Matrix Algebra I

1 Terminology

Matrix algebra is widely used in econometrics. The course will follow the appendix of Green(2003). Besides, the course will use Curtis(2012) as the supplement. There are many books which contain much books contain more advanced theory about matrix algebra, such as Magnus and Neudecker(2002).

Definition. A *field* is a mathematical system F consisting of a nonempty set F together with two operations, addition and multiplication\, which assign to each pair of elements $\alpha, \beta \in F$ uniquely determined elements $\alpha + \beta$ and $\alpha\beta$ of F, such that the following conditions are satisfied, for $\alpha, \beta, \gamma \in F$.

1. $\alpha + \beta = \beta + \alpha$, $\alpha\beta = \beta\alpha$.
2. $\alpha + (\beta + \gamma) = (\alpha + \beta) + \gamma$, $(\alpha\beta)\,\gamma = \alpha\,(\beta\gamma)$.
3. $\alpha\,(\beta + \gamma) = \alpha\beta + \alpha\gamma$.
4. $\exists 0 \in F$, such that$\alpha + 0 = \alpha$,$\forall \alpha \in F$.
5. $\forall \alpha \in F$, $\exists -\alpha \in F$, such that $\alpha + (-\alpha) = 0$.
6. $\exists 1 \in F$, such that $1 \neq 0$, $\alpha 1 = \alpha$, $\forall \alpha \in F$.
7. $\forall \alpha \in F$, $\alpha \neq 0$, $\exists \alpha^{-1} \in F$, such that $\alpha\alpha^{-1} = 1$.

Exercise. 1. Examine that the real number system $\mathbb{R}$ is a field.
2. Verify that $Z = \{\dots, -2, -1, 0, 1, 2, \dots\}$ is not a field.

Definition. A $n \times 1$ *vector* $\boldsymbol{a}$ *over* F is an ordered set of n elements arranged either in a column. Let a_i be the ith component of the vecor $\boldsymbol{a}$, then$\boldsymbol{a}$ can be denote as $\boldsymbol{a} = (a_1, a_2, \dots, a_n)'$, $a_i \in F$ for $i = 1, 2, \dots, n$.

Since the field of real number $\mathbb{R}$ is a specific example of field, the results holds over the general field F are also applicable over the field $\mathbb{R}$.

Definition. A $m \times n$ *matix* A *over field* F is a rectangular array of elements in m rows and n columns with each element belongs to F, and $m \times n$ is its dimension.

If we denote the element at ith row and jth column as a_{ij}, the matrix $A \in F^{m\times n}$ can be represented as:

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} = (a_{ij})\,.$$

Sometimes, we write the matrix in a vector form, that is, $\boldsymbol{A} = (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)$, $\boldsymbol{a}_i = (a_{1i}, a_{2i,}, \ldots, a_{mi})'$.

In view of the preceding, a row vector is also a matrix with one row, whereas a column vector is a matrix with one column. A matrix can also be viewed as a set of column vectors or as a set of row vectors.

Based on the different properties of the matrix, there are some terminology defining a corresponding type of matrix. Given the matrix $m \times n$ matrix $\boldsymbol{A}$,

- **Symmetric matrix**: for $i = 1, 2, ..., n$ and $j = 1, 2, ..., n$, $a_{ij} = a_{ji}$, $m = n$.
- **Square matrix**: the dimension of the matrix satisfyes $m = n$.
- **Diagonal matrix**: a diagonal matrix is a square matrix whose only nonzero elements appear on the main diagonal, that is, moving from upper left to lower right.
- **Identity matrix**: a diagonal matrix with the same value, which is one, on the main diagonal. It is always denote by $\boldsymbol{I}$.
- **Scalar matrix**: a diagonal matrix with the same value, λ, in all diagonal elements, which can be written as $\boldsymbol{A} = \lambda \boldsymbol{I}$.
- **Triangular matrix**: one that has only zeros either above or below the main diagonal. If the zeros are above the diagonal, the matrix is **lower triangular**. Otherwise, it is **upper triangular**.
- **Idempotent Matrix**: An idempotent matrix $\boldsymbol{A}$, is the one that satisfies $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{A}$.

2 Algebraic Manipulation and Application

2.1 Basic Algebraic Manipulation for Matrices

Similar to the algebra manipulation of the scalar variable, there are also corresponding manipulation for the matrix. They are defined as follows.

- **Equality of matrix**: $\boldsymbol{A} = \boldsymbol{B}$ if and only if $a_{ij} = b_{ij}$, for all i and j.
- **Transposition**: Given the $m \times n$ matrix $\boldsymbol{A}$, its transposition can be written as $\boldsymbol{A}'$, and

$$\boldsymbol{A}' = \begin{pmatrix} a_{11} & a_{m1} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{pmatrix} = (a_{ji})\,.$$

If $\boldsymbol{A}$ is symmetric, $\boldsymbol{A}' = \boldsymbol{A}$. For a vector $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)'$, $\boldsymbol{a}' = (a_1, a_2, \ldots, a_n)$.

- **Addition**: Given two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ with same dimension $m \times n$, the addition of this two matrix is defined as

$$\boldsymbol{A} + \boldsymbol{B} = (a_{ij} + b_{ij}) \ for \ i = 1, 2, \ldots, m, j = 1, 2, \ldots, n. \tag{1}$$

  Matrices cannot be added unless they have the same dimensions, in which case they are said to be **conformable for addition**. A **zero matrix** or **null matrix** is one whose elements are all zero. In the addtion of matrices, the zero matrix plays the same role as the scalar 0 in scalar addition; that is,

$$\boldsymbol{A} + \boldsymbol{0} = \boldsymbol{A}.$$

Exercise. Based on (1), examine the following properties.

1. Commutative: $\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{B} + \boldsymbol{A}$.
2. Associative:$(\boldsymbol{A} + \boldsymbol{B}) + \boldsymbol{C} = \boldsymbol{A} + (\boldsymbol{B} + \boldsymbol{C})$.
3. $(\boldsymbol{A} + \boldsymbol{B})' = \boldsymbol{A}' + \boldsymbol{B}'$.

- **Multiplication**: For a $m \times K$ matrix $\boldsymbol{A}$ and a $K \times n$ matrix $\boldsymbol{B}$, the product matrix, $\boldsymbol{C} = \boldsymbol{AB}$, is an $m \times n$ matrix with

$$\underbrace{\boldsymbol{C}}_{m \times n} = \underbrace{\boldsymbol{A}}_{m \times K} \underbrace{\boldsymbol{B}}_{K \times n} \Leftrightarrow c_{ij} = \sum_{k=1}^{K} a_{ik} b_{kj}. \tag{2}$$

  To multiply two matrices, the number of columns in the first must be the same as the number of rows in the second, in which case they are **conformable for multiplication**. Mulplication of matrices is generally not commutative. In some cases, $\boldsymbol{AB}$ may exist, even if $\boldsymbol{AB}$ and $\boldsymbol{BA}$ do have the same dimensions, they will not be equal(You can examin this based on (2).

Example. If

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}, \ \boldsymbol{B} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix},$$

then,

$$\boldsymbol{AB} = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

However,

$$\boldsymbol{BA} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix},$$

hence,

$$\boldsymbol{AB} \neq \boldsymbol{BA}.$$

- **Scalar Multiplication**: the operation of multiplying every element of the matrix by a given scalar. For scalar $c \in F$ and matrix $A \in$,

$$c\boldsymbol{A} = (ca_{ij}), \ c \in R. \tag{3}$$

Exercise. Please examine the following properties,

1. Associative law: $(\boldsymbol{AB})\,\boldsymbol{C} = \boldsymbol{A}\,(\boldsymbol{BC})$.
2. Distributive law: $\boldsymbol{A}\,(\boldsymbol{B}+\boldsymbol{C}) = \boldsymbol{AB}+\boldsymbol{AC}$.
3. Transpose of a product: $\boldsymbol{(AB)'} = \boldsymbol{B'A'}, \boldsymbol{(ABC)'} = \boldsymbol{C'B'A'}$.

2.2 Some Useful Applications

Based the above algbric manipulations, the product of a $m \times n$ matrix X and a $n \times 1$ vector β is written

$$\mathbf{y} = \boldsymbol{X\beta} = \sum_{i=1}^{n} \boldsymbol{X}_i \beta_i,$$

where $\boldsymbol{X}_i$ is the ith column vector of $\boldsymbol{X}$; the number of elements in $\boldsymbol{\beta}$ must equal the number of the columns in $\boldsymbol{X}$; and the result $\mathbf{y}$ is a vector with number of elements equal to the number of rows in X, $m \times 1$.

Example. The equations

$$\begin{aligned} 5 &= 4a + 2b + c \\ 4 &= 2a + 6b + c \\ 1 &= a + b \end{aligned}$$

can be written as,

$$\begin{pmatrix} 5 \\ 4 \\ 1 \end{pmatrix} = \begin{pmatrix} 4 & 2 & 1 \\ 2 & 6 & 1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} \Leftrightarrow \boldsymbol{y} = \boldsymbol{X\beta},$$

which is the usual form of regression model in econometrics. And we can see that the model can be represented in a concise form now.

We can easily derive that for any matrix or vector $\boldsymbol{A}$, $\boldsymbol{AI} = \boldsymbol{A}, \boldsymbol{IA} = \boldsymbol{A}$, $A\mathbf{0} = \mathbf{0}, \mathbf{0}A = \mathbf{0}$. only when $\boldsymbol{A}$ is a square matrix, $\boldsymbol{IA} = \boldsymbol{AI} = \boldsymbol{A}, \mathbf{0}\boldsymbol{A} = \boldsymbol{A}\mathbf{0} = \mathbf{0}$.

Usually we always need to sum up the values of a vector of a column vector in a matrix. When the dimension of the target vector is very high, we can use matrix to simplify the representation. Define a $n \times 1$ vector,

$$\iota = \underbrace{(1, 1, \ldots, 1)'}_{n \text{ elements}}.$$

For a vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)\,, \boldsymbol{y} = (y_1, y_2, \ldots, y_n)$, we can have following useful representation,

- $\sum_{i=1}^{n} x_i = \iota' x$.
- If $x = \lambda\iota, \lambda \in R$, then $\sum_{i=1}^{n} x_i = \iota'\iota\lambda = n\lambda$.
- Given $\lambda \in \mathbb{R}$, $\sum_{i=1}^{n} \lambda x_i = \lambda \sum_{i=1}^{n} x_i = \lambda\iota'\mathbf{x}$. In particular, if $\lambda = 1/n$, then $\bar{x} = \iota'\mathbf{x}/n$.

- $\sum_{i=1}^{n} x_i^2 = \mathbf{x}'\mathbf{x}$.
- $\sum_{i=1}^{n} x_i y_i = \mathbf{x}'\mathbf{y}$.

Furthermore, given the matrix $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1' \\ \boldsymbol{X}_2' \\ \vdots \\ \boldsymbol{X}_n' \end{pmatrix}$, $\boldsymbol{X}_i = (x_{i1}, x_{i2}, \ldots, x_{in})'$ *for* $i = 1, 2, \ldots, m$,

we have,

$$\boldsymbol{X'X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n) \begin{pmatrix} \boldsymbol{X}_1' \\ \boldsymbol{X}_2' \\ \vdots \\ \boldsymbol{X}_n' \end{pmatrix} = \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i'.$$

A fundamental matrix in statistics is the one that is used to transform data to deviations from their mean, that is, $\mathbf{x} - \iota\overline{x}$. It can be written as

$$\mathbf{x} - \iota\overline{x} = I\mathbf{x} - \frac{\iota\iota'}{n}x = \left(I - \frac{\iota\iota'}{n}\right)\mathbf{x} = \boldsymbol{M}^0\mathbf{x}. \tag{4}$$

Henceforth, the symbol $\boldsymbol{M}^0$will be used only for this matrix. Its diaonal elements are all $(1 - 1/n)$, and its off-diagonal elements are $-1/n$, which implies that the matrix are **symmetric**. The matrix $\boldsymbol{M}^0$ is primarily useful in computing sums of squared deviations. Some conputations are simplified by the following results, which is left as an excercise.

Exercise. Prove the following equations,

1. $\boldsymbol{M}^0\iota = \mathbf{0}$.
2. (**Idempotent Matrix**)$\boldsymbol{M}^{0\prime}\boldsymbol{M}^0 = \boldsymbol{M}^0$.

Hence, $\iota' M^0 = \mathbf{0}'$ and further,

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \iota' \left(\boldsymbol{M^0}\mathbf{x}\right) = \left(\iota'\boldsymbol{M}^0\right)\mathbf{x} = \mathbf{0}'\mathbf{x} = 0.$$

Thus the sum of the squared deviations about the mean is,

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = (\mathbf{x} - \bar{x}\iota)'(\mathbf{x} - \bar{x}\iota) = \left(\boldsymbol{M}^0 x\right)'\boldsymbol{M}^0\mathbf{x} = \mathbf{x}'\boldsymbol{M}^{0\prime}\boldsymbol{M}^0\mathbf{x} = \mathbf{x}'\boldsymbol{M}^0\mathbf{x}. \tag{5}$$

For two vectors x and y,

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \left(\boldsymbol{M}^0\mathbf{x}\right)'\left(\boldsymbol{M}^0\mathbf{y}\right) = \mathbf{x}'\boldsymbol{M}^0\mathbf{y}. \tag{6}$$

So,

$$\begin{aligned} \begin{pmatrix} \sum_{i=1}^{n}(x_i - \bar{x})^2 & \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\ \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) & \sum_{i=1}^{n}(y_i - \bar{y})^2 \end{pmatrix} &= \begin{pmatrix} \mathbf{x}'\boldsymbol{M}^0\mathbf{x} & \mathbf{x}'\boldsymbol{M}^0\mathbf{y}. \\ \mathbf{y}'\boldsymbol{M}^0\mathbf{x}. & \mathbf{y}'\boldsymbol{M}^0\mathbf{y} \end{pmatrix} \\ &= \left(\boldsymbol{M}^0\boldsymbol{Z}\right)'\left(\boldsymbol{M}^0\boldsymbol{Z}\right) \\ &= \boldsymbol{Z}'\boldsymbol{M}^0\boldsymbol{Z}, \end{aligned}$$

where $\boldsymbol{Z} = (\mathbf{x}, \mathbf{y})$.

3 Geometry Interpretation

3.1 Vector Space, Basis and Linear Independence

As we have said, the vector can be treated as matrix, thus the addition and scalar multiplication of the vector is the same as defined for the matrix.

Definition. Let F be an arbitrary field. A *vector space* V *over* F is a nonempty set V of objects $\boldsymbol{v}$, called *vectors*, together with two operatons, one of which assigns to each pair of vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ a vector $\boldsymbol{v}+\boldsymbol{w}$ called *sum* of $\boldsymbol{v}$ and $\boldsymbol{w}$, and the other of which assigns to each element $\alpha \in F$ and each vector $\boldsymbol{v} \in V$ a vector $\alpha\boldsymbol{v}$ called the *product* of $\boldsymbol{v}$ by the element $\alpha \in F$. The operations are assumed to satisfy the following axioms, for $\alpha, \beta \in F$ and $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in V$.

1. $\boldsymbol{u}+(\boldsymbol{v}+\boldsymbol{w})=(\boldsymbol{u}+\boldsymbol{v})+\boldsymbol{w}$ and $\boldsymbol{u}+\boldsymbol{v}=\boldsymbol{v}+\boldsymbol{u}$.
2. $\exists \mathbf{0} \in V$, such that $\boldsymbol{u}+\mathbf{0}=\boldsymbol{u}$, $\forall \boldsymbol{u} \in V$.
3. $\forall \boldsymbol{u} \in V$, $\exists -\boldsymbol{u} \in V$, such that $\boldsymbol{u}+(-\boldsymbol{u})=\mathbf{0}$.
4. $\alpha(\boldsymbol{u}+\boldsymbol{v})=\alpha\boldsymbol{u}+\alpha\boldsymbol{v}$.
5. $(\alpha+\beta)\boldsymbol{v}=\alpha\boldsymbol{v}+\beta\boldsymbol{v}$.
6. $(\alpha\beta)\boldsymbol{u}=\alpha(\beta\boldsymbol{u})$.
7. $1\boldsymbol{u}=\boldsymbol{u}$.

The usual common vector space is the $\mathbb{R}^n$, which is defined in the following.

Definition. The *vector space* $\mathbb{R}^n$ over the field of real numbers $\mathbb{R}$ is the algebraic system consisting of all $n \times 1$ vector $\boldsymbol{a}$ with $a_i \in \mathbb{R}$. $\forall \boldsymbol{a}=(a_1,\ldots,a_n), \boldsymbol{b}=(b_1,\ldots,b_n) \in \mathbb{R}^n$, $\boldsymbol{a}=\boldsymbol{b}$ if and only if $a_i=b_i$, for $i=1,2\ldots,n$. And it includes operations *sum*,$\boldsymbol{a}+\boldsymbol{b}=(a_1+b_1,\ldots,a_n+b_n) \in \mathbb{R}^n$ and *product*, $\forall \lambda \in \mathbb{R}$, $\lambda\boldsymbol{a}=(\lambda a_1,\ldots,\lambda a_n) \in \mathbb{R}^n$.

Exercise. Examine the space $\mathbb{R}^n$ is a vector space.

Example. For a special case, $\mathbb{R}^2$, given two vectors $\boldsymbol{\alpha}=(1,0)'$, $\boldsymbol{\beta}=(0,1)'$, $\forall \gamma \in \mathbb{R}^2$, we can always find $a \in R$ and $b \in R$ that satisfy $\boldsymbol{\gamma}=a\boldsymbol{\alpha}+b\boldsymbol{\beta}$. As it suggests, any vector in $\mathbb{R}^2$ can be represented as a linear combination of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

Definition. A *subspace* S of the vector space V is a nonempty set of vectors in V such that:

- If $\boldsymbol{a}$ and $\boldsymbol{b}$ are in S, then $\boldsymbol{a}+\boldsymbol{b} \in S$.
- If $\boldsymbol{a} \in S$ and $\lambda \in F$, then $\lambda\boldsymbol{a} \in S$.

The definition for a vector space are clearly satisfied for any subspace S of a vector space V.

Definition. Let $(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)$ be a set of vectors in V. A vector $\boldsymbol{b} \in V$ is said to be a *linear combination* of $(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)$ if $\exists \lambda_1, \ldots, \lambda_n \in F$ and not all equal to zero, such that

$$\boldsymbol{b}=\lambda_1\boldsymbol{a}_1+\cdots+\lambda_n\boldsymbol{a}_n.$$

Exercise. If S is a subspace of V containing the vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$, then every linear combination of $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ belongs to S.

Definition. Given a set of vectors $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)$, $S = \{\boldsymbol{a} : \boldsymbol{a} = \sum_{i=1}^{n} \lambda_i \boldsymbol{X}_i, \lambda_i \in F\}$ is a subspace *spanned* by $\boldsymbol{X}$, denote $S = \langle \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \rangle$ or $S = \langle \boldsymbol{X} \rangle$. A subspace S is called *finitely spanned* if $\exists \boldsymbol{s}_1, \ldots, \boldsymbol{s}_k$ such that $S = \langle \boldsymbol{s}_1, \ldots, \boldsymbol{s}_k \rangle$.

Lemma. *Let $(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)$ be a set of vectors in V, for $n \geq 1$. Some vector $\boldsymbol{a}_i$ can be expressed as a linear combination of the remaining vectors $(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_{i-1}, \boldsymbol{a}_{i+1}, \ldots, \boldsymbol{a}_n)$ if and only if $\exists \mu_1, \ldots, \mu_m \in F$, not all equal zero, such that $\sum_{i=1}^{n} \mu_i \boldsymbol{a}_i = 0$.*

Definition. A set of vectors $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_n)$ is *linearly independent* if and only if the only solution to

$$\lambda_1 \boldsymbol{\beta}_1 + \lambda_2 \boldsymbol{\beta}_2 + \cdots + \lambda_n \boldsymbol{\beta}_n = 0,\ \lambda_i \in F,$$

is $\lambda_1 = \lambda_2 = \cdots = \lambda_n = 0$. It is *linearly dependent* if and only if $\{\lambda_i\}$ not all equal to zero.

Lemma. *Let S be a subspace of a vector space V over a field F, such that S is spanned by n vectors $(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)$. Suppose $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m)$ are vectors in S with $m > n$. Then the vectors $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m)$ are linearly dependent.*

Proof. For $n = 1$, $S = \langle \boldsymbol{a} \rangle$, $\boldsymbol{b}_i = \lambda_i \boldsymbol{a}$, for $i = 1, 2, \ldots, m$. Then, at least one $\lambda_i \neq 0$, otherwise, $\boldsymbol{b}_i = 0, i = 1, \ldots, m$. Thus, assume $\lambda_j \neq 0$,

$$\lambda_j \boldsymbol{b}_1 + 0\boldsymbol{b}_2 + \cdots + (-\lambda_1)\, \boldsymbol{b}_j + 0\boldsymbol{b}_{j+1} + \cdots + 0\boldsymbol{b}_n = \lambda_j \lambda_1 \boldsymbol{a} - \lambda_1 \lambda_j \boldsymbol{a} = \boldsymbol{0},$$

which means that $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m)$ are linearly dependent.

Now suppose the theorem is true for subspaces spanned by $n-1$ vectors, and consider m distinct vectors$(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m)$ in S, with $m > n$. Then, $\boldsymbol{b}_i = \sum_{j=1}^{n} \lambda_{ij} \boldsymbol{a}_j$.

Without loss of generality, let $\lambda_{i1} = 0$, then the terms involving $\boldsymbol{a}_1$ are all missing for $\boldsymbol{b}_j$. $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m)$ belongs to the subspace $\langle \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n \rangle$. Since $m > n > n-1$, the induction hypothesis implies that $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m)$ are linear dependent.

Assume for some i, $\lambda_{i1} \neq 0$, more specifically, let $i = 1$. Then the coefficient of $\boldsymbol{a}_1$ in $\boldsymbol{b}_2 - \lambda_{21}\lambda_{11}^{-1}\boldsymbol{b}_1$ is $\lambda_{21} - \lambda_{21}\lambda_{11}^{-1}\lambda_{11} = 0$, so that, for each j, $\boldsymbol{b}_j - \lambda_{j1}\lambda_{11}^{-1}\boldsymbol{b}_1$ belongs to $\langle \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n \rangle$. And from the induction hypothesis, $m - 1 > n - 1$, $\exists \mu_2, \mu_3, \ldots, \mu_m \in F$, not all zero such that

$$\mu_2 \left(\boldsymbol{b}_2 - \lambda_{21}\lambda_{11}^{-1}\boldsymbol{b}_1\right) + \cdots + \mu_m \left(\boldsymbol{b}_m - \lambda_m \lambda_{11}^{-1}\boldsymbol{b}_1\right) = 0.$$

It can be rewritten as

$$\left[\left(-\mu_2 \lambda_{21}\lambda_{11}^{-1}\right) + \cdots + \left(-\mu_m \lambda_{m1}\lambda_{11}^{-1}\right)\right] \boldsymbol{b}_1 + \mu_2 \boldsymbol{b}_2 + \cdots + \mu_m \boldsymbol{b}_m = 0,$$

since $\mu_2, \mu_3, \ldots, \mu_m$ are not all zero, then $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m)$ are linearly dependent. □

Lemma. *Let S be a subspace of a vector space V, and suppose that $(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)$ and $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m)$ are both sets of generators of S, which are linearly independet. Then $n = m$.*

Exercise. Prove the above lemma.

Definition. A finite set of vectors A set of vectors $(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)$ is said to be a *basis* of a vector space V if $(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)$ is a linearly independent set of generators of V.

Obviously, in the above case, $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is a basis for $\mathbb{R}^2$. There are also other basis for this vector space. Generally, for $\mathbb{R}^n$, a specific basis is that $E = (\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n)$ with $\boldsymbol{e}_i = \left(0, 0, \ldots, \underbrace{1}_{ith}, \ldots, 0\right)'$. Then $\forall \boldsymbol{\beta} \in \mathbb{R}^n$, $\exists b \in \mathbb{R}^n$, $\boldsymbol{\beta} = Eb$. For a specific vector space, the basis is not unique. However, given the basis, the representation of a given vector is unique.

Definition. Suppose V is a vector space which has a basis $(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)$. The uniquely determined number of basis vectors, n, is called the *dimension* of the vector space V ($\dim V = n$).

Example. By definition, the space spanned by a basis for $\mathbb{R}^n$ is $\mathbb{R}^n$. An implication of this is that if a and b bare a basis for $\mathbb{R}^2$ and c is another vector in $\mathbb{R}^2$, the space spanned by (a, b, c) is, again, $\mathbb{R}^2$.

And consider the set of three coordinate vectors whose third element is zero. In particular,

$$\boldsymbol{a} = (a_1, a_2, 0)' \text{ and } \boldsymbol{b} = (b_1, b_2, 0)'.$$

If the spanned vector space $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \left\{ \boldsymbol{x} : \boldsymbol{x} = (x_1, x_2, x_3) \in \mathbb{R}^3, x_3 = 0 \right\}$. This area is a **subspace**, a two-dimensional subspace.

Exercise. Let $(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)$ be a basis for a vector space V. Prove that $\boldsymbol{a}_i \neq \boldsymbol{0}$ for $i = 1, \ldots, n$.

Lemma. *If $(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)$ is linearly dependent and if $(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_{n-1})$ is linearly independent, then $\boldsymbol{a}_n$ is a linear combination of $(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_{n-1})$.*

Proposition. *Let $V = \langle \boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n \rangle$ be a finitely generated vector space with genenrators $(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)$. Then a basis for V can be selected from among the set of generators $(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)$. In other words, a set of generators for a finitely generated vector space always contains a basis.*

3.2 Rank of a Matrix

We view a matrix as a set of column vectors. The number of columns in the matrix equals the number of vectors in the set, and the number of rows equals the number of coordinates in each column vector.

Definition. The *column space* of a matrix is the vector space that is spanned by its column vectors.The *row space* of a matrix is the vector space that is spanned by its row vectors.

If the matrix $\boldsymbol{X} = (x_{ij})_{m \times n}$, it can be written as $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)$, where $\boldsymbol{X}_i = (x_{1i}, x_{2i}, \ldots, x_{mi})' \in \mathbb{R}^{m \times 1}$. Then the column space of the matrix $\boldsymbol{X}$ is $\langle \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \rangle$. The column space might have n dimensions. But, as we have seen, it might have fewer dimensions; the column vectors migth be linearly dependent, or there might be fewer rows than n.

Example. The matrix

$$A = \begin{pmatrix} 1 & 5 & 6 \\ 2 & 6 & 8 \\ 7 & 1 & 8 \end{pmatrix}$$

has three vectors from $\mathbb{R}^3$, but the third is the sum of the first two, so the column space of the matrix cannot have three dimensions. Nor does it have only one, since the three columns are not all scalar multiplies of one another. Hence, the column space is a two-dimensional subspace of $\mathbb{R}^3$. Similar to the row space, it can easily find the dimension of the row space is also two.

Definition. Let $\boldsymbol{A}$ be an $m \times n$ matrix with column vector $\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n\}$. The *rank* of the matrix,$rank\,(\boldsymbol{A})$, is defined as the dimension of the column space $\langle \boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n \rangle$.

Actually, we can also conclude that the rank of the matrix is equal to the dimension of the row sapce. And it holds regardless of the actual row and column rank, which means $rank\,(\boldsymbol{A}) = rank\,(\boldsymbol{A}')$.

Definition. For an $m \times n$ matrix $\boldsymbol{A}$, if $rank\,(\boldsymbol{A}) = m$, then the matrix is said to have *full row rank*. Similarly, if $rank\,(\boldsymbol{A}) = n$, it has *full column rank*.

Furthermore, we have,

$$rank\,(\boldsymbol{A}) \leq \min(row\ number, column\ number). \tag{7}$$

Proposition. *For the rank of the matrix, the following results hold, let* $\boldsymbol{A}$ *be* $m \times K$ *matrix and* $\boldsymbol{B}$*,* $K \times n$*.*

$$rank\,(\boldsymbol{AB}) \leq \min\left(rank\,(\boldsymbol{A}), rank\,(\boldsymbol{B})\right).$$

Proof. For $\boldsymbol{C} = \boldsymbol{AB}$, if $\boldsymbol{C} = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n)$, $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K)$, then $\boldsymbol{c}_i = \sum_{i=1}^{K} b_{i1}\boldsymbol{a}_i$, for $i = 1, \ldots, n$, which means the column space is spanned by the column vectors of $\boldsymbol{A}$. Hence, $rank\,(\boldsymbol{C}) \leq rank\,(\boldsymbol{A})$. Apply the same logic to the rows of $\boldsymbol{C}$, which are all linear combinations of the rows of $\boldsymbol{B}$, $rank\,(\boldsymbol{C}) \leq rank\,(\boldsymbol{B})$. Therefore,

$$rank\,(\boldsymbol{AB}) \leq \min\left(rank\,(\boldsymbol{A}), rank\,(\boldsymbol{B})\right). \tag{8}$$

□

Further we have ,

- If $\boldsymbol{B}$ *is* $K \times K$ matrix of rank K*,* then $rank\,(\boldsymbol{AB}) = rank\,(\boldsymbol{A})$.
- $rank\,(\boldsymbol{A}) = rank\,(\boldsymbol{A'A}) = rank\,(\boldsymbol{AA'})$.

3.3 Determinant of Matrix

The determinant is only defined for square matrix. It is a function of the elements of the matrix. There are various definitions, most of which are not useful for our work. Determinants figure into our results in several ways, however, that we can enumerate before we need formally to define the computations.

In this section, the determinant is defined as a rule which assigns to matrix a number which tells something about the its behavior.

Definition. Let F be an arbitrary field. A *determinant* is a function which assigns to a matrix $\boldsymbol{A} = (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)$ in $F^{n\times n}$ an element of F, $D\,(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)\,(or, |\boldsymbol{A}|, |(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)|) : F^{n\times n} \mapsto F$, such that

- $|(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{i-1}, \boldsymbol{a}_i + \boldsymbol{a}_j, \boldsymbol{a}_{i+1} \ldots, \boldsymbol{a}_n)| = |(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)|$, for $i \leq n$ and $j \neq i$.
- $|(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{i-1}, \lambda\boldsymbol{a}_i, \boldsymbol{a}_{i+1} \ldots, \boldsymbol{a}_n)| = \lambda |(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)|$, for all $\lambda \in F$.
- $|(\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n)| = 1$, if $\boldsymbol{e}_i = \left(0, 0, \ldots, \underbrace{1}_{ith}, \ldots, 0\right)$.

Proposition. *Let D and D' be two functions satisfying the definition of determinant, then for all $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n \in F^n$,*

$$D(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n) = D'(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)$$

The proposition establishes the uniqueness of determinant.

Proposition. *There exists a function $D(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)$ satisfying the definition of determinant. Further,*

$$\begin{aligned} D(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n) &= \sum_{j_1=1}^{n} \cdots \sum_{j_n=1}^{n} a_{1j_1} \ldots a_{nj_n} D(\boldsymbol{e}_{j_1}, \boldsymbol{e}_{j_2}, \ldots, \boldsymbol{e}_{j_n}) \\ &= \sum_{j_1, \ldots, j_n} a_{1j_1} \ldots a_{nj_n} D(\boldsymbol{e}_{j_1}, \boldsymbol{e}_{j_2}, \ldots, \boldsymbol{e}_{j_n}). \end{aligned}$$

where it is understood that the sum is taken over the $n!$ possible choices of $\{j_1, \ldots, j_n\}$ in which all the j_i's are distinct.

Example. There are some settings in which the value of the determinant is also of interest, so we now consider some algebraic results. Consider a matrix,

$$\boldsymbol{D} = \begin{pmatrix} d_1 & 0 & \ldots & 0 \\ 0 & d_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & d_n \end{pmatrix}.$$

The determinant is simply the product of the diagonal values, which is,

$$|\boldsymbol{D}| = \prod_{i=1}^{n} d_i.$$

A special case is the identity matrix, $|\boldsymbol{I}| = 1$. Let $\boldsymbol{C} = diag(c_1, c_2, \ldots, c_n)$, then,

$$|\boldsymbol{DC}| = \prod_{i=1}^{n} c_i d_i = \prod_{i=1}^{n} c_i \prod_{j=1}^{n} d_j,$$

which can be written as,

$$|\boldsymbol{DC}| = |\boldsymbol{D}| \, |\boldsymbol{C}|. \tag{9}$$

And then we have

$$|c\boldsymbol{D}| = c^n |\boldsymbol{D}|. \tag{10}$$

Actually, we have the following result.

Proposition. *Let* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *be* $n \times n$ *matrices, then*

- $|\boldsymbol{A}| = |\boldsymbol{A}'|$.
- $|\boldsymbol{AB}| = |\boldsymbol{A}|\,|\boldsymbol{B}|$.

For matrices more than two dimensions, the determinant can be obtained by using an *expansion by cofactors.* Using any row, say, i, we obtain

$$\boldsymbol{A} = \sum_{j=1}^{n} a_{ij}\,(-1)^{i+k}\,|\boldsymbol{A_{ij}}|\,,$$

where $\boldsymbol{A}_{ij}$ is called *cofactors,* the matrix obtained from $\boldsymbol{A}$ by deleting row i and column j. The determinant of the $\boldsymbol{A}_{ij}$ is called a **minor** of $\boldsymbol{A}$.

Exercise. Calculate the following resutls,

1. $\begin{vmatrix} a & c \\ b & d \end{vmatrix} = ?$.
2. $\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = ?$

Combined with the rank of the matrix, the following proposition holds.

Proposition. *For a* $n \times n$ *matrix* $\boldsymbol{A}$, $|\boldsymbol{A}| \neq 0$ *if and only if rank* $(\boldsymbol{A}) = n$, *which means* $\boldsymbol{A}$ *is full of rank.*

3.4 A Least Squares Problem

For a regression model,

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e},$$

$\boldsymbol{y}$ is the $m \times 1$ vector of independent varables, the $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ with $\boldsymbol{X}_i \in F^{m\times 1}$, $\boldsymbol{\beta}$ the $n \times 1$ parameter that we are interested, $\boldsymbol{e}$ is the $n \times 1$ residual term vector.

Based on above sections, let $\langle \boldsymbol{X} \rangle$ be the column space spanned by column vectors, if $\boldsymbol{y} \in \langle \boldsymbol{X} \rangle$, $\exists \boldsymbol{\beta} \in F^{n\times 1}$, $\boldsymbol{y} = \sum_{i=1}^{n} \beta_i \boldsymbol{X}_i = \boldsymbol{X\beta}$, which means that $\boldsymbol{e} = \boldsymbol{0}$.

Suppose, however, that $\boldsymbol{y}$ is not in the column space of $\boldsymbol{X}$. Then there is no $\boldsymbol{\beta}$ satisfying $\boldsymbol{y} = \boldsymbol{X\beta}$. But we can still we can find that $\boldsymbol{X\beta} \in \langle \boldsymbol{X} \rangle$, and $\boldsymbol{e}$ is the difference, or 'residual'. We are interested in finding $\boldsymbol{\beta}$ such that $\boldsymbol{y}$ is as close as possible to $\boldsymbol{X\beta}$ in the sense that $\boldsymbol{e}$ is as short as possible.

Definition. The length, or *norm*, of a vector $\boldsymbol{e}$ is

$$\|\boldsymbol{e}\| = \sqrt{\boldsymbol{e'e}}.$$

The problem is to find the $\|\boldsymbol{e}\| = \|\boldsymbol{y} - \boldsymbol{X\beta}\|$ is as small as possible. The solution is that $\boldsymbol{\beta}$ that makes $\boldsymbol{e}$ perpendicular, or orthogonal, to $\boldsymbol{X\beta}$, which represents the column space $\langle \boldsymbol{X} \rangle$.

Definition. Two nonzero vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are *orthogonal*, written $\boldsymbol{a} \perp \boldsymbol{b}$, if and only if

$$\boldsymbol{a}'\boldsymbol{b} = \boldsymbol{b}'\boldsymbol{a} = \boldsymbol{0}.$$

Returning once again to our fitting problem, we find that the $\boldsymbol{\beta}$ we seek is that for which

$$\boldsymbol{e} \perp \boldsymbol{X}\boldsymbol{\beta}.$$

Expanding this set of equation gives the requirement

$$\begin{aligned} (\boldsymbol{X}\boldsymbol{\beta})'\,\boldsymbol{e} &= \boldsymbol{0} \\ &= \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}'\left(\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}\right). \end{aligned}$$

Generally, $\boldsymbol{\beta} \neq \boldsymbol{0}$, then the set of equations becomes

$$\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}\,. \tag{11}$$

The means of solving such a set of equations is the subject of next section. Actually, the linear combination $\boldsymbol{X}\boldsymbol{\beta}$ is called the *projection* of $\boldsymbol{y}$ into the column space of $\boldsymbol{X}$, $\langle \boldsymbol{X} \rangle$. We use the cosine function in order to measure how close of $\boldsymbol{y}$ to the column space $\langle \boldsymbol{X} \rangle$, that is,

Theorem 1. *The angle θ between two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ satisfies*

$$\cos\theta = \frac{\boldsymbol{a}'\boldsymbol{b}}{\|\boldsymbol{a}\| \cdot \|\boldsymbol{b}\|}.$$

$\cos\theta \in [0, 1]$. *The larger, $\boldsymbol{a}$ is closer to $\boldsymbol{b}$.*

References

[1] Curtis, C. (2012). Linear algebra: an introductory approach. Springer Science & Business Media.

[2] Greene, W. H. (2003). Econometric analysis. Pearson Education India.

[3] Magnus, J. R., & Neudecker, H. (2002). Matrix differential calculus with applications in statistics and econometrics.

Matrix Algebra II

1 Solution of a System of Linear Equations

Consider the set of n linear equations,

$$\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y} \tag{1}$$

in which $\boldsymbol{\beta}$ is the $n \times 1$ unknown parameters vector. $\boldsymbol{X}$ is a known matrix of data, $m \times n$, and $\boldsymbol{y}$ is a specified vector of values. We are interested in knowing whether a solution *exists*; if so, then how to *obtain* it; and finally, if it does exist, then whether it is *unique*.

1.1 Systems of Linear Equations

For most of our applications, we shall consider only square systems of equations, that is, those in which $\boldsymbol{X}$ is a *square matrix*. In what follows, therefore, we take m to equal n. Since the number of rows in $\boldsymbol{X}$ is the number of equations, whereas the number of columns in $\boldsymbol{X}$ is the number of variables, this case is the familiar one of "*n equations with n unknowns*."

There are two types of systems of equations.

Definition. A *homogeneus system* is of the form $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}$.

By definition, a nonzero solution to such a system will exist if and only if $\boldsymbol{X}$ does not have full rank. If so, then for at least one column of $\boldsymbol{X}$, we can write the preceding as

$$\boldsymbol{X}_k = -\sum_{j \neq k} \frac{\boldsymbol{\beta}_j}{\boldsymbol{\beta}_k} \boldsymbol{X}_j.$$

This means, as we know, that the columns of $\boldsymbol{X}$ are linearly dependent and that $|\boldsymbol{X}| = 0$.

Definition. A *nonhomogeneous system* of equations is of the orm $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y}$, where $\boldsymbol{y}$ is a nonzero vector.

The vector $\boldsymbol{y}$ is chosen arbitrariliy and is to be expressed as a linear combination of the columns of $\boldsymbol{X}$. Since $\boldsymbol{\beta}$ has n elements, this solution will exist only if the columns of $\boldsymbol{X}$ span the entire n-dimensional space, $\mathbb{R}^n$. Equivalently, we shall require that the columns of $\boldsymbol{X}$ be linearly independent or that $|\boldsymbol{X}|$ not be equal to zero.

1.2 Inverse Matrices

Definition. Given a $n \times n$ matrix $\boldsymbol{A}$, if there is a $n \times n$ matrix $\boldsymbol{B}$ satisfies $\boldsymbol{AB} = \boldsymbol{BA} = \boldsymbol{I}$, then $\boldsymbol{B}$ is the *inverse matrix* of $\boldsymbol{A}$. Denote it as $\boldsymbol{B} = \boldsymbol{A}^{-1}$. And $\boldsymbol{A}$ is also called *nonsingular matrix*.

Example. We consider the 2×2 matrix here, let

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix},$$

then,

$$\boldsymbol{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix} = \frac{1}{|\boldsymbol{A}|} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}.$$

For equation system $\boldsymbol{X\beta} = \boldsymbol{y}$, suppose that we could find the inverse matrix $\boldsymbol{X}^{-1}$, then the following would be obtained:

$$\boldsymbol{X}^{-1}\boldsymbol{X\beta} = \boldsymbol{I\beta} = \boldsymbol{\beta} = \boldsymbol{X^{-1}y}.$$

Proposition. *If the inverse exists, then it must be unique.*

Proof. Suppose it is not true and let $\boldsymbol{B}$ and $\boldsymbol{C}$ be the different inverse matrices of the matrix $\boldsymbol{A}$. Then $\boldsymbol{CAB} = \boldsymbol{CAB}$, but $\boldsymbol{B} = \boldsymbol{IB} = (\boldsymbol{CA})\,\boldsymbol{B} = \boldsymbol{CAB} = \boldsymbol{C}\,(\boldsymbol{AB}) = \boldsymbol{CI} = \boldsymbol{C}$, which would be a contradiction if $\boldsymbol{C}$ did not equal $\boldsymbol{B}$. □

We shall use a^{ij} to indicate the ijth element of $\boldsymbol{A}^{-1}$. The general formula for computing an inverse matrix is

$$a^{ij} = \frac{(-1)^{j+i}\,|\boldsymbol{A}_{ji}|}{|\boldsymbol{A}|} = \frac{C_{ji}}{|\boldsymbol{A}|} \tag{2}$$

where C_{ji} is the jith cofactor of $\boldsymbol{A}$. Then we can write,

$$\boldsymbol{A}^{-1} = \frac{1}{|\boldsymbol{A}|}\boldsymbol{C}',\ \boldsymbol{C} = (C_{ij}).$$

It follows, therefore, that for $\boldsymbol{A}$ to be nonsingular, $|\boldsymbol{A}|$ must be nonzero.

Proposition. *If a $n \times n$ matrix $\boldsymbol{A}$ is invsersible if and only if $|\boldsymbol{A}| \neq 0$ or $rank\,(\boldsymbol{A}) = n$.*

Exercise. Some computational results involving inverses are

1. $\left|\boldsymbol{A}^{-1}\right| = \frac{1}{|\boldsymbol{A}|}$;
2. $\left(\boldsymbol{A}^{-1}\right)^{-1} = \boldsymbol{A}$;
3. $\left(\boldsymbol{A}^{-1}\right)' = (\boldsymbol{A}')^{-1}$;
4. If $\boldsymbol{A}$ is symmetric, then $\boldsymbol{A}^{-1}$ is symmetric;
5. $(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$;
6. $(\boldsymbol{ABC})^{-1} = \boldsymbol{C}^{-1}\,(\boldsymbol{AB})^{-1} = \boldsymbol{C}^{-1}\boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$

Recall that for a data matrix $\boldsymbol{X}$, $\boldsymbol{X'X}$ is the sum of the outer products of the rows $\boldsymbol{X}$. The following result, which is called an *updating formula*, shows how to compute the new **S** that would result when a new row $\boldsymbol{b}$ is addded to $\boldsymbol{X}$: $\tilde{\boldsymbol{X}} = \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{b}' \end{pmatrix}$, then

$$\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)^{-1} = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{b}\boldsymbol{b}'\right)^{-1} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} - \left(\frac{1}{1 + \boldsymbol{b}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{b}}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{b}\boldsymbol{b}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{3}$$

Note it results from the following general formulas:

$$\left(\boldsymbol{A} \pm \boldsymbol{b}\boldsymbol{c}'\right)^{-1} = \boldsymbol{A}^{-1} \mp \left(\frac{1}{1 \pm \boldsymbol{c}'\boldsymbol{A}^{-1}\boldsymbol{b}}\right)\boldsymbol{A}^{-1}\boldsymbol{b}\boldsymbol{c}'\boldsymbol{A}^{-1}; \tag{4}$$

$$\left(\boldsymbol{A} \pm \boldsymbol{B}\boldsymbol{C}\boldsymbol{B}'\right)^{-1} = \boldsymbol{A}^{-1} \mp \boldsymbol{A}^{-1}\boldsymbol{B}\left(\boldsymbol{C}^{-1} \pm \boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{B}'\right)^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}. \tag{5}$$

1.3 Solving the Least Squares Problem

We now have the tool needed to solve the least squares problem posed before. The problem is in equation (**??**):

$$\boldsymbol{X'y} = \boldsymbol{X'X}\boldsymbol{\beta}\,.$$

Then if $\boldsymbol{X'X}$ nonsingular, then

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X'y}.$$

1.4 The Generalized Inverse of a Matrix

Inverse matrices are fundamental in econometrics. Although we shall not require them much in our treatment in this book, there are more general forms of matrices than we have considered thus far. A *general inverse* of a matrix $\boldsymbol{A}^{+}$ that satisfies the following requirements:

- $\boldsymbol{A}\boldsymbol{A}^{+}\boldsymbol{A} = \boldsymbol{A}$.
- $\boldsymbol{A}^{+}\boldsymbol{A}\boldsymbol{A}^{+} = \boldsymbol{A}^{+}$.
- $\boldsymbol{A}^{+}\boldsymbol{A}$ is symmetric.
- $\boldsymbol{A}\boldsymbol{A}^{+}$ is symmetric.

A unique $\boldsymbol{A}^{+}$ can be found for any matrix, whether $\boldsymbol{A}$ is singular or not, or even if $\boldsymbol{A}$ is not square.

The unique matrix that satisfies all four requirements is called the *Moore-Penrose inverse* or *pseudoinverse* of $\boldsymbol{A}$. If $\boldsymbol{A}$ happens to be square and nonsingular, then the generalized inverse will be the familiar ordinary inverse. But if $\boldsymbol{A}^{-1}$ does not exist, then $\boldsymbol{A}^{+}$can still be computed.

Example. For a special case,

$$\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y},$$

where $\boldsymbol{X}$ has m rows, $n < m$ columns, and column rank equal $r = n$. Then, since $\boldsymbol{X}'\boldsymbol{X}$ is $n \times n$ and

$$rank\left(\boldsymbol{X}'\boldsymbol{X}\right) = rank\left(\boldsymbol{X}\right) = n.$$

$\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ exists. Then the Moore-Penrose inverse of $\boldsymbol{X}$ is $\boldsymbol{X}^{+} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$, which can be verified by multiplication. Then the solution can be written as

$$\boldsymbol{\beta} = \boldsymbol{X}^{+}\boldsymbol{y}.$$

Recall this was the solution to the least squares problem obtained before. If $\boldsymbol{y}$ lies in the column space of $\boldsymbol{X}$, this vector will be zero, but otherwise, it will not.

2 Partitioned Matrices

In formulating the elements of a matrix, it is sometimes useful to group some of the elements in *submatrices*. For example, we may want to test the *structural change,* or make the *partitioned regression.* One way to partition the matrix is like the following.

$$\underbrace{\boldsymbol{A}}_{m\times n} = \begin{pmatrix} \underbrace{\boldsymbol{A}_{11}}_{m_1\times n_1} & \underbrace{\boldsymbol{A}_{12}}_{m_1\times n_2} \\ \underbrace{\boldsymbol{A}_{21}}_{m_2\times n_1} & \underbrace{\boldsymbol{A}_{22}}_{m_2\times n_2} \end{pmatrix},$$

where $n = n_1+n_2$, $m = m_1+m_2$. $\boldsymbol{A}$ is a partitioned matrix. The subscripts of the submatrices are defined in the same fashion as those for the elements of a matrix. A common special case is the *block diagonal matrix*,

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_{22} \end{pmatrix},$$

where $\boldsymbol{A}_{11}$ and $\boldsymbol{A}_{22}$ are square matrices.

2.1 Manipulation of Partitioned Matrices

Similar to the algebraic manipulation of matrices element by element, we can define the minipulation for partitioned matrices. In all the following manipulation, the matrices must be conformable for the operations involved.

- **Addition**:

$$\boldsymbol{A} + \boldsymbol{B} = \begin{pmatrix} \boldsymbol{A}_{11} + \boldsymbol{B}_{11} & \boldsymbol{A}_{11} + \boldsymbol{B}_{12} \\ \boldsymbol{A}_{11} + \boldsymbol{B}_{21} & \boldsymbol{A}_{22} + \boldsymbol{B}_{22} \end{pmatrix}.$$

- **Multiplication**:

$$\begin{aligned} \boldsymbol{AB} &= \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix}\begin{pmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{A}_{11}\boldsymbol{B}_{11} + \boldsymbol{A}_{12}\boldsymbol{B}_{21} & \boldsymbol{A}_{11}\boldsymbol{B}_{12} + \boldsymbol{A}_{12}\boldsymbol{B}_{22} \\ \boldsymbol{A}_{21}\boldsymbol{B}_{11} + \boldsymbol{A}_{22}\boldsymbol{B}_{21} & \boldsymbol{A}_{21}\boldsymbol{B}_{12} + \boldsymbol{A}_{22}\boldsymbol{B}_{22} \end{pmatrix} \end{aligned}$$

- **Determinant**:

$$\left| \begin{array}{cc} \boldsymbol{A}_{11} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_{22} \end{array} \right| = |\boldsymbol{A}_{11}|\,|\boldsymbol{A_{22}}|\,,$$

$$\left| \begin{array}{cc} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{array} \right| = |\boldsymbol{A}_{11}|\left|\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\right| = |\boldsymbol{A}_{22}|\left|\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\right|.$$

- **Inverse**:

$$\left(\begin{array}{cc} \boldsymbol{A}_{11} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_{22} \end{array} \right)^{-1} = \left(\begin{array}{cc} \boldsymbol{A}_{11}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_{22}^{-1} \end{array} \right),$$

$$\left(\begin{array}{cc} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{array} \right)^{-1} = \left(\begin{array}{cc} \boldsymbol{A}_{11}^{-1}\left(\boldsymbol{I} + \boldsymbol{A}_{12}\boldsymbol{F}_2\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\right) & -\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{F}_2 \\ -\boldsymbol{F}_2\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} & \boldsymbol{F}_2 \end{array} \right),$$

  where

$$\boldsymbol{F}_2 = \left(\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\right)^{-1}.$$

  The upper left block could also be written as

$$\boldsymbol{F}_1 = \left(\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\right)^{-1}.$$

Example. *(Deviations from Means)*For a matrix,

$$\boldsymbol{A} = \left(\begin{array}{cc} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{array} \right) = \left(\begin{array}{cc} \iota'\iota & \iota'\boldsymbol{x} \\ \boldsymbol{x}'\iota & \boldsymbol{x}'\boldsymbol{x} \end{array} \right),$$

the inverse of the lower right-hand element of $\boldsymbol{A}^{-1}$ is

$$\begin{aligned} \boldsymbol{F}_2 &= \left(\boldsymbol{x}'\boldsymbol{x} - \left(\boldsymbol{x}'\iota\right)\left(\iota'\iota\right)^{-1}\left(\iota'\boldsymbol{x}\right)\right)^{-1} \\ &= \left[\boldsymbol{x}'\left(\boldsymbol{I} - \frac{\iota\iota'}{n}\right)\boldsymbol{x}\right]^{-1} = \left(\boldsymbol{x}'\boldsymbol{M}^0\boldsymbol{x}\right)^{-1}. \end{aligned}$$

Now suppose that we replace $\boldsymbol{x}$ with $\boldsymbol{X}$**,** a matrix with several columns. We seek the lower right block of $\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}$, where $\boldsymbol{Z} = (\iota, \boldsymbol{X})$. Tha analogous result is

$$\left(\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\right)_{22} = \left(\boldsymbol{X}'\boldsymbol{M}^0\boldsymbol{X}\right)^{-1}.$$

- **Kronecker Product**: For general matrix $\boldsymbol{A}$ and $\boldsymbol{B}$,

$$\boldsymbol{A} \otimes \boldsymbol{B} = \left(\begin{array}{cccc} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \ldots & a_{1n}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \ldots & a_{2n}\boldsymbol{B} \\ & & \ldots & \\ a_{m1}\boldsymbol{B} & a_{m2}\boldsymbol{B} & \ldots & a_{mn}\boldsymbol{B} \end{array} \right).$$

  Notice that there is no requirement for conformability in this operation. The Kronecker product can be computed for any pair of matrices. If $\boldsymbol{A}$ is $m_1 \times n_1$ and $\boldsymbol{B}$ is $m_2 \times n_2$, then $\boldsymbol{A} \otimes \boldsymbol{B}$ is $m_1 m_2 \times n_1 n_2$. And furthermore, if $\boldsymbol{A}$ is $m \times m$ and $\boldsymbol{B}$ is $n \times n$,

  - $|\boldsymbol{A} \otimes \boldsymbol{B}| = |\boldsymbol{A}|^m |\boldsymbol{B}|^n$;
  - $(\boldsymbol{A} \otimes \boldsymbol{B})' = \boldsymbol{A}' \otimes \boldsymbol{B}'$;
  - $tr\,(\boldsymbol{A} \otimes \boldsymbol{B}) = tr\,(\boldsymbol{A})\,tr\,(\boldsymbol{B})$;
  - $(\boldsymbol{A} \otimes \boldsymbol{B})\,(\boldsymbol{C} \otimes \boldsymbol{D}) = \boldsymbol{AC} \otimes \boldsymbol{BD}$.

3 Characteristic Roots and Vectors

Definition. For a square matrix $\boldsymbol{A}$ over field F, the *characteristic vectors* $\boldsymbol{c}$ over field F and *characteristic roots* $\lambda \in F$ are those that satisfy

$$\boldsymbol{A}\boldsymbol{c} = \lambda \boldsymbol{c}. \tag{6}$$

If $\boldsymbol{c}$ is any solution vector, then $k\boldsymbol{c}$ is also for any value of k. To remove the indeterminancy, $\boldsymbol{c}$ is *nomalized* so that $\boldsymbol{c}'\boldsymbol{c} = 1$.

Proposition. *$\lambda \in F$ is a charateristic root of $n \times n$ matrix $\boldsymbol{A}$ if and only if the determinant $|\boldsymbol{A} - \lambda \boldsymbol{I}| = 0$.*

Proof. The (6) can be rewritten as $(\boldsymbol{A} - \lambda \boldsymbol{I})\,\boldsymbol{c} = \boldsymbol{0}$. This equation is a *homogeneous system* that has nonzero solution if and only if the matrix $(\boldsymbol{A} - \lambda \boldsymbol{I})$ is singular or has a zero determinant. Therefore, if λ is a solution, then

$$|\boldsymbol{A} - \lambda \boldsymbol{I}| = 0. \tag{7}$$

This polynomial in λ is the *characteristic equation* of $\boldsymbol{A}$. □

Example. If

$$\boldsymbol{A} = \begin{pmatrix} 5 & 1 \\ 2 & 4 \end{pmatrix},$$

then

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

The two solutions are $\lambda = 6$ and $\lambda = 3$.

In solving the characteristic equation, there is no guarantee that the characteristic roots will be real.

Example. If

$$\boldsymbol{A} = \begin{pmatrix} 5 & 1 \\ -2 & 4 \end{pmatrix},$$

then

$$|\boldsymbol{A} - \lambda \boldsymbol{I}| = \begin{vmatrix} 5-\lambda & 1 \\ -2 & 4-\lambda \end{vmatrix} = \lambda^2 - 9\lambda + 22.$$

The two solutions are $\lambda = \frac{9 \pm \sqrt{7}i}{2}$.

The same result can emerge in the general $n \times n$ case.

However, the characteristic roots of a symmetric matrix are real. This result will be convenient because most of our applications will involve the characteristic roots and vectors of symmetric matrices.

For an $n \times n$ matrix, the characteristic equation is an nth-order polynomial in λ. Its solutions may be n distinct values, as in the preceding example, or may contain repeated values of λ, and may contain some zeros as well.

A $n \times n$ sysmmetric matrix has n distinct characteristic vectors, $\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_n$. The corresponding characteristic roots, $\lambda_1, \lambda_2, \ldots, \lambda_n$, although real, need not be distinct, which means there may be many characteristic belonging to a given characteristic root. The characteristic vectors of a symmetric matrix are orthogonal, which implies that for every $i \neq j$, $\boldsymbol{c}_i \boldsymbol{c}_j = 0$. It is convenient to collect the n-characteristic vectors in a $n \times n$ matrix whose ith columns is the $\boldsymbol{c}_i$corresponding to λ_i,

$$\boldsymbol{C} = (\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_n),$$

and the n-characteristic roots in the same order, in a diagonal matrix,

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ & & \ldots & \\ 0 & 0 & 0 & \lambda_n \end{pmatrix}.$$

Then, the full set of equations

$$\boldsymbol{A}\boldsymbol{c}_i = \lambda_i \boldsymbol{c}_i$$

is contained in

$$\boldsymbol{A}\boldsymbol{C} = \boldsymbol{C}\boldsymbol{\Lambda}.$$

Since the vectors are orthogonal and $\boldsymbol{c}_i'\boldsymbol{c}_i = 1$, we have

$$\boldsymbol{C}'\boldsymbol{C} = \boldsymbol{I}. \tag{8}$$

It implies that

$$\boldsymbol{C}' = \boldsymbol{C}^{-1}. \tag{9}$$

Consequently,

$$\boldsymbol{C}\boldsymbol{C}' = \boldsymbol{C}'\boldsymbol{C} = \boldsymbol{I} \tag{10}$$

as well, so the rows as well as the columns of $\boldsymbol{C}$ are orthogonal.

Definition. The *diagonalization* of matrix $\boldsymbol{A}$ is

$$\boldsymbol{C}'\boldsymbol{A}\boldsymbol{C} = \boldsymbol{C}'\boldsymbol{C}\boldsymbol{\Lambda} = \boldsymbol{I}\boldsymbol{\Lambda} = \boldsymbol{\Lambda}. \tag{11}$$

Furthermore, we also have,

Definition. The *spectral decomposition* of $\boldsymbol{A}$ is

$$\boldsymbol{A} = \boldsymbol{C}\boldsymbol{\Lambda}\boldsymbol{C}' = \sum_{i=1}^{n} \lambda_i \boldsymbol{c}_i \boldsymbol{c}_i'. \tag{12}$$

In this representation, the $n \times n$ matrix $\boldsymbol{A}$ is written as a sum of n rank one matrices. This sum is also called the *eigenvalue decomposition* of $\boldsymbol{A}$.

3.1 Rank of the Matrix

The diagonalization result enables us to obtain the rank of a matrix very easily.

Theorem 1. *For any matrix* $\boldsymbol{A}$ *and nonsingular matrices* $\boldsymbol{B}$ and $\boldsymbol{C}$, the rank of $\boldsymbol{BAC}$ is equal to the rank of $\boldsymbol{A}$.

Proof. $rank\,(\boldsymbol{BAC}) = rank\,((\boldsymbol{BA})\,\boldsymbol{C}) = rank\,(\boldsymbol{BA})$. And $rank\,(\boldsymbol{BA}) = rank\,(\boldsymbol{A'B'}) = rank\,(\boldsymbol{A'}) = rank\,(\boldsymbol{A})$. □

Corollary. *For a symmetric matrix* $\boldsymbol{A}$, $\boldsymbol{A} = \boldsymbol{C\Lambda C'}$, $rank\,(\boldsymbol{A}) = rank\,(\boldsymbol{\Lambda})$.

Proof. Since $\boldsymbol{C}$ and $\boldsymbol{C'}$ are nonsingular and square matrices, then,

$$rank\,(\boldsymbol{A}) = rank\left(\boldsymbol{C\Lambda C'}\right) = rank\,(\boldsymbol{\Lambda})\,.$$

□

Finding the rank of $\boldsymbol{\Lambda}$ is trivial. Since $\boldsymbol{\Lambda}$ is a diagonal matrix, its rank is just the number of nonzero values on its diagonal. By extending this result, we can prove the following theorems.

Theorem 2. *The rank of a symmetric matrix is the number of nonzero characteristic roots it contains.*

Note how this result enters the spectral decomposition given above. If any of the characteristic roots are zero, then the number of the rank one matrices in the sum is reduced correspondingly. It would appear that this simple rule will not be useful if $\boldsymbol{A}$ is not square. But recall that

$$rank\,(\boldsymbol{A}) = rank\left(\boldsymbol{A'A}\right).$$

Since $\boldsymbol{A'A}$ is always square, we can use it instead of $\boldsymbol{A}$**.** Indeed, we can use it even if $\boldsymbol{A}$ is not square, which leads to a fully general result.

Theorem 3. *The rank of any matrix* $\boldsymbol{A}$ *equals the number of nonzero characteristic roots in* $\boldsymbol{A'A}$.

Since the row rank and column rank of a matrix are equal, we should be able to apply this theorem to $\boldsymbol{AA'}$ as well. This process, however, requires an additional result.

Theorem 4. *The nonzero characteristic roots of* $\boldsymbol{AA'}$ *are the same as those of* $\boldsymbol{A'A}$.

3.2 Condition Number of a Matrix

As the preceding might suggest, therre is a discrete difference between full rank and short rank matrices. In analyzing data matrices, however, we shall often encounter cases in which a matrix is not quite short ranked, because it has all nonzero roots, but it is close. That is, by some measure, we can come very close to being able to write one column as a linear combination of the others. This case is important; it will be examined in the discussion of multicollinearity in econometrics I. Our definitions of rank and determinant will fail to indicate

this possibility, but an alternative measure, the *condition number*, is designed for that purpose. Formally, the condition number for a square matrix $\boldsymbol{A}$ is

$$\gamma = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}},$$

where λ_{max} is maximum characteristic root and λ_{min} the minimum one. For nonsquare matrices $\boldsymbol{X}$, such as the data matrix in the example, we use $\boldsymbol{A} = \boldsymbol{X}'\boldsymbol{X}$. As a further refinement, because the charecteristic roots are affected by the scaling of the columns of $\boldsymbol{X}$, we scale the columns to have length 1 by dividing each column by its norm.

For this measure, the smallest root is close to zero compared with the largest means that this matrix is nearly singular. Matrices with large condition numbers are difficult to invert accurately.

3.3 Trace of a Matrix

Definition. The **trace** of a square $n \times n$ matrix is the sum of its diaonal elements:

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

Exercise. Examine the following properties of trace manipulation.

1. $tr\left(c\boldsymbol{A}\right) = c \cdot tr\left(\boldsymbol{A}\right)\ c \in \mathbb{R}$.
2. $tr\left(\boldsymbol{A}'\right) = tr\left(\boldsymbol{A}\right)$.
3. $tr\left(\boldsymbol{A} + \boldsymbol{B}\right) = tr\left(\boldsymbol{A}\right) + tr\left(\boldsymbol{B}\right)$.
4. $tr\left(\boldsymbol{I}_n\right) = n$.
5. $tr\left(\boldsymbol{AB}\right) = tr\left(\boldsymbol{BA}\right)$.
6. $\boldsymbol{a}'\boldsymbol{a} = tr\left(\boldsymbol{a}'\boldsymbol{a}\right) = tr\left(\boldsymbol{a}\boldsymbol{a}'\right)$.
7. $tr\left(\boldsymbol{A}'\boldsymbol{A}\right) = \sum_{i=1}^{n} \boldsymbol{a}_i'\boldsymbol{a}_i = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2$.
8. The permutation rule can be extended to any *cyclic* permutation in a product

$$tr\left(\boldsymbol{ABCD}\right) = tr\left(\boldsymbol{BCDA}\right) = tr\left(\boldsymbol{CDAB}\right) = tr\left(\boldsymbol{DABC}\right). \tag{13}$$

Theorem 5. *The trace of a matrix equals the sum of its characteristic roots.*

Proof. For a matrix $\boldsymbol{A}$,

$$tr\left(\boldsymbol{A}\right) = tr\left(\boldsymbol{C\Lambda C}'\right) = tr\left(\boldsymbol{C}'\boldsymbol{C\Lambda}\right) = tr\left(\boldsymbol{\Lambda}\right). \tag{14}$$

Since $\boldsymbol{\Lambda}$ is diagonal with the roots of $\boldsymbol{A}$ on its diagonal, the general result is the following. □

3.4 Determinant of a Matrix

Recalling how tedious the calculation of a determinant promised to be, we find that the following is particularly useful. Since

$$\boldsymbol{C}'\boldsymbol{A}\boldsymbol{C} = \boldsymbol{\Lambda},$$

$$\begin{aligned} |\boldsymbol{\Lambda}| &= \left|\boldsymbol{C}'\boldsymbol{A}\boldsymbol{C}\right| = \left|\boldsymbol{C}'\right| |\boldsymbol{A}| \, |\boldsymbol{C}| \\ &= \left|\boldsymbol{C}'\right| |\boldsymbol{C}| \, |\boldsymbol{A}| = \left|\boldsymbol{C}'\boldsymbol{C}\right| |\boldsymbol{A}| \\ &= |\boldsymbol{A}| \, . \end{aligned} \tag{15}$$

Since $|\boldsymbol{\Lambda}|$ is just the product of its diagonal elements, the following is implied.

Theorem 6. *The determinant of a matrix equals the product of its characteristic roots.*

Notice that we get the expected result if any of these roots is zero. Since the determinant is the product of the roots, it follow that *a matrix is singular if and only if its determinant is zero and, in return, if and only if it has at lest one zero characteristic root.*

3.5 Powers of a Matrix

We often use expressions involving powers of matrices, such as $\boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}^2$. For positive integer powers, these expressions can be computed by repeated multiplication. But this does not show how to handle a problem such as finding a $\boldsymbol{B}$ such that $\boldsymbol{B}^2 = \boldsymbol{A}$, that is, the square root of a matrix. The characteristic roots and vectors provide a solution. Consider first,

$$\boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}^2 = \boldsymbol{C}\boldsymbol{\Lambda}\boldsymbol{C}'\boldsymbol{C}\boldsymbol{\Lambda}\boldsymbol{C}' = \boldsymbol{C}\boldsymbol{\Lambda}^2\boldsymbol{C}' \tag{16}$$

Two results follow. Since $\boldsymbol{\Lambda}$ is a diagonal matrix whose nonzero elements are the squares of those in $\boldsymbol{\Lambda}$, the following is implied.

For any symmetric matrix, the characteristic roots of $\boldsymbol{A}^2$ are the squares of those of $\boldsymbol{A}$, and the charecteristic vectors are the same.

The proof is obtained by observing that (16) is the spectral decomposition of the matrix $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{A}$. Since $\boldsymbol{A}^3 = \boldsymbol{A}^2\boldsymbol{A}$ and so on, it is easy to extend to any positive integer. By convention, for any $\boldsymbol{A}$, $\boldsymbol{A}^0 = \boldsymbol{I}$. Thus, for any symmetric matrix $\boldsymbol{A}$, $\boldsymbol{A}^k = \boldsymbol{C}\boldsymbol{\Lambda}^k\boldsymbol{C}'$for $k = 1, 2 \ldots$. Hence, the chracteristic roots of $\boldsymbol{A}^k$ is λ^k if the characteristic roots of $\boldsymbol{A}$ is λ, whereas the charecteristic vectors remain the same as those of $\boldsymbol{A}$.

If $\boldsymbol{A}$ is nonsingular, so that its roots λ_i are nonzero, then this proof can be extended to negative powers as well. If $\boldsymbol{A}^{-1}$ exists, then,

$$\boldsymbol{A}^{-1} = \left(\boldsymbol{C}\boldsymbol{\Lambda}\boldsymbol{C}'\right)^{-1} = \left(\boldsymbol{C}'\right)^{-1} \left(\boldsymbol{\Lambda}\right)^{-1} \left(\boldsymbol{C}\right)^{-1} = \boldsymbol{C}\boldsymbol{\Lambda}^{-1}\boldsymbol{C}',$$

where we have used the earlier result, $\boldsymbol{C}' = \boldsymbol{C}^{-1}$. This gives an important result that is useful for analyzing inverse matrices.

Theorem 7. *For any nonsingular symmetric matrix $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{\Lambda}\boldsymbol{C}'$, $\boldsymbol{A}^k = \boldsymbol{C}\boldsymbol{\Lambda}^k\boldsymbol{C}'$, for $k = \ldots, -2, -1, 0, 1, 2, \ldots$.*

We now turn to the general problem of how to compute the square root of a matrix. In the scalar case, the value would have to be nonnegative. The matrix analog to this requirement is that all the characteristic roots are nonnegative. Consider, then, the candidate

$$\boldsymbol{A}^{1/2} = \boldsymbol{C}\boldsymbol{\Lambda}^{1/2}\boldsymbol{C}' = \boldsymbol{C}diag\left(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}\right)\boldsymbol{C}'.$$

This equation satisfies the requirement for a square root, since

$$\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2} = \boldsymbol{C}\boldsymbol{\Lambda}^{1/2}\boldsymbol{C}'\boldsymbol{C}\boldsymbol{\Lambda}^{1/2}\boldsymbol{C}' = \boldsymbol{C}\boldsymbol{\Lambda}\boldsymbol{C}' = \boldsymbol{A}.$$

If we continue in this fashion, we can define the powers of a matrix more generally, still assuming that all the chareacteristic roots are nonnegative.

If all the roots are strictly positive, we can go one step further and extend the result to any real power.

Theorem 8. *For the characteristic roots of matrix $\boldsymbol{A}$ is positive, then $\boldsymbol{A}^r = \boldsymbol{C}\boldsymbol{\Lambda}^r\boldsymbol{C}'$, for any real number r.*

Example. *(Idempotent Matrices)*

For an idempotent matrix $\boldsymbol{A}$, $\boldsymbol{A}^2 = \boldsymbol{A}$. Therefore for any nonnegative number k, $\boldsymbol{A}^k = \boldsymbol{A}$. Further, if λ is a characteristic root of an indempotent matrix, then $\lambda = \lambda^k$, from which $\lambda = 1$ or $\lambda = 0$. *If $\boldsymbol{A}$ is a symmetric idempotent matrix, all its roots are one or zero.*

Assume that all the roots of $\boldsymbol{A}$ are one. Then $\boldsymbol{\Lambda} = \boldsymbol{I}$, and $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{\Lambda}\boldsymbol{C}' = \boldsymbol{C}\boldsymbol{C}' = \boldsymbol{I}$. If the roots are not all one, then one or more are zero, which means $|\boldsymbol{A}| = |\boldsymbol{\Lambda}| = 0$. Hence, we have the following results for symmetric idempotent matrices: (1) *The only full rank, symmetric idempotent matrix is the identity matrix $\boldsymbol{I}$;*(2)*All symmetric idempotent matrices except the identity matrix are singular.*

The diagonal matrix only contains value one or zero, then

$$rank\left(\boldsymbol{A}\right) = rank\left(\boldsymbol{A}'\boldsymbol{A}\right) = rank\left(\boldsymbol{A}\boldsymbol{A}\right) = rank\left(\boldsymbol{\Lambda}\right) = tr\left(\boldsymbol{\Lambda}\right) = tr\left(\boldsymbol{A}\right).$$

Example. *(Factoring a Matrix)*

In some applications, we shall require a matrix $\boldsymbol{P}$ such that

$$\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{A}^{-1}.$$

One choice is

$$\boldsymbol{P} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{C}'$$

so that,

$$\boldsymbol{P}'\boldsymbol{P} = \left(\boldsymbol{\Lambda}^{-1/2}\boldsymbol{C}'\right)'\boldsymbol{\Lambda}^{-1/2}\boldsymbol{C}' = \boldsymbol{C}\boldsymbol{\Lambda}^{-1}\boldsymbol{C}',$$

are desired. Thus, the spectral decomposition of $\boldsymbol{A}$, $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{\Lambda}\boldsymbol{C}'$ is a useful result for this kind of computation.

There are other method that we can use, such as the *Cholesky factorization* and *singular value decomposition.*

4 Quadratic Forms and Definite Matrices

Many optimization problem involves double sums of the form

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

This quadratic form can be written as

$$q = \boldsymbol{x}' \boldsymbol{A} \boldsymbol{x},$$

where $\boldsymbol{A}$ is a symmetric matrix. In general, q may be positive, negative or zero; it depends on $\boldsymbol{A}$ and $\boldsymbol{x}$. There are some matrices, however, for which q will be positive regardless of $\boldsymbol{x}$, and others for which q will always be negative.

Definition. For a given matrix $\boldsymbol{A}$,

- If $\boldsymbol{x}' \boldsymbol{A} \boldsymbol{x} > (<) \, 0$ for all nonzero $\boldsymbol{x}$, then $\boldsymbol{A}$ is positive(negative) definite.
- If $\boldsymbol{x}' \boldsymbol{A} \boldsymbol{x} \geq (\leq) \, 0$ for all nonzero $\boldsymbol{x}$, then $\boldsymbol{A}$ is nonegative definite or positive semidefinite(nonpositive definite).

Recall that a symmetric matrix can be decomposed into

$$\boldsymbol{A} = \boldsymbol{C} \boldsymbol{\Lambda} \boldsymbol{C}'.$$

Therefore, the quadratic form can be written as

$$\boldsymbol{x}' \boldsymbol{A} \boldsymbol{x} = \boldsymbol{x}' \boldsymbol{C} \boldsymbol{\Lambda} \boldsymbol{C}' \boldsymbol{x}.$$

Let $\boldsymbol{y} = \boldsymbol{C}' \boldsymbol{x}$. Then

$$\boldsymbol{x}' \boldsymbol{A} \boldsymbol{x} = \boldsymbol{y}' \boldsymbol{\Lambda} \boldsymbol{y} = \sum_{i=1}^{n} \lambda_i y_i^2. \tag{17}$$

If λ_i is postitive for all i, then regardless of $\boldsymbol{y}$, that is, regardless of $\boldsymbol{x}$, q will be positive. Continuing this line of reasoning, we obtain the following theorem.

Theorem 9. *Let $\boldsymbol{A}$ be a symmetric matrix.*

- *If all the chareacteristic roots of $\boldsymbol{A}$ are positive(negative), then $\boldsymbol{A}$ is **postitive definite(negative definite)**.*
- *If some of the roots are zero, then $\boldsymbol{A}$ is **nonnegative(nonpositive) definite** if the remainder are positive.*
- *If $\boldsymbol{A}$ has both negative and positive roots, then $\boldsymbol{A}$ is **indefinite.***

Exercise. Examine the theorem implies a number of related results.

1. If $\boldsymbol{A}$ is nonnegative definite, then $|\boldsymbol{A}| \geq 0$.
2. If $\boldsymbol{A}$ is positive definite, so is $\boldsymbol{A}^{-1}$.

3. The identity matrix $\boldsymbol{I}$ is positive definite.

4. If $\boldsymbol{A}$ is $m \times n$ matrix with full column rank and $m > n$, then $\boldsymbol{A}'\boldsymbol{A}$ is positive definite and $\boldsymbol{A}\boldsymbol{A}'$ is nonnegative definite.

5. If $\boldsymbol{A}$ is positive definite and $\boldsymbol{B}$ is a nonsingular matrix, then $\boldsymbol{B}'\boldsymbol{A}\boldsymbol{B}$ is positive definite.

6. For the idempotent symmetric matrices,

7. Every symmetric idempotent matrix is nonnegative definite.

8. If $\boldsymbol{A}$ is symmetric and idempotent, $n \times n$ with rank J, then every quadratic form in $\boldsymbol{A}$ can be written as $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \sum_{j=1}^{J} y_j^2$, where $\boldsymbol{y}$ has relationship with $\boldsymbol{x}$.

4.1 Comparing Matrices

Derivations in econometrics often focus on whether one matrix is "larger" than another. As a starting point, the two matrices must have the same dimensions.

Definition. For two $n \times n$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$,

$$d = \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} - \boldsymbol{x}'\boldsymbol{B}\boldsymbol{x} = \boldsymbol{x}'\left(\boldsymbol{A} - \boldsymbol{B}\right)\boldsymbol{x}.$$

If d is always positive for any nonzero vector, $\boldsymbol{x}$, then by this criterion, we can say that $\boldsymbol{A}$ is *larger* than $\boldsymbol{B}$. The reverse would apply if d is always negative.

It follows from the definition that *if $d > 0$ for all nonzero $\boldsymbol{x}$, then $\boldsymbol{A}-\boldsymbol{B}$ is positive definite.*

If d is only greater than or equal to zero, then $\boldsymbol{A}-\boldsymbol{B}$ is nonnegative definite. The ordering is not complete. For some pairs of matrices, d could have either sign, depending on $\boldsymbol{x}$. In this case, there is no simple comparison.

A particular case of the general result which we will encounter frequently is: *if $\boldsymbol{A}$ is positive definite and $\boldsymbol{B}$ is nonnegative definite, then $\boldsymbol{A} + \boldsymbol{B} \geq \boldsymbol{A}$.*

Example. For the updating formula,

$$\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}} = \boldsymbol{X}'\boldsymbol{X} + \boldsymbol{b}\boldsymbol{b}' \geq \boldsymbol{X}'\boldsymbol{X}.$$

Furthermore, we have the following theorem.

Theorem 10. *If $\boldsymbol{A}$ and $\boldsymbol{B}$ are two positive definite matrices with the same dimensions and if every characteristic root of $\boldsymbol{A}$ is larger than (at least as large as) the corresponding characteristic root of $\boldsymbol{B}$ when bothsets of roots are ordered from largest to smallest, then $\boldsymbol{A} - \boldsymbol{B}$ is positive(nonnegative) definite.*

Example. For the inverses of symmetric matrices, the result analogous to a familiar result for scalars is: if $\boldsymbol{A} > \boldsymbol{B}$, for any nonzero vector $\boldsymbol{x}$,

$$d = \boldsymbol{x}'\left(\boldsymbol{B}^{-1} - \boldsymbol{A}^{-1}\right)\boldsymbol{x} > 0 \Rightarrow \boldsymbol{B}^{-1} > \boldsymbol{A}^{-1}. \tag{18}$$

5 Matrix Calculus

After the introductory sections above, now we will turn to the matrix calculus.

5.1 Differentiation

For a scalar function, $f(x) : \mathbb{R} \to \mathbb{R}$, let $y = f(x)$. If it is differentiable, then we denote

$$f'(x) = \frac{dy}{dx},\ f''(x) = \frac{d^2y}{dx^2}.$$

Then for a scalar-valued function, $f(\boldsymbol{x}) : \mathbb{R}^n \mapsto \mathbb{R}$, let $y = f(\boldsymbol{x})$. The vector of partial derivatives, or gradient vector, or simply gradient, is

$$\boldsymbol{g}(\boldsymbol{x}) = \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \begin{pmatrix} \partial y/\partial x_1 \\ \partial y/\partial x_2 \\ \vdots \\ \partial y/\partial x_1 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix}.$$

Notice that it is a column vector. The shape of the derivative is determined by the denominator of the derivative.

A second derivatives matrix or Hessian is computed as

$$\begin{aligned} \boldsymbol{H}(\boldsymbol{x}) &= \begin{pmatrix} \frac{\partial^2 y}{\partial x_1 \partial x_1} & \frac{\partial^2 y}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 y}{\partial x_1 \partial x_n} \\ \frac{\partial^2 y}{\partial x_2 \partial x_1} & \frac{\partial^2 y}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 y}{\partial x_2 \partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 y}{\partial x_n \partial x_1} & \frac{\partial^2 y}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 y}{\partial x_n \partial x_n} \end{pmatrix} \\ &= \begin{pmatrix} \frac{\partial}{\partial x_1}\left(\frac{\partial y}{\partial \boldsymbol{x}}\right) & \frac{\partial}{\partial x_2}\left(\frac{\partial y}{\partial \boldsymbol{x}}\right) & \cdots & \frac{\partial}{\partial x_n}\left(\frac{\partial y}{\partial \boldsymbol{x}}\right) \end{pmatrix} \\ &= \frac{\partial}{\partial \boldsymbol{x}'}\left(\frac{\partial y}{\partial \boldsymbol{x}}\right) \\ &= \frac{\partial^2 y}{\partial \boldsymbol{x}' \partial \boldsymbol{x}}. \end{aligned}$$

If function is defined as $f(\boldsymbol{X}) : \mathbb{R}^{m \times n} \mapsto \mathbb{R}$, the gradient is defined as:

$$\frac{\partial f(\boldsymbol{X})}{\partial \boldsymbol{X}} = \left(\frac{\partial f(\boldsymbol{X})}{\partial x_{ij}}\right)_{m \times n}. \tag{19}$$

The following are some results for common function.

- For a linear function, $y = \boldsymbol{a}'\boldsymbol{x} = \sum_{i=1}^n a_i x_i$, the gradient is

$$\boldsymbol{g}(\boldsymbol{x}) = \frac{\partial y}{\partial \boldsymbol{x}} = \begin{pmatrix} \partial y/\partial x_1 \\ \partial y/\partial x_2 \\ \vdots \\ \partial y/\partial x_n \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \boldsymbol{a}. \tag{20}$$

$$\boldsymbol{H} = \boldsymbol{0}.$$

- For a set of linear functions $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$, we have $y_i = \boldsymbol{X}_i'\boldsymbol{\beta}$, where $\boldsymbol{X}_i'$ is the ith row of $\boldsymbol{X}$. Therefore, for each i,

$$\frac{\partial y_i}{\partial \boldsymbol{\beta}} = \boldsymbol{X}_i$$

Therefore,

$$\begin{pmatrix} \partial y_1/\partial \boldsymbol{\beta}' \\ \partial y_2/\partial \boldsymbol{\beta}' \\ \vdots \\ \partial y_n/\partial \boldsymbol{\beta}' \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1' \\ \boldsymbol{X}_2' \\ \vdots \\ \boldsymbol{X}_n' \end{pmatrix} = \boldsymbol{X}.$$

This can be written as $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{\beta}} = \boldsymbol{X}'$, whereas the more familiar form will be

$$\frac{\partial \boldsymbol{X}\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = \boldsymbol{X}'. \tag{21}$$

- For a quadratic form,

$$\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j a_{ij}, \tag{22}$$

which is very common in econometrics. The gradiant w.r.t. $\boldsymbol{x}$,

$$\frac{\partial \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \begin{pmatrix} \frac{\partial \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\partial x_1} \\ \frac{\partial \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\partial x_2} \\ \vdots \\ \frac{\partial \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\partial x_n} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} x_i a_{1i} + \sum_{i=1}^{n} x_i a_{i1} \\ \sum_{i=1}^{n} x_i a_{2i} + \sum_{i=1}^{n} x_i a_{i2} \\ \cdots \\ \sum_{i=1}^{n} x_i a_{ni} + \sum_{i=1}^{n} x_i a_{in} \end{pmatrix} = \left(\boldsymbol{A} + \boldsymbol{A}'\right)\boldsymbol{x}. \tag{23}$$

For the gradient w.r.t. $\boldsymbol{X}$**,** the coefficient on a_{ij} is $x_i x_j$. Therefore,

$$\frac{\partial \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\partial a_{ij}} = x_i x_j.$$

The square matrix whose ijth element is $x_i x_j$ is $\boldsymbol{x}\boldsymbol{x}'$, so

$$\frac{\partial \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{A}} = \boldsymbol{x}\boldsymbol{x}'. \tag{24}$$

- For a determinant form, $f(\boldsymbol{A}) = \log|\boldsymbol{A}|$, the gradient for each term is

$$\frac{\partial |\boldsymbol{A}|}{\partial a_{ij}} = (-1)^{i+j}|\boldsymbol{A}_{ij}|,$$

where $|\boldsymbol{A}_{ij}|$ is the ijth minor in $\boldsymbol{A}$. The ijth element of the inverse of $\boldsymbol{A}$ is

$$b_{ij} = \frac{(-1)^{i+j}|\boldsymbol{A}_{ji}|}{|\boldsymbol{A}|},$$

which implies that,

$$\frac{\partial f(\boldsymbol{A})}{\partial a_{ij}} = \frac{(-1)^{i+j}|\boldsymbol{A}_{ij}|}{|\boldsymbol{A}|} = b_{ji},$$

hence,

$$\frac{\partial f(\boldsymbol{A})}{\partial \boldsymbol{A}} = \left(\boldsymbol{A}^{-1}\right)' = \boldsymbol{A}'^{-1} \tag{25}$$

Those above are some basic calculus. For more examples, please refer to Magnus and Neudecker(2002).

5.2 Taylor Expansion

In econometrics, Taylor series approximation is frequently used. A Taylor series is a polynominal approximation to the target function. For example for function $f(x)$, let x_0 be an arbitrarily chosen expansion point

$$f(x) \approx f(x_0) + \sum_{n=1}^{P} \frac{1}{n!} \frac{d^n f(x_0)}{dx^i} (x - x_0)^n \tag{26}$$

The choice of the number of terms is arbitrary; the more that are used, the more accurate the approximation will be. The approximation used most frequently in econometrics is the following approximations.

- **Linear approximation**:

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0), \tag{27}$$

which can be written as

$$\begin{aligned} f(x) &\approx \left[f(x_0) - f'(x_0) x_0\right] + f'(x_0) x \\ &= \alpha + \beta x \end{aligned}$$

- **Quadratic approximation**:

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2. \tag{28}$$

It can also be written as

$$\begin{aligned} f(x) &\approx \left[f(x_0) - f'(x_0) x_0 + \frac{1}{2} f''(x_0) x_0^2\right] + \left[f'(x_0) - f''(x_0) x_0\right] x + \frac{1}{2} f''(x_0) x^2 \\ &= \alpha + \beta x + \gamma x^2 \end{aligned}$$

Extend the above results to single-valued function, we have

- **Linear approximation**

$$\begin{aligned} f(\boldsymbol{x}) &\approx f(\boldsymbol{x}_0) + \sum_{n=1}^{n} f_i(\boldsymbol{x}_0)(x_i - x_{0i}) \\ &= f(\boldsymbol{x}_0) + \frac{\partial f(\boldsymbol{x}_0)}{\partial \boldsymbol{x}'} (\boldsymbol{x} - \boldsymbol{x}_0). \\ &= \left[f(\boldsymbol{x}_0) - \boldsymbol{g}(\boldsymbol{x}_0)' \boldsymbol{x}_0\right] + \boldsymbol{g}(\boldsymbol{x}_0)' \boldsymbol{x} \\ &= \alpha + \boldsymbol{\beta}' \boldsymbol{x}. \end{aligned} \tag{29}$$

- **Quadratic approximation:**

$$f(\boldsymbol{x}) = f(\boldsymbol{x}_0) + \sum_{n=1}^{n} f_i(\boldsymbol{x}_0)(x_i - x_{0i}) + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial f(\boldsymbol{x}_0)}{\partial x_i \partial x_j} (x_i - x_{0i})(x_j - x_{0j}) \tag{30}$$

The second-oder terms in the expansion is

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial f\left(\boldsymbol{x}_0\right)}{\partial x_i \partial x_j}\left(x_i - x_{0i}\right)\left(x_j - x_{0j}\right) = \frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{x}_0\right)' \boldsymbol{H}\left(\boldsymbol{x}_0\right)\left(\boldsymbol{x} - \boldsymbol{x}_0\right),$$

to the preceding one. Collecting all the terms, we have

$$f\left(\boldsymbol{x}\right) \approx \alpha + \boldsymbol{\beta}'\boldsymbol{x} + \frac{1}{2}\boldsymbol{x}'\boldsymbol{H}\left(\boldsymbol{x}_0\right)\boldsymbol{x}, \tag{31}$$

where

$$\alpha = f\left(\boldsymbol{x}_0\right) - \boldsymbol{g}\left(\boldsymbol{x}_0\right)'\boldsymbol{x}_0 + \frac{\boldsymbol{x}_0'\boldsymbol{H}\left(\boldsymbol{x}_0\right)\boldsymbol{x}_0}{2},\ \boldsymbol{\beta} = \boldsymbol{g}\left(\boldsymbol{x}_0\right) - \boldsymbol{H}\left(\boldsymbol{x}_0\right)\boldsymbol{x}_0.$$

5.3 Transformations

If $\boldsymbol{y}$ is a column vector of functions, $\boldsymbol{y} = f\left(\boldsymbol{x}\right)$. The inverse trasformation is $\boldsymbol{x} = f^{-1}\left(\boldsymbol{y}\right)$ if $f^{-1}\left(\boldsymbol{y}\right)$ exists. Then the *Jacobian* matrix of the transformation from $\boldsymbol{y}$ to $\boldsymbol{x}$ is

$$\boldsymbol{J} = \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}'} = \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_n} \\ & & \cdots & \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{pmatrix}. \tag{32}$$

Then the absolute value of the determinant of $\boldsymbol{J}$, $abs\left(|\boldsymbol{J}|\right)$ is the *Jacobian* determinant of the transformation from $\boldsymbol{y}$ to $\boldsymbol{x}$. In the nonsingular case,

$$abs\left(|\boldsymbol{J}|\right) = abs\left(\left|\boldsymbol{A}^{-1}\right|\right) = \frac{1}{abs\left(|\boldsymbol{A}|\right)}.$$

In the singular case, the matrix of the partial derivatives will be singular and the determinant of the Jacobian will be zero.

Example. For linear function $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} = f\left(\boldsymbol{x}\right)$, the inverse transformation is $\boldsymbol{x} = f^{-1}\left(\boldsymbol{y}\right)$. If $\boldsymbol{A}$ is nonsingular, then,

$$\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{y}$$

If $\boldsymbol{A}$ is singular, then there is no inverse transformation. Let $\boldsymbol{J}$ be the matrix of partial derivatives of the inverse function:

$$\boldsymbol{J} = \left(\frac{\partial x_i}{\partial y_j}\right).$$

The absolute value of the determinant of $\boldsymbol{J}$ is

$$abs\left(|\boldsymbol{J}|\right) = abs\left(\left|\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}}\right|\right).$$

And it is the *Jacobian determinant* of the transformation from $\boldsymbol{y}$ to $\boldsymbol{x}$. In the nonsingular case,

$$abs\left(|\boldsymbol{J}|\right) = abs\left(\left|\boldsymbol{A}^{-1}\right|\right) = \frac{1}{abs\left(|\boldsymbol{A}|\right)}.$$

In the singular case, the matrix of the partial derivatives will be singular and the determinant of the Jacobian will be zero, which means the transformation from $\boldsymbol{x}$ to $\boldsymbol{y}$ are functionally dependent.

Clearly, if the vector $\boldsymbol{x}$ is given, then $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ can be computed from $\boldsymbol{x}$. Whether $\boldsymbol{x}$ can be deduced from $\boldsymbol{y}$ is another question. Evidently, it depends on the Jacobian. If the jacobian is not zero, then the inverse transformations exist, and we can obtain $\boldsymbol{x}$. If not, then we cannot obtain $\boldsymbol{x}$.

5.4 Optimization and Constrained Optimization

The optimization is very important in econometrics. There are many estimators obtained by minimizing or maximizing some specific criterion function. In this section, we assume that the criterion function are secondly diffentiable.

Consider finding the x where $f(x)$ is maximized or minimized. Since $f'(x)$ is the slope of the $f(x)$, either optimum must occur where $f'(x) = 0$. Otherwise, the function will be increasing or decreasing at x. This result implies the *first-order or necessary condition for an optimum:*

$$\frac{dy}{dx} = 0 \tag{33}$$

For a maximun, the function must be *concave*; for a minimum, it must be convex. The *sufficient condition for an optimun* is*:*

$$\text{For a maximun, } \frac{d^2y}{dx^2} < 0;$$

$$\text{For a minimum, } \frac{d^2y}{dx^2} > 0.$$

Some functions, such as the sine and cosine functions, have many ***local optima***, that is, many minima and maxima.

Example. The function $f(x) = \frac{\cos(x)}{1+x^2}$, the damped cosine wave, does as well but differs in that although it has many local maxima, it has one, at $x = 0$, at which $f(x)$ is greater than it is at any other point. We can see those properties from the following figure.

Thus, $x = 0$ is the *global maximun***,** whereas the other maxima are only *local maxima.*

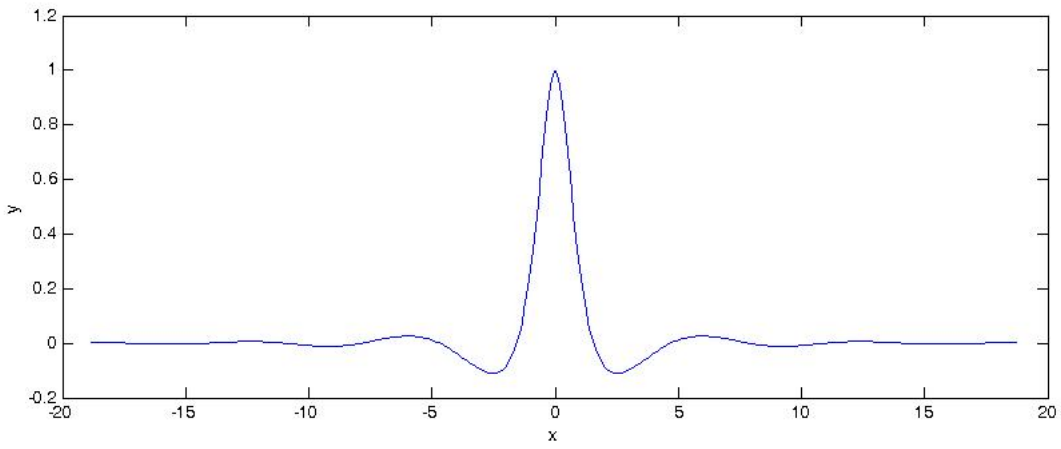
Certain functions have only a single optimum, such as the gllobally convex or globally concave. These function are very common in econometrics and some theoretical economics modelling. Further, the globally concave functions usually have a maximun and globally convex functions have a minimum.

For maximizing or minimizing a function of several variables, $f(\boldsymbol{x})$, the first-order conditions are,

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \boldsymbol{0}. \tag{34}$$

The result is interpreted in the same manner as the necessary condition in the univariate case. At the optimum, it must be true that no small change in any variable leads to an

Figure 1: The function $f(x) = \frac{\cos(x)}{1+x^2}$



improvement in the function value. Meanwhile, the second-order condition for an optimum in the multivariate case is that, at the optimizing value,

$$\boldsymbol{H} = \frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}'} \tag{35}$$

must be positive definite for a minimum and negative definite for a maximum.

In a single-variable problem, the second-order can usually be verified by inspection. This situation will not generally be true in the multivariate case. As discussed earlier, checking the definiteness of a matrix is, general, a difficult problem. For most of the problems encountered in econometrics, however, the second-order condition will be implied by the structure of the problem. That is, the matrix $\boldsymbol{H}$ will usually be of such a form that it is always definite.

Example. Consider the problem,

$$\max_{\boldsymbol{x} \in \mathbb{R}^n} R = \boldsymbol{a}'\boldsymbol{x} - \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}.$$

The first-order condition is

$$\frac{\partial R}{\partial \boldsymbol{x}} = \boldsymbol{a} - \left(\boldsymbol{A} + \boldsymbol{A}'\right)\boldsymbol{x} = \boldsymbol{0}.$$

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

$$\boldsymbol{x} = \left(\boldsymbol{A} + \boldsymbol{A}'\right)^{-1}\boldsymbol{a}.$$

And the second-order condition is,

$$\frac{\partial R}{\partial \boldsymbol{x} \partial \boldsymbol{x}'} = -\left(\boldsymbol{A} + \boldsymbol{A}'\right).$$

Then we can verify the charateristic roots of $-\left(\boldsymbol{A} + \boldsymbol{A}'\right)$ are all negative or positive.

Exercise. Solve the following problem,

$$\min_{\boldsymbol{\beta} \in F^{n \times 1}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Example. It is often necessary to solve an optimization problem subject to some constraints on the solution. For such problems, the common used method is the **Lagrange multipliers.**

Basics of Probability Theory

1 Brief Introduction

The subject of probability theory is the foundation upon which all of statistics is built, providing a means for modeling populations, experiments, or almost anything else that could be considered a random phenomenon. Before we define probability, we have to introduce the set theory. For more details, please refer to Resnick(1998).

1.1 Set Theory

Definition 1. If A is any set(whose element may be numbers or any other objects), then

If x is a member(or an element) of A, then $x \in A$.

If x is not a memeber of A, then $x \notin A$.

The set which contains no element will be called the *empty set,* $\emptyset$. If a set has at least one element, it is called *nonempty*.

If A and B are sets, and if every element of A is an element of B, we say that A is a subset of B, and write $A \subset B$, or $B \supset A$. If, in addition, there is an element of B which is not in A, then A is said to be a *proper subset* of B.

If $A \subset B$ and $B \subset A$, we write $A = B$. Otherwise, $A \neq B$.

In order to draw conclusions related to a random phenomenon, the first step is to identify the following definitions.

Definition 2. An *experiment* is repeatable process that has a well defined set of possible outcomes. The set, Ω, of all possible outcomes of a particular experiment is called the *sample space* for the experiment. An *event* is a subset of the sample space, $A \subseteq \Omega$.

Exercise 1. For each of the following experiments, describe the sample space.

1. Toss a coin four times.
2. Count the number of insect-damaged leaves on a plant.
3. Measure the lifetime(in hours) of a particular brand of light bulb.
4. Record the weights of 10-day-old rats.

Remember that probability is a mathematical construct and when we apply probability ideas to real situations we always make assumptions (e.g.. fairness of a dice, etc.). Hence, probability statements are statements about a mathematical model not statements about reality.

Given any two events(or sets), we have the following elementary set operations:

- *Union*: The union of A and B, $A \cup B = \{x : x \in A \text{ or } x \in B\}$.
- *Intersection*: The intersection of A and B, $A \cap B = \{x : x \in A \text{ and } x \in B\}$.
- *Complementation*: The complementation of A, $A^c = \{x : x \notin A\}$.

Example 1. Consider the experiment of selecting a card at random from a standard deck and noting its suit: clubs(C), diamonds(D), hearts(H), or spades(S). The sample space is

$$\Omega = \{C, D, H, S\},$$

and possible events are

$$A = \{C, D\},\ B = \{D, H, S\}.$$

From the events we can form

$$A \cup B = \{C, D, H, S\},\ A \cap B = \{D\},\ A^C = \{H, S\}.$$

Furthermore, notice that $A \cup B = \Omega$ and $(A \cup B)^c = \emptyset$.

Theorem 1. *For any three events A, B and C, defined on a sample space, Ω, then we have*

1. *Commutativity: $A \cup B = B \cup A$, $A \cap B = B \cap A$.*
2. *Associativity: $A \cup (B \cup C) = (A \cup B) \cup C$,$A \cap (B \cap C) = (A \cap B) \cap C$.*
3. *Distributive Laws: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.*
4. *DeMorgan's Law: $(A \cup B)^c = A^c \cap B^c$, $(A \cap B)^c = A^c \cup B^c$.*

Exercise 2. Do the following excercise.

1. Prove Theorem 1.
2. Verify the following identities.
   (a) $A \backslash B = A \backslash (A \cap B) = A \cap B^c$.
   (b) $B = (B \cap A) \cup (B \cap A^c)$.
   (c) $B \backslash A = B \cap A^c$.
   (d) $A \cup B = A \cup (B \cap A^c)$.

The operations of union and intersection can be extended to infinite collections of sets as well, which is called *countable union* and *countable intersection*. If $A_1, A_2, \ldots$ is a collection of sets, all defined on a sample space, Ω, then

$$\bigcup_{i=1}^{\infty} A_i = \{x \in \Omega : x \in A_i \text{ for some } i\},$$

$$\bigcap_{i=1}^{\infty} A_i = \{x \in \Omega : x \in A_i \text{ for all } i\}.$$

Exercise 3. Let $\Omega = (0, 1]$, $A_i = [1/i, 1]$. Verify the following identiy

1. $\bigcup_{i=1}^{\infty} \left[\frac{1}{i}, 1\right] = (0, 1]$.
2. $\bigcap_{i=1}^{\infty} \left[\frac{1}{i}, 1\right] = \{1\}$.

Definition 3. Two events A and B are *disjoint(or mutually exclusive)* if $A \cap B = \emptyset$. The events $A_1, A_2 \ldots$ are *pairwise disjoint* if $A_i \cap A_j = \emptyset$ for all $i \neq j$.

Example 2. Consider the collection $A_i = [i, i + 1)$, $i = 0, 1, \ldots$, then it is consist of pairwise disjoint sets. Further, $\bigcup_{i=1}^{\infty} A_i = [0, \infty)$.

Definition. If $A_1, \ldots$ are pairwise disjoint and $\bigcup_{i=1}^{\infty} A_i = \Omega$, then the collection $A_1, A_2, \ldots$ forms a *partition* of Ω.

2 Mathematical Probability

2.1 Axiomatic Foundations

For each event A in the sample space Ω we want to associate with A a number between zero and one that will be called the probability of A, denote by $P(A)$. It would seem natural to define the domain of P as all subsets of Ω; that is, for each $A \subset \Omega$ we define $P(A)$ as the probability that A occurs.

Definition 4. A collection of subsets of Ω is called a $\sigma-$*algebra*, denoted by $\mathcal{B}$, if it satisfies the following three properties:

- $\Omega \in \mathcal{B}$.
- If $A \in \mathcal{B}$, then $A^c \in \mathcal{B}$.
- If $A_1, A_2, \cdots \in \mathcal{B}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$.

The pair consisting of a set and a $\sigma-$algebra defined on that set is referred to as a *measurable space*.

Exercise 4. Do the following exercise,

1. Prove if $A_1, A_2, \cdots \in \mathcal{B}$, $\mathcal{B}$ is a $\sigma-$algebra, then $\bigcap_{i=1}^{\infty} A_i \in \mathcal{B}$.
2. Prove if $\mathcal{B}_1$ and $\mathcal{B}_2$ are both $\sigma-$algebra, then $\mathcal{B}_1 \cap \mathcal{B}_2$ is also a $\sigma-$algebra.

Example 3. Consider a set Ψ together with non-trivial subset $A \subset \Psi$.

1. The *trivial* $\sigma-$*algebra*: $\mathcal{B} = \{\emptyset, \Omega\}$.
2. The *power set.* $\mathcal{B} = \mathcal{P}(\Omega)$ or $\{0, 1\}^{\Psi}$, the class of all subsets of Ω.

The sample space Ω is the set of all possible outcomes of an experiment and we can define events as subsets of Ω containing outcomes that are of interest. Now we can generate a measurable space $(\Omega, \mathcal{F})$, where $\mathcal{F}$ is a σ-algebra defined on Ω. Here $\mathcal{F}$ is a collection of subsets of Ω and we interpret the elements of $\mathcal{F}$ as being events. Thus if $A \in \mathcal{F}$ then A is an event. We say "event A occur". Since probability is always associated with events so $\mathcal{F}$ will be the domain for probability measure.

Definition 5. Given a sample space Ω and an associated $\sigma-$algebra $\mathcal{F}$, the *probability measure* on $(\Omega, \mathcal{F})$ is a function $P : \mathcal{F} \to [0, 1]$ with the following properties,

- $P(A) \geq 0$ for $\forall A \in \mathcal{F}$.
- $P(\Omega) = 1$.
- If $A_1, \cdots \in \mathcal{F}$ are disjoint then $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$.

Then $(\Omega, \mathcal{F}, P)$ is called a *probability space.*

2.2 The Calculus of Probabilities

From the definition of the probability, we can build up many properties of probability measure, which are quite helpful in calculation of more complicated probabilities.

Proposition 1. *If P is a probability measure on $(\Omega, \mathcal{F})$, and A is any set in $\mathcal{F}$, then,*

1. $P(\emptyset) = 0$, *where $\emptyset$ is the empty set.*
2. $P(A) \leq 1$.
3. $P(A^c) = 1 - P(A)$.

For Proposition 1, it is very easy to prove based on the definition of probability measure.

Theorem 2. *If P is a probability measure on $(\Omega, \mathcal{F})$, A and B is any set in $\mathcal{F}$, then,*

1. $P(B \cap A^c) = P(B) - P(A \cap B)$.
2. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
3. *If $A \subset B$, then $P(A) \leq P(B)$.*

Exercise 5. Prove Proposition 1 and Theorem 2.

Proposition 2. *Consider a probability space $(\Omega, \mathcal{F}, P)$,*

1. $P(A) = \sum_{i=1}^{\infty} P(A \cap C_i)$ *for any partition $C_1, C_2, \ldots$.*
2. *If $A_1, \cdots \in \mathcal{F}$, then, $P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i)$.*

Theorem 3. *For general, Consider a probability space $(\Omega, \mathcal{F}, P)$, with $A_1, \ldots, A_n \in \mathcal{F}$, then,*

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) + \cdots + (-1)^{n+1} P(A_1 \cap \cdots \cap A_n).$$

Proposition 3. *(Probability limit of sequence of sets)*

1. *If $\{A_i\}$ is an increasing sequence of sets $A_1 \subset A_2 \subset \ldots$, then, $\lim_{n\to\infty} P(A_n) = P\left(\bigcup_{i=1}^{\infty} A_i\right)$,*
2. *If $\{A_i\}$ is an increasing sequence of sets $A_1 \supset A_2 \supset \ldots$, then, $\lim_{n\to\infty} P(A_n) = P\left(\bigcap_{i=1}^{\infty} A_i\right)$.*

Exercise 6. Prove Proposition 2, Theorem 3 and Proposition 3.

3 Conditional Probability and Independence

All of the probabilities that we have dealt with thus far have been unconditional probabiliteis. A sample space was defined and all probabilities were calculated with respect to that sample space. In many instances, however, we are in a position to update the sample space based on new information. In such cases, we want to be able to update probability calculations or to calculate *conditional probabilites.*

Definition 6. Consider a probability space $(\Omega, \mathcal{F}, P)$. If A and B are events in $\mathcal{F}$, and $P(B) > 0$, then the *conditional probability* of A given B, written $P(A|B)$, is

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

Example 4. Some trivial properties of conditional probabilities can be obtain.

1. We can notice that $P(A|B) = 1$ if $A \subset B$. The intuition is that our original sample space, Ω, has been updated to B. All further occurrences are then calibarated with respect to their relation to B.

2. In particular, note what happens to conditional probabilites of disjoint sets. Suppose A and B are disjoint, so $P(A \cap B) = 0$. It then follows that $P(A|B) = P(B|A) = 0$.

3. We have $P(A \cap B) = P(A|B) P(B)$, and similarly, $P(A \cap B) = P(B|A) P(A)$, then $P(A|B) = P(B|A) \frac{P(A)}{P(B)}$.

Proposition 4. *(law of total probability) Consider a probability space $(\Omega, \mathcal{F}, P)$ with a partition $\{C_1, C_2 \cdots\}$ of Ω. For all $A \in \mathcal{F}$, $P(A) = \sum_{i=1}^{\infty} P(A|B_i) P(B_i)$.*

Theorem 4. *(Bayes' rule) Let $A_1, A_2, \ldots,$ be a partition of the sample space, and let $B \in \mathcal{F}$. Then for each $i = 1, 2, \ldots,$*

$$P(A_i|B) = \frac{P(B|A_i) P(A_i)}{\sum_{j=1}^{\infty} P(B|A_j) P(A_j)}.$$

In some cases, it may happen that the occurence of a particular event, B, has no effect on the probability of another event, A. Symbolically, we are saying that

$$P(A|B) = P(A).$$

If this holds, then by Bayes' rule and Example 4, we have,

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

so the occurrence of A has no effect on B. Moreover, since $P(B|A) P(A) = P(A \cap B)$, it then follows that

$$P(A \cap B) = P(A) P(B).$$

Definition 7. For a probability space $(\Omega, \mathcal{F}, P)$, if $A, B \in \mathcal{F}$, A and B are *statistically independent* if

$$P(A \cap B) = P(A) P(B).$$

Based on the independence, we introduce the mutual independence.

Definition 8. Consider a probability space$(\Omega, \mathcal{F}, P)$ and a set of events $\{A_1, A_2, \ldots, A_n\}$. We say that $\{A_1, \ldots, A_n\}$ are *mutually independent* if any subcollection $A_{i_1}, \ldots, A_{i_k}$, we have

$$P\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} P\left(A_{i_j}\right).$$

Exercise 7. Consider the experiment of tossing a fair coin three times. List the sample space Ω. Let H_i be the event that ith toss is a head. Prove that H_1, H_2 and H_3 are mutually independent.

Probability Distributions

1 Random Variables

In many experiments it is easier to deal with a summary variable X than with the original probability structure.

Example 1.1. In an opinion poll, we might decide to ask 50 people whether they agree or disagreee with a certain issue. If we record a 1 for agree and 0 for disagree, the sample space for this experiment has 2^{250} elements, each an ordered string of 1s and 0s of length 50. We should be able to reduce this to a reasonable size! It may be that the only quantity of interest is the number of people who agree out of 50 and, if we define a variable X =number of 1s recorded out of 50, we have captured the essence of the problem. Note that the sample space for X is the set of integers $\{0, 1, 2, \ldots, 50\}$ and is much easier to deal with than the original sample space.

Usually in defining the quantity X, we have to defined a mapping from the original sample space to a new sample space, usually a set of real numbers.

Suppose Ω is the sample space and $\mathcal{F}$ is the corresponding $\sigma-$algebra generated by Ω. Let $\nleqslant'$ be the another set and $\mathcal{F}'$ be the $\sigma-$algebra generated by Ω'. Hence, $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ are two measurable spaces. Suppose a function,

$$X : \Omega \to \Omega',$$

meaning X is a function with domain Ω and range $\nleqslant'$.Then X determins a function,

$$X^{-1} : \mathcal{P}(\Omega') \to \mathcal{P}(\Omega),$$

where $\mathcal{P}(\Omega')$ and $\mathcal{P}(\Omega)$ are the power sets for Ω' and Ω, respectively. It can be defined by

$$X^{-1}(A') = \{\omega \in \Omega : X(\omega) \in A'\}, \ for\ A' \in \Omega',$$

which you can see from the Figure 1.

Definition 1.2. The map X is called *measurable* if $X^{-1}(\mathcal{F}') \subset \mathcal{F}$. Then X is also called a *random element* of Ω'. The following notation is used to represent X,

$$X : (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}').$$

Definition 1.3. Consider a probability space $(\Omega, \mathcal{F}, P)$, and suppose

$$X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})),$$

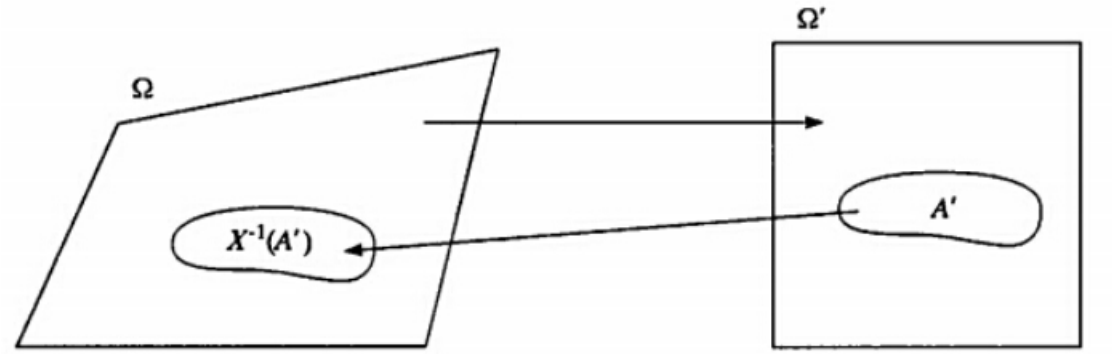
where $\mathcal{B}(\mathbb{R})$ is the $\sigma-$algebra generated by $\mathbb{R}$, then X is called *random variable.*

Although we define the random variable, we must check that our probability function, which is defined on the original sample space, can be used for the random varaible.

Suppose we have a sample space $S = \{s_1, \ldots, s_n\}$ with a probability function P and we define a random variable X with a range $\mathcal{X} = \{x_1, \ldots, x_m\}$. We can define a probability function P_X on $\mathcal{X}$ in the following way. We will observe $X = x_i$ if and only if the outcome of the random experiment is an $s_j \in S$ such that $X(s_j) = x_i$. Thus,

$$P_X(X = x_i) = P(\{s_j \in S : X(s_j) = x_i\}).$$

Figure 1: Inverse



Definition 1.4. Consider a probability space $(\Omega, \mathcal{F}, P)$, for $X : (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$, define a set function, $P \circ X^{-1}(A') = P\left(X^{-1}(A')\right)$. Then $P \circ X^{-1}$ is a probability function on $(\Omega', \mathcal{F}')$, and it is called *induced probability* or the *distribution* of X.

Hence, based on the discussion above, we don't have to discuss the probability based on the general set now. We can turn to the probability based on the measural space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, which we are very familiar with.

Exercise 1.5. Consider the experiment of tossing a fair coin three times again. Let X =the number of heads obtained in the three tosses. Find the range of X , $\mathcal{X}$ and the induced probability function on $\mathcal{X}$.

2 Distribution Functions

With every random variable X, we associate a function called the cumulative distribution function of X.

Definition 2.1. The *cumulative distribution function* or *cdf* of a random variable X, denoted by $F_X(x)$, is defined by

$$F_X(x) = P_X(X \leq x), \text{ for all } x.$$

Exercise 2.2. *(Cintinuation of Exercise 1.5)*Consider the experiment of tossing three fair coins, and let X =number of heads observed. Define and draw the cdf of X.

Theorem 2.3. *The function $F(x)$ is a cdf if and only if the following three conditions hold:*

- $\lim_{x\to-\infty} F(x) = 0$ *and* $\lim_{x\to+\infty} F(x) = 1$.
- $F(x)$ *is a nondecreasing function of* x.
- $F(x)$ *is right-continuous; that is, for every number* x_0, $\lim_{x\downarrow x_0} F(x) = F(x_0)$.

Exercise 2.4. Suppose we do an experiment that consists of tossing a coin until a head appears. Let p =probability of a head on any given toss, and X =number of tosses required to get a head. Define the cdf of X and verify the three properties of this cdf.

We may state that F_X completely determines the probability distribution of a random variable X. This is true if $P(x \in A)$ is defined only for events in $\mathcal{B}^1$, the smallest $\sigma-$algebra containing all the intervals of real numbers of the form $(a, b), [a, b), [a, b]$ and $(a, b]$. If probability is defined for a

larger class of events, it is possible for two random variables to have the same distribution function but not the same probability for every event. In our usual applications, we are concerned only with events that are intervals, countable unions or intersections of intervals. Hence, we do not consider such pathological cases.

Definition 2.5. The random variables X and Y are *identically distributed* if, for every set $A \in \mathcal{B}^1$, $P(X \in A) = P(Y \in A)$.

Note that two random variables that are identically distributed are not necessarily equal.

Theorem 2.6. *The following two statements are equivalent:*

1. *The random variables X and Y are identically distributed.*
2. *$F_X(x) = F_Y(x)$ for every x.*

Exercise 2.7. *(Cintinuation of Exercise 2.2)* Define Y =the number of tails observed. Check that X and Y are identically distributed.

3 Density and Mass Functions

Exercise 3.1. *(logistic distribution)* Check the function $F_X = \frac{1}{1+e^{-x}}$ is a cdf.

Combined with Exercise 2.4, whether a cdf is continuous or has jumps corresponds to the associated random variable being continuous or not.

Definition 3.2. A random variable X is *continuous* if $F_X(x)$ is a continuous function of x A random variable X is *discrete* if $F_X(x)$ is a step function of x, which means X only takes values that are in some countable subset $\{x_1, x_2, \dots\}$ of $\mathbb{R}$..

Associatedwith a random variable X and its cdf $F_X(x)$ is another function, called either the *probability density function(pdf)* or *probability mass function(pmf).* The terms pdf and pmf refer to the continuous and discrete cases, respectively.

Definition 3.3. The *probability mass function(pmf)* of a discrete random variable X ,$f_X(x) : \mathbb{R} \to [0, 1]$, is given by

$$f_X(x) = P(X = x), \text{ for all } x,$$

which is the probability at the point x.

Proposition 3.4. *Properties of pmf:*

- $0 \le f_X(x) \le 1$, $\forall x$.
- $f_X(x) = 0$, *if* $x \notin \{x_1, x_2, \dots\}$.
- $\sum_x f_X(x) = 1$.

For the relationship between the cdf of X and pmf of X,

$$F_X(x) = \sum_{u:u \le x} f_X(u).$$

Definition 3.5. The *probability density function(pdf)*, $f_X(x)$, of a continuous random variable is the integrable function $f_X(x) : \mathbb{R} \to [0, \infty)$ that satiesfies

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

It can also be written as

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

Note that the expression "X has a distribution given by $F_X(x)$" is abbreviated symbolically by "$X \sim F_X(x)$".

For continuous random variable, since $\{X = x\} \subset \{x - \epsilon < X \leq x\}$, for any $\epsilon > 0$, then

$$P(X = x) \leq P(x - \epsilon < X \leq x) = F_X(x) - F_X(X - \epsilon),$$

for any $\epsilon > 0$. Therefore,

$$0 \leq P(X = x) \leq \lim_{\epsilon \downarrow 0} [F_X(x) - F_X(X - \epsilon)] = 0,$$

by the continuity of F_X. It implies that

$$P(a < x < b) = P(a \leq x \leq b) = P(a \leq x < b) = P(a < x \leq b).$$

Proposition 3.6. *Properties of pdf:*

- $0 \leq f_X(x) \leq 1$, $\forall x$.
- $\int_{-\infty}^{\infty} f_X(x) = 1$.

Exercise 3.7. Check the following function is a pdf,

1. *(Normal distribution)* $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right)$, $x \in \mathbb{R}$.
2. *(Exponential distribution)* $f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & Otherwise \end{cases}$.
3. *(Cauchy distribution)* $f_X(x) = \frac{1}{\pi(1+x^2)}$ for $x \in \mathbb{R}$.

4 Expected Value and Variance

The expected value, or expectation, of a random variable is merely its average value, where we speak of "average" value as one that is weighted according to the probability distribution. By weighting the values of the random variable according to the probability distribution, we hope to obtain a number that summarizes a typical or expected value of an observation of the random variable.

Definition 4.1. If X is a random variable with sample space $\mathcal{X}$($\mathbb{R}$ or subsets of $\mathbb{R}$), the *expected value* or *mean* of the function $g(X) : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y}$ is new sample space($\mathbb{R}$ or subsets of $\mathbb{R}$), is

$$E_X[g(X)] = \begin{cases} \int_{\mathcal{X}} g(x) f_X(x)\, dx & \text{If } X \text{ is continuous} \\ \sum_{x \in \mathcal{X}} g(x) f_X(x) = \sum_{x \in \mathcal{X}} g(x) P(X = x) & \text{If } X \text{ is discrete,} \end{cases}$$

provided that the integral or sum exist, that is, $E_X[|g(X)|] < \infty$. In particular, the *mean* of the random variable X is defined as,

$$E_X[X] = \begin{cases} \int_{\mathcal{X}} x f_X(x)\, dx & \text{If } X \text{ is continuous} \\ \sum_{x \in \mathcal{X}} x f_X(x) = \sum_{x \in \mathcal{X}} x P(X = x) & \text{If } X \text{ is discrete,} \end{cases}$$

provided that the integral or sum exist, that is, $E_X[|X|] < \infty$.

Note that usually we just write it as $E[g(X)]$ and $E(X)$.

Exercise 4.2. Calculate the mean of the follow distributin,

1. $X \sim Exp(\lambda)$, with $f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{Otherwise} \end{cases}$.
2. Let pdf of X be $f_X(x) = \frac{1}{\pi(1+x^2)}$.
3. Let pdf of X be $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$, $x \in \mathbb{R}$.

Theorem 4.3. *Let X be a random variable and let a, b, c be constants. Then for any functions $g_1(x)$ and $g_2(x)$ whose expectations exist,*

- $E(ag_1(x) + bg_2(x) + c) = aE[g_1(x)] + bE[g_2(x)] + c$.
- *If $g_1(x) \geq 0$ for all x, then $E[g_1(x)] \geq 0$.*
- *If $g_1(x) \geq g_2(x)$ for all x, then $E[g_1(x)] \geq E[g_2(x)]$.*
- *If $a \leq g_1(x) \leq b$ for all x, then $a \leq E[g_1(x)] \leq b$.*

Exercise 4.4. Prove Theorem 4.3.

Exercise 4.5. Prove $\min_b E\left[(X-b)^2\right] = E(X - EX)^2$.

Definition 4.6. If X is a random variable with sample space $\mathcal{X}$($\mathbb{R}$ or subsets of $\mathbb{R}$), the variance of X is $Var(X) = E_X\left[(X - EX)^2\right]$, that is,

$$Var(X) = \begin{cases} \int_{\mathcal{X}} [x - E(X)]^2 f_X(x)\, dx & \text{If } X \text{ is continuous} \\ \sum_{x \in \mathcal{X}} [x - E(X)]^2 f_X(x) = \sum_{x \in \mathcal{X}} [x - E(X)]^2 P(X = x) & \text{If } X \text{ is discrete,} \end{cases}$$

whenever the sum or integral is fintie.

Exercise 4.7. Calculate the variance of the follow distributin,

1. $X \sim Exp(\lambda)$, with $f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{Otherwise} \end{cases}$.
2. Let pdf of X be $f_X(x) = \frac{1}{\pi(1+x^2)}$.
3. Let pdf of X be $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$, $x \in \mathbb{R}$.

Theorem 4.8. *If X is a random variable with finite variance, then for any constants a and b,*

$$Var(aX + b) = a^2 Var(X).$$

Exercise 4.9. Prove Theorem 4.8.

5 Higher Moments and Moment Generating Functions

The various moments of a distribution are an important class of expectations.

Definition 5.1. For each integer, the nth *moment* of X (or $F_X(x)$), μ'_n, is

$$\mu'_n = E(X^n).$$

The nth *central moment* of X, μ_n, is

$$\mu_n = E(X - \mu)^n,$$

where $\mu = \mu'_1 = E(X)$.

We now introduce a new function that is associated with aprobability distribution, the *moment generating function(mgf)*. As its name suggests, the mgf can be used to generate moments. In practise, it is easier in many cases to calculate moments directly than to use the mgf. However, the main use of the mgf is not to generate moments, but to help characterizing a distribution. This property can lead to some extremely powerful results when used properly.

Definition 5.2. Let X be a random variable with cdf F_X. The *moment generating function(mgf)* of X, denoted by $M_X(t)$, is

$$M_X(t) = E\left[e^{tX}\right] = \begin{cases} \int_{\mathcal{X}} e^{tx} f_X(x)\,dx & \text{If } X \text{ is continuous} \\ \sum_{x\in\mathcal{X}} e^{tx} f_X(x) = \sum_{x\in\mathcal{X}} e^{tx} P(X=x) & \text{If } X \text{ is discrete,} \end{cases},$$

provided that the expectation exists for t in some neighborhood of 0. That is, $\exists h > 0$, $\forall t \in (-h, h)$, $E\left[e^{tX}\right]$ exists.

If the expectation does not exist in a neighborhood of 0, we say that the moment generating function does not exist.

Theorem 5.3. *If X has mgf $M_X(t)$, then*

$$E(X^n) = M_X^{(n)}(0),$$

where

$$M_X^{(n)}(0) = \left.\frac{d^n}{dt^n} M_X(t)\right|_{t=0}.$$

That is, the nth moment is equal to the nth derivative of $M_X(t)$ evaluated at $t = 0$.

Exercise 5.4. Prove Theorem 5.3.

Exercise 5.5. Calculate the mgf of the follow distributin. Use the mgf to calculate the mean and variance.

1. $X \sim Exp(\lambda)$, with $f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{Otherwise} \end{cases}$.
2. Let pdf of X be $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$, $x \in \mathbb{R}$.

Proposition 5.6. *For any constants a and b, the mgf of the random variable $aX + b$ is given by*

$$M_{aX+b}(t) = e^{bt} M_X(at).$$

Exercise 5.7. Prove Proposition 5.6.

As we mentioned, the usefulness of the mgf stem from the fact that, in many cases, the mgf can characterize a distribution.

Theorem 5.8. *Let $F_X(x)$ and $F_Y(y)$ be two cdfs all of whose moment exist. If the mgf exist and $M_X(t) = M_Y(t)$ for all t in some neighborhood of 0, then $F_X(u) = F_Y(u)$ for all u.*

Further, we can have the following theorem.

Theorem 5.9. *(Convergence of mgfs) Suppose $\{X_i, i = 1, 2, \dots\}$ is a sequence of random variables, each with mgf $M_{X_i}(t)$. Furthermore, suppose that*

$$\lim_{i\to\infty} M_{X_i}(t) = M_X(t), \text{ for all } t \text{ in a neighborhood of 0,}$$

and $M_X(t)$ is an mgf. Then there is a unique cdf F_X whose moments are determened by $M_X(t)$ and, for all x where $F_X(x)$ is continuous, we have

$$\lim_{i\to\infty} F_{X_i}(t) = F_X(t).$$

That is, convergence, for $|t| < h$, of mgfs to an mgf implies convergence of cdfs(Convergence in distribution).

6 Differentiating and Integral

In some cases, we will encounter such a problem like

$$\frac{d}{d\theta}\int_{a(\theta)}^{b(\theta)} f(x,\theta)\,dx,$$

where $-\infty < a(\theta), b(\theta) < \infty$ for all θ. Sometimes we are desired to interchange the order of integration and differentiation.

The following theorem provides a rule for that problem.

Theorem 6.1. *(Leibnitz's rule) If $f(\theta)$, $a(\theta)$ and $b(\theta)$ are differentiable with respect to θ, then*

$$\frac{d}{d\theta}\int_{a(\theta)}^{b(\theta)} f(x,\theta)\,dx = f(b(\theta),\theta)\frac{d}{d\theta}b(\theta) - f(a(\theta),\theta)\frac{d}{d\theta}a(\theta) + \int_{a(\theta)}^{b(\theta)} \frac{\partial f(x,\theta)}{\partial\theta}dx.$$

Thus if $a(\theta)$ and $b(\theta)$ are constant, we have a special case of Leibnitz's rule:

$$\frac{d}{d\theta}\int_{a}^{b} f(x,\theta)\,dx = \int_{a}^{b} \frac{\partial f(x,\theta)}{\partial\theta}dx.$$

Thus, in general, if we have the integral of a differentiable function over a fintie range, differentiation of the integral poses no problem.

However, for infinite range, problems can arise. For such problems, we use the following theorem to verify the interchange. The following theorems are all corollaries of Lebesgue's Dominated Convergence Theorem.

Theorem 6.2. *Suppose the function $h(x,y)$ is continuous at y_0 for each x, and there exists a function $g(x)$ satisfying*

- *$|h(x,y)| \le g(x)$ for all x and y,*
- *$\int_{-\infty}^{+\infty} g(x)\,dx < \infty$.*

Then

$$\lim_{y\to y_0}\int_{-\infty}^{+\infty} h(x,y)\,dx = \int_{-\infty}^{+\infty}\lim_{y\to y_0} h(x,y)\,dx.$$

The key condition in this theorem is the existence of a dominating function $g(x)$, with a finite integral, which ensures the integrals connot be too badly behaved. Let

$$h(x,y) = \frac{f(x,\theta+\delta) - f(x,\theta)}{\delta},$$

then we have the following theorem.

Theorem 6.3. *Suppose $f(x,\theta)$ is differentiable at $\theta = \theta_0$, that is,*

$$\lim_{\delta\to 0}\frac{f(x,\theta_0+\delta) - f(x,\theta_0)}{\delta} = \frac{\partial}{\partial\theta}f(x,\theta)\bigg|_{\theta=\theta_0}$$

exists for every x, and there exists a function $g(x,\theta_0)$ and a constant $\delta_0 > 0$ such that

- *$\left|\frac{f(x,\theta_0+\delta)-f(x,\theta_0)}{\delta}\right| \le g(x,\theta_0)$, for all x and $|\delta| \le \delta_0$.*
- *$\int_{-\infty}^{+\infty} g(x,\theta_0)\,dx < \infty$.*

Then

$$\frac{d}{d\theta}\int_{-\infty}^{+\infty} f(x,\theta)\,dx\bigg|_{\theta=\theta_0} = \int_{-\infty}^{+\infty}\left[\frac{\partial}{\partial\theta}f(x,\theta)\bigg|_{\theta=\theta_0}\right]dx.$$

In Theorem 6.3, the statement is for one value of θ, θ_0. Typically, if $f(x,\theta)$ is differentiable at all θ, not joust one value θ_0, we have the following corollary.

Corollary 6.4. *Suppose $f(x,\theta)$ is differentiable in θ and there exists a function $g(x,\theta)$ such that*

- $\left|\frac{\partial}{\partial\theta}f(x,\theta)\big|_{\theta=\theta'}\right| \le g(x,\theta)$, *for all θ' such that $|\theta'-\theta| \le \delta_0$.*
- $\int_{-\infty}^{+\infty} g(x,\theta)\,dx < \infty$, *for all θ' such that $|\theta'-\theta| \le \delta_0$.*

Then

$$\frac{d}{d\theta}\int_{-\infty}^{+\infty} f(x,\theta)\,dx\bigg|_{\theta=\theta_0} = \int_{-\infty}^{+\infty}\left[\frac{\partial}{\partial\theta}f(x,\theta)\bigg|_{\theta=\theta_0}\right]dx.$$

Random Variables

1 Univariate Random Variable

1.1 Transformation for Univariate Distribution

If X is a random variable, then any function of X, say $g(X)$, is also a random variable. Often $g(X)$ is of interest itself and we write $Y = g(X)$ to denote the new random variable $g(X)$. Formally, $g(X)$ is defined as

$$g(X) : \mathcal{X} \to \mathcal{Y}.$$

Then for any set $A \subset \mathcal{Y}$, we have

$$P(Y \in A) = P(g(X) \in A) = P(\{x \in \mathcal{X} : g(x) \in A\}) = p\left(X \in g^{-1}(A)\right).$$

When transformations are made, it is important to keep track of the sample spaces of the random variables. We define

$$\mathcal{X} = \{x : f_X(x) > 0\} \text{ and } \mathcal{Y} = \{y : y = g(x), \text{for some } x \in \mathcal{X}\}. \tag{1.1}$$

The pdf of X is positive only on the set $\mathcal{X}$ and is 0 elesewhere, $\mathcal{X}$ is called *support* of a distribution. This terminology can also apply to a pmf.

Theorem 1.1. *Let X have cdf $F_X(x)$, let $Y = g(X)$, and let $\mathcal{X}$ and $\mathcal{Y}$ be defined as (1.1).*

- *If g is an increasing function on $\mathcal{X}$, $F_Y(y) = F_X\left(g^{-1}(y)\right)$ for $y \in \mathcal{Y}$.*
- *If g is a decreasing function on $\mathcal{X}$ and X is a continuous random variable, $F_Y(y) = 1 - F_X\left(g^{-1}(y)\right)$ for $y \in \mathcal{Y}$.*

Exercise 1.2. Prove Theorem 1.1.

Theorem 1.3. *Let X have a pdf $f_X(x)$ and let $Y = g(X)$, where g is a monotone function. Let $\mathcal{X}$ and $\mathcal{Y}$ be defined as (1.1). Suppose that $f_X(x)$ is conituous on $\mathcal{X}$ and that $g^{-1}(y)$ has a continuous derivative on $\mathcal{Y}$. Then the pdf of Y is given by*

$$f_Y(y) = \begin{cases} f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right| & y \in \mathcal{Y} \\ 0 & otherwise \end{cases}.$$

Exercise 1.4. Prove Theorem 1.3.

1.2 Normal Distribution

A fundamental distribution in statistical inference. Not only it provides a reasonable model for many quantities that practitioners are interested in it is also the approximate distribution of many statistics like sample mean. A key feature of the normal distribution is that it is completely characterised by its mean and variance.

Definition 1.5. The pdf of the *normal distribution* with mean μ and variance σ^2 is given by

$$f\left(x|\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\mathbf{x}-\mu)^{\mathbf{2}}}{2\sigma^2}\right),-\infty<x<\infty. \tag{1.2}$$

If $X\sim N\left(\mu,\sigma^2\right)$, then the random variable $Z=\left(X-\mu\right)/\sigma$ has $N\left(0,1\right)$ distribtuion also known as the *standard normal*. It is easily established by the result in Section 1.1,

$$f\left(Z\leq z\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}e^{-\frac{t^2}{2}}dt.$$

then the pdf of Z is $f\left(z\right)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)$.

Example 1.6. As we have talked, $f\left(z\right)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)$ is a pdf, then it should have

$$\int_{-\infty}^{+\infty}\exp\left(-\frac{z^2}{2}\right)dz=\sqrt{2\pi}.$$

Notice that the ingegrand above is symmetric around 0, implying that $\mathrm{m}\int_0^{+\infty}\exp\left(-\frac{z^2}{2}\right)dz=\int_{-\infty}^{0}\exp\left(-\frac{z^2}{2}\right)dz$, which means

$$\int_{0}^{+\infty}\exp\left(-\frac{z^2}{2}\right)dz=\sqrt{\frac{\pi}{2}}.$$

It can be shown in the following.

$$\begin{aligned}\left[\int_0^{+\infty}\exp\left(-\frac{z^2}{2}\right)dz\right]^2 &= \int_0^{+\infty}\exp\left(-\frac{u^2}{2}\right)du\int_0^{+\infty}\exp\left(-\frac{v^2}{2}\right)dv\\ &= \int_0^{+\infty}\int_0^{+\infty}\exp\left(-\frac{u^2}{2}-\frac{v^2}{2}\right)dudv.\end{aligned}$$

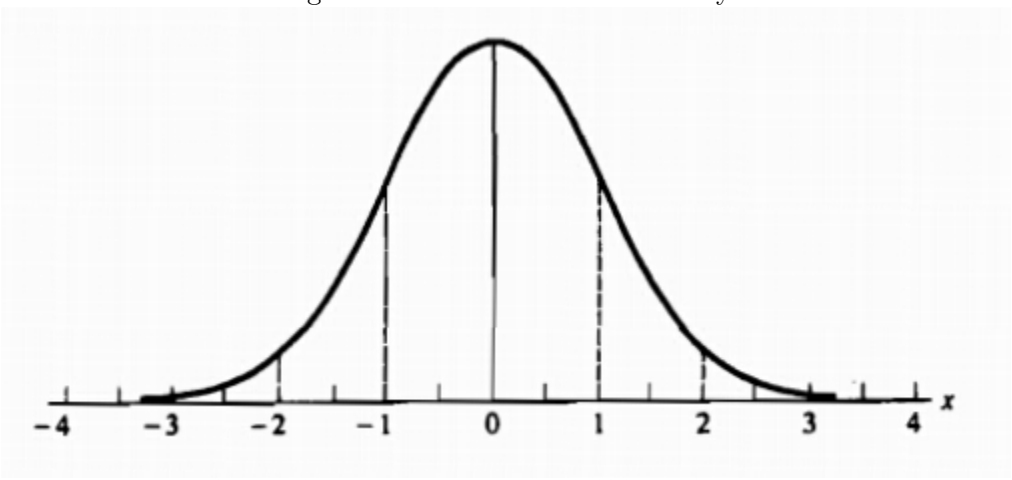
Define $u=r\cos\theta$, and $v=r\sin\theta$,

$$abs\left(\left|\frac{\partial\left(u,v\right)}{\partial\left(r,\theta\right)}\right|\right)=abs\left(\left|\begin{array}{cc}\cos\theta & \sin\theta\\ -r\sin\theta & r\cos\theta\end{array}\right|\right)=r,$$

then,

$$\begin{aligned}\int_0^{+\infty}\int_0^{+\infty}\exp\left(-\frac{u^2}{2}-\frac{v^2}{2}\right)dudv &= \int_0^{\frac{\pi}{2}}\int_0^{+\infty}\exp\left(-\frac{r^2}{2}\right)abs\left(\left|\frac{\partial\left(u,v\right)}{\partial\left(r,\theta\right)}\right|\right)drd\theta\\ &= \int_0^{\frac{\pi}{2}}\int_0^{+\infty}\exp\left(-\frac{r^2}{2}\right)rdrd\theta\\ &= \frac{\pi}{2}\int_0^{+\infty}\exp\left(-\frac{r^2}{2}\right)rdr\\ &= \frac{\pi}{2}.\end{aligned}$$

Figure 1: Standard Normal Density



Since the normal distribution is determined by its mean and variance, we can use the result of last class to derive the following properties. If $X \sim N\left(\mu, \sigma^2\right)$, then $Y = aX + b \sim N\left(a\mu + b, a^2\sigma^2\right)$.

Straightforward calculus show that the normal pdf (1.2) has its maximum at $x = \mu$ and inflection points(where the curve changes from concave to convex) at $\mu \pm \sigma$. Furthermore, the probability content within 1,2, or 3 standard deviation from the mean is

$$P\left(|X - \mu| \leq \sigma\right) = P\left(|Z| < 1\right) = .6826,$$

$$P\left(|X - \mu| \leq 2\sigma\right) = P\left(|Z| < 2\right) = .9544,$$

$$P\left(|X - \mu| \leq 2\sigma\right) = P\left(|Z| < 3\right) = .9974.$$

Figure 1.1 shows the features of the nomal pdf.

Exercise 1.7. Examine the mgf of normal distribution $N\left(\mu, \sigma^2\right)$ is $M_X\left(t\right) = \exp\left(\mu t + \frac{\sigma^2}{2}t^2\right)$.

1.3 Gamma Distribution

The gamma family of distributions is a flexible family of distributions on $[0, \infty)$. At first, we talk about the *gamma function*.

Definition 1.8. The *gamma function* is a function $\Gamma\left(\alpha\right) : \mathbb{R}^+ \to \mathbb{R}^+$ such that

$$\Gamma\left(\alpha\right) = \int_0^\infty t^{\alpha-1}e^{-t}dt.$$

One can easily verify that $\Gamma\left(\alpha + 1\right) = \alpha\Gamma\left(a\right)$, In particular, when $\alpha = n$, $\Gamma\left(n\right) = n \times (n-1) \times \cdots \times 1 = n!$.

Based on the gamma function it immediately follows that

$$f\left(t|\alpha\right) = \frac{t^{\alpha-1}e^{-t}}{\Gamma\left(\alpha\right)},\ 0 < t < \infty,$$

is a pdf. Further, we let $X = \beta T$, from the transformation rule, we have,

$$f_X(x|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}}, 0 < x < \infty, \alpha > 0, \beta > 0. \tag{1.3}$$

Definition 1.9. A random variable X with pdf (1.3) is called that X follows gamma distribution with parameter α and β, denoted as $X \sim Gamma(\alpha,\beta)$.

The α is called shape parameter since it most influences the peakedness of the distribution, while the parameter β is called the scale parameter, since most of its influence i son the spread of the distribution.

If $X \sim Gamma(\alpha,\beta)$, we have

- *Moment generating function:* $M_X(t) = \left(\frac{1}{1-\beta t}\right)^\alpha, t < \frac{1}{\beta}$.
- $E(X) = \alpha\beta$, $Var(X) = a\beta^2$.

Exercise 1.10. Verify the properties of the gamma distribution.

1.4 Chi-square Distribution

Chi-square distribution is a very widely used distribution in econometrics. Actually it is a special case of gamma distribution family. If we set $\alpha = p/2$, where p is an integer, and $\beta = 2$, then the gamma pdf becomes

$$f(x|p) = \frac{1}{\Gamma(p/2)\,2^{p/2}} x^{p/2-1} e^{-x/2},\ 0 < x < \infty. \tag{1.4}$$

Definition 1.11. A random variable X with pdf (1.4) is called that X follows Chi-square distribution with p degrees of freedom, denoted as $X \sim \chi^2(p)$.

We know that Chi-square distribution is special distribution in the gamma distribution family. Hence, it shares the properties of the gamma distribution. If $X \sim \chi^2(p)$, then

- *Moment generating function:*$M_X(t) = \left(\frac{1}{1-2t}\right)^{\frac{p}{2}}, t < \frac{1}{2}$.
- $E(X) = p$, $Var(X) = 2p$.

1.5 Beta Distribution

Before we talk about the beta distribution, we have to talk about the beta function.

Definition 1.12. The beta function is defined as $B(\alpha,\beta) : \mathbb{R}^+ \times \mathbb{R}^+ \to [0,+\infty)$,

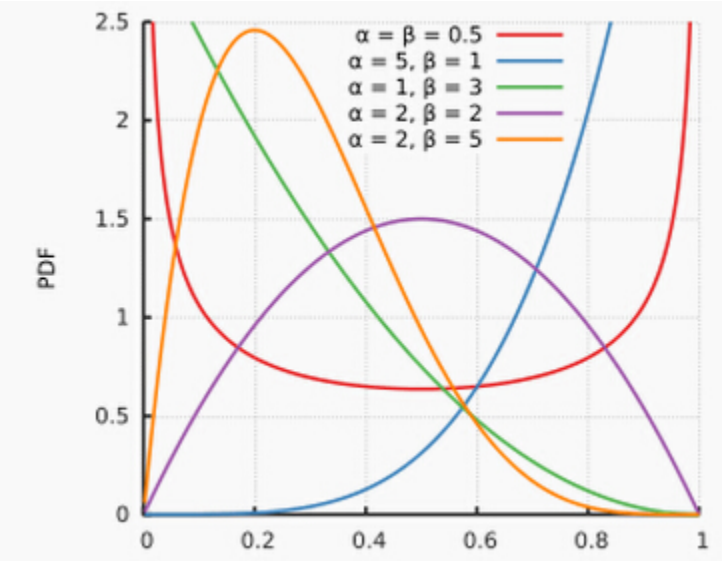
$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx.$$

The beta function is related to the gamma function through the following identity,

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

This equation is very useful in dealing with the beta function since we can take advantage of the properties of the gamma function.

Figure 2: Beta Densities



Definition 1.13. The beta family of distributions is continuous family on $(0,1)$ indexed by two parameters (α, β),

$$f(x|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, 0 < x < 1, \alpha > 0, \beta > 0.$$

The beta distribution is one of the few common 'named' distributions that give probability 1 to a finite interval, here taken to be $(0,1)$. As the parameters α and β vary, the beta distribution takes on many shapes, as shown in Figure 2.

Further, for the beta distribution, we have

- $E(X^n) = \frac{1}{B(\alpha,\beta)} \int_0^1 x^{\alpha+n-1}(1-x)^{\beta-1} dx = \frac{B(\alpha+n,\beta)}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+n)\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+n)\Gamma(\alpha)}$.
  $E(X) = \frac{\alpha}{\alpha+\beta}$, $Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.
- *Moment generating function:* $M_X(t) = 1 + \sum_{k=1}^{\infty} \left(\prod_{r=0}^{k-1} \frac{\alpha+r}{\alpha+\beta+r} \right) \frac{t^k}{k!}$.

2 Multivariate Random Variable

What we have talked about are all univariate random variable, actually we always encounter multivariate case. A multivariate model attempts to capture the nature of any dependencies (e.g. dependence between income and expenditure).

Definition 2.1. An *n-dimensional random variables* is a function from a sample space Ω into $\mathbb{R}^d$, the n-dimensional Euclidean space. That is, $\boldsymbol{X} : \Omega \to \mathbb{R}^d$. And the probability space change from $(\Omega, \mathcal{F}, P')$ to $\left(\mathbb{R}^d, \sigma\left(\mathbb{R}^d\right), P\right)$.

2.1 Joint and Marginal Distributions

Similar to the univariate case, for multivariate random variables, we have the *joint distribution function, joint density/mass function, moment generating function.*

Definition 2.2. If $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)'$ is a random variable, the joint cumulative distribution is a function $F_{\boldsymbol{X}} : \mathbb{R}^d \to [0,1]$ given by, $F_{\boldsymbol{X}}(\boldsymbol{x}) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)$, where $\boldsymbol{x} = (x_1, \ldots, x_n)$.

Most ideas associated with multivariate distributions are illustrated using the bivariate case. The generalisations to n−dimensions are usually obvious.

Theorem 2.3. *Suppose that X and Y are random variables. If $F_{X,Y}$ is the joint distribution function of X and Y then $F_{X,Y}$ has the following properties.*
$F_{X,Y}(-\infty, y) = \lim_{x \to -\infty} F_{X,Y}(x,y) = 0$.
$F_{X,Y}(y, -\infty) = \lim_{y \to -\infty} F_{X,Y}(x,y) = 0$.
$F_{X,Y}(\infty, \infty) = \lim_{x \to \infty, y \to \infty} F_{X,Y}(x,y) = 1$.
Right continuous in x and y: $\lim_{h \to 0^+} F_{X,Y}(x+h, y) = F_{X,Y}(x,y)$ *and* $\lim_{h \to 0^+} F_{X,Y}(x, y+h) = F_{X,Y}(x,y)$.

There are also discrete random variables and continuous random variabls. Correspondingly we have the following definitions.

Definition 2.4. Let $\boldsymbol{Z} = (X,Y)'$ be a discrete bivariate random vector. Then the function $f(x,y) : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ defined by $f(x,y) = P(X = x, Y = y)$ is called the *joit probability mass function* or *joint pmf* of $\boldsymbol{Z}$. Sometime we also write $f_{\boldsymbol{Z}}(x,y) = P(X = x, Y = y)$.

Apparently, we have,

- $0 \leq f_{\boldsymbol{Z}}(x,y) \leq 1$.
- $\sum_{(x,y) \in \mathbb{R}^2} f_{\boldsymbol{Z}}(x,y) = 1$.
- $P((X,Y) \in A) = \sum_{(x,y) \in A} f_{\boldsymbol{Z}}(x,y)$.

Definition 2.5. Let $\boldsymbol{Z} = (X,Y)'$ be a discrete bivariate random vector. Then the function $f(x,y) : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ is called a *joint probability density function* or *joint pdf* of the *continuous bivariate random vector* $\boldsymbol{Z}$ if , for every $A \subset \mathbb{R} \times \mathbb{R}$,

$$P(\boldsymbol{Z} \in A) = \int \int_A f(x,y)\, dxdy.$$

We sometimes denote the $\mathbb{R} \times \mathbb{R}$ as $\mathbb{R}^2$. And we have,

- $0 \leq f_{\boldsymbol{Z}}(x,y) \leq 1$.
- $\int \int_{\mathbb{R}^2} f(x,y)\, dxdy = 1$.
- $\frac{\partial^2 F(x,y)}{\partial x \partial y} = f(x,y)$.

Definition 2.6. Let $\boldsymbol{Z} = (X, Y)'$ be a discrete bivariate random vector. Function $g(x, y)$ is defined as a real-valued function, $g(X, Y) : \mathbb{R}^2 \to \mathbb{R}$. The expectation of $g(X, Y)$ is

$$E[g(X, Y)] = \begin{cases} \sum_{(x,y)\in\mathbb{R}^2} g(x, y) f(x, y) & \text{if } \boldsymbol{Z} \text{ is discrete} \\ \int_{\mathbb{R}} \int_{\mathbb{R}} g(x, y) f(x, y) \, dxdy & \text{if } \boldsymbol{Z} \text{ is continuous} \end{cases}.$$

Definition 2.7. Let $\boldsymbol{Z} = (X, Y)'$ be a discrete bivariate random vector. Denote $t = (t_1, t_2)'$. The *moment generating function* of $\boldsymbol{Z}$ is defined as

$$M_{\boldsymbol{Z}}(t) = E_{\boldsymbol{Z}}\left(e^{t'\boldsymbol{Z}}\right) = \begin{cases} \sum_{(x,y)\in\mathbb{R}^2} \exp(t_1 X + t_2 Y) f(x, y) & \text{if } \boldsymbol{Z} \text{ is discrete} \\ \int_{\mathbb{R}} \int_{\mathbb{R}} \exp(t_1 X + t_2 Y) f(x, y) \, dxdy & \text{if } \boldsymbol{Z} \text{ is continuous} \end{cases}.$$

Proposition 2.8. *Let $\boldsymbol{Z} = (X, Y)'$ be a bivariate random vector. If $\boldsymbol{Z}$ is discrete with joit pmf $f_{X,Y}(x, y)$. Then the marginal pmf of X, $f_X(x) = P(X = x)$ is given by*

$$f_X(x) = \sum_{y\in\mathbb{R}} f_{X,Y}(x, y),$$

and the marginal pmf of Y is defined similarly.

If $\boldsymbol{Z}$ is continuous, the marginal probability density functions of X is definded as

$$f_X(x) = \int_{\mathbb{R}} f(x, y) \, dy, \; -\infty < x < \infty.$$

Exercise 2.9. Define a joint pdf by

$$f(x, y) = \begin{cases} kxy^2 & x \in (0, 1), y \in (0, 1) \\ 0 & \text{otherwise} \end{cases}.$$

Calculate

- k.
- $P(X + Y > 1)$.
- $f_X(x)$, $P\left(\frac{1}{2} < X < \frac{3}{4}\right)$.

2.2 Transformation for Multivariate Distributions

The transformation can be extended to multivariate distributions. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$be a random vector with pdf $f_{\boldsymbol{X}}(x_1, \ldots, x_n)$. Let $\mathcal{A} = \{\boldsymbol{x} : f_{\boldsymbol{X}}(\boldsymbol{x}) > 0\}$. Consider a new random vector $(U_1, \ldots, U_n)$, defined by $U_1 = g_1(X_1, \ldots, X_n)$, $U_2 = g_2(X_1, \ldots, X_n) \ldots U_n = g_n(X_1, \ldots, X_n)$. Suppose that $A_0, A_1, \ldots, A_k$ form a partition of $\mathcal{A}$ with these properties. The set A_0, which may be empty satisfies $P((X_1, \ldots, X_n) \in A_0) = 0$. The transformation $(U_1, \ldots, U_n) = (g_1(\boldsymbol{X}), \ldots, g_n(\boldsymbol{X}))$ is a one-to-one transformation from A_i to $\mathcal{B}$ for each $i = 1, 2, \ldots, k$. Then for each i, the inverse functions from $\mathcal{B}$ to A_i can be found. Denote the ith inverse by $x_1 = h_{1i}(u_1, \ldots, u_n), \ldots, x_n = h_{ni}(u_1, \ldots, u_n)$. This ith inverse gives, for $(u_1, \ldots, u_n) \in \mathcal{B}$, the unique $(x_1, \ldots, x_n) \in A_i$ such that $(u_1, \ldots, u_n) =$

$(g_1(x_1,\ldots,x_n),\ldots,g_n(x_1,\ldots,x_n))$. Let J_i denote the Jacobian computed from the ith inverse. That is,

$$J_i = \begin{vmatrix} \frac{\partial x_1}{\partial u_1} & \frac{\partial x_1}{\partial u_2} & \cdots & \frac{\partial x_1}{\partial u_n} \\ \frac{\partial x_2}{\partial u_1} & \frac{\partial x_2}{\partial u_2} & \cdots & \frac{\partial x_2}{\partial u_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial u_1} & \frac{\partial x_n}{\partial u_2} & \cdots & \frac{\partial x_n}{\partial u_n} \end{vmatrix} = \begin{vmatrix} \frac{\partial h_{1i}(\boldsymbol{u})}{\partial u_1} & \frac{\partial h_{1i}(\boldsymbol{u})}{\partial u_2} & \cdots & \frac{\partial h_{1i}(\boldsymbol{u})}{\partial u_n} \\ \frac{\partial h_{2i}(\boldsymbol{u})}{\partial u_1} & \frac{\partial h_{2i}(\boldsymbol{u})}{\partial u_2} & \cdots & \frac{\partial h_{2i}(\boldsymbol{u})}{\partial u_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_{ni}(\boldsymbol{u})}{\partial u_1} & \frac{\partial h_{ni}(\boldsymbol{u})}{\partial u_2} & \cdots & \frac{\partial h_{ni}(\boldsymbol{u})}{\partial u_n} \end{vmatrix},$$

the determinant of an $n \times n$ matrix. Assuming that these Jacobians do not vanish identically on $\mathcal{B}$, we have the following representation of the joint pdf, $f_U(\boldsymbol{u})$ for $\boldsymbol{u} \in \mathcal{B}$:

$$f_{\boldsymbol{U}}(\boldsymbol{u}) = \sum_{i=1}^{k} f_{\boldsymbol{X}}(h_{1i}(\boldsymbol{u}),\ldots,h_{ni}(\boldsymbol{u}))|J_i|. \tag{2.1}$$

Example 2.10. Let $X = (X_1, X_2, X_3, X_4)$ with $f_{\boldsymbol{X}}(\boldsymbol{x}) = 24\exp(-x_1 - x_2 - x_3 - x_4)$,$0 < x_1 < x_2 < x_3 < x_4 < \infty$. Consider the transformation,

$$U_1 = X_1, U_2 = X_2 - X_1, U_3 = X_3 - X_2, U_4 = X_4 - X_3.$$

This transformation maps the set $\mathcal{A}$ onto the set $\mathcal{B} = \{\boldsymbol{u}; 0 < u_i < \infty, i = 1,2,3,4\}$. The transformation is one-to-one, so $k = 1$, and the inverse is

$$X_1 = U_1, X_2 = U_1 + U_2, X_3 = U_1 + U_2 + U_3, X_4 == U_1 + U_2 + U_3 + U_4.$$

The Jacobian of the inverse is

$$J = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{vmatrix} = 1.$$

Since the matrix is triangular, the determinant is equal the the product of the diagonal elements. Thus, by (2.1), we obtain

$$f_{\boldsymbol{U}}(\boldsymbol{u}) = 24\exp(-4u_1 - 3u_2 - 2u_3 - u_4),$$

on $\mathcal{B}$.

2.3 Conditional Distributions and Independence

As we have learned in the previous classes, the conditional probability plays an important rolein econometrics. And since the original sample space is very difficult to deal with, we change the sample space to the real line. Based on the multivariate random variables, we define some useful definitions. We still consider the bivariate case.

Definition 2.11. Let (X, Y) be a discrete bivariate random vector with joint pmf $f(x,y)$and marginal pmfs $f_X(x)$ and $f_Y(y)$. For any x such that $P(X = x) = f_X(x) > 0$, the *conditional pmf* of Y given that $X = x$ is the function of y denoted by $f(y|x) = P(Y = y|X = x) = \frac{f(x,y)}{f_X(x)}$.

Definition 2.12. Let (X, Y) be a continuous bivariate random vector with joint pdf $f(x, y)$ and marginal pdfs $f_X(x)$ and $f_Y(y)$. For any x such that $f_X(x) > 0$, the *conditional pdf of* Y *given that* $X = x$ is the function of y denoted by $f(y|x)$ and defined by

$$f(y|x) = \frac{f(x, y)}{f_X(x)}.$$

Besides, remember that $f(y|x)$ is pdf or pmf and use it in the same way that have previously used unconditional pdfs or pmfs.

Definition 2.13. If $g(Y)$ is a funcion of Y, then the *conditional expected value of* $g(Y)$ *given that* $X = x$ is denoted by $E[g(Y)|x]$ and is given by

$$E[g(Y)] = \begin{cases} \sum_{y \in \mathbb{R}} g(y) f(y|x) & \text{if } \boldsymbol{Z} \text{ is discrete} \\ \int_{\mathbb{R}} g(y) f(y|x)\, dxdy & \text{if } \boldsymbol{Z} \text{ is continuous} \end{cases}.$$

Of course, the conditional expected value has all of the properties of the usual expected value.

Theorem 2.14. *If* X *and* Y *are any two random variables, then*

$$E(X) = E[E(X|Y)].$$

It can be interpreted as

$$E(X) = \int_{\mathbb{R}} E(X|Y = y) f_Y(y)\, dy = \int_{\mathbb{R}} \left[\int_{\mathbb{R}} x f(x|y)\, dx \right] f_Y(y)\, dy = \int_{\mathbb{R}} \int_{\mathbb{R}} x f(x, y)\, dxdy,$$

where $E(X|Y = y)$ is a function of y, then we can interpret it by the definition of expected value.

Theorem 2.15. *(Conditional variance identity) For any two random variables* X *and* Y*,*

$$Var(X) = E(Var(X|Y)) + Var(E(X|Y)),$$

provided that the expectation exists.

Proof. By definition,

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

□

Again, we also have talked about the independence of probability, here is the similar definition.

Definition 2.16. Let (X, Y) be a bivriate random vector with joint pdf or pmf $f(x, y)$ and marginal pdfs or pmfs $f_X(x)$ and $f_Y(y)$. Then X and Y are called *independent random variables* if, for every $x \in \mathbb{R}$ and $y \in \mathbb{R}$,

$$f(x, y) = f_X(x) f_Y(y).$$

Theorem 2.17. *Let X and Y be independent random variables.*

- *For any $A \subset \mathbb{R}$ and $B \subset \mathbb{R}$, $P(X \in A, Y \in B) = P(X \in A) P(Y \in B)$; that is, events $\{X \in A\}$ and $\{Y \in B\}$ are independent events.*
- *Let $g(x)$ be a function only of x and $h(y)$ be a function only of y. Then*

$$E[g(X) h(Y)] = E[g(X)] E[h(Y)].$$

Exercise 2.18. Prove Theorem 2.17.

Theorem 2.19. *Let X and Y be independent random variables with mgf $M_X(t)$ and $M_Y(t)$. Then the mgf of the random variable $Z = X + Y$ is given by*

$$M_Z(t) = M_X(t) M_Y(t).$$

Exercise 2.20. Prove Theorem 2.19.

2.4 Covariance and Correlation

In the univariate case we used mean as a measure of central tendency and the variance as a measure of spread. In the multivariate case, in the last section, we have discussed the absence or presence of a relationship between two random variables, independence or nonindependence. In this section, we define some tools to measure the relationship.

Definition 2.21. The *covariane of X and Y is the number defined as*

$$Cov(X, Y) = E[(X - EX)(Y - EY)].$$

Theorem 2.22. *For any two random variables X and Y,*

- $Cov(X, Y) = E(XY) - E(X) E(Y)$.
- $Var(aX \pm bY) = a^2 Var(X) + b^2 Var(Y) \pm 2ab(X, Y)$.
- $Cov(aX, bY) = abCov(X, Y)$.

Definition 2.23. The *correlation coeficient of X and Y is the number defined by*

$$\rho_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y},$$

where $\sigma_X = \sqrt{Var(X)}$ and $\sigma_Y = \sqrt{Var(Y)}$.

Correlation measures the degree of linear association between two variables. While the magnitude of the covariance between X and Y depends on the variances of X and Y , correlation always satisfies $-1 \leq \rho_{XY} \leq 1$.

Theorem 2.24. *For any random variable X and Y,*

- $-1 \leq \rho_{XY} \leq 1$.
- $|\rho_{XY}| = 1$ *if and only if* $\exists a \neq 0, b$ *such that* $P(Y = aX + b) = 1$.*If* $\rho_{XY} = 1$,*then* $a > 0$, *and if* $\rho_{XY} = -1$, *then* $a < 0$.

2.5 More about Univariate Distributions

In this section, we will use the tools introduced above to examine some properties of specific distributions and the relationship between distributions.

Theorem 2.25. *Let $X_1, \ldots, X_n$ be mutually independent random variables with mgfs $M_{X_1}(t), \ldots, M_{X_n}(t)$. Let $Z = X_1 + X_2 + \cdots + X_n$. Then the mgf of Z is*

$$M_Z(t) = M_{X_1}(t) \cdot \cdots \cdot M_{X_n}(t).$$

In particular, if $X_1, \ldots, X_n$ all have the same distribution with mgf $M_X(t)$, then,

$$M_Z(t) = (M_X(t))^n.$$

This theorem is an extension of Theorem 2.19.

Proposition 2.26. *Suppose $X_1, \ldots, X_n$ are mutually independent random variables, and $X_i \sim Gamma(\alpha_i, \beta)$. If $Z = X_1 + \cdots + X_n$, $Z \sim Gamma(\sum_{i=1}^n \alpha_i, \beta)$.*

Corollary 2.27. *Suppose $X_1, \ldots, X_n$ are mutually independent random variables, and $X_i \sim \chi^2(p_i)$. If $Z = X_1 + \cdots + X_n$, $Z \sim \chi^2(\sum_{i=1}^n p_i)$.*

For Chi-square distribution, we have,

- If $X \sim N(0,1)$, then $Y = X^2 \sim \chi^2(1)$.
- If $X \sim N(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim \chi^2(1)$.
- If $X_1, X_2, \ldots, X_n$ are mutually independent, $X_i \sim N(0,1)$, $\sum_{i=1}^n X_i^2 \sim \chi^2(n)$.

Exercise 2.28. Examine the properties of Chi-square distribution.

Definition 2.29. If $Z \sim N(0,1)$ and $X \sim \chi^2(p)$, X and Z are independent. Then the ratio

$$\frac{Z}{\sqrt{X/p}} \sim t(p),$$

where $t(p)$ is called t distribution with p degree of freedom. The density of $t(p)$ is

$$f(x|p) = \frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\sqrt{p\pi}}\left(1 + \frac{x^2}{p}\right)^{-\frac{p+1}{2}}, -\infty < x < \infty.$$

Further, we can verify that for $t(p)$, $E(X) = 0, p > 1$, $Var(X) = \frac{p}{p-2}, p > 2$.

Definition 2.30. If X_1 and X_2 are two independent chi-square distribution with degrees of freedom parameters p_1 and p_2, then the ratio

$$\frac{X_1/p_1}{X_2/P_2} \sim F(p_1, p_2),$$

where $F(p_1, p_2)$ is called the F distribution with degree p_1 and p_2. Further, the density of $F(p_1, p_2)$ is

$$f(x|p_1, p_2) = \frac{\Gamma\left(\frac{p_1+p_2}{2}\right)}{\Gamma\left(\frac{p_1}{2}\right)\Gamma\left(\frac{p_2}{2}\right)} \left(\frac{p_1}{p_2}\right)^{\frac{p_1}{2}} \frac{x^{(p_1-2)/2}}{\left(1+\frac{p_1 x}{p_2}\right)^{\frac{p_1+p_2}{2}}}.$$

2.6 Extension to General Case

The extension of the results for bivariate distributions to more than two variables is direct.

Let $\boldsymbol{X} = (X_1, \ldots, X_n)'$ be a random vector, then the mean vector is

$$\boldsymbol{\mu} = E(\boldsymbol{X}) = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{pmatrix}.$$

The *covariance matrix* is defined as

$$Var(\boldsymbol{X}) = E\left[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})'\right] = E\left[\boldsymbol{X}\boldsymbol{X}'\right] - \boldsymbol{\mu}\boldsymbol{\mu}'.$$

Usually we denote it as Σ. That is $Var(\boldsymbol{X}) = \Sigma$. Further, based on the properties of the expected value operation, let $\boldsymbol{a} = (a_1, \ldots, a_n)'$,

$$E\left(\boldsymbol{a}'\boldsymbol{X}\right) = \boldsymbol{a}' E(\boldsymbol{X}) = \boldsymbol{a}'\boldsymbol{\mu}. \tag{2.2}$$

$$\begin{aligned} Var\left(\boldsymbol{a}'\boldsymbol{X}\right) &= E\left[\left(\boldsymbol{a}'\boldsymbol{X} - \boldsymbol{a}'\boldsymbol{\mu}\right)^2\right] \\ &= E\left[\left(\boldsymbol{a}'\boldsymbol{X} - \boldsymbol{a}'\boldsymbol{\mu}\right)\left(\boldsymbol{a}'\boldsymbol{X} - \boldsymbol{a}'\boldsymbol{\mu}\right)'\right] \\ &= \boldsymbol{a}' E\left[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})'\right]\boldsymbol{a} \\ &= \boldsymbol{a}' Var(\boldsymbol{X})\,\boldsymbol{a}. \end{aligned} \tag{2.3}$$

If we let $Var(\boldsymbol{X}) = (\sigma_{ij})$, $Var(\boldsymbol{a}'\boldsymbol{X}) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \sigma_{ij}$.

For the case $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$, where $\boldsymbol{A}$ is a $m \times n$ matrix. Then $Y_i = \boldsymbol{a}_i'\boldsymbol{X}$ since $\boldsymbol{A} = (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)'$, from (2.2) and (2.3),

$$E(Y_i) = \boldsymbol{a}_i'\boldsymbol{\mu},\ Var(Y_i) = \boldsymbol{a}_i' Var(\boldsymbol{X})\,\boldsymbol{a}_i.$$

Further,

$$E(\boldsymbol{Y}) = \begin{pmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_m) \end{pmatrix} = \begin{pmatrix} \boldsymbol{a}_1'\boldsymbol{\mu} \\ \boldsymbol{a}_2'\boldsymbol{\mu} \\ \vdots \\ \boldsymbol{a}_m'\boldsymbol{\mu} \end{pmatrix} = \begin{pmatrix} \boldsymbol{a}_1' \\ \boldsymbol{a}_2' \\ \vdots \\ \boldsymbol{a}_m' \end{pmatrix} \boldsymbol{\mu} = \boldsymbol{A}\boldsymbol{\mu}.$$

We can also have

$$\begin{aligned} Cov\left(\boldsymbol{a}_i'\boldsymbol{X}, \boldsymbol{a}_j'\boldsymbol{X}\right) &= E\left[\left(\boldsymbol{a}_i'\boldsymbol{X} - \boldsymbol{a}_i'\boldsymbol{\mu}\right)\left(\boldsymbol{a}_j'\boldsymbol{X} - \boldsymbol{a}_j'\boldsymbol{\mu}\right)'\right] \\ &= \boldsymbol{a}_i' E\left[\left(\boldsymbol{X}-\boldsymbol{\mu}\right)\left(\boldsymbol{X}-\boldsymbol{\mu}\right)'\right]\boldsymbol{a}_j \\ &= \boldsymbol{a}_i' Var\left(\boldsymbol{X}\right)\boldsymbol{a}_j \\ &= \boldsymbol{a}_i'\Sigma\boldsymbol{a}_j. \end{aligned}$$

Since $\boldsymbol{a}_i'\Sigma\boldsymbol{a}_j$ is the ijth element of $\boldsymbol{A}\Sigma\boldsymbol{A}'$,

$$Var\left(\boldsymbol{A}\boldsymbol{X}\right) = \boldsymbol{A}\Sigma\boldsymbol{A}'.$$

2.7 Multivariate Normal Distribution and Chi-square Distribution

Definition 2.31. If $\boldsymbol{X} = (X_1, \ldots, X_n)'$ follows a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance Σ, then the joint density is

$$f\left(\boldsymbol{x}|\Sigma, \boldsymbol{\mu}\right) = (2\pi)^{-\frac{n}{2}}\left|\Sigma\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)'\Sigma^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\right\},\ \boldsymbol{x}\in\mathbb{R}^n.$$

It is denoted as $\boldsymbol{X}\sim N\left(\boldsymbol{\mu}, \Sigma\right)$.

Theorem 2.32. *Let* $\boldsymbol{X} = (\boldsymbol{X}_1', \boldsymbol{X}_2')'$, *then correspondingly,* $\boldsymbol{\mu} = (\boldsymbol{\mu}_1', \boldsymbol{\mu}_2')'$, $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. *And then the marginal distributions are*

$$\boldsymbol{X_1}\sim N\left(\boldsymbol{\mu}_1, \Sigma_{11}\right),$$
$$\boldsymbol{X_1}\sim N\left(\boldsymbol{\mu}_2, \Sigma_{22}\right).$$

The conditional distribution of $\boldsymbol{X_1}$*given* $\boldsymbol{X}_2$ *is normal as well:*

$$\boldsymbol{X}_1|\boldsymbol{x}_2\sim N\left(\boldsymbol{\mu}_{1.2}, \Sigma_{11.2}\right),$$

where

$$\boldsymbol{\mu}_{1.2} = \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}\left(\boldsymbol{x}_2 - \boldsymbol{\mu}_2\right),$$
$$\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

The result is similar to $\boldsymbol{X}_2|\boldsymbol{x}_1$.

Proposition 2.33. *If* $\boldsymbol{X}\sim N\left(\boldsymbol{\mu}, \Sigma\right)$, *then* $\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}\sim N\left(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\Sigma\boldsymbol{A}'\right)$.

Proof. Since the mgf of $\boldsymbol{X}$ is

$$M_{\boldsymbol{X}}\left(\boldsymbol{t}\right) = \exp\left(\boldsymbol{t}'\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}'\Sigma\boldsymbol{t}\right),$$

$$\begin{aligned} M_Z\left(\boldsymbol{t}\right) &= E\left[\exp\left(\boldsymbol{t}'\left(\boldsymbol{A}\boldsymbol{X}+\boldsymbol{b}\right)\right)\right] \\ &= E\left[\exp\left(\boldsymbol{t}'\boldsymbol{A}\boldsymbol{X} + \boldsymbol{t}'\boldsymbol{b}\right)\right] \\ &= \exp\left(\boldsymbol{t}'\boldsymbol{b}\right)E\left\{\exp\left[\left(\boldsymbol{A}'\boldsymbol{t}\right)'\boldsymbol{X}\right]\right\} \\ &= \exp\left(\boldsymbol{t}'\boldsymbol{b}\right)\exp\left(\boldsymbol{t}'\boldsymbol{A}\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}'\boldsymbol{A}\Sigma\boldsymbol{A}'\boldsymbol{t}\right) \\ &= \exp\left(\boldsymbol{t}'\left(\boldsymbol{A}\boldsymbol{\mu}+\boldsymbol{b}\right) + \frac{1}{2}\boldsymbol{t}'\boldsymbol{A}\Sigma\boldsymbol{A}'\boldsymbol{t}\right), \end{aligned}$$

thus,

$$\boldsymbol{Z} \sim N\left(\boldsymbol{A\mu} + \boldsymbol{b}, \boldsymbol{A\Sigma A'}\right).$$

□

Corollary 2.34. *If* $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{I})$, and $\boldsymbol{C}$ is a square matrix such that $\boldsymbol{C'C} = \boldsymbol{I}$, then $\boldsymbol{C'X} \sim N(\boldsymbol{0}, \boldsymbol{I})$.

Theorem 2.35. *If* $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{I})$, $\boldsymbol{A}$ is a idempotent and symmetric matrix, then $\boldsymbol{X'AX} \sim \chi^2(rank(\boldsymbol{A}))$.

Proof. $\boldsymbol{X'AX} = \boldsymbol{X'C\Lambda C'X}$ after the diagnalization. And we know that $\boldsymbol{Y} = \boldsymbol{C'X} \sim N(\boldsymbol{0}, \boldsymbol{I})$. And $\boldsymbol{A}$ is projection, which is symmetric and idempotent, such that there are only 1 or 0 in the diagnal line of$\boldsymbol{\Lambda}$. Thus,

$$\boldsymbol{Y'\Lambda Y} = \sum \lambda_i y_i^2 \sim \chi^2(rank(\boldsymbol{A})).$$

□

Corollary 2.36. *If* $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{I}_n)$, $\sum (x_i - \bar{\boldsymbol{x}}) \sim \chi^2(n-1)$.

Theorem 2.37. *If* $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{I})$, $\boldsymbol{X'AX}$ and $\boldsymbol{X'BX}$ are two idempotent quadratic forms in $\boldsymbol{X}$, then $\boldsymbol{X'AX}$ and $\boldsymbol{X'BX}$ are independent if $\boldsymbol{AB} = \boldsymbol{0}$.

Proof. We know that

$$\boldsymbol{X'AX} = \boldsymbol{X'A'AX} = \boldsymbol{X}_1\boldsymbol{X}_1', \ \boldsymbol{X'BX} = \boldsymbol{X'B'BX} = \boldsymbol{X}_2\boldsymbol{X}_2'.$$

Then

$$E\left(\boldsymbol{X}_1\boldsymbol{X}_2'\right) = \boldsymbol{A}E\left(\boldsymbol{XX'}\right)\boldsymbol{B'} = \boldsymbol{AB}.$$

Since $\boldsymbol{AX}$ and $\boldsymbol{BX}$ are also normal distributed random vector. Their zero covariance matrix implies that they are statistically independent. Thus, if $\boldsymbol{AB} = \boldsymbol{0}$, then $\boldsymbol{AX}$ and $\boldsymbol{BX}$ are independent. □

Proposition 2.38. *If* $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{I})$, $\boldsymbol{X'AX}$ and $\boldsymbol{X'BX}$ are two idempotent quadratic forms in $\boldsymbol{X}$, $\boldsymbol{AB} = \boldsymbol{0}$,

$$\frac{\boldsymbol{X'AX}/rank(\boldsymbol{A})}{\boldsymbol{X'BX}/rank(\boldsymbol{B})} \sim F(rank(\boldsymbol{A}), rank(\boldsymbol{B})).$$

Surely some of cases above can be easily extended to general case that $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$.

Estimation and Inference

1 Basic Concepts of Random Samples

Often, the data collected in an experiment consist of several observations on a variable of interest. We usually present a model for data collection that is often used to describe the situation. In assumptions of many econometric models, we always assume the data satisfies some similar conditions, for example, in classical regression, the data collected are assumed to be *independent and identical* ; in time series, we always construct the *data generating process* for the data.

Definition 1.1. The random variables $X_1, \ldots, X_n$ are called *random sample of size n from the population* $f(x)$ if $X_1, \ldots, X_n$ are mutually independent random variables and the marginal pdf or pmf of each X_i is the same function $f(x) : \mathbb{R} \to [0, +\infty)$. Alternatively, $X_1, \ldots, X_n$ are called *independent and identically distributed(i.i.d) random variables with pdf or pmf* $f(x)$.

As we have emphasized, the random variable can help us to transform the original sample place to the real number system. Here, the random sampling model describes a type of situation in which the variable of interest has a probability distribution described by $f(x)$. Since the random variables $X_1, \ldots, X_n$ are mutually independent, the joint pdf or pmf is

$$f(x_1, \ldots, x_n) = f(x_1) f(x_2) \ldots f(x_n) = \prod_{i=1}^{n} f(x_i).$$

Further, since they are identically distributed, each marginal density function is $f(x)$.

In particular, if the population pdf or pmf is a member of a parametric family, with pdf or pmf given by $f(x|\theta)$, then the joint pdf or pmf is

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta),$$

where the same parameter value θ is used in each term in the product. By considering different possible values of θ, we can study how a random sample would behave for different populations.

2 Data Reduction

When a sample $X_1, \ldots, X_n$ is drawn, some summary of the values is usually computed. Any well-defined summary may be expressed mathematically as a function $S(x_1, \ldots, x_n)$ whose domain includes the sample space of the random vector $(X_1, \ldots, X_n)$. Since the random

sample $X_1, \ldots, X_n$ has a simple probabilistic structure, the disribution of $\boldsymbol{Y} = S(X_1, \ldots, X_n)$, which is also a random variable, is particularly tractable.

Since the distribution of Y is always derived from the distribution of the variables in the random sample, it is called the *sampling distribution* of Y.

Definition 2.1. Let $X_1, \ldots, X_n$ be a random sample of size n from a population and let $S(x_1, \ldots, x_n) : \mathbb{R}^n \to \mathbb{R}_+^d \cup \{0\}$. Then $\boldsymbol{Y} = S(X_1, \ldots, X_n)$ is called a statistic. The probability distribution of a statistic Y is called the *sampling distribution* of Y.

Remark 2.2. The definition of a statistic is very board, with the only restriction being that a statistic cannot be a function of a parameter. For example, if $X_i \sim f(x|\theta)$,$i = 1, \ldots, n$, then $S_1 = X_1$ is a statistic but $S_1 = X_1 + \theta$ is not.

Example 2.3. Before attempting to estimate parameters of a population or fit models to data, we normally examine the data themselves. . Given $X_i \sim F(x)$ for $i = 1, \ldots, n$, where $F(x)$ is a probability distribution, we define the following widely used descriptive statistics.

- Sample mean: $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$.
- Median: M =middle ranked observation.
- Sample midrange: midrange = $\frac{\text{Maximum-Minimum}}{2}$.
- Sample variance:$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$.
- Standard deviation:S is the square root of S^2.

Lemma 2.4. *Let $X_1, \ldots, X_n$ be a random sample from a population and let $g(x)$ be a funxtion such that $E[g(x)]$ and $Var[g(X)]$ exist. Then*

$$E\left(\sum_{i=1}^{n} g(X_i)\right) = nE(g(X_1)), \tag{1}$$

and

$$Var\left(\sum_{i=1}^{n} g(X_i)\right) = nVar[g(X_1)]. \tag{2}$$

Proof. Note that

$$E\left(\sum_{i=1}^{n} g(X_i)\right) = \sum_{i=1}^{n} E[g(X_i)] = nE(g(X_1)).$$

Since the X_is are identically distributed, the second equality is true because $E[g(X_i)]$ is the same for all i. *Note that the independence of $X_1, \ldots, X_n$ is not needed for (1) to hold.* To prove (2),

$$\begin{aligned} Var\left(\sum_{i=1}^{n} g(X_i)\right) &= E\left[\sum_{i=1}^{n} g(X_i) - E\left(\sum_{i=1}^{n} g(X_i)\right)\right]^2 \\ &= E\left[\sum_{i=1}^{n} (g(X_i) - E[g(X_i)])^2\right]. \end{aligned}$$

Then there is n^2 terms. For each term that $i \neq j$,

$$E\left\{\left(g\left(X_i\right) - E\left[g\left(X_i\right)\right]\right)\left(g\left(X_j\right) - E\left[g\left(X_j\right)\right]\right)\right\} = Cov\left(g\left(X_i\right), g\left(X_j\right)\right) = 0$$

due tot he independence. Hence,

$$Var\left(\sum_{i=1}^{n} g\left(X_i\right)\right) = \sum_{i=1}^{n} E\left[\left(g\left(X_i\right) - E\left[g\left(X_i\right)\right]\right)^2\right] = nVar\left[g\left(X_1\right)\right].$$

□

Theorem 2.5. *Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Then,*

- $E\left[\bar{X}\right] = \mu$.
- $Var\left[\bar{X}\right] = \frac{\sigma^2}{n}$.
- $E\left[S^2\right] = \sigma^2$.

Exercise 2.6. Prove Theorem 2.5.

Remark 2.7. Pay attention to that the expectation and variance are all based on the probability measure of the random sample $X_1, \ldots, X_n$.

Theorem 2.8. *Let $X_1, \ldots, X_n$ be a random sample from a population with mgf $M_X(t)$. Then the mgf of sample mean is*

$$M_{\bar{X}}(t) = \left[M_X\left(\frac{t}{n}\right)\right]^n.$$

Exercise 2.9. Prove Theorem 2.8.

2.1 The Sufficiency Principle

Usually we want to use the information in a sample $X_1, \ldots, X_n$ to make inferences about an unkown parameter θ. If the sample size n is large, then the oberved sample $x_1, \ldots x_n$ is a long list of numbers that may be hard to interpret. Hence in this way, we also want to use some statistic to summarizes the information. Remerber, no matter how you summarize the sample, all the randomness comes from the sample $X_1, \ldots, X_n$.

Any statistic, $T(\boldsymbol{X})$, $\boldsymbol{X} = (X_1, \ldots, X_n)'$, defines a form of data reduction or data summary. An experimenter who uses only the observed value of the statistic, $T(\boldsymbol{X})$, rather than the entire observed sample, $\boldsymbol{X}$, will treat as equal two samples, $\boldsymbol{x}$ and $\boldsymbol{y}$, that satisfy $T(\boldsymbol{x}) = T(\boldsymbol{y})$ even though the actual sample values may be different in some ways.

In this section, we study the sufficiency principle when we are doing data reduction, which promotes a method of reduction that does not discard information about θ while achieving some summarization of the data.

Definition 2.10. *SUFFICIENCY PRINCIPLE:* If $T(\boldsymbol{X})$ is a sufficient statistic for θ, then any inference about θ should depend on the sample $\boldsymbol{X}$ only through the value $T(\boldsymbol{X})$. That is, if $\boldsymbol{x}$ and $\boldsymbol{y}$ are two sample points such that $T(\boldsymbol{x}) = T(\boldsymbol{y})$, then the inference about θ should be the same whether $\boldsymbol{X} = \boldsymbol{x}$ or $\boldsymbol{X} = \boldsymbol{y}$ is observed.

Definition 2.11. Let $\boldsymbol{X}$ be a vector containing a random sample of size n. A statistic $T(\boldsymbol{X})$ is a *sufficient statistic for* θif the conditional distribution of the sample $\boldsymbol{X}$ given the value of $T(\boldsymbol{X})$ does not depend on θ.

Remark 2.12. When $T(\boldsymbol{X})$ has a continuous distribution, then $P_\theta(T(\boldsymbol{X}) = t) = 0$ for all values of t. For this case, a more sophisticated notion of conditional probability is needed. A discussion of this can be found in Lehmann(1986). In the following, we will use the discrete case to illustrate the concepts.

As we have emphasized, the randomness comes from $\boldsymbol{X}$. And $X_i \sim F(x|\theta), i = 1, \ldots, n$. Therefore, we use $P_\theta(\cdot)$ denote the probability measure of $\boldsymbol{X}$ condion onθ. If $P_\theta(T(\boldsymbol{X}) = t) = P_\theta(\boldsymbol{X} : T(\boldsymbol{X}) = t) > 0$, conider the probability $P_\theta(\boldsymbol{X} = \boldsymbol{x}|T(\boldsymbol{X}) = t)$. By the definition, if $T(\boldsymbol{X})$ is a sufficient statistic, this conditional probability is the same for all values of θ so we have omitted the subscript, that is,

$$P_\theta(\boldsymbol{X} = \boldsymbol{x}|T(\boldsymbol{X}) = t) = P(\boldsymbol{X} = \boldsymbol{x}|T(\boldsymbol{X}) = t).$$

The sufficient statistic captures all the information about θ in this sense. Consider the first case, we only observes $\boldsymbol{X} = \boldsymbol{x}$ and we can compute $T(\boldsymbol{X}) = T(\boldsymbol{x})$, since we define $T(\cdot)$ and have the observations $\boldsymbol{x}$. Then we can use the $\boldsymbol{X} = \boldsymbol{x}$ and $T(\boldsymbol{X}) = T(\boldsymbol{x})$ to make the inference of θ.

For the second case, assume $T(\boldsymbol{X}) = T(\boldsymbol{x})$ is a sufficient statistic and we cannot oberve $\boldsymbol{X}$ but only know $T(\boldsymbol{X}) = T(\boldsymbol{x})$, then we know $P(\boldsymbol{X} = \boldsymbol{y}|T(\boldsymbol{X}) = T(\boldsymbol{x}))$, a probability distribution on a sample space $A_{T(\boldsymbol{z})} = \{\boldsymbol{y} : T(\boldsymbol{y}) = T(\boldsymbol{x})\}$, because this can be computed from the model without knowledge of the true value of θ. Then we can define a new random variable satisfies $P(\boldsymbol{Y} = \boldsymbol{y}|T(\boldsymbol{X}) = T(\boldsymbol{x})) = P(\boldsymbol{X} = \boldsymbol{y}|T(\boldsymbol{X}) = T(\boldsymbol{x}))$. We donot have to worry about the value of θ. Further, $\boldsymbol{X}$ and $\boldsymbol{Y}$ have the same unconditional distribution. Since $\{\boldsymbol{Y} = \boldsymbol{x}\}$ and $\{\boldsymbol{X} = \boldsymbol{x}\}$ are both subsets of the event $\{T(\boldsymbol{X}) = T(\boldsymbol{x})\}$,

$$\begin{aligned} P_\theta(\boldsymbol{X} = \boldsymbol{x}) &= P_\theta(\boldsymbol{X} = \boldsymbol{x} \, and \, T(\boldsymbol{X}) = T(\boldsymbol{x})) \\ &= P(\boldsymbol{X} = \boldsymbol{y}|T(\boldsymbol{X}) = T(\boldsymbol{x}))\, P_\theta(T(\boldsymbol{X}) = T(\boldsymbol{x})) \\ &= P(\boldsymbol{Y} = \boldsymbol{y}|T(\boldsymbol{X}) = T(\boldsymbol{x}))\, P_\theta(T(\boldsymbol{X}) = T(\boldsymbol{x})) \\ &= P_\theta(\boldsymbol{Y} = \boldsymbol{y}). \end{aligned}$$

Then for both cases, we all have equivalent knowledge about the random sample $\boldsymbol{X}$. For the first case, we can observe $\boldsymbol{X} = \boldsymbol{x}$ and the information about θ contained in $\boldsymbol{X} = \boldsymbol{x}$. For the second case, we do not have more knowledge about θ than the first case. But we can still have have equivalent information about random sample $\boldsymbol{X}$. All the knowledge about θ is contained in the statistic $T(\boldsymbol{X})$.

Remark 2.13. We know that if $T(\boldsymbol{X})$ is a sufficient statistic for θ, fix $\boldsymbol{x}$ and t, for any θ, $P_\theta(\boldsymbol{X} = \boldsymbol{x}|T(\boldsymbol{X}) = T(\boldsymbol{x}))$ is the same. Further,

$$\begin{aligned} P_\theta(\boldsymbol{X} = \boldsymbol{x}|T(\boldsymbol{X}) = t) &= \frac{P_\theta(\boldsymbol{X} = \boldsymbol{x} \, and \, T(\boldsymbol{X}) = T(\boldsymbol{x}))}{P_\theta(T(\boldsymbol{X}) = T(\boldsymbol{x}))} \\ &= \frac{P_\theta(\boldsymbol{X} = \boldsymbol{x})}{P_\theta(T(\boldsymbol{X}) = T(\boldsymbol{x}))} \\ &= \frac{f(\boldsymbol{x}|\theta)}{q(T(\boldsymbol{x})|\theta)}, \end{aligned}$$

where $f(\boldsymbol{x}|\theta)$ is the joint pmf of the sample $\boldsymbol{X}$ and $q(t|\theta)$ is the pmf of $T(\boldsymbol{X})$. Thus, $T(\boldsymbol{X})$ is a sufficient statistic for θ is and only if, for every $\boldsymbol{x}$, the above ratio of pmfs is constant as a function of θ.

Theorem 2.14. *If $f(\boldsymbol{x}|\theta)$ is the joint pdf or pmf of $\boldsymbol{X}$ and $q(t|\theta)$ is the pdf or pmf of $T(\boldsymbol{X})$, then $T(\boldsymbol{X})$ is a sufficient statistic for θ if, for every $\boldsymbol{x}$ in the sample space, the ratio $\frac{f(\boldsymbol{x}|\theta)}{q(T(\boldsymbol{x})|\theta)}$ is constant as afunction of θ.*

Exercise 2.15. Let $X_1, \ldots, X_n$ be i.i.d. $N\left(\mu, \sigma^2\right)$, where σ^2 is known. Show that $T(\boldsymbol{X}) = \bar{\boldsymbol{X}}$ is the sufficient statistic for μ.

Theorem 2.16. *Let $f(\boldsymbol{x}|\theta)$ denote the joint pdf or pmf of a sample $\boldsymbol{X}$. A statistic $T(\boldsymbol{X})$ is a sufficient statistic for θ if and only if there exist function $g(t|\theta)$ and $h(\boldsymbol{x})$ such that ,for all sample points $\boldsymbol{x}$ and all parameter points θ,*

$$f(\boldsymbol{x}|\theta) = g(T(\boldsymbol{x})|\theta)\, h(\boldsymbol{x}).$$

Exercise 2.17. Prove Theorem 3.7.

Exercise 2.18. Assume $X_1, \ldots, X_n$ are i.i.d. $N\left(\mu, \sigma^2\right)$ and both μ and σ^2 are unknown. Show that $T(\boldsymbol{X}) = \left(\bar{X}, S^2\right)$ is a sufficient statistic for $\left(\mu, \sigma^2\right)$.

3 Point Estimation

Usually we want to use the information in a sample $X_1, \ldots, X_n$ to make inferences about an unkown parameter θ. If the sample size n is large, then the oberved sample $x_1, \ldots x_n$ is a long list of numbers that may be hard to interpret. Hence in this way, we also want to use some statistic to summarizes the information. Remerber, no matter how you summarize the sample, all the randomness comes from the sample $X_1, \ldots, X_n$.

Any statistic, $T(\boldsymbol{X})$, $\boldsymbol{X} = (X_1, \ldots, X_n)'$, defines a form of data reduction or data summary. An experimenter who uses only the observed value of the statistic, $T(\boldsymbol{X})$, rather than the entire observed sample, $\boldsymbol{X}$, will treat as equal two samples, $\boldsymbol{x}$ and $\boldsymbol{y}$, that satisfy $T(\boldsymbol{x}) = T(\boldsymbol{y})$ even though the actual sample values may be different in some ways.

Definition 3.1. If $X_1, \ldots, X_n$ is sampled from density function $f(x|\boldsymbol{\theta})$. Any statistic $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ with the same dimension of $\boldsymbol{\theta}$ is a *point estimate.*

Definition 3.2. An *estimator, $T(\boldsymbol{X})$,* is a rule or strategy for using the data to estimate the parameter, which is defined before the data is drawn.

Remark 3.3. What is the difference between estimators and estimates? The estimator is a ramdom variable from whose distribution the estimates are sampled; an estimator captures our uncertainty about the values of our estimates. Properties of an estimator determines how useful it is. A good point estimator will have most of its mass concentrated around the value that we're trying to estimate. This will ensure that the estimator has a high probability of yielding a value close to the parameter value.

Definition 3.4. *Finite sample properties* of estimators are those attributes that can be compared regardless of the sample size.

Some estimation problems involve characteristics that are not known in finite samples. In these instances, estimators are compared on the basis on their large sample, or *asymptotic properties.*

3.1 Methods of Finding Estimatros

3.1.1 Method of Moments

Let $X_1, \dots X_n$ be a sample from a population with pdf $f(x|\boldsymbol{\theta}), \boldsymbol{\theta} = (\theta_1, \dots, \theta_k)'$. Method of moments estimators are found by equating the first k sample moents to the correponding k population moments, and solving the resulting system of simultaneous equations. More precisely, define

$$m_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j,\ \mu_j = E_{\boldsymbol{\theta}}\left[X^j\right]\ j = 1, \dots k.$$

Then let $m_j = \mu_j, j = 1, \dots, k$. We have k equations for k unkonwn parameters.

3.1.2 Maximum Likelihood Estimators

Definition 3.5. Let $f(\boldsymbol{x}|\boldsymbol{\theta})$ denote the joint pdf or pmf of the sample $\boldsymbol{X} = (X_1, \dots, X_n)'$. Then given that $\boldsymbol{X} = \boldsymbol{x}$ is observed, the function of $\boldsymbol{\theta}$ defined by

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = f(\boldsymbol{x}|\boldsymbol{\theta})$$

is called the *likelihood function.* The log-likelihood function is

$$\ell(\boldsymbol{\theta}|\boldsymbol{x}) = \log L(\boldsymbol{\theta}|\boldsymbol{x}).$$

Remark 3.6. When $X_1, \dots X_n$ are i.i.d., the likelihood can be written as

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}),\ \ell(\boldsymbol{\theta}|\boldsymbol{x}) = \sum_{i=1}^{n} \log f(x_i|\boldsymbol{\theta}).$$

If not, sometimes we can use the decomposition, $f(\boldsymbol{x}|\boldsymbol{\theta}) = f(x_n|\boldsymbol{x}_{1:n-1}, \boldsymbol{\theta}) f(x_{n-1}|\boldsymbol{x}_{1:n-2}, \boldsymbol{\theta}) \dots f(x_1|\boldsymbol{\theta})$, $\boldsymbol{x}_{1:j} = (x_1, \dots, x_{j-1})'$, which is widely used in time series analysis, then,

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = f(x_1|\boldsymbol{\theta}) \prod_{i=2}^{n} f(x_i|\boldsymbol{x}_{1:i-1}, \boldsymbol{\theta}),\ \ell(\boldsymbol{\theta}|\boldsymbol{x}) = \sum_{i=1}^{n} \log f(x_i|\boldsymbol{x}_{1:i-1}\boldsymbol{\theta}) + \log f(x_1|\boldsymbol{\theta}).$$

Likelihood is used as a measure of the plausibility of parameter values for a given sample hence it is reasonable to take as a point estimator the value that maximises the likelihood function.

Definition 3.7. For each sample point $\boldsymbol{x}$, let $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ be a parameter value at which $L(\boldsymbol{\theta}|\boldsymbol{x})$ attains its maximum as a function of $\boldsymbol{\theta}$, with $\boldsymbol{x}$ held fixed. A *maximum likelihood estimator(MLE)* of the parameter $\boldsymbol{\theta}$ based on a sample $\boldsymbol{X}$ is m$\hat{\boldsymbol{\theta}}(\boldsymbol{X})$.

Remark 3.8. Intuitivel, the MLE is the parameter point for which the observed sample is most likely. In general, the MLE is a good point estimator, possessing some of the optimality properties discussed later.

The way we obtain the MLE is always differentiable the likelihood with respect to $\boldsymbol{\theta}$, that is

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}|\boldsymbol{x}) = 0. \tag{3}$$

And then check the second derivative. Any way the procedure is a case in optimization application.

We proceed by defining for $\tau(\boldsymbol{\theta})$ the *induced likelihood function* L^*, given by

$$L*(\eta|\boldsymbol{x}) = \sup_{\{\boldsymbol{\theta}:\tau(\boldsymbol{\theta})=\eta\}} L(\boldsymbol{\theta}|\boldsymbol{x}).$$

The value $\hat{\eta}$ that maximizes $L^*(\eta|\boldsymbol{x})$ will be called the MLE of $\eta = \tau(\boldsymbol{\theta})$, and it can be seen that the maxima of L^* and L conincide.

Theorem 3.9. *(Invariance property of MLEs) If $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, then for any function $\tau(\boldsymbol{\theta})$, the MLE of $\tau(\boldsymbol{\theta})$ is $\tau\left(\hat{\boldsymbol{\theta}}\right)$.*

Proof. Let $\hat{\eta}$ denote the value that maximizes $L^*(\eta|\boldsymbol{x})$. We must show that $L^*(\hat{\eta}|\boldsymbol{x}) = L^*[\tau(\boldsymbol{\theta})|\boldsymbol{x}]$. Then

$$\begin{aligned} L^*(\hat{\eta}|\boldsymbol{x}) &= \sup_{\eta} \sup_{\{\boldsymbol{\theta}:\tau(\boldsymbol{\theta})=\eta\}} L(\boldsymbol{\theta}|bx) \\ &= \sup_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\boldsymbol{x}) \\ &= L\left(\hat{\boldsymbol{\theta}}|\boldsymbol{x}\right), \end{aligned}$$

where the second equality follows because the iterated maximization is equal to the unconditional maximization over $\boldsymbol{\theta}$, which is attained at $\hat{\boldsymbol{\theta}}$. Furthermore,

$$\begin{aligned} L\left(\hat{\boldsymbol{\theta}|\boldsymbol{x}}\right) &= \sup_{\{\boldsymbol{\theta}:\tau(\boldsymbol{\theta})=\tau(\hat{\boldsymbol{\theta}})\}} L(\boldsymbol{\theta}|\boldsymbol{x}) \\ &= L^*\left[\tau\left(\hat{\boldsymbol{\theta}}\right)|\boldsymbol{x}\right]. \end{aligned}$$

Hence, the string of equalities shows that $L^*(\hat{\eta}|\boldsymbol{x}) = L^*\left(\tau\left(\hat{\boldsymbol{\theta}}\right)|\boldsymbol{x}\right)$ and that $\tau\left(\hat{\boldsymbol{\theta}}\right)$ is the MLE of $\tau(\boldsymbol{\theta})$. □

There are many other methods to find estimator for parametric models, such as Markov Chain Monte Carlo, Expectation maximization and so on.

3.2 Methods of Evaluating Estimators

In order to identify good estimators, we may insist that the center of the distribution of the estimator is close to the parameter ϑ. The usual measure of central tendency is the mean. If the mean of an estimator is equal to the parameter, then the estimator is unbiased.

Definition 3.10. An estimator of parameter $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, is *unbiased* if

$$Bias_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}\right) = E_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) = \mathbf{0},$$

i.e., $E_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}\right) = \boldsymbol{\theta}$.

Unbiasesdness is a desirable attribute, but many unbiased estimators are poor uses of data. A second criterion used to choose among unbiased estimators is efficiency.

Definition 3.11. An unbiased estimator $\hat{\boldsymbol{\theta}}_1$ is more *efficient* than another unbiased estimator $\hat{\boldsymbol{\theta}}_2$ if

$$Var\left(\hat{\boldsymbol{\theta}}_1\right) - Var\left(\hat{\boldsymbol{\theta}}_2\right)$$

is a positive definite matrx.

In discussing efficiency, we have restricted the discussion to unbiased estimators. But focusing on unbiasedness may still preclude a tolerably biased estimator with a much smaller variance. A criterion that recognizes this possible tradeoff is the *mean-squared error.*

Definition 3.12. The *mean-square error* of an estimator $\hat{\boldsymbol{\theta}}$ is

$$MSE\left(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}\right) = E_{\boldsymbol{\theta}}\left[\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)'\right].$$

Remark 3.13. It is obvious that

$$\begin{aligned} E_{\boldsymbol{\theta}}\left[\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)'\right] &= E_{\boldsymbol{\theta}}\left[\left(\hat{\boldsymbol{\theta}} - E_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}\right) + E_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\theta}\right)\left(\hat{\boldsymbol{\theta}} - E_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}\right) + E_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\theta}\right)'\right] \\ &= E_{\boldsymbol{\theta}}\left[\left(\hat{\boldsymbol{\theta}} - E_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}\right)\right)\left(\hat{\boldsymbol{\theta}} - E_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}\right)\right)'\right] + \left(E_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\theta}\right)\left(E_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}\right) - \boldsymbol{\theta}\right)' \\ &= Var\left(\hat{\boldsymbol{\theta}}\right) + Bias_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}\right) Bias_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}\right)'. \end{aligned}$$