f_{\beta\beta^{\mathrm{T}}}\left(z_i;\beta_0\right)}{f\left(z_i;\beta_0\right)}\right]=\Omega$$

By WLLN,

$$\nabla_{\beta\beta^{\mathrm{T}}}M_n\left(\beta\right)\xrightarrow{\mathbb{P}}\Omega$$

By the F.O.C. and mean-value theorem, we have

$$0=\nabla_\beta M_n\left(\hat{\beta}_n\right)=\nabla_\beta M_n\left(\beta_0\right)+\nabla_{\beta\beta^{\mathrm{T}}}M_n\left(\bar{\beta}\right)\left(\hat{\beta}_n-\beta_0\right)$$

Then by Slutzky Theorem,

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}_n-\beta_0\right)&=-\left[\nabla_{\beta\beta^{\mathrm{T}}}M_n\left(\bar{\beta}\right)\right]^{-1}\nabla_\beta M_n\left(\beta_0\right)\\&\xrightarrow{d}\mathcal{N}\left(0,\Omega^{-1}\Omega\left(\Omega^{-1}\right)^{\mathrm{T}}\right)\\&=\mathcal{N}\left(0,\Omega^{-1}\right)\end{aligned}$$

□

8.2.5 Examples

Example 8.17. Uniform Distribution A case that MLE is not asymptotic normal when M_n is not differentiable).

Let $X_1,\cdots,X_n\overset{\text{i.i.d.}}{\sim}U\left[0,\theta\right]$, θ is fixed but unknown, and we want to estimate θ.

The log-likelihood function is

$$M_n = \ln \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{I}_{\{X_i \in [0,\theta]\}} = \sum_{i=1}^{n} \ln \mathbb{I}_{\{X_i \in [0,\theta]\}} - n \ln \theta$$

Let $X_{(1)}, \cdots, X_{(n)}$ be the ordered sample. Firstly,

$$\hat{\theta}_{\text{MLE}} \leqslant X_{(n)}$$

If $\hat{\theta}_{\text{MLE}} > X_{(n)}$, M_n is not at its maximum, since as θ grows from $X_{(n)}$, $\sum_{i=1}^{n} \ln \mathbb{I}_{\{X_i \in [0,\theta]\}}$ is unchanged, but $-n \ln \theta$ is decreasing. Second,

$$\hat{\theta}_{\text{MLE}} \geqslant X_{(n)}$$

If $\hat{\theta}_{\text{MLE}} < X_{(n)}$, then $X_{(n)} \notin [0, \theta]$, leading to $M_n = -\infty$. Therefore,

$$\hat{\theta}_{\text{MLE}} = X_{(n)}$$

For $x < 0$, we have

$$\begin{aligned} \mathbb{P}\left\{n\left(X_{(n)} - \theta\right) \leqslant x\right\} &= \mathbb{P}\left\{\max_{1 \leqslant i \leqslant n} X_i \leqslant \theta + \frac{x}{n}\right\} = \prod_{i=1}^{n} \mathbb{P}\left\{X_i \leqslant \theta + \frac{x}{n}\right\} \\ &= \left(1 + \frac{x}{n\theta}\right)^n \to e^{\frac{x}{\theta}} \end{aligned}$$

In short, [2]

$$-n\left(X_{(n)} - \theta\right) \xrightarrow{d} \exp\left(\frac{1}{\theta}\right)$$

Note that the rate of convergence is n instead of $\sqrt{n}$.

Example 8.18. Binary Choice Models.

Consider such a model:

$$y_i = \mathbb{I}_{\left\{\beta_0^{\mathrm{T}} x_i - \varepsilon_i \geqslant 0\right\}}$$

where $\varepsilon_i \perp x_i$ and $\varepsilon_i \overset{\text{i.i.d.}}{\sim}$ a CDF $\Lambda(\cdot)$. Then

$$\begin{aligned} \mathbb{P}\left\{y_i = 1 \middle| x_i\right\} &= \mathbb{P}\left\{\varepsilon_i \leqslant \beta_0^{\mathrm{T}} x_i \middle| x_i\right\} = \Lambda\left(\beta_0^{\mathrm{T}} x_i\right) \\ \mathbb{P}\left\{y_i = 0 \middle| x_i\right\} &= 1 - \Lambda\left(\beta_0^{\mathrm{T}} x_i\right) \end{aligned}$$

[2] Because let $y = -x$,

$$\mathbb{P}\left\{n\left(X_{(n)} - \theta\right) \leqslant -y\right\} \to e^{\frac{-y}{\theta}}$$

or (note that $X_{(n)}$ is continuous)

$$1 - \mathbb{P}\left\{-n\left(X_{(n)} - \theta\right) \leqslant y\right\} \to -e^{\frac{-y}{\theta}}$$

then

$$\mathbb{P}\left\{-n\left(X_{(n)} - \theta\right) \leqslant y\right\} \to 1 - e^{\frac{-y}{\theta}}$$

Then

$$\mathbb{P}\{y_i = y|\, x_i = x;\beta\} = \Lambda\left(\beta^{\mathrm{T}}x\right)^y\left[1-\Lambda\left(\beta^{\mathrm{T}}x\right)\right]^{1-y} := f\left(y|\, x;\beta\right)$$

The MLE is

$$\hat{\beta}_{\mathrm{MLE}} = \arg\max_{\beta\in B}\frac{1}{n}\sum_{i=1}^{n}\left\{y_i\ln\Lambda\left(\beta^{\mathrm{T}}x_i\right)+(1-y_i)\ln\left(1-\Lambda\left(\beta^{\mathrm{T}}x_i\right)\right)\right\}$$

- In Probit model,

$$\Lambda(z) = \int_0^z \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\mathrm{d}x$$

- In Logit model,

$$\Lambda(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$

Now we show the identification of the model: Assume that $\mathbb{E}\left[x_ix_i^{\mathrm{T}}\right]$ is positive definite, or $\mathbb{E}\left[x_ix_i^{\mathrm{T}}\right] > 0$, then

$$\left(\beta-\beta_0\right)^{\mathrm{T}}\mathbb{E}\left[x_ix_i^{\mathrm{T}}\right]\left(\beta-\beta_0\right) > 0,\ \text{for } \beta\neq\beta_0$$

or

$$\mathbb{E}\left[\left(\beta-\beta_0\right)^{\mathrm{T}}x_i\left(\left(\beta-\beta_0\right)x_i\right)^{\mathrm{T}}\right] > 0,\ \text{for } \beta\neq\beta_0$$

then

$$\mathbb{E}\left[\left(\beta-\beta_0\right)^{\mathrm{T}}x_i\right]^2 > 0,\ \text{for } \beta\neq\beta_0$$

then

$$\exists x,\ \text{such that } \mathbb{P}\left\{\beta^{\mathrm{T}}x\neq\beta_0^{\mathrm{T}}x\right\} > 0$$

Assume Λ is strictly increasing, then

$$\exists x,\ \text{such that } \mathbb{P}\left\{\Lambda\left(\beta^{\mathrm{T}}x\right)\neq\Lambda\left(\beta_0^{\mathrm{T}}x\right)\right\} > 0$$

then

$$\int\int\left[\sqrt{f\left(y|\, x;\beta\right)f\left(x\right)}-\sqrt{f\left(y|\, x;\beta_0\right)f\left(x\right)}\right]^2\mathrm{d}x\mathrm{d}y > 0,\ \text{for } \beta\neq\beta_0$$

Example 8.19. Establish the asymptotic distributions of the MLEs for Probit and Logit models.

Firstly,

$$\Omega := \mathbb{E}\left[\frac{f_\beta\left(z_i;\beta_0\right)f_{\beta^{\mathrm{T}}}\left(z_i;\beta_0\right)}{f^2\left(z_i;\beta_0\right)}\right]$$

By Theorem 8.16,

$$\sqrt{n}\left(\hat{\beta}_n-\beta_0\right)\xrightarrow{d}\mathcal{N}\left(0,\Omega^{-1}\right)$$

where

$$\frac{\partial\ln f\left(z_i;\beta\right)}{\partial\beta} = y_i\frac{1}{\Lambda\left(\beta^{\mathrm{T}}x_i\right)}\left.\frac{\mathrm{d}\Lambda\left(z\right)}{\mathrm{d}z}\right|_{z=\beta^{\mathrm{T}}x_i}x_i-\left(1-y_i\right)\frac{1}{1-\Lambda\left(\beta^{\mathrm{T}}x_i\right)}\left.\frac{\mathrm{d}\Lambda\left(z\right)}{\mathrm{d}z}\right|_{z=\beta^{\mathrm{T}}x_i}x_i$$

Note that

$$\frac{\partial\ln f\left(z_i;\beta\right)}{\partial\beta} = \frac{1}{f\left(z_i;\beta\right)}\frac{\partial f\left(z_i;\beta\right)}{\partial\beta}$$

- For Probit model,

$$\begin{aligned}\frac{\partial \ln f\left(z_i;\beta\right)}{\partial \beta} &= \left[y_i \frac{1}{\Phi\left(\beta^{\mathrm{T}} x_i\right)} - \left(1-y_i\right)\frac{1}{1-\Phi\left(\beta^{\mathrm{T}} x_i\right)}\right]\phi\left(\beta^{\mathrm{T}} x_i\right) x_i \\ &= \frac{y_i - \Phi\left(\beta^{\mathrm{T}} x_i\right)}{\Phi\left(\beta^{\mathrm{T}} x_i\right)\left[1-\Phi\left(\beta^{\mathrm{T}} x_i\right)\right]}\phi\left(\beta^{\mathrm{T}} x_i\right) x_i\end{aligned}$$

  then

$$\Omega_{\text{Probit}} = \mathbb{E}\left\{\left[\frac{y_i - \Phi\left(\beta_0^{\mathrm{T}} x_i\right)}{\Phi\left(\beta_0^{\mathrm{T}} x_i\right)\left[1-\Phi\left(\beta_0^{\mathrm{T}} x_i\right)\right]}\phi\left(\beta_0^{\mathrm{T}} x_i\right)\right]^2 x_i x_i^{\mathrm{T}}\right\}$$

- For Logit model,

$$\frac{\partial \ln f\left(z_i;\beta\right)}{\partial \beta} = y_i\left[1-\Lambda\left(\beta^{\mathrm{T}} x_i\right)\right] x_i - \left(1-y_i\right)\Lambda\left(\beta^{\mathrm{T}} x_i\right) x_i = \left[y_i - \Lambda\left(\beta^{\mathrm{T}} x_i\right)\right] x_i$$

  then

$$\Omega_{\text{Logit}} = \mathbb{E}\left\{\left[y_i - \Lambda\left(\beta^{\mathrm{T}} x_i\right)\right]^2 x_i x_i^{\mathrm{T}}\right\}$$

8.3 Generalized Method of Moments

8.3.1 Implementation, Asymptotic Variance and Efficiency

Remark 8.20. GMM is based on

$$\mathbb{E}\left[g\left(Z,\beta_0\right)\right] = 0$$

We can estimate β_0 by solving

$$\frac{1}{n}\sum_{i=1}^{n} g\left(z_i,\beta\right) = 0$$

Now the question is the uniqueness of the solution $\hat{\beta}_n$. Under certain conditions, ruling out multicollinearity-like conditions, we generally have

- Just identified: $\beta \in \mathbb{R}^k, g\left(z_i,\beta\right) \in \mathbb{R}^k$, can be uniquely solved.
- Under identified: $\beta \in \mathbb{R}^k, g\left(z_i,\beta\right) \in \mathbb{R}^j, j < k$, have infinite number of solutions.
- Over identified: $\beta \in \mathbb{R}^k, g\left(z_i,\beta\right) \in \mathbb{R}^j, j > k$, in general, no solution.

Example 8.21. Just identification.

F.O.C. of OLS is

$$\frac{1}{n}\sum_{i=1}^{n} x_i\left(y_i - x_i^{\mathrm{T}}\beta\right) = 0$$

Let $g\left(z_i,\beta\right) = x_i\left(y_i - x_i^{\mathrm{T}}\beta\right)$, we have GMM $\frac{1}{n}\sum_{i=1}^{n} g\left(z_i,\beta\right) = 0$. Thus,

$$\hat{\beta}_n = \left(\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}\right)^{-1}\sum_{i=1}^{n} x_i y_i$$

This is the case of just identification, since we have k unknowns $\beta = \left(\beta_1,\cdots,\beta_k\right)^{\mathrm{T}}$ and k equations $\frac{1}{n}\sum_{i=1}^{n} x_i\left(y_i - x_i^{\mathrm{T}}\beta\right) = 0_{k\times 1}$.

Example 8.22. Overidentification.

- IV estimation.
$$y_i = x_i^{\mathrm{T}}\beta_i + \varepsilon_i, \mathbb{E}(\varepsilon_i x_i) \neq 0, \mathbb{E}(\varepsilon_i z_i) = 0$$
but $\dim(x_i) < \dim(z_i)$.

- A strict conditional moment condition of a non-linear model.
$$y_i = m(x_i, \beta) + \varepsilon_i, \mathbb{E}(\varepsilon_i | x_i) = 0$$
Then, for any (measurable) function q, the number of these function could be infinite, we have
$$\mathbb{E}[q(x_i)\varepsilon_i] = \mathbb{E}[q(x_i)\mathbb{E}[\varepsilon_i | x_i]] = 0$$
Then the moment conditions are much larger than $k = \dim(\beta)$.
Example: Arrellano and Bond estimator for dynamic panel.
$$y_{it} = \delta y_{i,t-1} + x_{it}^{\mathrm{T}}\beta + \mu_i + v_{it}, i = 1, \cdots, n; t = 1, \cdots, T$$
we assume strict exogeneity:
$$\mathbb{E}[v_{it} | x_{i1}, \cdots, x_{iT}, y_{i,t-1}, \cdots, y_{i1}, y_{i0}, \mu_i] = 0, t = 1, \cdots, T$$
Conditional mean restriction is
$$\mathbb{E}[\Delta v_{it} | x_{i1}, \cdots, x_{iT}, y_{i,t-2}, \cdots, y_{i1}, y_{i0}, \mu_i] = 0$$
is implied by
  - $\mathbb{E}[v_{i,t-1} | x_{i1}, \cdots, x_{iT}, y_{i,t-2}, \cdots, y_{i1}, y_{i0}, \mu_i] = 0$ by assumption.
  - $\mathbb{E}[v_{it} | x_{i1}, \cdots, x_{iT}, y_{i,t-2}, \cdots, y_{i1}, y_{i0}, \mu_i] = 0$. [3]

Parameters of interest are $\delta \in \mathbb{R}^1, \beta \in \mathbb{R}^k$, so the dimension of unknown parameter is $k+1$.
But we have $T(T-1) + \frac{T(T-1)}{2}$ potential moment conditions, which can be larger than $k+1$: for each $t = 2, \cdots, T$, we have
  - $t-1$ conditions:
$$\mathbb{E}[y_{i,t-2}\Delta v_{it}] = 0, \mathbb{E}[y_{i,t-3}\Delta v_{it}] = 0, \cdots, \mathbb{E}[y_{i0}\Delta v_{it}] = 0$$

[3] Since
$$\mathbb{E}[\mathbb{E}[v_{it} | x_{i1}, \cdots, x_{iT}, y_{i,t-1}, y_{i,t-2}, \cdots, y_{i1}, y_{i0}, \mu_i] | x_{i1}, \cdots, x_{iT}, y_{i,t-2}, \cdots, y_{i1}, y_{i0}, \mu_i] = 0$$

– T conditions:

$$\mathbb{E}\left[x_{i1}\Delta v_{it}\right]=0,\cdots,\mathbb{E}\left[x_{iT}\Delta v_{it}\right]=0$$

In total, we have $\sum_{t=2}^{T}\left(T-t+1\right)=T\left(T+1\right)-\sum_{t=1}^{T-1}t=T\left(T+1\right)+\frac{T(T+1)}{2}$ conditions.

Theorem 8.23. (Asymptotic Variance of GMM) Consider the GMM estimator

$$\hat{\beta}_n:=\hat{\beta}_{\text{GMM}}=\underset{\beta\in B}{\arg\min}\left(\frac{1}{n}\sum_{i=1}^{n}g\left(z_i,\beta\right)\right)^{\mathrm{T}}\hat{W}_{p\times p}\left(\frac{1}{n}\sum_{i=1}^{n}g\left(z_i,\beta\right)\right)$$

Suppose that $\hat{W}\xrightarrow{\mathbb{P}}W$, where W is positive definite, and $\hat{\beta}_n\xrightarrow{\mathbb{P}}\beta_0$, and that $\mathbb{E}\left[g\left(z_i,\beta_0\right)\right]=0$.

Then

$$\sqrt{n}\left(\hat{\beta}_n-\beta_0\right)\xrightarrow{d}\mathcal{N}\left(0,\left(G_0^{\mathrm{T}}WG_0\right)^{-1}G_0^{\mathrm{T}}W\Omega_0WG_0\left(G_0^{\mathrm{T}}WG_0\right)^{-1}\right)$$

where $G_0:=\mathbb{E}\left[\frac{\partial g}{\partial\beta}\left(z_i,\beta\right)\right]$, $\Omega_0=\operatorname{Var}\left[g\left(z_i,\beta_0\right)\right]=\mathbb{E}\left[g\left(z_i,\beta_0\right)g\left(z_i,\beta_0\right)^{\mathrm{T}}\right]$.

If $W=\Omega_0^{-1}$, then

$$\sqrt{n}\left(\hat{\beta}_n-\beta_0\right)\xrightarrow{d}\mathcal{N}\left(0,\left(G_0^{\mathrm{T}}WG_0\right)^{-1}\right)$$

Proof. See Example 7.29. □

Theorem 8.24. Define

$$\begin{aligned}V_W&=\left(G_0^{\mathrm{T}}WG_0\right)^{-1}G_0^{\mathrm{T}}W\Omega WG_0\left(G_0^{\mathrm{T}}WG_0\right)^{-1}\\V_{\Omega^{-1}}&=\left(G_0^{\mathrm{T}}WG_0\right)^{-1}\end{aligned}$$

Then for any W with nonsingular $G^{\mathrm{T}}WG$, it holds that $V_W-V_{\Omega^{-1}}$ is positive semi-definite, or

$$V_W\geqslant V_{\Omega^{-1}}$$

In other words, when $W=\Omega^{-1}$, the GMM estimator is the most efficient.

Proof. Let $A=WG_0\left(G_0^{\mathrm{T}}WG_0\right)^{-1}$, and $B=\Omega^{-1}G_0\left(G_0^{\mathrm{T}}\Omega^{-1}G_0\right)^{-1}$, then

$$\begin{aligned}V_W&=A^{\mathrm{T}}\Omega A\\V_{\Omega^{-1}}&=B^{\mathrm{T}}\Omega B\end{aligned}$$

Thus,

$$\begin{aligned}V_W&=\left(A-B+B\right)^{\mathrm{T}}\Omega\left(A-B+B\right)\\&=\left(A-B\right)^{\mathrm{T}}\Omega\left(A-B\right)+B^{\mathrm{T}}\Omega\left(A-B\right)+\left(A-B\right)^{\mathrm{T}}\Omega B+V_{\Omega^{-1}}\end{aligned}$$

where we have

$$\begin{aligned}B^{\mathrm{T}}\Omega\left(A-B\right)&=\left(\Omega^{-1}G_0\left(G_0^{\mathrm{T}}\Omega^{-1}G_0\right)^{-1}\right)^{\mathrm{T}}\Omega\left(WG_0\left(G_0^{\mathrm{T}}WG_0\right)^{-1}-\Omega^{-1}G_0\left(G_0^{\mathrm{T}}\Omega^{-1}G_0\right)^{-1}\right)\\&=\left(G_0^{\mathrm{T}}\Omega^{-1}G_0\right)^{-1}G_0^{\mathrm{T}}WG_0\left(G_0^{\mathrm{T}}WG_0\right)^{-1}-\left(G_0^{\mathrm{T}}\Omega^{-1}G_0\right)^{-1}G_0^{\mathrm{T}}\Omega^{-1}G_0\left(G_0^{\mathrm{T}}\Omega^{-1}G_0\right)^{-1}\\&=\left(G_0^{\mathrm{T}}\Omega^{-1}G_0\right)^{-1}-\left(G_0^{\mathrm{T}}\Omega^{-1}G_0\right)^{-1}\\&=0\end{aligned}$$

Therefore,

$$V_W - V_{\Omega^{-1}} = (A-B)^{\mathrm{T}} \Omega (A-B)$$

where Ω is positive semi-definite, then so is $V_W - V_{\Omega^{-1}}$. □

Example 8.25. (The intuition of the efficiency of GMM) Suppose we have two samples:

- Sample 1: $x_1, x_2, \cdots, x_n$, where $x_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\beta_0, 1)$
- Sample 2: $y_1, y_2, \cdots, y_n$, where $y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\beta_0, 2)$

And x_i and y_j are pairwise independent. Now we consider different method to estimate.

1. $\bar{x}$: $\operatorname{Var}(\bar{x}) = \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^n x_i\right) = \frac{1}{n}$.
2. $\bar{y}$: $\operatorname{Var}(\bar{y}) = \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^n y_i\right) = \frac{2}{n}$.
3. How about $\frac{1}{2}\bar{x} + \frac{1}{2}\bar{y}$? The variance is

$$\operatorname{Var}\left(\frac{1}{2}\bar{x} + \frac{1}{2}\bar{y}\right) = \frac{1}{4}\frac{1}{n} + \frac{1}{4}\frac{2}{n} = \frac{3}{4n} < \min\left\{\frac{1}{n}, \frac{2}{n}\right\}$$

4. We can also formulate the GMM estimator: Let $g(z_i, \beta) = \begin{pmatrix} x_i - \beta \\ y_i - \beta \end{pmatrix}$, then $\mathbb{E}g(z_i, \beta) = 0$. This moment condition is over identified, since we have one unknown and two equations.

   Let

$$\Omega^{-1} := \left[\operatorname{Var}(g(z_i, \beta_0))\right]^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$$

   then

$$g_n(\beta) = \frac{1}{n}\sum_{i=1}^n g(z_i, \beta) = \begin{pmatrix} \bar{x} - \beta \\ \bar{y} - \beta \end{pmatrix}$$

   So the GMM estimator is

$$\begin{aligned} \hat{\beta}_{\text{GMM}} &= \arg\min_{\beta} g_n(\beta)^{\mathrm{T}} \Omega^{-1} g_n(\beta) \\ &= \arg\min_{\beta} (\bar{x} - \beta, \bar{y} - \beta) \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} \bar{x} - \beta \\ \bar{y} - \beta \end{pmatrix} \\ &= \arg\min_{\beta} \left\{(\bar{x} - \beta)^2 + \frac{1}{2}(\bar{y} - \beta)^2\right\} \end{aligned}$$

   Then

$$2\left(\bar{x} - \hat{\beta}_{\text{GMM}}\right) + \bar{y} - \hat{\beta}_{\text{GMM}} = 0 \Longrightarrow \hat{\beta}_{\text{GMM}} = \frac{2}{3}\bar{x} + \frac{1}{3}\bar{y}$$

   and

$$\operatorname{Var}\left(\frac{2}{3}\bar{x} + \frac{1}{3}\bar{y}\right) = \frac{4}{9n} + \frac{2}{9n} = \frac{2}{3n} < \frac{3}{4n}$$

Now we only check that GMM with $W = \Omega^{-1}$ is most efficient unbaised estimator of the form $a\bar{x} + b\bar{y}$.

Unbiasedness gives us

$$\mathbb{E}\left[a\bar{x} + b\bar{y}\right] = a\beta_0 + b\beta_0 = \beta_0 \Longrightarrow a + b = 1$$

we need to solve

$$\min_a \operatorname{Var}\left(a\bar{x} + (1-a)\,\bar{y}\right) = \min_a \left\{\frac{a^2}{n} + \frac{2\,(1-a)^2}{n}\right\} \Longrightarrow a = \frac{2}{3}$$

Example 8.26. With the assumption $\mathbb{E}\left[\varepsilon_i^2 \middle| z_i\right] = \sigma^2$.

1. What is the most efficient GMM estimator?
2. What is the minimum variance-covariance matrix?

Solution. Linear model:

$$y_i = x_i^{\mathrm{T}}\beta_0 + \varepsilon_i$$

The moment condition is

$$\mathbb{E}\left[x_i\varepsilon_i\right] = \mathbb{E}\left[x_i\left(y_i - x_i^{\mathrm{T}}\beta_i\right)\right] = 0$$

The GMM estimator is

$$\hat{\beta}_{\mathrm{GMM}} = \arg\min_{\beta} \left(\sum_{i=1}^{n} x_i\varepsilon_i\right)^{\mathrm{T}} \hat{W}_{k\times k} \left(\sum_{i=1}^{n} x_i\varepsilon_i\right)$$

where $\hat{W} \xrightarrow{\mathbb{P}} W$. Then

$$\sqrt{n}\left(\hat{\beta}_{\mathrm{GMM}} - \beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \left(G^{\mathrm{T}}WG\right)^{-1} G^{\mathrm{T}}W\Omega WG\left(G^{\mathrm{T}}WG\right)^{-1}\right)$$

where

$$\Omega = \operatorname{Var}\left(g_i\left(z_i, \beta\right)\right) = \operatorname{Var}\left(x_i\varepsilon_i\right) = \mathbb{E}\left[x_i x_i^{\mathrm{T}}\right]\sigma^2$$
$$G = \mathbb{E}\left[\left.\frac{\partial g_i\left(z_i,\beta\right)}{\partial\beta}\right|_{\beta=\beta_0}\right] = \mathbb{E}\left[x_i x_i^{\mathrm{T}}\right]$$

When $W = \Omega^{-1}$, the GMM estimator is the most efficient, or

$$\sqrt{n}\left(\hat{\beta}_{\mathrm{GMM}} - \beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \left(G^{\mathrm{T}}WG\right)^{-1}\right)$$

where

$$\left(G^{\mathrm{T}}WG\right)^{-1} = \left(\mathbb{E}\left[x_i x_i^{\mathrm{T}}\right]\mathbb{E}\left[x_i x_i^{\mathrm{T}}\right]^{-1}\frac{1}{\sigma^2}\mathbb{E}\left[x_i x_i^{\mathrm{T}}\right]\right)^{-1} = \sigma^2\mathbb{E}\left[x_i x_i^{\mathrm{T}}\right]^{-1}$$

which is just the variance of OLS estimator, implying OLS is the most efficient unbaised estimator. □

Example 8.27. Unbiasedness of GMM? GMM can be biased!

Consider the model:

$$y_i = x_i \log \beta_0 + \varepsilon_i, \mathbb{E}\left[y_i - x_i \log \beta_0\right] = 0$$

where $y_i, x_i \in \mathbb{R}$. Let $\hat{W} = I$, then

$$\hat{\beta}_{\text{GMM}} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - x_i \log \beta\right)^2 \Longrightarrow \widehat{\log \beta} = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}$$

or

$$\hat{\beta} = \exp\left\{\frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}\right\}$$

then

$$\mathbb{E}\hat{\beta} = \mathbb{E}\exp\left\{\frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}\right\} < \exp\left\{\mathbb{E}\left[\frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}\right]\right\} = e^{\log \beta_0} = \beta_0$$

Example 8.28. An economic example of GMM from Hansen and Singleton (1982).

A representative consumer who maximizes utility:

$$\mathbb{E}_0\left[\sum_{t=0}^{\infty} \beta^t U\left(C_t\right)\right]$$

subject to

$$c_t + P_t Q_t \leqslant R_t Q_{t-1} + W_t$$

where P_t is the asset price, Q_t is the asset holding, R_t is the payoff, W_t is the wage.

Assume

$$U'\left(C_t\right) = C_t^{-\gamma}$$

The F.O.C. is

$$P_t C_t^{-\gamma} = \beta \mathbb{E}_t\left[R_{t+1} C_{t+1}^{-\gamma}\right]$$

or

$$\mathbb{E}_t\left[\beta \frac{R_{t+1}}{P_t}\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma}\right] = 0$$

Let $Z_t = \frac{R_{t+1}}{P_t}, Y_{t+1} = \frac{C_{t+1}}{C_t}$, Let X_t belongs to information set at t, then:

$$\mathbb{E}_t\left[X_t\left(\beta Z_{t+1} Y_{t+1}^{\gamma-1} - 1\right)\right] = 0$$

which requires GMM estimation.

8.3.2 Overidentification Test

Theorem 8.29. Suppose we want to test the following hypothesis:

$$H_0 : \mathbb{E}\left[g\left(w_i, \beta_0\right)\right] = 0$$

Under H_0, the most efficient GMM estimator $\hat{\beta}_n$ satisfies :

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \left(G^{\mathrm{T}}\Omega^{-1}G\right)^{-1}\right)$$

Then, under H_0,

$$J_n = \left[\sqrt{n}g_n\left(\hat{\beta}_n\right)\right]^{\mathrm{T}} \hat{\Omega}^{-1} \left[\sqrt{n}g_n\left(\hat{\beta}_n\right)\right] \xrightarrow{d} \chi^2\left(p-k\right)$$

where $W = \Omega^{-1}$ is $p \times p$, β_0 is $k \times 1$, and

$$\hat{\Omega}^{-1} = \frac{1}{n}\sum_{i=1}^{n} g_i\left(z_i, \hat{\beta}_n\right) g_i\left(z_i, \hat{\beta}_n\right)^{\mathrm{T}} \xrightarrow{\mathbb{P}} \Omega = \mathrm{Var}\left(g_i\left(z_i, \beta_0\right)\right) = \mathbb{E}\left[g_i\left(z_i, \beta_0\right) g_i\left(z_i, \beta_0\right)^{\mathrm{T}}\right]$$

Proof. By Taylor's expansion,

$$g_n\left(\hat{\beta}_n\right) = g_n\left(\beta_0\right) + G_n\left(\bar{\beta}\right) \cdot \left(\hat{\beta}_n - \beta_0\right)$$

where $\bar{\beta}$ is between $\hat{\beta}_n$ and β_0. Then

$$\sqrt{n}g_n\left(\hat{\beta}_n\right) = \sqrt{n}g_n\left(\beta_0\right) + G_n\left(\bar{\beta}\right) \cdot \sqrt{n}\left(\hat{\beta}_n - \beta_0\right)$$

By Example 7.29, we have

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) = \left[G_n\left(\hat{\beta}_n\right)^{\mathrm{T}} \hat{\Omega}^{-1} G_n\left(\hat{\beta}_n\right)\right]^{-1} G_n\left(\hat{\beta}_n\right)^{\mathrm{T}} \hat{\Omega}\sqrt{n}g_n\left(\beta_0\right)$$

then

$$\begin{aligned}
\sqrt{n}g_n\left(\hat{\beta}_n\right) &= \sqrt{n}g_n\left(\beta_0\right) + G_n\left(\bar{\beta}\right)\left[G_n\left(\hat{\beta}_n\right)^{\mathrm{T}} \hat{\Omega}^{-1} G_n\left(\hat{\beta}_n\right)\right]^{-1} G_n\left(\hat{\beta}_n\right)^{\mathrm{T}} \hat{\Omega}\sqrt{n}g_n\left(\beta_0\right) \\
&= \left[I_{p\times p} - G_n\left(\bar{\beta}\right)\left[G_n\left(\hat{\beta}_n\right)^{\mathrm{T}} \hat{\Omega}^{-1} G_n\left(\hat{\beta}_n\right)\right] G_n\left(\hat{\beta}_n\right)^{\mathrm{T}} \hat{\Omega}\right] \sqrt{n}g_n\left(\beta_0\right) \\
&= \left[I - G\left(G^{\mathrm{T}}\Omega^{-1}G\right)G\Omega\right]\sqrt{n}g_n\left(\beta_0\right) + o_p\left(1\right)
\end{aligned}$$

Then

$$\begin{aligned}
J_n &= \left(\left[I - G\left(G^{\mathrm{T}}\Omega^{-1}G\right)G\Omega\right]\sqrt{n}g_n\left(\beta_0\right)\right)^{\mathrm{T}} \hat{\Omega}^{-1}\left(\left[I - G\left(G^{\mathrm{T}}\Omega^{-1}G\right)G\Omega\right]\sqrt{n}g_n\left(\beta_0\right)\right) + o_p\left(1\right) \\
&= \sqrt{n}g_n\left(\beta_0\right)^{\mathrm{T}}\left[\Omega^{-1} - \Omega^{-1}G\left(G^{\mathrm{T}}\Omega^{-1}G\right)^{-1}G^{\mathrm{T}}\Omega^{-1}\right]\sqrt{n}g_n\left(\beta_0\right) + o_p\left(1\right) \\
&= \left[\sqrt{n}\Omega^{-\frac{1}{2}}g_n\left(\beta_0\right)\right]^{\mathrm{T}}\left[I - \Omega^{-\frac{1}{2}}G\left(G^{\mathrm{T}}\Omega^{-1}G\right)^{-1}G^{\mathrm{T}}\Omega^{-\frac{1}{2}}\right]\left[\sqrt{n}\Omega^{-\frac{1}{2}}g_n\left(\beta_0\right)\right] + o_p\left(1\right)
\end{aligned}$$

where

$$\sqrt{n}\Omega^{-\frac{1}{2}}g_n\left(\beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, I_{p\times p}\right)$$

then

$$B := \Omega^{-\frac{1}{2}}G \Longrightarrow I - \Omega^{-\frac{1}{2}}G\left(G^{\mathrm{T}}\Omega^{-1}G\right)^{-1}G^{\mathrm{T}}\Omega^{-\frac{1}{2}} = I - B\left(B^{\mathrm{T}}B\right)^{-1}B^{\mathrm{T}} = M_B$$

□

8.3.3 IV, 2SLS, and GMM

Remark 8.30. For the endogenous problem,

$$y_i = \beta_0^{\mathrm{T}} x_i + \varepsilon_i$$

where $\mathbb{E}\left[x_i\varepsilon_i\right]_{k\times 1} \neq 0$, but we have $\mathbb{E}\left[z_i\varepsilon_i\right]_{p\times 1} = 0$ The moment condition is

$$\mathbb{E}\left[z_i\left(y_i - \beta_0^{\mathrm{T}} x_i\right)\right]_{p\times 1} = 0$$

where we have p equations and k unknowns. We have learnt that 2SLS can find an estimation of β_0, now we use the GMM to solve it.

$$\hat{\beta}_{\mathrm{GMM}} = \arg\min_{\beta} M_n(\beta) = \arg\min_{\beta} \left(\frac{1}{n}\sum_{i=1}^{n} z_i\varepsilon_i\right)^{\mathrm{T}} \hat{W}_{p\times p} \left(\frac{1}{n}\sum_{i=1}^{n} z_i\varepsilon_i\right)$$

Note that

$$\begin{aligned}
M_n(\beta) &= \left(\frac{1}{n}\sum_{i=1}^{n} z_i\left(y_i - \beta^{\mathrm{T}} x_i\right)\right)^{\mathrm{T}} \hat{W}_{p\times p} \left(\frac{1}{n}\sum_{i=1}^{n} z_i\left(y_i - \beta^{\mathrm{T}} x_i\right)\right) \\
&= \left(\frac{1}{n}\sum_{i=1}^{n} y_i z_i^{\mathrm{T}} - \beta^{\mathrm{T}}\frac{1}{n}\sum_{i=1}^{n} x_i z_i^{\mathrm{T}}\right) \hat{W}_{p\times p} \left(\frac{1}{n}\sum_{i=1}^{n} z_i y_i - \frac{1}{n}\sum_{i=1}^{n} z_i x_i^{\mathrm{T}}\beta\right) \\
&= \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} y_i z_i^{\mathrm{T}}\right) \hat{W}_{p\times p} \left(\frac{1}{n}\sum_{i=1}^{n} z_i y_i\right)}_{\text{constant}} \underbrace{- 2\beta^{\mathrm{T}} \left(\frac{1}{n}\sum_{i=1}^{n} z_i^{\mathrm{T}} x_i\right) \hat{W}_{p\times p} \left(\frac{1}{n}\sum_{i=1}^{n} y_i z_i\right)}_{\text{scalar}} \\
&\quad + \beta^{\mathrm{T}} \left(\frac{1}{n}\sum_{i=1}^{n} x_i z_i^{\mathrm{T}}\right) \hat{W}_{p\times p} \left(\frac{1}{n}\sum_{i=1}^{n} z_i x_i^{\mathrm{T}}\right) \beta
\end{aligned}$$

Then F.O.C. implies that

$$-2\left(\frac{1}{n}\sum_{i=1}^{n} z_i^{\mathrm{T}} x_i\right) \hat{W}_{p\times p} \left(\frac{1}{n}\sum_{i=1}^{n} y_i z_i\right)_{p\times 1} + 2\left(\frac{1}{n}\sum_{i=1}^{n} x_i z_i^{\mathrm{T}}\right) \hat{W}_{p\times p} \left(\frac{1}{n}\sum_{i=1}^{n} z_i x_i^{\mathrm{T}}\right) \hat{\beta}_n = 0$$

Then

$$\begin{aligned}
\hat{\beta}_{\mathrm{GMM}} &= \left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i z_i^{\mathrm{T}}\right) \hat{W}_{p\times p} \left(\frac{1}{n}\sum_{i=1}^{n} z_i x_i^{\mathrm{T}}\right)\right]^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} z_i^{\mathrm{T}} x_i\right) \hat{W}_{p\times p} \left(\frac{1}{n}\sum_{i=1}^{n} y_i z_i^{\mathrm{T}}\right) \\
&= \left[\left(\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z}\right) \hat{W} \left(\frac{1}{n}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{X}\right)\right]^{-1} \left(\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z}\right) \hat{W} \left(\frac{1}{n}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Y}\right) \\
&= \left[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z}\right) \hat{W} \left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{X}\right)\right]^{-1} \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z}\right) \hat{W} \left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Y}\right)
\end{aligned}$$

Example 8.31. IV and 2SLS are special cases of GMM.

- IV: if $p = k$, and assume that $\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{X}$ is invertible, then the GMM estimator agrees with the IV estimator.

$$\hat{\beta}_{\mathrm{IV}} = \left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Y}\right) = \left(\frac{1}{n}\sum_{i=1}^{n} z_i x_i^{\mathrm{T}}\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} y_i z_i\right)$$

- 2SLS: If $W = \left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\right)^{-1}$, then the GMM estimator agrees with the 2SLS estimator.

$$\begin{aligned}\hat{\beta}_{\mathrm{2SLS}} &= \left[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z}\right)\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{X}\right)\right]^{-1}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z}\right)\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Y}\right) \\ &= \left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}_{Z}\boldsymbol{X}\right]^{-1}\left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}_{Z}\boldsymbol{Y}\right]^{-1}\end{aligned}$$

Note that if we assume $\mathbb{E}\left[\varepsilon_i^2 \middle| z_i\right] = \sigma^2$, and let

$$\Omega := \mathbb{E}\left[g\left(z_i,\beta\right)g\left(z_i,\beta\right)^{\mathrm{T}}\right] = \mathbb{E}\left[z_i z_i^{\mathrm{T}}\right]\sigma^2$$

then we can get the most efficient estimator by letting $W = \Omega^{-1} = \mathbb{E}\left[z_i z_i^{\mathrm{T}}\right]^{-1}\sigma^{-2}$.

8.3.4 Conditional Moment Restrictions and Optimal Instruments

Remark 8.32. Conditional moment restriction has the form

$$\mathbb{E}\left[h\left(w_i,\theta_0\right)\middle| x_i\right] = 0$$

For simplicity, we assume $h\left(w_i,\theta_0\right) \in \mathbb{R}^1$. By Law of iterated expectation, we have, for any measurable function a,

$$\mathbb{E}\left[a\left(x_i\right)h\left(w_i,\theta_0\right)\right] = 0$$

Based on this unconditional moment restrictions, we can do GMM. Some times we need to choose a, which is a general type of *instrument*, to make the estimator efficient and consistent.

By the asymptotic variance of GMM estimator, we have

$$\mathrm{Var}\left(\lim_{n\to\infty}\hat{\theta}_n\right) = \left(G_a^{\mathrm{T}}\Omega_a^{-1}G_a\right)^{-1}$$

where $G_a = \mathbb{E}\left[a\left(x_i\right)\frac{\partial h(w_i,\theta_0)}{\partial\theta^{\mathrm{T}}}\right]$, and $\Omega_a = \mathrm{Var}\left[a\left(x_i\right)h\left(w_i,\theta_0\right)\right] = \mathbb{E}\left[a\left(x_i\right)a\left(x_i\right)^{\mathrm{T}}h\left(w_i,\theta_0\right)^2\right]$.

Theorem 8.33. For any a, it holds that

$$\left(G_a^{\mathrm{T}}\Omega_a^{-1}G_a\right)^{-1} \geqslant \left(G_*^{\mathrm{T}}\Omega_*^{-1}G_*\right)^{-1}$$

where $G_* = \mathbb{E}\left[a^*\left(x_i\right)\frac{\partial h(w_i,\theta_0)}{\partial\theta^{\mathrm{T}}}\right]$, $\Omega_* = \mathbb{E}\left[a^*\left(x_i\right)a^*\left(x_i\right)^{\mathrm{T}}h\left(w_i,\theta_0\right)^2\right]$, and

$$a^*\left(x_i\right) = \mathbb{E}\left[\left.\frac{\partial h\left(w_i,\theta_0\right)}{\partial\theta^{\mathrm{T}}}\right| x_i\right]\mathbb{E}\left[h\left(w_i,\theta_0\right)^2\middle| x_i\right]^{-1}$$

we call a^* the *optimal IV*. Note that $\dim\left(a^*\left(x_i\right)\right) = \dim\left(\theta\right)$. And thus, the optimal IV estimator is defined as method of moments estimator such that

$$\frac{1}{n}\sum_{i=1}^{n}a^*\left(x_i\right)h\left(w_i,\hat{\theta}_*\right) = 0$$

Proof. Pick any a, denote

$$\begin{aligned}h_i &:= h\left(w_i,\theta_0\right) \\ a_i &:= a\left(x_i\right) \\ H_i &:= \mathbb{E}\left[\left.\frac{\partial h\left(w_i,\theta_0\right)}{\partial\theta^{\mathrm{T}}}\right| x_i\right] \\ V_i &:= \mathbb{E}\left[h\left(w_i,\theta_0\right)^2\middle| x_i\right]\end{aligned}$$

Then

$$V_a^{-1} := G_a^{\mathrm{T}}\Omega_a^{-1}G_a = \mathbb{E}\left[a_i H_i\right]^{\mathrm{T}}\mathbb{E}\left[a_i a_i^{\mathrm{T}}V_i\right]^{-1}\mathbb{E}\left[a_i H_i\right]$$
$$V_*^{-1} := G_*^{\mathrm{T}}\Omega_*^{-1}G_* = \mathbb{E}\left[a^* H_i\right]^{\mathrm{T}}\mathbb{E}\left[a^*\left(a_i^*\right)^{\mathrm{T}}V_i\right]^{-1}\mathbb{E}\left[a_i^* H_i\right] = \mathbb{E}\left[H_i^{\mathrm{T}}V_i^{-1}H_i\right]$$

Denote $m_i := G_a^{\mathrm{T}}\Omega_a^{-1}a_i h_i$, and $m_i^* := a_i^* h_i$, then then

$$\mathbb{E}\left[m_i m_i^{\mathrm{T}}\right] = G_a^{\mathrm{T}}\Omega_a^{-1}G_a = V_a^{-1}$$
$$\mathbb{E}\left[m_i^*\left(m_i^*\right)^{\mathrm{T}}\right] = \mathbb{E}\left[a_i^* V_i^{-1}a_i^{*\mathrm{T}}\right] = V_*^{-1}$$
$$\mathbb{E}\left[m_i\left(m_i^*\right)^{\mathrm{T}}\right] = G_a^{\mathrm{T}}\Omega_a^{-1}\mathbb{E}\left[a_i V_i\left(a_i^*\right)^{\mathrm{T}}\right] = V_a^{-1}$$

then

$$\begin{aligned} V_a - V_* &= V_a V_a^{-1} V_a - V_* \\ &= \mathbb{E}\left[m_i\left(m_i^*\right)^{\mathrm{T}}\right]^{-1}\mathbb{E}\left[m_i m_i^{\mathrm{T}}\right]\mathbb{E}\left[m_i\left(m_i^*\right)^{\mathrm{T}}\right]^{-1} - \mathbb{E}\left[m_i^*\left(m_i^*\right)^{\mathrm{T}}\right] \\ &= \mathbb{E}\left[RR^{\mathrm{T}}\right] \end{aligned}$$

is positive semi-definite where

$$R = \mathbb{E}\left[m_i\left(m_i^*\right)^{\mathrm{T}}\right]^{-1}\left[m_i - \mathbb{E}\left[m_i\left(m_i^*\right)^{\mathrm{T}}\right]\mathbb{E}\left[m_i\left(m_i^*\right)^{\mathrm{T}}\right]^{-1}m_i^*\right]$$

□

Example 8.34. Best instrument for OLS.

For OLS,

$$h\left(w_i,\theta_0\right) = x_i\left(y_i - x_i^{\mathrm{T}}\theta_0\right)$$

The best instrument is

$$\begin{aligned} a^*\left(x_i\right) &= \mathbb{E}\left[\left.\frac{\partial h\left(w_i,\theta_0\right)}{\partial\theta}\right| x_i\right]\mathbb{E}\left[\left.h\left(w_i,\theta_0\right)h\left(w_i,\theta_0\right)^{\mathrm{T}}\right| x_i\right]^{-1} \\ &= \mathbb{E}\left[\left.x_i x_i^{\mathrm{T}}\right| x_i\right]\mathbb{E}\left[\left.x_i x_i^{\mathrm{T}}\left(y_i - x_i^{\mathrm{T}}\theta_0\right)^2\right| x_i\right]^{-1} \\ &= x_i x_i^{\mathrm{T}}\left(x_i x_i^{\mathrm{T}}\right)^{-1}\sigma^{-2} \\ &= I_k\sigma^{-2} \end{aligned}$$

does not depend on x_i. This means that OLS is efficient enough.

Chapter 9

Panel Data Models

9.1 Introduction

Remark 9.1. Why we study panel data?

- Panel data is a natural data structure of observations of individuals for multiple time periods.
- Advantages
  - Allowing for broader forms of heterogeneity (cross both individual and times).
  - Controlling for unobserved time-invariant without IV.
- Previous cross section model:

$$y_i = \alpha + x_i^{\mathrm{T}}\beta + u_i, i = 1, 2, \cdots n$$

  Change the indices in the subscript

$$y_{it} = \alpha + x_{it}^{\mathrm{T}}\beta + u_{it}, i = 1, 2, \cdots n, t = 1, 2, \cdots, T$$

  But there is one issue, we have two directions (time and individuals) but the constant α is the same. So, we may change α into two components in each direction, one is individual effect α_i, the other is trend γ_t:

$$y_{it} = \alpha_i + \gamma_t + x_{it}^{\mathrm{T}}\beta + u_{it}, i = 1, 2, \cdots n, t = 1, 2, \cdots, T$$

Remark 9.2. This chapter we focus on

$$y_{it} = \alpha_i + x_{it}^{\mathrm{T}}\beta + u_{it}, i = 1, 2, \cdots n, t = 1, 2, \cdots, T$$

There are two ways to treat α_i:

1. Random Effect (RE) Model: we can treat the individual as an **unobserved variable**, and combine it into error terms.

$$y_{it} = x_{it}^{\mathrm{T}}\beta + (u_{it} + \alpha_i)$$

2. Fixed Effect (FE) Model: we can treat the individual as an **unknown parameter**.

The estimation methods of FE and RE models are different, because the FE and RE models have different assumptions.

9.2 Random Effect Model

Remark 9.3. In the cross-section linear model,

$$y_i = \alpha + x_i^{\mathrm{T}}\beta + u_i, i = 1, 2, \cdots n$$

we assume that $\mathbb{E}[u_i x_i] = 0$ (or a stronger condition $\mathbb{E}[u_i | x_i] = 0$).

Likewise in RE model

$$y_{it} = x_{it}^{\mathrm{T}}\beta + (u_{it} + \alpha_i)$$

we may assume $\mathbb{E}[(u_{it} + \alpha_i) x_{it}] = 0$, but usually we impose

$$\mathbb{E}[u_{it} x_{it}] = 0$$

at first place, so we need subsequently to impose

$$\mathbb{E}[\alpha_i x_{it}] = 0$$

or a stronger condition $\mathbb{E}[\alpha_i | x_{it}] = 0$.

Remark 9.4. In practice, we can first conduct the Hausman test to check if $\mathbb{E}[\alpha_i x_{it}] = 0$ holds, i.e., if α_i is exogenous.

If $\mathbb{E}[\alpha_i x_{it}] \neq 0$, then the RE model fails, we need to develop a FE model.

Remark 9.5. Now suppose we have $\mathbb{E}[\alpha_i x_{it}] = 0$.

In principle, we can use the OLS to estimate it, since the model meet the assumptions by OLS.

However, we have a more efficient estimator (GLS estimator) because of the **heteroskedasticity**. In OLS, we impose $\mathbb{E}[u_i^2 | x_i] = \sigma_u^2$.

Remark 9.6. Let's recall the Generalized Least Square (GLS) estimator for cross-section model:

$$y_i = x_i^{\mathrm{T}}\beta_0 + e_i$$

Let $\boldsymbol{e} = (e_1, e_2, \cdots, e_n)^{\mathrm{T}}$, and $\boldsymbol{D} := \mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}} \middle| \boldsymbol{X}\right]$

$$\hat{\beta}_{\mathrm{OLS}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \Longrightarrow \hat{\beta}_{\mathrm{OLS}} - \beta_0 = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{e}$$

Then,

$$\begin{aligned}
\mathrm{Var}\left(\hat{\beta}_{\mathrm{OLS}} \middle| \boldsymbol{X}\right) &= \mathbb{E}\left[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \middle| \boldsymbol{X}\right] \\
&= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}} \middle| \boldsymbol{X}\right]\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \\
&= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}
\end{aligned}$$

- If e_i is homoskedastic, then $\boldsymbol{D} = \sigma^2 \boldsymbol{I}$, and

$$\operatorname{Var}\left(\hat{\beta}_{\mathrm{OLS}} \middle| \boldsymbol{X}\right) = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \sigma^2 = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \sigma^2$$

- If e_i is heteroskedastic, Aitken's theorem tells us

$$\hat{\beta}_{\mathrm{GLS}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{y}$$

  is the most efficient estimator, where $\boldsymbol{D} = \mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}} \middle| \boldsymbol{X}\right]$.

  Note that

$$\begin{aligned}
\operatorname{Var}\left(\hat{\beta}_{\mathrm{GLS}} \middle| \boldsymbol{X}\right) &= \operatorname{Var}\left(\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{e} \middle| \boldsymbol{X}\right) \\
&= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{D}\boldsymbol{D}^{-1}\boldsymbol{X} \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{X}\right)^{-1} \\
&= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{X} \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{X}\right)^{-1} \\
&= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{X}\right)^{-1}
\end{aligned}$$

Theorem 9.7. (Aitken's Theorem) For the cross-section model:

$$y_i = x_i^{\mathrm{T}}\beta_0 + e_i$$

Let

$$\tilde{\beta} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{y}$$

where $\boldsymbol{D} = \mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}} \middle| \boldsymbol{X}\right]$, then

$$\operatorname{Var}\left(\hat{\beta}_{\boldsymbol{A}} \middle| \boldsymbol{X}\right) \geqslant \operatorname{Var}\left(\tilde{\beta} \middle| \boldsymbol{X}\right)$$

where $\hat{\beta}_{\boldsymbol{A}}$ is any linear estimator obtained by multiply $\boldsymbol{y}$ with matrix $\boldsymbol{A}$, or $\hat{\beta}_{\boldsymbol{A}} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{y}$.

In other words, it's optimal to set

$$\boldsymbol{A}_*^{\mathrm{T}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}^{-1}$$

Proof. Pick and $\boldsymbol{A}$ such that $\hat{\beta}_{\boldsymbol{A}}$ is unbaised, or

$$\beta_0 = \mathbb{E}\left[\hat{\beta}_{\boldsymbol{A}} \middle| \boldsymbol{X}\right] = \mathbb{E}\left[\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} \middle| \boldsymbol{X}\right] = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}\beta_0$$

or

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{I}$$

Then

$$\operatorname{Var}\left(\hat{\beta}_{\boldsymbol{A}} \middle| \boldsymbol{X}\right) = \operatorname{Var}\left[\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} \middle| \boldsymbol{X}\right] = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{A}$$

Let

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{A} = \left(\boldsymbol{C} + \boldsymbol{A}_*\right)^{\mathrm{T}} \boldsymbol{D} \left(\boldsymbol{C} + \boldsymbol{A}_*\right) = \boldsymbol{C}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{C} + \boldsymbol{A}_*^{\mathrm{T}}\boldsymbol{D}\boldsymbol{A}_* \geqslant \boldsymbol{A}_*^{\mathrm{T}}\boldsymbol{D}\boldsymbol{A}_*$$

it follows that

$$\begin{aligned}\boldsymbol{A}_*^{\mathrm{T}}\boldsymbol{D}\boldsymbol{C} &= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{D}\left(\boldsymbol{A}-\boldsymbol{A}_*\right)\\ &= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}_*\right)\\ &= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{I}-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{X}\right)^{-1}\right)\\ &= 0\end{aligned}$$

we should note that $\boldsymbol{D}$ is positive semi-definite, so the inequality holds. □

Remark 9.8. For random effect model, with the heteroskedasticity, OLS is also unbiased, but GLS is more efficient.

However, we cannot observe the error term thus we cannot get the matrix $\boldsymbol{D}$ in advance. Therefore, we need to adopt the following two steps to address the problem:

1. Estimate $\boldsymbol{D}$ by OLS.
2. Run GLS.

The procedure is called Feasible Least Square (FLS).

Remark 9.9. Estimation of random effect model.

$$y_{it} = x_{it}^{\mathrm{T}}\beta + (\alpha_i + u_{it})$$

Note that $\mathrm{Var}\left(\boldsymbol{\alpha}+\boldsymbol{u}\right)$ is no longer a diagonal matrix. We need to use GLS (or FLS in practice) to estimate it.

We should first figure out $\boldsymbol{D} := \mathrm{Var}\left(\boldsymbol{\alpha}+\boldsymbol{u}\right)$. We further assume that

- $\mathrm{Var}\left(\alpha_i + u_{it}\right) = \sigma_\alpha^2 + \sigma_u^2$.
- Fix t, for $i \neq j$, $\mathrm{Cov}\left(\alpha_i + u_{it}, \alpha_j + u_{jt}\right) = 0$.
- Fix i, for $t \neq s$, $\mathrm{Cov}\left(\alpha_i + u_{it}, \alpha_i + u_{is}\right) = \sigma_\alpha^2$.

Therefore, $\boldsymbol{D}$ is

$$\begin{bmatrix}\boldsymbol{d}_1 & 0 & \cdots & 0\\ 0 & \boldsymbol{d}_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \boldsymbol{d}_n\end{bmatrix}_{nT\times nT}$$

where

$$\boldsymbol{d}_1 = \cdots = \boldsymbol{d}_n = \boldsymbol{d} = \begin{bmatrix}\sigma_\alpha^2+\sigma_u^2 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2\\ \sigma_\alpha^2 & \sigma_\alpha^2+\sigma_u^2 & \cdots & \sigma_\alpha^2\\ \vdots & \vdots & \ddots & \vdots\\ \sigma_\alpha^2 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2+\sigma_u^2\end{bmatrix}_{T\times T}$$

Then we can do GLS to get

$$\tilde{\beta} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}^{-1}\boldsymbol{y}$$

9.3 Fixed Effect Model

Remark 9.10. How to estimate a fixed effect model?

$$y_{it} = \alpha_i + x_{it}^{\mathrm{T}}\beta + u_{it}$$

We need to handle individual fixed effect, which cannot be treated as an error term.

We can include α_i into the model, by Dummy variables regression.

And we have two methods to get rid of α_i:

- Within estimator— demean.
- Difference estimator— take difference.

Remark 9.11. Dummy variables regression.

We write the fixed effect model in matrix form:

$$y_{it} = \alpha_i + x_{it}^{\mathrm{T}}\beta + u_{it} \Longleftrightarrow \boldsymbol{y} = \boldsymbol{D}\alpha + \boldsymbol{X}\beta + \boldsymbol{u}$$

where

- $\boldsymbol{y}_{nT\times 1}$ is a vector of dependent variables.
- $\boldsymbol{u}_{nT\times 1}$ is a vector of error terms.
- β is an k-dimensional unknown parameter.
- $\boldsymbol{D}_{nT\times n}$ is a vector of dummy variables.
- $\boldsymbol{X}_{nT\times k}$ is a vector of covariates.
- $\alpha_{n\times 1}$ is a vector of individual fixed effect

In particular,

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{e}_1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{e}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{e}_n \end{bmatrix}_{nT\times n}, \text{ where } \boldsymbol{e} := \boldsymbol{e}_1 = \boldsymbol{e}_2 = \cdots = \boldsymbol{e}_n = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}_{T\times 1}$$

Then

$$\boldsymbol{D} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \Longrightarrow \boldsymbol{D}_{nT\times n}\alpha_{n\times 1} = \begin{pmatrix} \alpha_1 \\ \alpha_1 \\ \vdots \\ \alpha_1 \\ \alpha_2 \\ \alpha_2 \\ \vdots \\ \alpha_2 \\ \vdots \\ \alpha_n \\ \alpha_n \\ \vdots \\ \alpha_n \end{pmatrix}_{nT\times 1}$$

And

$$\boldsymbol{X} = \begin{pmatrix} x_{11}^{\mathrm{T}} \\ x_{12}^{\mathrm{T}} \\ \vdots \\ x_{1T}^{\mathrm{T}} \\ x_{21}^{\mathrm{T}} \\ x_{22}^{\mathrm{T}} \\ \vdots \\ x_{2T}^{\mathrm{T}} \\ \vdots \\ x_{n1}^{\mathrm{T}} \\ x_{n2}^{\mathrm{T}} \\ \vdots \\ x_{nT}^{\mathrm{T}} \end{pmatrix}_{nT\times k} = \begin{pmatrix} x_{11}^{(1)} & x_{11}^{(2)} & \cdots & x_{11}^{(k)} \\ x_{12}^{(1)} & x_{12}^{(2)} & \cdots & x_{12}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1T}^{(1)} & x_{1T}^{(2)} & \cdots & x_{1T}^{(k)} \\ x_{21}^{(1)} & x_{21}^{(2)} & \cdots & x_{21}^{(k)} \\ x_{22}^{(1)} & x_{22}^{(2)} & \cdots & x_{22}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{2T}^{(1)} & x_{nT}^{(2)} & \cdots & x_{2T}^{(k)} \\ & & \vdots & \\ x_{n1}^{(1)} & x_{n1}^{(2)} & \cdots & x_{n1}^{(k)} \\ x_{n2}^{(1)} & x_{n2}^{(2)} & \cdots & x_{n2}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{nT}^{(1)} & x_{2T}^{(2)} & \cdots & x_{nT}^{(k)} \end{pmatrix}_{nT\times k}$$

One can check the notations.

Remark 9.12. We are only interested in β: Within estimator.

We can use Frisch–Waugh–Lovell theorem to solve β. This theorem consider a model:

$$\boldsymbol{Y} = \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \boldsymbol{u}$$

Define

$$\boldsymbol{M}_{\boldsymbol{X}_1} = \boldsymbol{I} - \boldsymbol{X}_1\left(\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1^{\mathrm{T}}$$

then

$$\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{Y} = \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2\beta_2 + \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{u}$$

Then, the OLS estimator of β_2 in $\boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{Y} = \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{X}_2\beta_2 + \boldsymbol{M}_{\boldsymbol{X}_1}\boldsymbol{u}$ will numerically equivalent to the OLS estimator of β_2 in $\boldsymbol{Y} = \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \boldsymbol{u}$.

For our cases, $\beta_1 = \alpha$, $\beta_2 = \beta$.

Moreover, we should note that

$$\begin{aligned}
\boldsymbol{M_D} &= \boldsymbol{I}_{nT\times nT} - \boldsymbol{D}\left(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}\right)^{-1}\boldsymbol{D}^{\mathrm{T}} \\
&= \boldsymbol{I}_{nT\times nT} - \begin{bmatrix} \boldsymbol{e} & 0 & \cdots & 0 \\ 0 & \boldsymbol{e} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{e} \end{bmatrix}_{nT\times n} \begin{bmatrix} T & 0 & \cdots & 0 \\ 0 & T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & T \end{bmatrix}_{n\times n}^{-1} \begin{bmatrix} \boldsymbol{e}^{\mathrm{T}} & 0 & \cdots & 0 \\ 0 & \boldsymbol{e}^{\mathrm{T}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{e}^{\mathrm{T}} \end{bmatrix}_{n\times nT} \\
&= \boldsymbol{I}_{nT\times nT} - \frac{1}{T}\begin{bmatrix} \boldsymbol{E}_1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{E}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{E}_n \end{bmatrix}_{nT\times nT}
\end{aligned}$$

where

$$\boldsymbol{E} := \boldsymbol{E}_1 = \cdots = \boldsymbol{E}_n = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}_{T\times T}$$

Applying Frisch–Waugh–Lovell, the dummy variables regression gives the **within estimator**.

It is equivalent to regressing $y_{it} - \bar{y}_i$ on $x_{it} - \bar{x}_i$, where $\bar{y}_i$ and $\bar{x}_i$ are the within-group averages of the individual i over $t = 1, 2, \cdots, T$.

$$\begin{aligned}
y_{it} &= \alpha_i + x_{it}^{\mathrm{T}}\beta + u_{it} \\
\bar{y}_i &= \alpha_i + \bar{x}_i^{\mathrm{T}}\beta + \bar{u}_i \\
y_{it} - \bar{y}_i &= 0 + \left(x_{it} - \bar{x}_i\right)^{\mathrm{T}}\beta + u_{it} - \bar{u}_i
\end{aligned}$$

Remark 9.13. Difference estimator.

For each individual, we take the first difference:

$$\begin{aligned}
\Delta y_{it} &= y_{it} - y_{i,t-1} \\
\Delta x_{it} &= x_{it} - x_{i,t-1} \\
\Delta u_{it} &= u_{it} - u_{i,t-1}
\end{aligned}$$

Then

$$\Delta y_{it} = \Delta x_{it}^{\mathrm{T}}\beta + \Delta u_{it}$$

Again, we get rid of the fixed effect individual effect. But we should assume that

$$\mathbb{E}\left[\Delta x_{it}^{\mathrm{T}}\Delta u_{it}\right] = 0, t = 2, 3, \cdots, T$$

9.4 Dynamic Panel

Remark 9.14. An example of dynamic panel data model.

$$y_{it} = y_{i,t-1}\gamma + x_{it}^{\mathrm{T}}\beta + \alpha_i + u_{it}, i = 1, 2, \cdots, n; t = 1, 2, \cdots, T$$

Even if $\mathbb{E}\left[x_{it}^{\mathrm{T}} u_i\right] = 0$ and $\mathbb{E}\left[x_{it}^{\mathrm{T}}\alpha_i\right] = 0$, we cannot estimate it with OLS consistently. Because

$$y_{i,t-1} = y_{i,t-2}\gamma + x_{i,t-1}^{\mathrm{T}}\beta + \alpha_i + u_{i,t-1}$$

then

$$\mathbb{E}\left[y_{i,t-1}^{\mathrm{T}}\alpha_i\right] \neq 0$$

Generally,

$$\begin{aligned} y_{it} &= y_{i,t-1}\gamma + x_{it}^{\mathrm{T}}\beta + \alpha_i + u_{it} \\ &= \gamma\left(y_{i,t-1}\gamma + x_{i,t-1}^{\mathrm{T}}\beta + \alpha_i + u_{i,t-1}\right) + x_{it}^{\mathrm{T}}\beta + \alpha_i + u_{it} \\ &= \gamma^n y_{i,t-n} + \sum_{\ell=0}^{n-1}\gamma^\ell x_{i,t-\ell}^{\mathrm{T}}\beta + \frac{1-\gamma^n}{1-\gamma}\alpha_i + \sum_{\ell=0}^{n-1}\gamma^\ell u_{i,t-\ell} \end{aligned}$$

- Solution 1 (Anderson-Hsiao Estimator): First differencing + IV estimator.
- Solution 2 (Arellano-Bond): First differencing + GMM estimation.

Chapter 10

Difference in Differences and Causal Inference

10.1 Introduction

Remark 10.1. Notations.

- Y: outcome.
  - $Y(0)$: Outcome without treatment.
  - $Y(1)$: Outcome with treatment.
- D: Treatment status (or policy choice).
- X: Other covariates (independent variables).

Example 10.2. Suppose A is treated by some policy, but B is not treated. There are two period 1 and 2. Thus, $Y_{A,2}(1)$ is observable, but $Y_{B,2}(1)$ is unobservable.

Table 10.1: An example for DID.

	A	B	Difference
Before Treatment	$Y_{A,1}(0)$	$Y_{B,1}(0)$	$Y_{B,1}(0) - Y_{A,1}(0)$
After Treatment	$Y_{A,2}(1)$	$Y_{B,2}(1)$	$Y_{B,2}(1) - Y_{A,2}(1)$
Difference	$Y_{A,2}(1) - Y_{A,1}(0)$	$Y_{B,2}(1) - Y_{B,1}(0)$	$[Y_{A,2}(1) - Y_{A,1}(0)] - [Y_{B,2}(1) - Y_{B,1}(0)]$

- The *difference estimator* $Y_{A,2}(1) - Y_{A,1}(0)$ is not a valid estimator of treatment effect.
- The *difference in differences estimator* $[Y_{A,2}(1) - Y_{A,1}(0)] - [Y_{B,2}(1) - Y_{B,1}(0)]$ should be the valid estimator.

10.2 Treatment Effect Framework

Example 10.3. Suppose that the average income of college graduates are 5000 EUR, and high school graduates is 4000 EUR. Now a government has to decide whether to invest more to allow universities to recruit more students, the cost (e.g., new teachers, classrooms, facilities) is about 700 EUR (monthly) If the treatment effect is $5000 - 4000 = 1000 > 700$, it looks a good deal.

$$\begin{array}{ccc} Y\,(\text{Income}) & \leftarrow & D\,(\text{College}) \\ & \nwarrow & \uparrow \\ & & X\,(\text{IQ}, \cdots) \end{array}$$

Suppose the true model:

$$y_i = \beta_0 D_i + \gamma_0 x_i + \varepsilon_i$$

If we misspecified:

$$y_i = \beta D_i + \varepsilon_i$$

then

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{i=1}^n D_i y_i}{\sum_{i=1}^n D_i^2} = \frac{\sum_{i=1}^n D_i \left(\beta_0 D_i + \gamma_0 x_i + \varepsilon_i\right)}{\sum_{i=1}^n D_i^2} \\ &= \beta_0 + \gamma_0 \frac{\frac{1}{n}\sum_{i=1}^n D_i x_i}{\frac{1}{n}\sum_{i=1}^n D_i^2} + \frac{\frac{1}{n}\sum_{i=1}^n D_i \varepsilon_i}{\frac{1}{n}\sum_{i=1}^n D_i^2} \\ &\xrightarrow{\mathbb{P}} \beta_0 + \gamma_0 \frac{\mathbb{E}\left[D_i x_i\right]}{\mathbb{E}\left[D_i^2\right]} + \frac{\mathbb{E}\left[D_i \varepsilon_i\right]}{\mathbb{E}\left[D_i^2\right]} = \beta_0 + \gamma_0 \frac{\mathbb{E}\left[D_i x_i\right]}{\mathbb{E}\left[D_i^2\right]} \end{aligned}$$

where $\mathbb{E}\left[D_i \varepsilon_i\right] = 0, \mathbb{E}\left[D_i^2\right] > 0$.

In two cases, the bias term $\gamma_0 \frac{\mathbb{E}[D_i x_i]}{\mathbb{E}\left[D_i^2\right]}$ is zero:

- $\gamma_0 = 0$.
- $\mathbb{E}\left[D_i x_i\right] = 0$, if x_i is demeaned, then this is equivalent with $\operatorname{Cov}\left(D_i, x_i\right) = 0$.

Address the issue:

- In **randomized experiment**
  - The policy maker can ensure that the treatment is independently and randomly assigned.
  - $D_i \perp x_i \Longrightarrow \operatorname{Cov}\left(D_i, x_i\right) = 0$.
- In an observational study (usual case)
  - The treatment choice is a consequence of a choice of individuals
  - A regression suffers from omitting variable biasedness (OVB) if these relevant characteristics are not controlled
  - We apply DID by assuming $D_i \perp y_i | X$

Remark 10.4. Treatment effect model.

- DID is an example of (linear) treatment effect model.
- Treatment effect model is one of the most important topics in economics (e.g., decisions of policy maker, program evaluation).

Recall that Y_i is the outcome.

- $Y_i(0)$: Outcome without treatment.
- $Y_i(1)$: Outcome with treatment.

Then

$$Y_i = DY_i(1) + (1 - D) Y_i(0)$$

For a fixed individual, we can only observe one of two counterfactual outcomes, we can observe either $Y(1)$ or $Y(0)$, but not both.

Example 10.5. Suppose we observe $N = 3$ individuals. The red ones cannot be observed, so

Table 10.2: Partially revealed example.

i	Before treatments are assigned Outcomes Potential outcomes	Treatments are assigned D_i
$i = 1$	$Y_1(1), Y_1(0)$	$D_i = 1$
$i = 2$	$Y_2(1), Y_2(0)$	$D_i = 2$
$i = 3$	$Y_3(1), Y_3(0)$	$D_i = 3$

the data is *partially revealed*.

Definition 10.6. The main object of interests are:

- Treatment effect of individual i :

$$\text{TE} = Y_i(1) - Y_i(0)$$

- Average treatment effect (ATE):

$$\text{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)]$$

- Average treatment effect for the treated (ATT):

$$\text{ATT} = \mathbb{E}[Y_i(1) - Y_i(0) | D_i = 1]$$

Remark 10.7. Some key points about ATE.

- ATE $= \mathbb{E}[Y_i(1) - Y_i(0)]$ is the expected difference of outcomes, if we only change the treatment from 0 to 1.
  - "Only" means we keep everything else constant.
  - "everything else" means the distribution of other observable and unobservable covariates
- The key inequality:

$$\mathbb{E}[Y(1)] \neq \mathbb{E}[Y|D=1]$$

  Because
  - $Y(1)$ is unobservable, we can estimate $\mathbb{E}[Y(1)]$ by

$$\frac{1}{N}\sum_{i=1}^{N} Y_i(1)$$

    which is infeasible to compute.
  - $Y|D=1$ is observable, we can estimate $\mathbb{E}[Y|D=1]$ by

$$\frac{1}{\#\{i|D_i=1\}}\sum_{i\in\{i|D_i=1\}} Y_i$$

Back to the Example 10.5, we assume each individual has equal probability. Then

Table 10.3: Partially revealed example.

i	Probability	Before treatments are assigned Outcomes Potential outcomes	Treatments are assigned D_i
$i=1$	$\frac{1}{3}$	$Y_1(1), Y_1(0)$	$D_i=1$
$i=2$	$\frac{1}{3}$	$Y_2(1), Y_2(0)$	$D_i=2$
$i=3$	$\frac{1}{3}$	$Y_3(1), Y_3(0)$	$D_i=3$

$$\mathbb{E}[Y(1)] = \frac{1}{3}[Y_1(1) + Y_2(1) + Y_3(1)] \neq \mathbb{E}[Y|D=1] = Y_1(1)$$

- We want to estimate

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)]$$

  - But we don't have a direct sample analogue of $\mathbb{E}[Y(1)]$ and $\mathbb{E}[Y(0)]$.
  - We do have a direct sample analogue of $\mathbb{E}[Y|D=1]$ and $\mathbb{E}[Y|D=0]$.
  - But $\mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0] \neq \mathbb{E}[Y(1) - Y(0)]$.

Back to the example,

$$\begin{aligned}\mathbb{E}[Y|D=1]-\mathbb{E}[Y|D=0] &= Y_1(1)-\frac{1}{2}[Y_2(0)+Y_3(0)]\\ &\neq \mathbb{E}[Y(1)-Y(0)]\\ &= \frac{1}{3}\sum_{i=1}^{3}Y_i(1)-\frac{1}{3}\sum_{i=1}^{3}Y_i(0)\end{aligned}$$

Remark 10.8. How can we find a sample analogue of ATE?

We impose the key assumption, which is known as *unconfoundedness* or *conditional exogeneity*:

$$\{Y(1),Y(0)\}\perp D|X$$

- Conditional on X, the counterfactual outcomes is independent of the treatment variables.
- With the unconfoundedness, by Law of iterated expectation, we have immediately

$$\begin{aligned}\text{ATE} &= \mathbb{E}[Y(1)-Y(0)]\\ &= \mathbb{E}[\mathbb{E}[Y(1)|X]]-\mathbb{E}[\mathbb{E}[Y(0)|X]]\\ &= \mathbb{E}[\mathbb{E}[Y(1)|X,D=1]]-\mathbb{E}[\mathbb{E}[Y(0)|X,D=0]]\end{aligned}$$

We should pay attention to the last line, the hidden condition is $\mathbb{E}[\mathbb{E}[Y(1)|X,D=1]]$ and $\mathbb{E}[\mathbb{E}[Y(0)|X,D=0]]$ are feasible, or

$$\mathbb{P}(D=1|X=x)\in(0,1),\forall x$$

To be succinct, define

$$P(x):=\mathbb{P}(D=1|X=x)$$

then we need to assume that

$$0<P(x)<1,\forall x$$

we call this assumption *overlap* condition.

Example 10.9. Treatment effect in linear models.

Under unconfoundedness, we have

$$\mathbb{E}[\mathbb{E}[Y(d)|X]]=\mathbb{E}[Y(d)],d=0,1$$

For the linear model

$$y_i=D\alpha+x_i^{\mathrm{T}}\beta+\varepsilon_i,\mathbb{E}(\varepsilon_i)=0$$

We compute the treatment effect:

$$\begin{aligned}\mathbb{E}[Y(1)] &= \mathbb{E}[\mathbb{E}[Y(1)|X]]=\mathbb{E}[\alpha+X\beta+\mathbb{E}[\varepsilon|X]]=\alpha+\mathbb{E}[X]\beta\\ \mathbb{E}[Y(0)] &= \mathbb{E}[\mathbb{E}[Y(0)|X]]=\mathbb{E}[X\beta+\mathbb{E}[\varepsilon|X]]=\mathbb{E}[X]\beta\\ \text{ATE} &= \mathbb{E}[Y(1)]-\mathbb{E}[Y(0)]=\alpha\end{aligned}$$

We also note that

$$\begin{aligned}\mathbb{E}[Y|D=1] &= \alpha + \mathbb{E}[X\beta+\varepsilon|D=1]\\ \mathbb{E}[Y|D=0] &= \mathbb{E}[X\beta+\varepsilon|D=0]\\ \mathbb{E}[Y|D=1]-\mathbb{E}[Y|D=0] &\neq \text{ATE}\end{aligned}$$

Moreover, we note different types of treatment effects:

- Treatment effect: $\text{TE} := Y(1) - Y(0) = \alpha$.
- Average treatment effect: $\text{ATE} := \mathbb{E}[Y(1) - Y(0)] = \alpha$.
- Conditional treatment effect: $\text{CATE} := \mathbb{E}[Y(1) - Y(0)|X] = \alpha$.

10.3 Randomized Experiments

Remark 10.10. Assume that the treatment status is independent of the potential outcomes, that is:

$$\{Y(1), Y(0)\} \perp D$$

then

$$\begin{aligned}\text{ATE} &= \mathbb{E}[Y(1) - Y(0)]\\ &= \mathbb{E}[Y(1)|D=1] - \mathbb{E}[Y(0)|D=0]\\ &= \mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0]\end{aligned}$$

Then we can estimate ATE by

$$\widehat{\text{ATE}} = \frac{1}{n_1}\sum_i Y_i D_i - \frac{1}{n_0}\sum_i Y_i(1-D_i)$$

where $n_1 = \sum_i D_i$ and $n_0 = \sum_i (1-D_i)$. Alternatively, we have

$$\begin{aligned}Y_i &= Y_i(1)\cdot D_i + Y_i(0)(1-D_i)\\ &= D_i\underbrace{(Y_i(1) - Y_i(0))}_{\tau_i} + Y_i(0)\\ &= \tau_i \cdot D_i + Y_i(0)\\ &= \underbrace{\mathbb{E}[Y_i(0)]}_{\alpha} + \tau\cdot D_i + \underbrace{(\tau_i - \tau)D_i + Y_i(0) - \mathbb{E}[Y_i(0)]}_{\varepsilon_i}\\ &= \alpha + \tau\cdot D_i + \varepsilon_i\end{aligned}$$

where we denote τ as ATE, also note that τ_i is a random variable. Then by independence, we can see that

$$\mathbb{E}[\varepsilon_i] = \mathbb{E}[(\tau_i - \tau)D_i + Y_i(0) - \mathbb{E}[Y_i(0)]] = 0$$

and

$$\mathbb{E}\left[\varepsilon_i D_i\right] = 0$$

Thus, by the identification result of OLS, we have

$$\tau = \frac{\operatorname{Cov}\left(Y_i, D_i\right)}{\operatorname{Var}\left(D_i\right)}$$

and the estimator of the ATE is given by:

$$\hat{\tau} = \frac{\widehat{\operatorname{Cov}}\left(Y_i, D_i\right)}{\widehat{\operatorname{Var}}\left(D_i\right)}$$

10.4 Estimation of ATE

10.4.1 Regression

Remark 10.11. For linear regression

$$Y_i = \beta D_i + \gamma^{\mathrm{T}} Y_i + \varepsilon_i$$

- $\mathbb{E}\left[Y_i \middle| D_i, Y_i\right] = \beta D_i + \gamma Y_i$.
- The key assumption for consistency is $\mathbb{E}\left[Y_i \varepsilon_i\right] = 0$, $\mathbb{E}\left[D_i \varepsilon_i\right] = 0$, or a stronger condition $\mathbb{E}\left[\varepsilon_i \middle| D_i, Y_i\right] = 0$.
- Certainly, we need other necessary Gauss-Markov assumptions.

For nonlinear regression

$$Y_i = m\left(D_i, Y_i, \beta_0\right) + \varepsilon_i$$

- $\mathbb{E}\left[Y_i \middle| D_i, Y_i\right] = m\left(D_i, Y_i, \beta_0\right)$ is the the best predictor of Y_i from Y_i and D_i.
- The key assumption for consistency is $\mathbb{E}\left[\varepsilon_i \middle| D_i, Y_i\right] = 0$.

Therefore, generally,

$$\begin{aligned}
\text{CATE} &= \mathbb{E}\left[Y(1) - Y(0) \middle| X_i = x\right] = m\left(1, x, \beta_0\right) - m\left(0, x, \beta_0\right) \\
\text{ATE} &= \mathbb{E}\left[\mathbb{E}\left[Y(1) - Y(0) \middle| X_i = x\right]\right] = \int \left[m\left(1, x, \beta_0\right) - m\left(0, x, \beta_0\right)\right] f_X(x)\, \mathrm{d}x
\end{aligned}$$

For linear model, we don't need to integrate CATE, since CATE = ATE = β in linear model.

Remark 10.12. For linear model,

$$\left\{Y(1), Y(0)\right\} \perp D \middle| X = x, \mathbb{E}\left[\varepsilon \middle| X\right] = 0 \Longrightarrow \mathbb{E}\left[X\varepsilon\right] = 0, \mathbb{E}\left[D\varepsilon\right] = 0$$

Proof. Firstly, $\mathbb{E}[X\varepsilon] = \mathbb{E}[X\mathbb{E}[\varepsilon|X]] = 0$ is self-evident.

Secondly, by linearity,

$$\{Y(1), Y(0)\} \perp D|X = x \Longleftrightarrow \left\{\beta \cdot 1 + \gamma^{\mathrm{T}} X + \varepsilon, \gamma^{\mathrm{T}} X + \varepsilon\right\} \perp D\Big|X = x \Longleftrightarrow \varepsilon \perp D|X$$

then

$$\mathbb{E}[D\varepsilon] = \mathbb{E}[\mathbb{E}[\varepsilon D|X]|X] = \mathbb{E}[\mathbb{E}[\varepsilon|X]\mathbb{E}[D|X]|X] = 0$$

□

Remark 10.13. For regression model, we don't need the overlap assumption $0 < P(x) < 1$.

10.4.2 Matching

Remark 10.14. Matching estimator.

- For every observed $Y_i(1)$, we construct (or estimate) its counterfactual outcome $\hat{Y}_i(0)$, to create the difference $Y_i(1) - \hat{Y}_i(0)$
- Similarly, for ever observed controlled outcome $Y_i(0)$, we construct (or estimate) its counterfactual outcome $\hat{Y}_i(1)$, to create a difference $\hat{Y}_i(1) - Y_i(0)$.
- In the end, we construct a sample analogue

$$\frac{1}{n}\sum_{i=1}^{n}\left[\tilde{Y}_i(1) - \tilde{Y}_i(0)\right]$$

where $\tilde{Y}_i(d) = \begin{cases} Y_i(d), & D_i = d \\ \hat{Y}_i(d), & D_i = 1 - d \end{cases}$.

Let $\mathcal{J}(i)$ be the matched set of observation i for $i = 1, 2, \cdots, n$, and $M_i = |\mathcal{J}(i)|$.
Then one intuitive way to construct $\tilde{Y}_i$ is

$$\tilde{Y}_i(d) = \begin{cases} Y_i(d), & D_i = d \\ \frac{1}{M_i}\sum_{j\in\mathcal{J}(i)} Y_i(d), & D_i = 1 - d \end{cases}$$

For matching estimator, the key is to find a suitable $\mathcal{J}(i)$, three possible ways are

1. Exact matching: When X is discrete and takes finite many values (e.g., binary), we can directly compare treated and untreated units having the same X values.

$$\mathcal{J}(i) = \{j = 1, \cdots, n : D_j = 1 - D_i \text{ and } X_j = X_i\}$$

2. M-nearest neighbor Matching.
   - If X is continuous variable, exact matching is impossible. We can match the treated (control) unit with control (treated) unit with closest M samples.

- The set of matches is determined by

$$\mathcal{J}_M(i) = \left\{ j = 1, \cdots, n : D_j = 1 - D_i \text{ and } \sum_{k:D_k=1-D_i} \mathbb{I}_{\{d(X_i,X_k) \leqslant d(X_i,X_j)\}} \leqslant M \right\}$$

where d is the distance measurement, for example, we can choose

  - Euclidean metric

$$d(X_i, X_j) = \|X_i - X_j\| = \sqrt{(X_i - X_j)^{\mathrm{T}} (X_i - X_j)}$$

  - Mahalanobis distance

$$d(X_i, X_j) = \sqrt{(X_i - X_j)^{\mathrm{T}} V^{-1} (X_i - X_j)}$$

where V is the (estimated) variance matrix of X.

3. Propensity score matching (later).

Remark 10.15. Matching estimator require unconfoundedness and overlap assumption.

10.4.3 Weighting and Propensity Score Matching

Example 10.16. Suppose $n = 10$, for the triple (Y, D, X), where $X = 1$ for all samples, and $D \in \{0, 1\}$ is i.i.d. assigned. We further suppose we have the sample:

Y	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}
D	1	0	0	1	1	0	0	1	0	0

Then we should estimate

$$\begin{aligned}
\widehat{\mathrm{ATE}} &= \hat{\mathbb{E}}[Y(1) - Y(0)] = \frac{1}{4}\sum_{D_i=1} Y_i - \frac{1}{6}\sum_{D_i=0} Y_i \\
&= \frac{1}{10}\sum_{D_i=1} \frac{Y_i}{4/10} - \frac{1}{10}\sum_{D_i=0} \frac{Y_i}{6/10} = \frac{1}{n}\sum_{D_i=1} \frac{Y_i}{\hat{\mathbb{P}}\{D_i = 1\}} - \frac{1}{n}\sum_{D_i=0} \frac{Y_i}{\hat{\mathbb{P}}\{D_i = 0\}} \\
&= \frac{1}{n}\sum_{i=1}^{n} \left[\frac{Y_i D_i}{\hat{\mathbb{P}}\{D_i = 1\}} - \frac{Y_i (1 - D_i)}{\hat{\mathbb{P}}\{D_i = 0\}} \right]
\end{aligned}$$

Example 10.17. Suppose $n = 20$, $X \in \{1, 2\}$, $D \in \{0, 1\}$ is i.i.d randomly assigned, but the assignment probability might differ for different X, i.e., D is randomly assigned conditional on each X.

We can calculate CATE as in the previous example, then integrate it to get ATE. For simplicity, suppose $\mathbb{P}\{X = 1\} = \mathbb{P}\{X = 2\} = \frac{1}{2}$, and we have samples:

- $X = 1$:
- $X = 2$:

Y	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}
D	1	0	0	1	1	0	0	1	0	0

Y	Y_{11}	Y_{12}	Y_{13}	Y_{14}	Y_{15}	Y_{16}	Y_{17}	Y_{18}	Y_{19}	Y_{10}
D	0	0	1	0	1	0	0	0	1	0

Then

$$\begin{aligned}\widehat{\mathrm{CATE}}\left(X=1\right)&=\hat{\mathbb{E}}\left[Y\left(1\right)-Y\left(0\right)\middle|X=1\right]=\frac{1}{4}\sum_{D_i=1,X_i=1}Y_i-\frac{1}{6}\sum_{D_i=0,X_i=1}Y_i\\&=\frac{1}{10}\sum_{i=1}^{10}\left(\frac{Y_iD_i}{\hat{\mathbb{P}}\left\{D_i=1\middle|X=1\right\}}-\frac{Y_i\left(1-D_i\right)}{\hat{\mathbb{P}}\left\{D_i=0\middle|X=1\right\}}\right)\\\widehat{\mathrm{CATE}}\left(X=2\right)&=\hat{\mathbb{E}}\left[Y\left(1\right)-Y\left(0\right)\middle|X=2\right]=\frac{1}{3}\sum_{D_i=1,X_i=2}Y_i-\frac{1}{7}\sum_{D_i=0,X_i=2}Y_i\\&=\frac{1}{10}\sum_{i=11}^{20}\left(\frac{Y_iD_i}{\hat{\mathbb{P}}\left\{D_i=1\middle|X=2\right\}}-\frac{Y_i\left(1-D_i\right)}{\hat{\mathbb{P}}\left\{D_i=0\middle|X=2\right\}}\right)\end{aligned}$$

We can estimate ATE by

$$\begin{aligned}\widehat{\mathrm{ATE}}&=\mathbb{E}\left[\widehat{\mathrm{CATE}}\left(X\right)\right]=\int\widehat{\mathrm{CATE}}\left(X\right)f_X\left(x\right)\mathrm{d}x\\&=\frac{1}{2}\widehat{\mathrm{CATE}}\left(X=1\right)+\frac{1}{2}\widehat{\mathrm{CATE}}\left(X=2\right)\\&=\frac{1}{20}\sum_{i=1}^{20}\left(\frac{Y_iD_i}{\hat{\mathbb{P}}\left\{D_i=1\middle|X_i\right\}}-\frac{Y_i\left(1-D_i\right)}{\hat{\mathbb{P}}\left\{D_i=0\middle|X_i\right\}}\right)\\&=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_iD_i}{\hat{\mathbb{P}}\left\{D_i=1\middle|X_i\right\}}-\frac{Y_i\left(1-D_i\right)}{1-\hat{\mathbb{P}}\left\{D_i=1\middle|X_i\right\}}\right)\end{aligned}$$

it holds only if we have a well-balanced sample, in this example, $\#\{X_i:X_i=1\}=\#\{X_i:X_i=2\}$.

Recall the key definition: Propensity score $P\left(x\right)=\mathbb{P}\left\{D=1\middle|X=x\right\}$. Then

$$\widehat{\mathrm{ATE}}=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_iD_i}{\hat{\mathbb{P}}\left\{D_i=1\middle|X_i\right\}}-\frac{Y_i\left(1-D_i\right)}{1-\hat{\mathbb{P}}\left\{D_i=1\middle|X_i\right\}}\right)=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_iD_i}{\hat{P}\left(X_i\right)}-\frac{Y_i\left(1-D_i\right)}{1-\hat{P}\left(X_i\right)}\right)$$

This is also called as Inverse probability weighting estimator (IPW).

At population level we have

$$\mathrm{ATE}:=\mathbb{E}\left[Y(1)-Y(0)\right]=\mathbb{E}\left[\frac{YD}{P\left(X\right)}-\frac{Y\left(1-D\right)}{1-P\left(X\right)}\right]$$

Proof. Note that $Y=Y\left(1\right)D+Y\left(0\right)\left(1-D\right)$ then

$$\begin{aligned}YD&=Y\left(1\right)D^2+Y\left(0\right)\left(1-D\right)D=Y\left(1\right)D\\Y\left(1-D\right)&=Y\left(1\right)\left(1-D\right)D+Y\left(0\right)\left(1-D\right)^2=Y\left(0\right)\left(1-D\right)\end{aligned}$$

Then

$$\begin{aligned}
&\mathbb{E}\left[\frac{YD}{P(X)}-\frac{Y(1-D)}{1-P(X)}\right]\\
=&\mathbb{E}\left[\mathbb{E}\left(\frac{YD}{P(X)}-\frac{Y(1-D)}{1-P(X)}\middle| X\right)\right]\\
=&\mathbb{E}\left[\frac{\mathbb{E}(YD|X)}{P(X)}-\frac{\mathbb{E}(Y(1-D)|X)}{1-P(X)}\right]\\
=&\mathbb{E}\left[\frac{\mathbb{E}(Y(1)D|X)}{P(X)}-\frac{\mathbb{E}(Y(0)(1-D)|X)}{1-P(X)}\right]\\
=&\mathbb{E}\left[\frac{\mathbb{E}(Y(1)|X)\mathbb{E}(D|X)}{P(X)}-\frac{\mathbb{E}(Y(0)|X)[1-\mathbb{E}(D|X)]}{1-P(X)}\right]\\
=&\mathbb{E}\left[\frac{\mathbb{E}(Y(1)|X)\mathbb{E}\left(\mathbb{I}_{\{D_i=1\}}\middle| X\right)}{P(X)}-\frac{\mathbb{E}(Y(0)|X)\left[1-\mathbb{E}\left(\mathbb{I}_{\{D_i=1\}}\middle| X\right)\right]}{1-P(X)}\right]\\
=&\mathbb{E}[\mathbb{E}(Y(1)|X)-\mathbb{E}(Y(0)|X)]\\
=&\mathbb{E}[Y(1)-Y(0)]
\end{aligned}$$

□

Remark 10.18. Benefit of weighting estimator.

- It is a non-parametric estimator.
- We do not assume any dependence structure among Y, X, and D.
  - No parametric linear or non-linear structures are imposed
  - However, we might need to specify a model of the propensity score or estimate it nonparametrically.
- We need unconfoundedness and overlap assumptions (Similar to matching estimator, we need some other mild regularity conditions[1]
- Compared to matching estimator
  - In general, weighting estimator is more efficient in the sense of having a smaller MSE
  - As a nonparametric estimator, under mild conditions, it can achieve $\sqrt{n}$ −convergence rate (like parametric model)

Theorem 10.19. (Rosenbaum and Rubin, 1983)

$$\{Y(1),Y(0)\}\perp D|X\Longrightarrow\{Y(1),Y(0)\}\perp D|P(X)$$

[1]For example,

- $\mathbb{E}[Y(d)]<\infty$ for $d=0,1$.
- $\mathbb{E}[Y(d)|X=x]$ is continuous in x.

Proof. Firstly, we show that

$$\mathbb{P}\left\{D=1|\,Y\left(d\right),P\left(X\right)\right\}=P\left(X\right):=\mathbb{P}\left\{D=1|\,X\right\}=\mathbb{E}\left[D|\,X\right]$$

by

$$\begin{aligned}\mathbb{P}\left\{D=1|\,Y\left(d\right),P\left(X\right)\right\}&=\mathbb{E}\left[D|\,Y\left(d\right),P\left(X\right)\right]\\&=\mathbb{E}\left[\mathbb{E}\left[D|\,Y\left(d\right),P\left(X\right),X\right]|\,Y\left(d\right),P\left(X\right)\right]\\&=\mathbb{E}\left[\mathbb{E}\left[D|\,Y\left(d\right),X\right]|\,Y\left(d\right),P\left(X\right)\right]\\&=\mathbb{E}\left[\mathbb{E}\left[D|\,X\right]|\,Y\left(d\right),P\left(X\right)\right]\\&=\mathbb{E}\left[P\left(X\right)|\,Y\left(d\right),P\left(X\right)\right]\\&=P\left(X\right)\end{aligned}$$

Second, we show $\mathbb{P}\left\{D=1|\,P\left(X\right)\right\}=P\left(X\right)$ by

$$\begin{aligned}\mathbb{P}\left\{D=1|\,P\left(X\right)\right\}&=\mathbb{E}\left[D|\,P\left(X\right)\right]\\&=\mathbb{E}\left[\mathbb{E}\left[D|\,P\left(X\right),X\right]|\,P\left(X\right)\right]\\&=\mathbb{E}\left[\mathbb{E}\left[D|\,X\right]|\,P\left(X\right)\right]\\&=\mathbb{E}\left[P\left(X\right)|\,P\left(X\right)\right]\\&=P\left(X\right)\end{aligned}$$

Then

$$\mathbb{P}\left\{D=1|\,Y\left(d\right),P\left(X\right)\right\}=\mathbb{P}\left\{D=1|\,P\left(X\right)\right\}$$

Therefore,

$$\mathbb{E}\left[D|\,Y\left(d\right),P\left(X\right)\right]=\mathbb{E}\left[D|\,P\left(X\right)\right]\Longrightarrow Y\left(d\right)\perp D|\,P\left(X\right)$$

□

Remark 10.20. Matching on propensity score (also a non-parametric method).

$$\mathcal{J}_M\left(i\right)=\left\{j=1,\cdots,n:D_j=1-D_i\text{ and }\sum_{k:D_k=1-D_i}\mathbb{I}_{\{|P(X_i)-P(X_k)|\leqslant|P(X_i)-P(X_j)|\}}\leqslant M\right\}$$

Compared to matching on covariates and weighting estimator,

- Under mild conditions, it is more efficient than matching on covariates
- As a nonparametric estimator, under mild conditions, it can achieve $\sqrt{n}-$convergence rate (like parametric model)
- It's bias is smaller than weighting estimator, but in general, it still less efficient than the latter in terms of mean square error (MSE)

10.5 Local ATE

Remark 10.21. Suppose that D_i is related to some unobserved characteristics, such that Conditional Independence assumption can not be satisfied. Still, we have

$$\begin{aligned} Y_i &= D_i Y_i(1) + (1 - D_i) Y_i(0) \\ &= \underbrace{\mathbb{E}[Y_i(0)]}_{\alpha} + \underbrace{\tau}_{\beta} \cdot D_i + \underbrace{[Y_i(1) - Y_i(0) - \tau] \cdot D_i + Y_i(0) - \mathbb{E}[Y_i(0)]}_{\varepsilon_i} \\ &= \alpha + \beta D_i + \varepsilon_i \end{aligned}$$

and $\mathbb{E}[\varepsilon_i] = 0$, but

$$\mathbb{E}[D_i \varepsilon_i] \neq 0$$

So, OLS can not be consistent. Now, suppose there is an IV Z_i such that

$$\{Y_i(0), Y_i(1)\} \perp Z_i \text{ and } \operatorname{Cov}(Z_i, D_i) \neq 0$$

The IV estimator of β is then

$$\hat{\beta}_{\mathrm{IV}} = \frac{\widehat{\operatorname{Cov}}(Z_i, Y_i)}{\widehat{\operatorname{Cov}}(Z_i, D_i)}$$

and

$$\hat{\beta}_{\mathrm{IV}} \xrightarrow{\mathbb{P}} \frac{\operatorname{Cov}(Z_i, Y_i)}{\operatorname{Cov}(Z_i, D_i)}$$

Let

$$D_i = Z_i \cdot D_i(1) + (1 - Z_i) \cdot D_i(0) = \begin{cases} D_i(1), & Z_i = 1 \\ D_i(0), & Z_i = 0 \end{cases}$$

where $Z_i \perp \{D_i(1), D_i(0)\}$. There are 4 types of individuals: (for example, we can choose whether to give a person a pill, and that person can choose whether to eat or not

1. $D_i(0) = 0, D_i(1) = 0$: always defiers (always not eat).
2. $D_i(0) = 1, D_i(1) = 1$: always compilers (always eat).
3. $D_i(0) = 0, D_i(1) = 1$: always compilers (If given, then eat; If not given, then not eat).
4. $D_i(0) = 1, D_i(1) = 0$: always compilers (If not given, then eat; If given, then not eat).

Some of them are impossible to happen. Now we show that

$$\frac{\operatorname{Cov}(Z_i, Y_i)}{\operatorname{Cov}(Z_i, D_i)} = \mathbb{E}[Y_i(1) - Y_i(0) | D_i(1) - D_i(0) = 1]$$

which is so-called local ATE.

Proof. We prove it by several steps.

- Firstly,

$$\begin{aligned}
&\operatorname{Cov}(Z_i, Y_i) \\
=&\mathbb{E}[Z_i Y_i] - \mathbb{E}[Z_i]\,\mathbb{E}[Y_i] \\
=&\mathbb{E}[Z_i \mathbb{E}[Y_i|\, Z_i]] - \mathbb{E}[Z_i]\,\mathbb{E}[\mathbb{E}[Y_i|\, Z_i]] \\
=&\mathbb{E}[Y_i|\, Z_i = 1]\,\mathbb{P}\{Z_i = 1\} - \mathbb{P}\{Z_i = 1\}\left(\mathbb{E}[Y_i|\, Z_i = 1]\,\mathbb{P}\{Z_i = 1\} + \mathbb{E}[Y_i|\, Z_i = 0]\,\mathbb{P}\{Z_i = 0\}\right) \\
=&\mathbb{P}\{Z_i = 1\}\left(\mathbb{E}[Y_i|\, Z_i = 1] - \mathbb{E}[Y_i|\, Z_i = 1]\,\mathbb{P}\{Z_i = 1\} - \mathbb{E}[Y_i|\, Z_i = 0]\,\mathbb{P}\{Z_i = 0\}\right) \\
=&\mathbb{P}\{Z_i = 1\}\,\mathbb{P}\{Z_i = 0\}\left(\mathbb{E}[Y_i|\, Z_i = 1] - \mathbb{E}[Y_i|\, Z_i = 0]\right)
\end{aligned}$$

  Likewise,

$$\operatorname{Cov}(Z_i, D_i) = \mathbb{P}\{Z_i = 1\}\,\mathbb{P}\{Z_i = 0\}\left(\mathbb{E}[D_i|\, Z_i = 1] - \mathbb{E}[D_i|Z_i = 0]\right)$$

  Assume that $\mathbb{P}\{Z_i = 1\}\,\mathbb{P}\{Z_i = 0\} \neq 0$, then

$$\frac{\operatorname{Cov}(Z_i, Y_i)}{\operatorname{Cov}(Z_i, D_i)} = \frac{\mathbb{E}[Y_i|\, Z_i = 1] - \mathbb{E}[Y_i|\, Z_i = 0]}{\mathbb{E}[D_i|\, Z_i = 1] - \mathbb{E}[D_i|\, Z_i = 0]}$$

- For the numerator,

$$\begin{aligned}
&\mathbb{E}[Y_i|\, Z_i = 1] - \mathbb{E}[Y_i|\, Z_i = 0] \\
=&\mathbb{E}[D_i(1)\, Y_i(1) + (1 - D_i(1))\, Y_i(0)|\, Z_i = 1] \\
&- \mathbb{E}[D_i(0)\, Y_i(1) + (1 - D_i(0))\, Y_i(0)|\, Z_i = 0] \\
=&\mathbb{E}[(D_i(1) - D_i(0))\,(Y_i(1) - Y_i(0))] \\
=&\mathbb{E}[\mathbb{E}[(D_i(1) - D_i(0))\,(Y_i(1) - Y_i(0))]|\, D_i(1) - D_i(0)] \\
=&\mathbb{E}[Y_i(1) - Y_i(0)|\, D_i(1) - D_i(0) = 1] \cdot \mathbb{P}(D_i(1) - D_i(0) = 1)
\end{aligned}$$

  where we assume that $\mathbb{P}(D_i(1) - D_i(0) = -1) = 0$. Similarly,

$$\begin{aligned}
\mathbb{E}[D_i|\, Z_i = 1] - \mathbb{E}[D_i Z_i = 0] &= \mathbb{E}[D_i(1)|\, Z_i = 1] - \mathbb{E}[D_i(0)|\, Z_i = 0] \\
&= \mathbb{E}[D_i(1) - D_i(0)] \\
&= \mathbb{P}\{D_i(1) - D_i(0) = 1\} - \mathbb{P}\{D_i(1) - D_i(0) = -1\} \\
&= \mathbb{P}\{D_i(1) - D_i(0) = 1\}
\end{aligned}$$

Finally, assume that $\mathbb{P}\{D_i(1) - D_i(0) = 1\} \neq 0$, then

$$\frac{\operatorname{Cov}(Z_i, Y_i)}{\operatorname{Cov}(Z_i, D_i)} = \mathbb{E}[(D_i(1) - D_i(0))\,(Y_i(1) - Y_i(0))]$$

□

10.6 DID Model

Example 10.22. Treatment effect in panel data models.

For a fixed effect model:

$$y_{it} = \alpha_i + \delta_t + x_{it}^{\mathrm{T}}\beta + u_{it}$$

If we define $x_{it} := D_{it} \in \{0, 1\}$ as a binary scalar, and $\phi := \beta$, then

$$y_{it} = \phi D_{it} + \delta_t + \alpha_i + u_{it}$$

where

$$D_{it} = \begin{cases} 1, & \text{if individual } i \text{ receives treatment in period } t \\ 0, & \text{o.w.} \end{cases}$$

What is the treatment effect (Assume the model only has two periods) ?

Firstly,

$$\Delta y_{it} := y_{it} - y_{i,t-1} = \phi \Delta D_{it} + (\delta_t - \delta_{t-1}) + \Delta u_{it}$$

where $\Delta D_{it} := D_{it} - D_{i,t-1} = \begin{cases} 1, & \text{if treated} \\ 0, & \text{if not treated} \end{cases}$, or we can write

$$\text{Treated group: } \Delta y_{it} = \phi \cdot 1 + (\delta_t - \delta_{t-1}) + \Delta u_{it}$$
$$\text{Control group: } \Delta y_{jt} = \phi \cdot 0 + (\delta_t - \delta_{t-1}) + \Delta u_{jt}$$

Within group average:

$$\overline{\Delta y_t^{tr}} = \phi + (\delta_t - \delta_{t-1}) + \overline{\Delta u_t^{tr}}$$
$$\overline{\Delta y_t^{ct}} = (\delta_t - \delta_{t-1}) + \overline{\Delta u_t^{ct}}$$

Note that there are only two periods, we can drop the subscript t, if we further assume $u \perp D$, then $\mathbb{E}[\varepsilon | D = 0] = \mathbb{E}[\varepsilon | D = 1] = 0$, and $\overline{\Delta u^{tr}} \approx 0$, $\overline{\Delta u^{ct}} \approx 0$. The treatment effect is the difference in differences:

$$\overline{\Delta y_t^{tr}} - \overline{\Delta y_t^{ct}} \approx \phi$$

Now all we need to do is to estimate ϕ.

Remark 10.23. What if the model has other covariates?

- The traditional difference in difference estimator is a special type of panel data model.
  - $T = 2$.
  - No other individual and time specific covariates.
  - With a treatment D_{it}, and $D_{i0} = 0$ for all i.
  - The time-fixed effect δ_t is the same for both treated and control. groups
- If we have other covariates, for example

  $$y_{it} = \phi D_{it} + x_{it}^{\mathrm{T}}\beta + \delta_t + \alpha_i + u_{it}$$

  We CANNOT use the following diff-in-diffs formula. We should solve this panel data model with standard methods: Two-way within estimator

Remark 10.24. Two way within estimator in diff and diff.

Consider the panel model:

$$y_{it} = x_{it}^{\mathrm{T}}\beta + \delta_t + \alpha_i + u_{it}$$

Now we need to get rid of BOTH fixed individual effect α_i and time effect δ_t.

First, we take a simpler model:

$$y_{it} = \delta_t + \alpha_i + u_{it}$$

Define:

- Average of the same individual (i) across different time periods: $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$.
- Average of the same time period (t) across different individuals: $\tilde{y}_i = \frac{1}{N}\sum_{i=1}^{N} y_{it}$.
- Full-sample average: $\bar{y} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}$.

Then

$$\begin{aligned}
\bar{y}_i &= \bar{\delta} + \alpha_i + \bar{u}_t \\
\tilde{y}_t &= \delta_t + \bar{\alpha} + \tilde{u}_i \\
\bar{y} &= \bar{\delta} + \bar{\alpha} + \bar{u}
\end{aligned}$$

Define

$$\ddot{y}_{it} = y_{it} - \bar{y}_i - \tilde{y}_t + \bar{y}$$

Therefore,

$$\begin{aligned}
\ddot{y}_{it} &= y_{it} - \bar{y}_i - \tilde{y}_t + \bar{y} \\
&= \delta_t + \alpha_i + u_{it} - \left(\bar{\delta} + \alpha_i + \bar{u}_t\right) - \left(\delta_t + \bar{\alpha} + \tilde{u}_i\right) + \bar{\delta} + \bar{\alpha} + \bar{u} \\
&= u_{it} - \bar{u}_t - \tilde{u}_i + \bar{u} \\
&:= \ddot{u}_{it}
\end{aligned}$$

We get rid of all the time and individual fixed effects.

Now back to

$$y_{it} = x_{it}^{\mathrm{T}}\beta + \delta_t + \alpha_i + u_{it}$$

we can likewise have

$$\ddot{y}_{it} = \ddot{x}_{it}^{\mathrm{T}}\beta + \ddot{u}_{it}$$

where $\ddot{x}_{it} = x_{it} - \bar{x}_i - \tilde{x}_t + \bar{x}$. If we further assume that $\mathbb{E}\left[\ddot{x}_{it}^{\mathrm{T}}\ddot{u}_{it}\right] = 0$, then we get β consistently estimated.

Moreover, if x_{it} contains a binary value $D_{it} \in \mathbb{R}$, then we can extract D_{it} out of x_{it} to get

$$\ddot{y}_{it} = \phi\ddot{D}_{it} + \ddot{x}_{it}^{\mathrm{T}}\beta + \ddot{u}_{it}$$

Remark 10.25. Summary: Identification of diff-in-diffs model.

1. Specification: $y_{it} = \phi D_{it} + x_{it}^{\mathrm{T}}\beta + \alpha_i + \delta_t + u_{it}$.
2. Estimation: Regress $\ddot{y}_{it} = \phi \ddot{D}_{it} + \ddot{x}_{it}^{\mathrm{T}}\beta + \ddot{u}_{it}$ with OLS.
3. Assumptions: Denote $\ddot{z}_{it} = \left(\ddot{D}_{it}, \ddot{x}_{it}^{\mathrm{T}}\right)^{\mathrm{T}}$,
   - $y_{it} = \phi D_{it} + x_{it}^{\mathrm{T}}\beta + \alpha_i + \delta_t + u_{it}$.
   - $\mathbb{E}\left[\ddot{z}_{it}\ddot{z}_{it}^{\mathrm{T}}\right] > 0$.
   - $\mathbb{E}\left[x_{it}\varepsilon_{is}\right] = 0, \forall t, s$.
   - $D_{it} \perp \varepsilon_{is} | \, x_{i1}, x_{i2}, \cdots, x_{iT}, \forall t, s$.
4. Then the coefficient ϕ equals average treatment effect (ATE).

10.7 Multiple Units and Time Periods

Remark 10.26. Goodman-Bacon (2021) JOE.

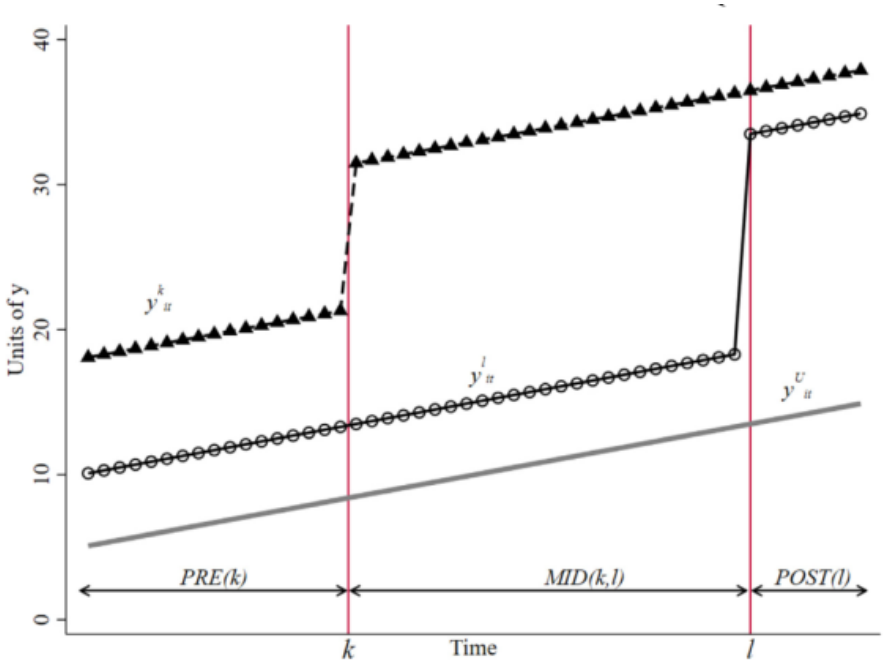

Figure 10.1. Goodman-Bacon (2021) JOE.

Chapter 11

Quantile Regression

11.1 Introduction

Definition 11.1. Suppose we have some CDF $F(\cdot)$. The *τ-th quantile of $F(\cdot)$*, defined for $\tau \in (0,1)$, is the inverse of the CDF:

$$Q_\tau = \inf \{c : F(c) \geqslant \tau\}$$

Suppose that $F(\cdot)$ is strictly increasing, then

$$Q_\tau = \{c : F(c) = \tau\}$$

Definition 11.2. When we want to estimate Q_τ from data, We use the *check function*, which is defined as:

$$\rho_\tau(u) = \left[\tau - \mathbb{I}_{\{u \leqslant 0\}}\right] \cdot u = \begin{cases} \tau u, & u > 0 \\ (\tau - 1)\, u, & u \leqslant 0 \end{cases}$$

Remark 11.3. We can see that:

- $\rho_\tau(u)$ is a piece-wise linear function.
- $\rho_\tau(u)$ is a convex function.
- $\rho_\tau(u)$ is NOT differentiable at $u = 0$.

Theorem 11.4. If Q_τ is well defined, then

$$Q_\tau = \arg\min_c \mathbb{E}\left[\rho_\tau(X - c)\right]$$

where $X \sim F$ is a continuous random variable.

Therefore, we can estimate Q_τ by

$$\hat{q}_\tau = \arg\min_c \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(x_i - c)$$

Proof. First,

$$\begin{aligned}\mathbb{E}\left[\rho_\tau\left(X-c\right)\right] &= \mathbb{E}\left[\left(\tau-\mathbb{I}_{\{X-c\leqslant 0\}}\right)\left(X-c\right)\right] \\ &= \tau\cdot\mathbb{E}X-c\tau-\int_{\{X-c\leqslant 0\}} x\mathrm{d}F\left(x\right)+c\mathbb{P}\left\{X-c\leqslant 0\right\} \\ &= \tau\cdot\mathbb{E}X-c\tau-\int_{-\infty}^{c} xf\left(x\right)\mathrm{d}x+cF\left(c\right)\end{aligned}$$

The F.O.C. is

$$\frac{\partial\mathbb{E}\left[\rho_\tau\left(X-c\right)\right]}{\partial c}=-\tau-cf\left(c\right)+F\left(c\right)+cf\left(c\right)=F\left(c\right)-\tau=0$$

The S.O.C. is

$$\frac{\partial^2\mathbb{E}\left[\rho_\tau\left(X-c\right)\right]}{\partial c^2}=f\left(c\right)\geqslant 0$$

Then

$$\left.\frac{\partial\mathbb{E}\left[\rho_\tau\left(X-c\right)\right]}{\partial c}\right|_{c=Q_\tau}=-\tau+F\left(Q_\tau\right)=0\Longrightarrow Q_\tau=\inf\left\{c:F\left(c\right)\geqslant\tau\right\}$$

□

11.2 Conditional Quantiles and Quantile Regressions

Definition 11.5. The *τ-th conditional quantile* of $X\sim F\left(\cdot\right)$ conditional on the event $\{Y=y\}$ is

$$Q_\tau\left(X\middle|\,Y=y\right)=\inf\left\{c:F_{X|Y}\left(c\middle|\,Y=y\right)\geqslant\tau\right\}$$

where $F_{X|Y}\left(c\middle|\,Y=y\right)=\mathbb{P}\left\{X\leqslant c\middle|\,Y=y\right\}$.

Remark 11.6. Now consider the model

$$y_i=\beta_0^{\mathrm{T}}\left(\tau\right)x_i+u_i$$

where $Q_\tau\left(u_i\middle|\,x_i=x\right)=0,\forall x$, and by definition $Q_\tau\left(u_i\middle|\,x_i=x\right)$ denotes the conditional quantile of u_i given $\{x_i=x\}$, or

$$Q_\tau\left(u_i\middle|\,x_i=x\right)=\inf\left\{c:F_{u_i|x_i}\left(c\middle|\,x_i=x\right)\geqslant\tau\right\}$$

where $F_{u_i|x_i}\left(c\middle|\,x_i=x\right)=\mathbb{P}\left\{u_i\leqslant c\middle|\,x_i=x\right\}$.

We can write

$$Q_\tau\left(y_i\middle|\,x_i=x\right)=Q_\tau\left(\beta_0\left(\tau\right)^{\mathrm{T}}x_i+u_i\middle|\,x_i=x\right)=\beta_0^{\mathrm{T}}\left(\tau\right)\cdot x$$

then

$$\beta_0^{\mathrm{T}}\left(\tau\right)=\frac{\partial Q_\tau\left(y_i\middle|\,x_i=x\right)}{\partial x^{\mathrm{T}}}$$

This is the partial effect of x_i on the conditional quantiles of y_i, which is allowed to vary across τ.

Example 11.7. Location-scale model.

$$y_i = \beta_0^{\mathrm{T}} x_i + \left(\gamma_0^{\mathrm{T}} x_i\right) \cdot \varepsilon_i$$

where $\varepsilon_i \overset{\text{i.i.d.}}{\sim} F_\varepsilon(\cdot)$, $\varepsilon_i \perp x_i$, and $\gamma_0^{\mathrm{T}} x_i \geqslant 0$.[1] Then

$$y_i = \underbrace{\left(\beta_0 + \gamma_0 Q_\tau\right)^{\mathrm{T}}}_{\beta_0(\tau)} x_i + \underbrace{\left(\gamma_0^{\mathrm{T}} x_i\right) \cdot \left(\varepsilon_i - Q_\tau\right)}_{u_i}$$

where Q_τ, is the τ-th quantile of ε_i. It can be shown that $Q_\tau\left(u_i \middle| x_i = x\right)$, actually,

$$\begin{aligned} Q_\tau\left(u_i \middle| x_i = x\right) &= Q_\tau\left(\gamma_0^{\mathrm{T}} x_i \varepsilon_i - \gamma_0^{\mathrm{T}} x_i Q_\tau \middle| x_i = x\right) \\ &= \gamma_0^{\mathrm{T}} x Q_\tau\left(\varepsilon_i\right) - \gamma_0^{\mathrm{T}} x Q_\tau \\ &= 0 \end{aligned}$$

Then, $\beta_0(\tau)$ can be estimate by QR:

$$\hat{\beta}_n = \arg\min_\beta \underbrace{\frac{1}{n} \sum_{i=1}^n \rho_\tau\left(y_i - \beta^{\mathrm{T}} x_i\right)}_{M_n(\beta)}$$

11.3 Consistency of Location-scale Model

Theorem 11.8. In location-scale model, If $\mathbb{E}\left[f_{y_i|x_i}\left(\beta^{\mathrm{T}} x_i\right) \cdot x_i x_i^{\mathrm{T}}\right] > 0$, then $\hat{\beta}_n$ is consistent, i.e., $\hat{\beta}_n \xrightarrow{\mathbb{P}} \beta_0$.

Proof. Firstly, since ρ_τ is convex, then $M_n(\beta)$ is also convex.

Second, we need to show that

$$M_n(\beta) \xrightarrow{\mathbb{P}} M_0(\beta) = \mathbb{E}\left[\rho_\tau\left(y_i - \beta^{\mathrm{T}} x_i\right)\right], \forall \beta \in B$$

This simplify comes from CLT, but it requires moment condition on y_i, and therefore on u_i. This can be avoided by defining

$$\begin{aligned} M_n(\beta) &= \frac{1}{n} \sum_{i=1}^n \left[\rho_\tau\left(y_i - \beta^{\mathrm{T}} x_i\right) - \rho_\tau\left(y_i - \beta_0^{\mathrm{T}} x_i\right)\right] \\ M_0(\beta) &= \mathbb{E}\left[\rho_\tau\left(y_i - \beta^{\mathrm{T}} x_i\right) - \rho_\tau\left(y_i - \beta_0^{\mathrm{T}} x_i\right)\right] \end{aligned}$$

[1] We should be careful about the sign, since we basically assume

$$\mathbb{P}\left\{u_i \leqslant 0 \middle| x_i\right\} = \tau \Longleftrightarrow Q_\tau\left(u_i \middle| x_i\right) = 0$$

but

$$\mathbb{P}\left\{-u_i \leqslant 0 \middle| x_i\right\} = \mathbb{P}\left\{u_i \geqslant 0 \middle| x_i\right\} = 1 - \tau$$

And this definition won't change $\hat{\beta}_n$.

$$
\begin{aligned}
&\left|\rho_\tau\left(y_i-\beta^{\mathrm{T}}x_i\right)-\rho_\tau\left(y_i-\beta_0^{\mathrm{T}}x_i\right)\right| \\
=&\left|\left(\tau-\mathbb{I}_{\{y_i\geqslant\beta^{\mathrm{T}}x_i\}}\right)\left(y_i-\beta^{\mathrm{T}}x_i\right)-\left(\tau-\mathbb{I}_{\left\{y_i\geqslant\beta_0^{\mathrm{T}}x_i\right\}}\right)\left(y_i-\beta_0^{\mathrm{T}}x_i\right)\right| \\
=&\left|\mathbb{I}_{\{y_i\geqslant\beta^{\mathrm{T}}x_i\}}\beta^{\mathrm{T}}x_i-\mathbb{I}_{\left\{y_i\geqslant\beta_0^{\mathrm{T}}x_i\right\}}\beta_0^{\mathrm{T}}x_i+\left(\mathbb{I}_{\left\{y_i\geqslant\beta_0^{\mathrm{T}}x_i\right\}}-\mathbb{I}_{\{y_i\geqslant\beta^{\mathrm{T}}x_i\}}\right)y_i+\tau\left(\beta_0-\beta\right)^{\mathrm{T}}x_i\right| \\
=&\left|\mathbb{I}_{\{y_i\geqslant\beta^{\mathrm{T}}x_i\}}\beta^{\mathrm{T}}x_i-\mathbb{I}_{\left\{y_i\geqslant\beta_0^{\mathrm{T}}x_i\right\}}\beta_0^{\mathrm{T}}x_i+\left(\mathbb{I}_{\left\{y_i\geqslant\beta_0^{\mathrm{T}}x_i\right\}}-\mathbb{I}_{\{y_i\geqslant\beta^{\mathrm{T}}x_i\}}\right)\left(\beta_0^{\mathrm{T}}x_i+u_i\right)+\tau\left(\beta_0-\beta\right)^{\mathrm{T}}x_i\right| \\
=&\left|\mathbb{I}_{\{y_i\geqslant\beta^{\mathrm{T}}x_i\}}\left(\beta-\beta_0\right)^{\mathrm{T}}x_i+\left(\mathbb{I}_{\left\{y_i\geqslant\beta_0^{\mathrm{T}}x_i\right\}}-\mathbb{I}_{\{y_i\geqslant\beta^{\mathrm{T}}x_i\}}\right)u_i+\tau\left(\beta_0-\beta\right)^{\mathrm{T}}x_i\right| \\
\leqslant&2\left|\left(\beta-\beta_0\right)^{\mathrm{T}}x_i\right| \\
\leqslant&2\left\|\beta-\beta_0\right\|\cdot\left\|x_i\right\|
\end{aligned}
$$

To show $M_n\left(\beta\right)\xrightarrow{\mathbb{P}}M_0\left(\beta\right)$, we only need moment conditions on x_i but not on u_i.

Third, we show β_0 uniquely minimize $M_0(\beta)$. Or we show that $\frac{\partial M_0}{\partial\beta}\left(\beta_0\right)=0$. Note that

$$
\begin{aligned}
M_0\left(\beta\right)&=\mathbb{E}\left[\rho_\tau\left(y_i-\beta^{\mathrm{T}}x_i\right)-\rho_\tau\left(y_i-\beta_0^{\mathrm{T}}x_i\right)\right] \\
&=\mathbb{E}\left[\left(\tau-\mathbb{I}_{\{y_i\leqslant\beta^{\mathrm{T}}x_i\}}\right)\left(y_i-\beta^{\mathrm{T}}x_i\right)-\left(\tau-\mathbb{I}_{\left\{y_i\leqslant\beta_0^{\mathrm{T}}x_i\right\}}\right)\left(y_i-\beta_0^{\mathrm{T}}x_i\right)\right] \\
&=-\tau\beta^{\mathrm{T}}\mathbb{E}\left[x_i\right]-\mathbb{E}\left[\mathbb{I}_{\{y_i\leqslant\beta^{\mathrm{T}}x_i\}}y_i\right]+\mathbb{E}\left[\mathbb{I}_{\{y_i\leqslant\beta^{\mathrm{T}}x_i\}}\beta^{\mathrm{T}}x_i\right]+\underbrace{\cdots}_{G(\beta_0)}
\end{aligned}
$$

Then $\frac{\partial M_0}{\partial\beta}$ has three terms:

- $-\frac{\partial\tau\beta^{\mathrm{T}}\mathbb{E}[x_i]}{\partial\beta}=-\tau\mathbb{E}\left[x_i\right]$.
- Note that

  $$\mathbb{E}\left[\mathbb{I}_{\{y_i\leqslant\beta^{\mathrm{T}}x_i\}}y_i\right]=\mathbb{E}\left[\mathbb{E}\left[\mathbb{I}_{\{y_i\leqslant\beta^{\mathrm{T}}x_i\}}y_i\middle|\,x_i\right]\right]=\int\left[\int_{-\infty}^{\beta^{\mathrm{T}}x_i}y\cdot f_{y_i|x_i}\left(y\right)\mathrm{d}y\right]f_{x_i}\left(x\right)\mathrm{d}x$$

  then

  $$-\frac{\partial\mathbb{E}\left[\mathbb{I}_{\{y_i\leqslant\beta^{\mathrm{T}}x_i\}}y_i\right]}{\partial\beta}=-\int\beta^{\mathrm{T}}x\cdot f_{y_i|x_i}\left(\beta^{\mathrm{T}}x\right)\cdot xf_{x_i}\left(x\right)\mathrm{d}x$$

- Note that

  $$\mathbb{E}\left[\mathbb{I}_{\{y_i\leqslant\beta^{\mathrm{T}}x_i\}}\beta^{\mathrm{T}}x_i\right]=\mathbb{E}\left[\beta^{\mathrm{T}}x_i\mathbb{E}\left[\mathbb{I}_{\{y_i\leqslant\beta^{\mathrm{T}}x_i\}}\middle|\,x_i\right]\right]=\int F_{y_i|x_i}\left(\beta^{\mathrm{T}}x\right)\beta^{\mathrm{T}}xf_{x_i}\left(x\right)\mathrm{d}x$$

  then

  $$\frac{\partial\mathbb{E}\left[\mathbb{I}_{\{y_i\leqslant\beta^{\mathrm{T}}x\}}\beta^{\mathrm{T}}x\right]}{\partial\beta}=\int f_{y_i|x_i}\left(\beta^{\mathrm{T}}x\right)\cdot x\cdot\beta^{\mathrm{T}}x\cdot f_{x_i}\left(x\right)+F_{y_i|x_i}\left(\beta^{\mathrm{T}}x\right)\cdot x\cdot f_{x_i}\left(x\right)\mathrm{d}x$$

Overall, note that

$$F_{y_i|x_i}\left(\beta_0^{\mathrm{T}}x_i\right)=\mathbb{P}\left\{y_i\leqslant\beta_0^{\mathrm{T}}x_i\middle|\,x_i\right\}=\mathbb{P}\left(u_i\leqslant0\middle|\,x_i\right)=F_{u_i|x_i}\left(0\right)=\tau$$

then

$$\begin{aligned}\frac{\partial M_0(\beta_0)}{\partial \beta} &= -\tau\mathbb{E}[x_i] + \int F_{y_i|x_i}\left(\beta_0^{\mathrm{T}} x_i\right)\cdot x_i \cdot f_{x_i}(x)\,\mathrm{d}x \\ &= -\tau\mathbb{E}[x_i] + \mathbb{E}\left[F_{y_i|x_i}\left(\beta_0^{\mathrm{T}} x_i\right)\cdot x_i\right] \\ &= -\tau\mathbb{E}[x_i] + \mathbb{E}\left[F_{u_i|x_i}(0)\cdot x_i\right] \\ &= 0\end{aligned}$$

Moreover,

$$\begin{aligned}\frac{\partial^2 M_0(\beta)}{\partial\beta\partial\beta^{\mathrm{T}}} &= \frac{\partial}{\partial\beta}\int F_{y_i|x_i}\left(\beta^{\mathrm{T}}x\right)\cdot x^{\mathrm{T}}\cdot f_{x_i}(x)\,\mathrm{d}x = \int f_{y_i|x_i}\left(\beta^{\mathrm{T}}x\right)\cdot xx^{\mathrm{T}}\cdot f_{x_i}(x)\,\mathrm{d}x \\ &= \mathbb{E}\left[f_{y_i|x_i}\left(\beta^{\mathrm{T}}x_i\right)\cdot x_i x_i^{\mathrm{T}}\right]\end{aligned}$$

is positive semi-definite. If $\mathbb{E}\left[f_{y_i|x_i}\left(\beta^{\mathrm{T}}x_i\right)\cdot x_i x_i^{\mathrm{T}}\right] > 0$, then β_0 uniquely minimizes $M_0(\beta)$. □

11.4 Asymptotic Distribution of Location-scale Model

Theorem 11.9. In location-scale model, assume that $u_i \perp x_i$, we have

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\tau(1-\tau)}{f_{u_i}^2(0)}\left(\mathbb{E}\left[x_i x_i^{\mathrm{T}}\right]\right)^{-1}\right)$$

Proof. Firstly, by F.O.C. and mean-value theorem,

$$0 = \nabla_\beta M_n\left(\hat{\beta}_n\right) = \nabla_\beta M_n(\beta_0) + \nabla^2_{\beta\beta^{\mathrm{T}}} M_n\left(\bar{\beta}\right)\left(\hat{\beta}_n - \beta_0\right)$$

and because

$$M_n\left(\bar{\beta}\right) \xrightarrow{\mathbb{P}} M_0(\beta_0)$$

we have [2]

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) &= -\sqrt{n}\left(\nabla^2_{\beta\beta^{\mathrm{T}}} M_n\left(\bar{\beta}\right)\right)^{-1}\nabla_\beta M_n(\beta_0) \\ &= -\left(\nabla^2_{\beta\beta^{\mathrm{T}}} M_n\left(\bar{\beta}\right)\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\tau - \mathbb{I}_{\left\{y_i \leqslant \beta_0^{\mathrm{T}} x_i\right\}}\right)x_i \\ &= -\left(\nabla^2_{\beta\beta^{\mathrm{T}}} M_0(\beta_0)\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\tau - \mathbb{I}_{\left\{y_i \leqslant \beta_0^{\mathrm{T}} x_i\right\}}\right)x_i + o_p(1)\end{aligned}$$

To apply CLT, we need to compute

$$\mathbb{E}\left[\tau - \mathbb{I}_{\left\{y_i \leqslant \beta_0^{\mathrm{T}} x_i\right\}}\right] = \mathbb{E}\left[\mathbb{E}\left[\tau - \mathbb{I}_{\left\{y_i \leqslant \beta_0^{\mathrm{T}} x_i\right\}}\middle| x_i\right]\right] = \tau - \mathbb{P}\left\{y_i \leqslant \beta_0^{\mathrm{T}} x_i \middle| x_i\right\} = 0$$

and [3]

$$\mathrm{Var}\left(\left(\tau - \mathbb{I}_{\left\{y_i \leqslant \beta_0^{\mathrm{T}} x_i\right\}}\right)\cdot x_i\right) = \mathrm{Var}\left(\mathbb{E}\left[\left(\tau - \mathbb{I}_{\left\{y_i \leqslant \beta_0^{\mathrm{T}} x_i\right\}}\right)\cdot x_i \middle| x_i\right]\right) + \mathbb{E}\left[\mathrm{Var}\left(\left(\tau - \mathbb{I}_{\left\{y_i \leqslant \beta_0^{\mathrm{T}} x_i\right\}}\right)\cdot x_i \middle| x_i\right)\right]$$

where

[2] Although $\left.\frac{\partial \rho_\tau(u)}{\partial u}\right|_{u\neq 0} = \tau - \mathbb{I}_{\{u\geqslant 0\}}$, we can ignore the undifferentiable point $u = 0$ in expectation.

[3] By law of total variance $\mathrm{Var}(Y) = \mathrm{Var}(\mathbb{E}[Y|X]) + \mathbb{E}[\mathrm{Var}(Y|X)]$.

- $\operatorname{Var}\left(\mathbb{E}\left[\left(\tau-\mathbb{I}_{\left\{y_i\leqslant\beta_0^{\mathrm{T}}x_i\right\}}\right)\cdot x_i\middle|\, x_i\right]\right)=\operatorname{Var}\left(\left[\tau-\mathbb{P}\left\{u_i\leqslant 0\middle|\, x_i\right\}\right]x_i\right)=0.$
- As for $\mathbb{E}\left[\operatorname{Var}\left(\tau-\mathbb{I}_{\left\{y_i\leqslant\beta_0^{\mathrm{T}}x_i\right\}}\middle|\, x_i\right)\right]$, we first compute

$$\begin{aligned}\operatorname{Var}\left(\left(\tau-\mathbb{I}_{\left\{y_i\leqslant\beta_0^{\mathrm{T}}x_i\right\}}\right)\cdot x_i\middle|\, x_i\right)&=\operatorname{Var}\left(\mathbb{I}_{\left\{y_i\leqslant\beta_0^{\mathrm{T}}x_i\right\}}\cdot x_i\middle|\, x_i\right)\\&=\left[\mathbb{E}\left(\mathbb{I}^2_{\left\{y_i\leqslant\beta_0^{\mathrm{T}}x_i\right\}}\middle|\, x_i\right)-\mathbb{E}\left(\mathbb{I}^2_{\left\{y_i\leqslant\beta_0^{\mathrm{T}}x_i\right\}}\middle|\, x_i\right)\right]x_ix_i^{\mathrm{T}}\\&=\left[\mathbb{P}\left\{y_i\leqslant\beta_0^{\mathrm{T}}x_i\middle|\, x_i\right\}-\mathbb{P}^2\left\{y_i\leqslant\beta_0^{\mathrm{T}}x_i\middle|\, x_i\right\}\right]x_ix_i^{\mathrm{T}}\\&=\tau\left(1-\tau\right)x_ix_i^{\mathrm{T}}\end{aligned}$$

  Then

$$\mathbb{E}\left[\operatorname{Var}\left(\left(\tau-\mathbb{I}_{\left\{y_i\leqslant\beta_0^{\mathrm{T}}x_i\right\}}\right)\cdot x_i\middle|\, x_i\right)\right]=\tau\left(1-\tau\right)\mathbb{E}\left[x_ix_i^{\mathrm{T}}\right]$$

Therefore,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\tau-\mathbb{I}_{\left\{y_i\leqslant\beta_0^{\mathrm{T}}x_i\right\}}\right)\xrightarrow{d}\mathcal{N}\left(0,\tau\left(1-\tau\right)\mathbb{E}\left[x_ix_i^{\mathrm{T}}\right]\right)$$

or

$$\sqrt{n}\left(\hat{\beta}_n-\beta_0\right)\xrightarrow{d}\mathcal{N}\left(0,\tau\left(1-\tau\right)\left(\nabla^2_{\beta\beta^{\mathrm{T}}}M_0\left(\beta_0\right)\right)^{-1}\mathbb{E}\left[x_ix_i^{\mathrm{T}}\right]\left(\nabla^2_{\beta\beta^{\mathrm{T}}}M_0\left(\beta_0\right)\right)^{-1}\right)$$

where

$$\nabla^2_{\beta\beta^{\mathrm{T}}}M_0\left(\beta_0\right)=\mathbb{E}\left[f_{y_i|x_i}\left(\beta_0^{\mathrm{T}}x_i\right)x_ix_i^{\mathrm{T}}\right]=\mathbb{E}\left[f_{u_i|x_i}\left(0\right)x_ix_i^{\mathrm{T}}\right]$$

By assumption, $u_i\perp x_i$, then

$$\nabla^2_{\beta\beta^{\mathrm{T}}}M_0\left(\beta_0\right)=f_{u_i}\left(0\right)\mathbb{E}\left[x_ix_i^{\mathrm{T}}\right]$$

then

$$\sqrt{n}\left(\hat{\beta}_n-\beta_0\right)\xrightarrow{d}\mathcal{N}\left(0,\frac{\tau\left(1-\tau\right)}{f^2_{u_i}\left(0\right)}\mathbb{E}\left[x_ix_i^{\mathrm{T}}\right]^{-1}\right)$$

□

Chapter 12

Regression Discontinuity Designs

12.1 Sharp Regression Discontinuity (SRD)

Remark 12.1. Let X_i be the *forcing variable* such that $D_i = \mathbb{I}_{\{X_i \geqslant c\}}$ where c is called the *threshold*.

We are interested in the average treatment effect at the discontinuity point

$$\tau_{\text{SRD}} = \mathbb{E}\left[Y_i(1) - Y_i(0) \middle| X_i = c\right]$$

Assume that $\mathbb{E}\left[Y_i(1) \middle| X_i = x\right]$ and $\mathbb{E}\left[Y_i(0) \middle| X_i = x\right]$ are both continuous in x (around $x = c$).

Since

$$\begin{aligned}
\mathbb{E}\left[Y_i \middle| X_i = x\right] =& \mathbb{E}\left[\mathbb{E}\left[Y_i \middle| D_i, X_i = x\right] \middle| X_i = x\right] \\
=& \mathbb{E}\left[Y_i \middle| D_i = 1, X_i = x\right] \cdot \mathbb{P}\left\{D_i = 1 \middle| X_i = x\right\} \\
&+ \mathbb{E}\left[Y_i \middle| D_i = 0, X_i = x\right] \cdot \mathbb{P}\left\{D_i = 0 \middle| X_i = x\right\}
\end{aligned}$$

then

$$\mathbb{E}\left[Y_i \middle| X_i = x\right] = \begin{cases} \mathbb{E}\left[Y_i(0) \middle| X_i = x\right], & x < c \\ \mathbb{E}\left[Y_i(1) \middle| X_i = x\right], & x \geqslant c \end{cases}$$

By the continuity assumption on the conditional means, we have:

$$\begin{aligned}
\tau_{\text{SRD}} &= \mathbb{E}\left[Y_i(1) - Y_i(0) \middle| X_i = c\right] \\
&= \mathbb{E}\left[Y_i(1) \middle| X_i = c\right] - \mathbb{E}\left[Y_i(0) \middle| X_i = c\right] \\
&= \lim_{x \downarrow c} \mathbb{E}\left[Y_i \middle| X_i = x\right] - \lim_{x \uparrow c} \mathbb{E}\left[Y_i \middle| X_i = x\right]
\end{aligned}$$

Remark 12.2. Here are some questions.

1. Show that in the setup of SRD, the conditional assumption $\left\{Y_i(1), Y_i(0)\right\} \perp D_i \middle| X_i$ is trivially satisfied.

   Proof. Fix X_i, if $X_i \geqslant c$, then $D_i = 1$; $X_i < c$, then $D_i = 0$. So we can regard D_i given X_i is a constant. Then $\left\{Y_i(1), Y_i(0)\right\} \perp D_i \middle| X_i$ is hold. □

2. Can we use matching-type or propensity score based methods to estimate the ATE?

Answer. No. □

12.2 Fuzzy Regression Discontinuity (FRD)

Definition 12.3. Assuming that there exists $c \in \mathbb{R}$ such that

$$\lim_{x \downarrow c} \mathbb{P}\left\{D_i = 1 \middle| X_i = x\right\} \neq \lim_{x \uparrow c} \mathbb{P}\left\{D_i = 1 \middle| X_i = x\right\}$$

this is called the *Fuzzy Regression Discontinuity Design* (FRD).

Example 12.4. Suppose

$$D_i = \mathbb{I}\left\{X_i + \gamma \cdot \mathbb{I}_{\{X_i \geqslant c\}} \geqslant Z_i\right\}$$

where $Z_i \perp X_i, Z_i \sim \Lambda(\cdot)$. Then

$$\begin{aligned}
\mathbb{P}\left\{D_i = 1 \middle| X_i = x\right\} &= \mathbb{P}\left\{X_i + \gamma \cdot \mathbb{I}_{\{X_i \geqslant c\}} \geqslant Z_i \middle| X_i = x\right\} = \mathbb{P}\left\{x + \gamma \cdot \mathbb{I}_{\{x \geqslant c\}} \geqslant Z_i \middle| X_i = x\right\} \\
&= \mathbb{P}\left\{x + \gamma \cdot \mathbb{I}_{\{x \geqslant c\}} \geqslant Z_i\right\} = \Lambda\left(x + \gamma \cdot \mathbb{I}_{\{x \geqslant c\}}\right) \\
&= \begin{cases} \Lambda(x), & x < c \\ \Lambda(x + \gamma), & x \geqslant c \end{cases}
\end{aligned}$$

Remark 12.5. Moreover, in FRD, because of the Z_i, the conditional independence assumption does not necessarily hold.

Suppose that we can write

$$Y_i = \alpha + \beta \cdot D_i + u_i$$

then

$$\mathbb{E}\left[Y_i \middle| X_i = x\right] = \alpha + \beta \cdot \mathbb{E}\left[D_i \middle| X_i = x\right] + \mathbb{E}\left[u_i \middle| X_i = x\right]$$

Then

$$\begin{aligned}
\mathbb{E}\left[Y_i \middle| X_i = c + \varepsilon\right] &= \alpha + \beta \cdot \mathbb{E}\left[D_i \middle| X_i = c + \varepsilon\right] + \mathbb{E}\left[u_i \middle| X_i = c + \varepsilon\right] \\
\mathbb{E}\left[Y_i \middle| X_i = c - \varepsilon\right] &= \alpha + \beta \cdot \mathbb{E}\left[D_i \middle| X_i = c - \varepsilon\right] + \mathbb{E}\left[u_i \middle| X_i = c - \varepsilon\right]
\end{aligned}$$

Then

$$\begin{aligned}
&\mathbb{E}\left[Y_i \middle| X_i = c + \varepsilon\right] - \mathbb{E}\left[Y_i \middle| X_i = c - \varepsilon\right] \\
=&\beta\left(\mathbb{E}\left[D_i \middle| X_i = c + \varepsilon\right] - \mathbb{E}\left[D_i \middle| X_i = c - \varepsilon\right]\right) + \mathbb{E}\left[u_i \middle| X_i = c + \varepsilon\right] - \mathbb{E}\left[u_i \middle| X_i = c - \varepsilon\right]
\end{aligned}$$

Assume that $\mathbb{E}\left[u_i \middle| X_i = x\right]$ is continuous at c, then we have

$$\beta = \frac{\lim_{\varepsilon \downarrow 0} \mathbb{E}\left[Y_i \middle| X_i = c + \varepsilon\right] - \mathbb{E}\left[Y_i \middle| X_i = c - \varepsilon\right]}{\lim_{\varepsilon \downarrow 0} \mathbb{E}\left[D_i \middle| X_i = c + \varepsilon\right] - \mathbb{E}\left[D_i \middle| X_i = c - \varepsilon\right]}$$

where by our assumption, $\lim_{\varepsilon \downarrow 0} \mathbb{E}\left[D_i \middle| X_i = c + \varepsilon\right] - \mathbb{E}\left[D_i \middle| X_i = c - \varepsilon\right] \neq 0$.

12.3 Examples

Example 12.6. Back to the ATE framework,

$$Y_i = \alpha + \beta \cdot D_i + u_i$$

where

$$\begin{aligned}
\alpha &= \mathbb{E}\left[Y_i\left(0\right)\right] \\
\beta &= \tau = \mathbb{E}\left[\tau_i\right] \\
u_i &= Y_i\left(0\right) - \mathbb{E}\left[Y_i\left(0\right)\right] + \left[\tau - \tau_i\right] D_i
\end{aligned}$$

Then

$$\mathbb{E}\left[u_i \middle| X_i = x\right] = \mathbb{E}\left[Y_i\left(0\right) - \mathbb{E}\left[Y_i\left(0\right)\right] \middle| X_i = x\right] + \mathbb{E}\left[\left(\tau - \tau_i\right) D_i \middle| X_i = x\right]$$

Assume the confounderness condition:

$$\tau_i \perp D_i \middle| X_i$$

Then

$$\begin{aligned}
&\mathbb{E}\left[u_i \middle| X_i = x\right] \\
=&\underbrace{\mathbb{E}\left[Y_i\left(0\right) - \mathbb{E}\left[Y_i\left(0\right)\right] \middle| X_i = x\right]}_{\text{continuous in } x} + \tau - \underbrace{\mathbb{E}\left[\tau_i \middle| X_i = x\right]}_{\text{continuous in } x} + \underbrace{\mathbb{E}\left[D_i \middle| X_i = x\right]}_{\text{not continuous in } x=c}
\end{aligned}$$

In this case, $\mathbb{E}\left[u_i \middle| X_i = x\right]$ is NOT continuous at $x = c$, so $\beta = \tau$ **can not be identified**.

Example 12.7. In this case,

$$Y_i = \alpha + \beta \cdot D_i + u_i$$

where

$$\begin{aligned}
\alpha &= \mathbb{E}\left[Y_i\left(0\right)\right] \\
\beta &= \tau\left(c\right) \\
\tau\left(x\right) &= \mathbb{E}\left[\tau_i \middle| X_i = x\right] = \mathbb{E}\left[Y_i(1) - Y_i(0) \middle| X_i = x\right] \\
u_i &= Y_i\left(0\right) - \mathbb{E}\left[Y_i\left(0\right)\right] + \left[\tau_i - \tau\left(c\right)\right]
\end{aligned}$$

Similarly, we can show that

$$\mathbb{E}\left[u_i \middle| X_i = x\right] = \mathbb{E}\left[Y_i\left(0\right) - \mathbb{E}\left[Y_i\left(0\right)\right] \middle| X_i = x\right] + \mathbb{E}\left[\tau_i - \tau\left(c\right) \middle| X_i = x\right] \cdot \mathbb{P}\left\{D_i = 1 \middle| X_i = x\right\}$$

Suppose $\tau(x) = \mathbb{E}\left[\tau_i \middle| X_i = x\right]$ is continuous at $x = c$, then

$$\lim_{x \to c} \mathbb{E}\left[\tau_i - \tau\left(c\right) \middle| X_i = x\right] = \lim_{x \to c} \tau\left(x\right) - \tau\left(c\right) = 0$$

Therefore, under the assumption that

- $\tau_i \perp D_i \middle| X_i$.

- $\mathbb{E}\left[Y_i\left(0\right)\middle|\, X_i = x\right], \mathbb{E}\left[Y_i\left(1\right)\middle|\, X_i = x\right]$ are both continuous at $x = c$.

Then it follows that

$$\beta = \frac{\lim_{\varepsilon\downarrow 0}\mathbb{E}\left[Y_i\middle|\, X_i = c+\varepsilon\right] - \mathbb{E}\left[Y_i\middle|\, X_i = c-\varepsilon\right]}{\lim_{\varepsilon\downarrow 0}\mathbb{E}\left[D_i\middle|\, X_i = c+\varepsilon\right] - \mathbb{E}\left[D_i\middle|\, X_i = c-\varepsilon\right]} = \frac{\tau\left(c\right)\left[1-0\right]}{1-0} = \mathbb{E}\left[\tau_i\middle|\, X_i = c\right]$$

In practice, the assumption $\tau_i \perp D_i\middle|\, X_i$ is hard to justify. Instead, we can assume that

- $D_i(x) = \mathbb{I}\left\{x_i + \gamma \mathbb{I}_{\{x_i \geqslant c\}} \geqslant Z_i\right\}$.
- $\{\tau_i, Z_i\} \perp X_i, X_i \in [c-\varepsilon, c+\varepsilon]$.

Since

$$Y_i = Y_i\left(1\right) D_i + \left(1 - D_i\right) Y_i\left(0\right) = Y_i\left(0\right) + \tau_i \cdot D_i$$

then

$$\mathbb{E}\left[Y_i\middle|\, X_i = x\right] = \mathbb{E}\left[Y_i\left(0\right)\middle|\, X_i = x\right] + \mathbb{E}\left[\tau_i D_i\middle|\, X_i = x\right]$$

thus,

$$\begin{aligned}
&\mathbb{E}\left[Y_i\middle|\, X_i = c+\varepsilon\right] - \mathbb{E}\left[Y_i\middle|\, X_i = c-\varepsilon\right] \\
=&\mathbb{E}\left[Y_i\left(0\right)\middle|\, X_i = c+\varepsilon\right] + \mathbb{E}\left[\tau_i D_i\middle|\, X_i = c+\varepsilon\right] - \mathbb{E}\left[Y_i\left(0\right)\middle|\, X_i = c-\varepsilon\right] - \mathbb{E}\left[\tau_i D_i\middle|\, X_i = c-\varepsilon\right] \\
\to&\mathbb{E}\left[\tau_i D_i\middle|\, X_i = c+\right] - \mathbb{E}\left[\tau_i D_i\middle|\, X_i = c-\right] \text{ as } x \to c
\end{aligned}$$

By the assumption that $D_i\left(x\right) = \mathbb{I}\left\{x_i + \gamma \mathbb{I}_{\{x_i \geqslant c\}} \geqslant Z_i\right\}$,

$$\begin{aligned}
&\mathbb{E}\left[\tau_i D_i\middle|\, X_i = c+\varepsilon\right] - \mathbb{E}\left[\tau_i D_i\middle|\, X_i = c-\varepsilon\right] \\
=&\mathbb{E}\left[\tau_i D_i\left(c+\varepsilon\right)\middle|\, X_i = c+\varepsilon\right] - \mathbb{E}\left[\tau_i D_i\left(c-\varepsilon\right)\middle|\, X_i = c-\varepsilon\right] \\
=&\mathbb{E}\left[\tau_i D_i\left(c+\varepsilon\right)\right] - \mathbb{E}\left[\tau_i D_i\left(c-\varepsilon\right)\right] \\
=&\mathbb{E}\left[\tau_i\left(D_i\left(c+\varepsilon\right) - D_i\left(c-\varepsilon\right)\right)\right] \\
=&\mathbb{P}\left\{D_i\left(c+\varepsilon\right) - D_i\left(c-\varepsilon\right) = 1\right\}\mathbb{E}\left[\tau_i\middle|\, D_i\left(c+\varepsilon\right) - D_i\left(c-\varepsilon\right) = 1\right]
\end{aligned}$$

where by "no defiers assumption", $\mathbb{P}\left\{D_i\left(c+\varepsilon\right) - D_i\left(c-\varepsilon\right) = -1\right\} = 0$.

Moreover,

$$\begin{aligned}
\mathbb{E}\left[D_i\middle|\, X_i = c+\varepsilon\right] - \mathbb{E}\left[D_i\middle|\, X_i = c-\varepsilon\right] &= \mathbb{E}\left[D_i\left(c+\varepsilon\right) - D_i\left(c-\varepsilon\right)\right] \\
&= \mathbb{P}\left\{D_i\left(c+\varepsilon\right) - D_i\left(c-\varepsilon\right)\right\}
\end{aligned}$$

Then

$$\begin{aligned}
\beta &= \frac{\lim_{\varepsilon\downarrow 0}\mathbb{E}\left[Y_i\middle|\, X_i = c+\varepsilon\right] - \mathbb{E}\left[Y_i\middle|\, X_i = c-\varepsilon\right]}{\lim_{\varepsilon\downarrow 0}\mathbb{E}\left[D_i\middle|\, X_i = c+\varepsilon\right] - \mathbb{E}\left[D_i\middle|\, X_i = c-\varepsilon\right]} \\
&= \mathbb{E}\left[\tau_i\middle|\, \underbrace{D_i\left(c+\varepsilon\right) - D_i\left(c-\varepsilon\right) = 1}_{\text{compliers}}\right]
\end{aligned}$$

which is the LATE for the compliers.