PRE-SESSIONAL MATHS & STATS REVIEW
STATISTICS REVIEW COURSE

September 2023

Chapter 1

Linear Algebra

1.1 Introduction

A matrix is an $n \times m$ array of numbers. So

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix}$$

where a_{ij} are scalar numbers and are called the elements of the matrix. A column vector is a matrix with one column and a row vector is a matrix with one row. By convention a vector is a column vector unless otherwise specified. Usually matrices are denoted by upper case letters and vectors by bold (usually lowercase) letters[1]. Note also the following notation:

$$\sum_{i=1}^{n} a_i = a_1 + a_2 + a_3 + \ldots + a_n$$

and

$$\prod_{i=1}^{n} a_i = a_1 \times a_2 \times a_3 \times \ldots \times a_n$$

1.2 Basic Matrix Operations

There are two basic matrix operations: Addition and multiplication

[1]There are some exceptions to this convention

1.2.1 Addition

Addition is straightforward. Matrices are added elementwise and matrices can only be added if they have the same number of rows and columns. So for

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix} \quad B = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \dots & \dots & \dots & \dots \\ b_{n1} & b_{n2} & \dots & b_{nm} \end{pmatrix}$$

$$A + B = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \dots & \dots & \dots & \dots \\ b_{n1} & b_{n2} & \dots & b_{nm} \end{pmatrix} =$$

$$\begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1m} + b_{1m} \\ a_{21} + b_{21} & a_{22} + b_{22} & \dots & a_{2m} + b_{2m} \\ \dots & \dots & \dots & \dots \\ a_{n1} + b_{n1} & a_{n2} + b_{n2} & \dots & a_{nm} + b_{nm} \end{pmatrix}$$

1.2.2 Multiplication

Two matrices A and B can be multipied and their product is denoted by AB if the number of columns of A is equal to the number of rows of B. Note that $AB \neq BA$. So that AB may be defined but BA may not. AB is defined as follows

$$A_{n\times m} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix} \quad B_{m\times k} = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1k} \\ b_{21} & b_{22} & \dots & b_{2k} \\ \dots & \dots & \dots & \dots \\ b_{m1} & b_{m2} & \dots & b_{mk} \end{pmatrix}$$

Then

$$AB = \begin{pmatrix} \sum_{j=1}^{m} a_{1j}b_{j1} & \sum_{j=1}^{m} a_{1j}b_{j2} & \dots & \sum_{j=1}^{m} a_{1j}b_{jk} \\ \sum_{j=1}^{m} a_{2j}b_{j1} & \sum_{j=1}^{m} a_{2j}b_{j2} & \dots & \sum_{j=1}^{m} a_{2j}b_{jk} \\ \dots & \dots & \dots & \dots \\ \sum_{j=1}^{m} a_{nj}b_{j1} & \sum_{j=1}^{m} a_{nj}b_{j2} & \dots & \sum_{j=1}^{m} a_{nj}b_{jk} \end{pmatrix}$$

For example, two 2×2 matrices we have

$$A_{2\times 2} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad B_{2\times 2} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \Rightarrow$$

$$AB = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{pmatrix}$$

Another example with actual numbers gives

$$A_{2\times 2} = \begin{pmatrix} 2 & 4 \\ 6 & 8 \end{pmatrix} \quad B_{2\times 2} = \begin{pmatrix} 1 & 3 \\ 5 & 7 \end{pmatrix} \Rightarrow$$

$$AB = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{pmatrix} = \begin{pmatrix} 2\times 1 + 4\times 5 & 2\times 3 + 4\times 7 \\ 6\times 1 + 8\times 5 & 6\times 3 + 8\times 7 \end{pmatrix} = \begin{pmatrix} 22 & 34 \\ 46 & 74 \end{pmatrix}$$

1.3 Some Operations on Individual matrices

1.3.1 Transpose of a matrix

The transpose of a $m \times n$ matrix $A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$ is

$$A' = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ a_{12} & a_{22} & \dots & a_{m2} \\ \dots & \dots & \dots & \dots \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{pmatrix}$$

1.3.2 Diagonal of a Matrix

A diagonal of an $m \times m$ matrix $A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$ consists of all elements a_{ij} for which $i - j$ is a given integer. The main diagonal consists of all the elements a_{ij} for which $i - j = 0$.

Example 1

$$A = \begin{pmatrix} 3 & 21 & 6 & 42 \\ 6 & -15 & 12 & -30 \\ -2 & -14 & 8 & 56 \\ -4 & 10 & 16 & -40 \end{pmatrix}$$

The diagonals of this matrix are $(3\ -15\ 8\ -40)$, $(21\ 12\ 56)$, $(6\ 30)$, 42, $(6\ -14\ 16)$, $(-2\ 10)$, -4. *The main diagonal is* $(3\ -15\ 8\ -40)$.

A $m \times m$ matrix, say $A = [a_{ij}]$, for which $a_{ij} = 0$ for $i \neq j$ is called a diagonal matrix.

Example 2

$$A = \begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & -15 & 0 & 0 \\ 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & -40 \end{pmatrix}$$

A is a diagonal matrix.

1.4 Some Special Matrices

- An $m \times m$ matrix, A, is idempotent if $AA = A$.
- An $m \times m$ matrix, $I = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$, with $a_{ii} = 1$ and $a_{ij} = 0$, for $i \neq j$, is called an identity matrix. Note that for all $m \times m$ matrices A, $A^0 = I$.
- An $m \times m$ matrix A is symmetric if $A' = A$.
- An $m \times m$ matrix, $A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$ is
  - lower triangular if $a_{ij} = 0$ for $i < j$
  - upper triangular if $a_{ij} = 0$ for $i > j$
- An $m \times m$ matrix A^{-1} is the inverse of the $m \times m$ matrix A if $A^{-1}A = AA^{-1} = I$.

1.5 Sums of Values

Some useful relationships

- $$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n = \mathbf{i}'\mathbf{x} = (11\ldots1)\begin{pmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{pmatrix}$$

- The arithmetic mean
$$\bar{x} = 1/n\sum_{i=1}^{n} x_i = 1/n\mathbf{i}'\mathbf{x}$$

- sums of squares and cross products
$$\sum_{i=1}^{n} x_i^2 = \mathbf{x}'\mathbf{x} = (x_1 x_2 \ldots x_n)\begin{pmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{pmatrix}$$
$$\sum_{i=1}^{n} x_i y_i = \mathbf{x}'\mathbf{y} = (x_1 x_2 \ldots x_n)\begin{pmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{pmatrix}$$
$$C = \mathbf{X}'_{k\times n}\mathbf{X}_{n\times k} = \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}'_i$$
where $\mathbf{x}_i$ is the i-th column of X. This matrix looks like
$$\begin{pmatrix} \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \ldots & \sum_{i=1}^n x_{i1}x_{ik} \\ \sum_{i=1}^n x_{i2}x_{i1} & \sum_{i=1}^n x_{i2}^2 & \ldots & \sum_{i=1}^n x_{i2}x_{ik} \\ \ldots & \ldots & \ldots & \ldots \\ \sum_{i=1}^n x_{in}x_{i1} & \sum_{i=1}^n x_{in}x_{i2} & \ldots & \sum_{i=1}^n x_{in}x_{ik} \end{pmatrix}$$

1.6 Idempotent Matrix $\mathbf{M} = \mathbf{I} - 1/n\mathbf{ii}'$

This matrix yields dedviations from the mean when applied. To see this note:

$$\bar{x} = 1/n\mathbf{i}'\mathbf{x} \Rightarrow \mathbf{i}\bar{x} = \begin{pmatrix} \bar{x} \\ \bar{x} \\ \ldots \\ \bar{x} \end{pmatrix} = 1/n\mathbf{ii}'\mathbf{x}$$

Define the column vector of deviations from the mean:

$$\begin{pmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \dots \\ x_n - \bar{x} \end{pmatrix} = [\mathbf{x} - \mathbf{i}\bar{x}] = [\mathbf{x} - 1/n\mathbf{i}\mathbf{i}'\mathbf{x}] = [\mathbf{I} - 1/n\mathbf{i}\mathbf{i}']\mathbf{x} = \mathbf{M}\mathbf{x}$$

Properties of M

- $$\mathbf{M}\mathbf{i} = \mathbf{0} = \mathbf{i}'\mathbf{M} = \mathbf{0}'$$
- $$\mathbf{M}' = \mathbf{M}$$
- $$\mathbf{M}^2 = \mathbf{M}\mathbf{M} = \mathbf{M}$$
- We can write

$$\sum_{i=1}^{n}(x_i - \bar{x}) = (x_1 - \bar{x}) + (x_2 - \bar{x}) + \ldots + (x_n - \bar{x}) =$$

$$(11\ldots1)\begin{pmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \dots \\ x_n - \bar{x} \end{pmatrix} = \mathbf{i}'\mathbf{M}\mathbf{x} = 0$$

- $$\sum_{i=1}^{n}(x_i - \bar{x})^2 = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2 =$$

$$[(x_1 - \bar{x}) \quad (x_2 - \bar{x}) \quad \ldots (x_n - \bar{x})]\begin{pmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \dots \\ x_n - \bar{x} \end{pmatrix} = \mathbf{x}'\mathbf{M}'\mathbf{M}\mathbf{x} = \mathbf{x}'\mathbf{M}\mathbf{x}$$

- $$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = (x_1-\bar{x})(y_1-\bar{y})+(x_2-\bar{x})(y_2-\bar{y})+\ldots+(x_n-\bar{x})(y_n-\bar{y}) =$$

$$[(x_1-\bar{x}) \;\; (x_2-\bar{x}) \;\; \ldots (x_n-\bar{x})] \begin{pmatrix} y_1-\bar{y} \\ y_2-\bar{y} \\ \cdots \\ y_n-\bar{y} \end{pmatrix} = \mathbf{x'M'My} = \mathbf{x'My}$$

1.7 The Determinant

The determinant of an $m \times m$ matrix $A = [a_{ij}]$, denoted by $|A|$ is defined as

$$|A| = \sum_{\sigma} sgn\ \sigma \prod_{i=1}^{m} a_{i\sigma(i)}$$

where the summation runs over all $m!$ permutations σ of the m items $\{1, \ldots, m\}$ and the sign of a permutation σ, $sgn\ \sigma$ is $+1$ or -1 according to whether the minimum number of transpositions, or pair-wise interchanges, necessary to achieve it starting from $\{1, \ldots, m\}$ is even or odd.

The above definition may look complex. Fortunately, there is another method, called the Laplace Expansion, of obtaining determinants. This is a recursive method. This means that in order to get the determinant of a, say, 4×4 matrix we must use lower order determinants of submatrices of our matrix.

This recursive method arises from the second definition of a determinant. Let A_{ij} denote the $m-1 \times m-1$ submatrix of the $m \times m$ matrix A obtained by deleting row i and column j of A. The determinant of A is given by

$$|A| = \sum_{j=1}^{m} (-1)^{j+i} a_{ij} |A_{ij}| = \sum_{i=1}^{m} (-1)^{j+i} a_{ij} |A_{ij}|$$

This is the Laplace expansion of a determinant.

Example 3

$$A = \begin{bmatrix} 3 & 2 & 6 \\ 6 & -1 & 1 \\ -2 & -1 & 8 \end{bmatrix}$$

$$|A| = \begin{vmatrix} 3 & 2 & 6 \\ 6 & -1 & 1 \\ -2 & -1 & 8 \end{vmatrix} = 3\begin{vmatrix} -1 & 1 \\ -1 & 8 \end{vmatrix} - 6\begin{vmatrix} 2 & 6 \\ -1 & 8 \end{vmatrix} - 2\begin{vmatrix} 2 & 6 \\ -1 & 1 \end{vmatrix} =$$

$$3(-8+1) - 6(16+6) - 2(2+6) = -169$$

1.8 Partitioned Matrices

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

A_{ij} is a submatrix of A.

- Addition and multiplication is as usual
- Determinants:

$$\begin{vmatrix} A_{11} & 0 \\ 0 & A_{22} \end{vmatrix} = |A_{11}||A_{22}|$$

$$\begin{vmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{vmatrix} = |A_{11}||A_{22} - A_{21}A_{11}^{-1}A_{12}|$$

- Partitioned inverse

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}^{-1} = \begin{pmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{pmatrix} = \begin{pmatrix} A_{11}^{-1}(I + A_{12}F_2A_{21}A_{11}^{-1}) & -A_{11}^{-1}A_{12}F_2 \\ -F_2A_{21}A_{11}^{-1} & F_2 \end{pmatrix}$$

where $F_2 = \left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1}$. This can be shown by verifying that

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}\begin{pmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}$$

1.9 Kronecker Products

$$C_{nl\times km} = A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \dots & a_{1k}B \\ a_{21}B & a_{22}B & \dots & a_{2k}B \\ \dots & \dots & \dots & \dots \\ a_{n1}B & a_{n2}B & \dots & a_{nk}B \end{pmatrix}$$

1.9.1 Some useful results for square matrices $A_{n\times n}$ and $B_{m\times m}$

A matrix is square if the number of columns equals the number of rows.

- $$(A \otimes B)(C \otimes D) = AC \otimes BD$$
- $$(A \otimes B)' = A' \otimes B'$$
- $$|A \otimes B| = |A|^n |B|^m$$
- $$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$$
- $$trace(A \otimes B) = trace(A) trace(B)$$

  where the trace of a matrix is the sum of its diagonal elements

Exercise 1 *Prove that*

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$$

1.10 The trace of a matrix

The trace of a matrix is the sum of its diagonal elements

1.10.1 Properties of the trace

- $$tr(cA) = ctr(A) \text{ c is scalar}$$
- $$tr(A') = tr(A)$$
- $$tr(A + B) = tr(A) + tr(B)$$

- $$tr(AB) = tr(BA)$$
- $$tr(A \otimes B) = tr(A)tr(B)$$

1.11 The Rank of a matrix

The rank of a matrix is the number of linearly independent columns of a matrix. Equivalently it is the number of linearly independent rows of a matrix. A set of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$, is a linearly independent set if there exists no set of numbers $a_1, \ldots, a_n$, with some $a_i \neq 0$, such that $a_1\mathbf{x}_1 + \ldots + a_n\mathbf{x}_n = 0$.

Example 4

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

Since

$$2\begin{pmatrix} 1 \\ 2 \end{pmatrix} - \begin{pmatrix} 2 \\ 4 \end{pmatrix} = 0$$

there is ony one linearly independent column and so the rank of A is 1.

1.12 Characteristic Roots and Vectors

Consider the square matrix A with

$$\mathbf{Ac} = \lambda\mathbf{c} \tag{1.1}$$

The pairs of solutions to this equation are the characteristic vectors $\boldsymbol{c}$ and the characteristic roots λ Note that since

$$\mathbf{Ac} = \lambda\mathbf{c} \Rightarrow \mathbf{A}k\mathbf{c} = k\lambda\mathbf{c}$$

$\mathbf{c}$ can be normalised so that $\mathbf{c}'\mathbf{c} = 1$. To solve (1.1), see that

$$\mathbf{Ac} = \lambda\mathbf{c} \Rightarrow (\mathbf{A} - \lambda\mathbf{I})\mathbf{c} = \mathbf{0}$$

So $\mathbf{c}$ has a nonzero solution only if $(A - \lambda I)$ is singular i.e. only if $|\mathbf{A} - \lambda\mathbf{I}| = \mathbf{0}$. This equation is called the characteristic equation

Example 5

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

$$0 = |A - \lambda I| \left| \begin{pmatrix} 5 & 1 \\ 2 & 4 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right| = 0$$

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

which is zero for $\lambda = 6, 3$. *To find the characteristic vectors*

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{c} = 0 \Rightarrow \begin{pmatrix} 5-\lambda & 1 \\ 2 & 4-\lambda \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

For $\lambda = 6$

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{c} = 0 \Rightarrow \begin{pmatrix} -1 & 1 \\ 2 & 2 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow c_1 = c_2$$

So we have one solution which justifies that $(\mathbf{A} - \lambda\mathbf{I})$ *is singular. For* $\lambda = 3$

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{c} = 0 \Rightarrow \begin{pmatrix} 2 & 1 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow c_1 = -1/2c_2$$

To identify $(c_1, c_2)'$ *we use the normalisation* ($\mathbf{c}'\mathbf{c} = 1$ *which gives, for* $\lambda = 6$, $c_1 = c_2 = \pm 1/\sqrt{2}$ *and for* $\lambda = 3$, $c_1 = \pm 1/\sqrt{5}$ *and* $c_2 = \pm(-2/\sqrt{5})$.

1.12.1 Useful applications of the characteristic roots and vectors

Note that

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

gives a matrix equation of the form

$$\mathbf{A}(\mathbf{c}_1, \ldots, \mathbf{c}_n) = (\mathbf{c}_1, \ldots, \mathbf{c}_n) \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \lambda_n \end{pmatrix} = \mathbf{AC} = \mathbf{C\Lambda}$$

Theorem 1 *If* $\mathbf{A}$ *is symmetric,* $\mathbf{c}_i'\mathbf{c}_i = 1$ *and* $\mathbf{c}_j'\mathbf{c}_i = 0$ $i \neq j$, *then* $\mathbf{C}' = \mathbf{C}^{-1}$

Proof 1

$$\mathbf{C}'\mathbf{C} = \begin{pmatrix} \mathbf{c}_1'\mathbf{c}_1 & \mathbf{c}_1'\mathbf{c}_2 & \dots & \mathbf{c}_1'\mathbf{c}_n \\ \mathbf{c}_2'\mathbf{c}_1 & \mathbf{c}_2'\mathbf{c}_2 & \dots & \mathbf{c}_2\mathbf{c}_n \\ \dots & \dots & \dots & \dots \\ \mathbf{c}_n'\mathbf{c}_1 & \mathbf{c}_n'\mathbf{c}_2 & \dots & \mathbf{c}_n'\mathbf{c}_n \end{pmatrix} = I$$

Thus

$$\mathbf{C}'\mathbf{C} = \mathbf{I} \Rightarrow \mathbf{C}' = \mathbf{C}^{-1}$$

- Diagonalisation of a symmetric matrix

$$\mathbf{C}'\mathbf{A}\mathbf{C} = \mathbf{C}'\mathbf{C}\mathbf{\Lambda} = \mathbf{I}\mathbf{\Lambda} = \mathbf{\Lambda}$$

- Spectral Decomposition of a matrix

$$\mathbf{A}\mathbf{C} = \mathbf{C}\mathbf{\Lambda} \Rightarrow \mathbf{A}\mathbf{C}\mathbf{C}^{-1} = \mathbf{C}\mathbf{\Lambda}\mathbf{C}^{-1} \Rightarrow \mathbf{A} = \mathbf{C}\mathbf{\Lambda}\mathbf{C}'$$

- Rank of a matrix
  If $\mathbf{A}$ is symmetric

$$rank(\mathbf{A}) = rank(\mathbf{\Lambda})$$

  To see this note

$$rank(\mathbf{C}'\mathbf{A}\mathbf{C}) = rank(\mathbf{\Lambda})$$

  But

$$rank(\mathbf{C}'\mathbf{A}\mathbf{C}) = rank(\mathbf{C}'\mathbf{A})$$

  where we use the result $rank(\mathbf{AB}) \leq min(rank(\mathbf{A}), rank(\mathbf{B}))$ when $\mathbf{B}$ is square and $\mathbf{A}$ is an $m \times n$ matrix But

$$rank(\mathbf{C}'\mathbf{A}) = rank(\mathbf{A}'\mathbf{C}) = rank(\mathbf{A}') = rank(\mathbf{A})$$

  If $\mathbf{A}$ is not symmetric we use

$$rank(\mathbf{A}) = rank(\mathbf{A}'\mathbf{A}) = rank(\mathbf{A}'\mathbf{A})$$

  and

$$rank(\mathbf{A}'\mathbf{A}) = rank(\mathbf{\Lambda})$$

  where $\mathbf{\Lambda}$ contains the characteristic roots of the symmetric matrix $\mathbf{A}'\mathbf{A}$.

- Power of a matrix
  For any symmetric matrix

$$\mathbf{A}^2 = \mathbf{AA} = \mathbf{C\Lambda C'C\Lambda C} = \mathbf{C\Lambda}^2\mathbf{C'}$$

  This generalises to $\mathbf{A}^k = \mathbf{C\Lambda}^k\mathbf{C'}$ for any real k.

- Factoring a matrix
  We want to find $\mathbf{P}$ such that

$$\mathbf{A}^{-1} = \mathbf{P'P}$$

  (i)

$$\mathbf{A}^{-1} = \mathbf{C\Lambda}^{-1}\mathbf{C'} = \mathbf{C\Lambda}^{-1/2}\mathbf{\Lambda}^{-1/2}\mathbf{C'} = \mathbf{P'}\mathbb{P}$$

  where $\mathbf{P} = \mathbf{\Lambda}^{-1/2}\mathbf{C'}$ (ii)

$$\mathbf{A} = \mathbf{LU}$$

  where $\mathbf{L}$ is a lower triangular matrix and $\mathbf{U}$ is an upper triangular matrix.

$$\mathbf{A}^{-1} = \mathbf{L}^{-1}\mathbf{U}^{-1}$$

- Trace of a matrix

$$tr(\mathbf{C'AC}) = tr(\mathbf{ACC'}) = tr(\mathbf{A}) = tr(\mathbf{\Lambda})$$

- Determinant of a matrix

$$|\mathbf{C'AC}| = |\mathbf{C'}||\mathbf{A}||\mathbf{C}| = |\mathbf{C'}||\mathbf{C}||\mathbf{A}| = |\mathbf{C'C}||\mathbf{A}| = |\mathbf{\Lambda}|$$

1.13 Quadratic Forms

Quadratic forms enable us to characterise a matrix as positive or negative definite. Consider the *quadratic form*

$$q = \mathbf{x'Ax}$$

e.g.

$$q = (x_1x_2)\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = a_{11}x_1^2 + a_{21}x_1x_2 + a_{12}x_1x_2 + a_{22}x_2^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} x_ix_ja_{ij}$$

If

$$\mathbf{x}'\mathbf{A}\mathbf{x} > (<)0 \Rightarrow \mathbf{A} \text{ is positive (negative) definite}$$

if

$$\mathbf{x}'\mathbf{A}\mathbf{x} \geq (\leq)0 \Rightarrow \mathbf{A} \text{ is nonnegative (nonpositive) definite}$$

Theorem 2 *A symmetric matrix* $\mathbf{A}$ *is positive (negative) definite if all the characteristic roots of* $\mathbf{A}$ *are positive (negative). If some of the roots are zero then the matrix is nonnegative (nonpositive)*

Proof 2 *Consider* $\mathbf{x}'\mathbf{A}\mathbf{x}$. *We know that a symmetric matrix can be written as*

$$\mathbf{C}'\mathbf{A}\mathbf{C} = \mathbf{\Lambda} \Rightarrow \mathbf{A} = \mathbf{C}\mathbf{\Lambda}\mathbf{C}'$$

Thus

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{x}'\mathbf{C}\mathbf{\Lambda}\mathbf{C}'\mathbf{x} \Rightarrow \mathbf{y}'\mathbf{\Lambda}\mathbf{y} = \sum_{i=1}^{k} \lambda_i y_i^2$$

Q.E.D.

Theorem 3 *If* $\mathbf{A}$ *is nonegative then* $|\mathbf{A}| \geq 0$

Proof 3 *If* $\mathbf{A}$ *is nonegative this means that*

$$\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0$$

But

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^{k} \lambda_i y_i^2$$

Therefore for all i $\lambda_i \geq 0$ *Then*

$$|A| = |\Lambda| = \prod_{i=1}^{n} \lambda_i \geq 0$$

Q.E.D.

1.14 Calculus and Matrix Algebra

Consider a scalar valued function of a vector $\mathbf{x} = (x_1, \dots x_n)'$

$$y = f(x_1, \dots, x_n) = f(\mathbf{x})$$

The vector of partial derivatives known as the gradient vector is

$$J_f \equiv \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}'} = \left[\frac{\partial f(\mathbf{x})}{\partial x_1}, \frac{\partial f(\mathbf{x})}{\partial x_1}, \dots, \frac{\partial f(\mathbf{x})}{\partial x_n}\right] = [f_1, f_2, \dots, f_n]$$

Generalising to vector functions we have a vector function of a vector

$$\mathbf{y} = \mathbf{f}(\mathbf{x})$$

where $\mathbf{x} = (x_1, \dots, x_n)'$, $\mathbf{y} = (y_1, \dots, y_m)'$. In general, the matrix of partial derivatives is known as the Jacobian matrix is given by

$$J_f = \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}'} = \begin{pmatrix} \frac{\partial \mathbf{f}_1(\mathbf{x})}{\partial x_1} & \frac{\partial \mathbf{f}_1(\mathbf{x})}{\partial x_2} & \dots & \frac{\partial \mathbf{f}_1(\mathbf{x})}{\partial x_n} \\ \frac{\partial \mathbf{f}_2(\mathbf{x})}{\partial x_1} & \frac{\partial \mathbf{f}_2(\mathbf{x})}{\partial x_2} & \dots & \frac{\partial \mathbf{f}_2(\mathbf{x})}{\partial x_n} \\ \dots & \dots & \dots & \dots \\ \frac{\partial \mathbf{f}_m(\mathbf{x})}{\partial x_1} & \frac{\partial \mathbf{f}_m(\mathbf{x})}{\partial x_2} & \dots & \frac{\partial \mathbf{f}_m(\mathbf{x})}{\partial x_n} \end{pmatrix}$$

Note that sometimes the transpose of the Jacobian matrix is presented, so that for a scalar function the Jacobian is a column vector rather than a row vector. So

$$J'_{\mathbf{f}} = \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} \begin{pmatrix} \frac{\partial \mathbf{f}_1(\mathbf{x})}{\partial x_1} & \frac{\partial \mathbf{f}_2(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial \mathbf{f}_m(\mathbf{x})}{\partial x_1} \\ \frac{\partial \mathbf{f}_1(\mathbf{x})}{\partial x_2} & \frac{\partial \mathbf{f}_2(\mathbf{x})}{\partial x_2} & \dots & \frac{\partial \mathbf{f}_m(\mathbf{x})}{\partial x_2} \\ \dots & \dots & \dots & \dots \\ \frac{\partial \mathbf{f}_1(\mathbf{x})}{\partial x_n} & \frac{\partial \mathbf{f}_2(\mathbf{x})}{\partial x_n} & \dots & \frac{\partial \mathbf{f}_m(\mathbf{x})}{\partial x_n} \end{pmatrix}$$

The second derivative matrix (known as Hessian matrix) for scalar functions is defined as:

$$H_f = [\frac{\partial (\frac{\partial y}{\partial x})'}{\partial x_1}, \frac{\partial (\frac{\partial y}{\partial x})'}{\partial x_2} \dots \frac{\partial (\frac{\partial y}{\partial x})'}{\partial x_n}] = \begin{pmatrix} \frac{\partial^2 y}{\partial x_1 \partial x_1} & \frac{\partial^2 y}{\partial x_1 \partial x_2} & \dots & \frac{\partial^2 y}{\partial x_1 \partial x_n} \\ \frac{\partial^2 y}{\partial x_2 \partial x_1} & \frac{\partial^2 y}{\partial x_2 \partial x_2} & \dots & \frac{\partial^2 y}{\partial x_2 \partial x_n} \\ \dots & \dots & \dots & \dots \\ \frac{\partial^2 y}{\partial x_n \partial x_1} & \frac{\partial^2 y}{\partial x_n \partial x_2} & \dots & \frac{\partial^2 y}{\partial x_n \partial x_n} \end{pmatrix}$$

Extending this to vector functions of vectors we have

$$H_{\mathbf{f}} = \begin{pmatrix} H_{f_1} \\ H_{f_2} \\ \cdots \\ H_{f_m} \end{pmatrix}$$

where f_i is the i-th argument of $\mathbf{f}$.

1.14.1 Some Examples

- Derivative of a linear function

$$y = \alpha'\mathbf{x} = a_1x_1 + \ldots + a_nx_n = \mathbf{x}'\alpha$$

$$\frac{\partial \alpha'\mathbf{x}}{\partial \mathbf{x}'} = [\frac{\partial \alpha'\mathbf{x}}{\partial x_1}, \frac{\partial \alpha'\mathbf{x}}{\partial x_2}, \ldots, \frac{\partial \alpha'\mathbf{x}}{\partial x_n}] = \alpha'$$

- Derivative of a set of linear functions

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_m \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + \ldots + a_{1n}x_n \\ a_{21}x_1 + \ldots + a_{2n}x_n \\ \cdots \\ a_{m1}x_1 + \ldots + a_{mn}x_n \end{pmatrix}$$

So

$$J = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix} = A$$

- Derivative of a quadratic form

$$q = \mathbf{x}'\mathbf{A}\mathbf{x}$$

If A is symmetric

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

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

If A is not symmetric

$$\frac{\partial q}{\partial \mathbf{x}'} = \mathbf{x}'(\mathbf{A}' + \mathbf{A})$$
$$\frac{\partial q}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}')\mathbf{x}$$
$$\frac{\partial q}{\partial \mathbf{A}} = \mathbf{x}\mathbf{x}'$$
$$\frac{\partial ln|\mathbf{A}|}{\partial \mathbf{A}} = A^{-1}$$

1.15 Transformations and Jacobian

Consider $y = f(x)$ then if f is a monotonic function

$$x = f^{-1}(y)$$

Denote the slope $\frac{dx}{dy} \equiv J$, e.g. For a linear function

$$y = a + bx \Rightarrow x = -a/b + (1/b)y$$

and

$$J = \frac{dx}{dy} = 1/b$$

Consider that now $\mathbf{y}$ is a column vector of functions, $\mathbf{y} = \mathbf{f}(\mathbf{x})$ then

$$\mathbf{J} = \frac{\partial \mathbf{f}^{-1}(\mathbf{x})}{\partial \mathbf{y}'} = \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_m} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_m} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_m} \end{pmatrix}$$

The absolute value of $|\mathbf{J}|$ is the Jacobian So e.g.

$$\mathbf{y} = \mathbf{A}\mathbf{x} \Rightarrow \mathbf{x} = \mathbf{A}^{-1}\mathbf{y} = \mathbf{C}\mathbf{y}$$

assuming that $\mathbf{A}$ is nonsingular Then

$$\mathbf{J} = \mathbf{A}^{-1}$$

and

$$abs(|\mathbf{J}|) = abs(|\mathbf{A}^{-1}|) = abs(1/|\mathbf{A}|) = 1/abs(|\mathbf{A}|)$$

1.16 Exercises

Exercise 2 *Exercise 3 of Chapter 2 of Greene, pp. 58*

Exercise 3 *Show that* $\mathbf{X}'\mathbf{M}\mathbf{X} = \mathbf{X}'\mathbf{X} - n\bar{x}\bar{x}$

Exercise 4 *Exercise 13 of Chapter 2 of Greene, pp. 59*

Exercise 5 *Exercise 23 of Chapter 2 of Greene, pp. 60*

Exercise 6 *Exercise 21 of Chapter 2 of Greene, pp. 60*

Exercise 7 *Consider the matrix*

$$A = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix} = \begin{pmatrix} \mathbf{i}'\mathbf{i} & \mathbf{i}'\mathbf{x} \\ \mathbf{x}'\mathbf{i} & \mathbf{x}'\mathbf{x} \end{pmatrix}$$

Show that the lower right-hand side element of A^{-1} *is*

$$F_2 = [\mathbf{x}'\mathbf{M}\mathbf{x}]^{-1} = \frac{1}{\sum_{i=1}(x_i - \bar{x})^2}$$

where $I - 1/n\mathbf{i}'\mathbf{i}$

Exercise 8 *Solve the Least Squares problem*

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{u}$$

i.e. minimise $\mathbf{u}'\mathbf{u}$ *with respect to* $\mathbf{b}$ *where* $\mathbf{X}$ *is* $n \times k$ *matrix and* $\mathbf{y}$ *is* $n \times 1$ *vector.*

Chapter 2

Random Variables and Probability Distributions

2.1 Introduction

• Consider the random variable X with probability distribution density function $f(x)$

TWO AXIOMS OF PROBABILITY

(i) Discrete random variable X

•
$$0 \leq Prob(X = x) \leq 1, \quad \text{where } x : \text{a value of } X$$

• $\sum_{X=-\infty}^{X=\infty} f(x) = 1$

(ii) Continuous random variable X

•
$$0 \leq Prob(x_1 \leq X \leq x_2) \leq 1 \;\; or \;\; \int_{x_1}^{x_2} f(x)dx = 1$$

• $\int_{-\infty}^{\infty} f(x)dx = 1$

• Cumulative distribution function (CDF): $F(x) = Prob(X \leq x)$ (i) Distrete random variable

$$F(x) = Prob(X \leq x) = \sum_{-\infty}^{x} f(X)$$

(i) Continuous random variable

$$F(x) = \int_{-\infty}^{x} f(x)dx$$

Properties of $F(x)$:

- $0 \leq F(x) \leq 1$
- $F(\infty) = 1$
- $F(-\infty) = 0$
- $f(x) = \frac{dF(x)}{dx}$

2.2 Specific Distributions

2.2.1 Normal Distribution

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

Properties:

- Mean (expectation): $E(x) = \int_{-\infty}^{\infty} xf(x; \mu, \sigma^2) = \mu$
- Variance: $Var(x) = E(x - \mu^2) = \int_{-\infty}^{\infty}(x - \mu)^2 f(x; \mu, \sigma^2) = \sigma^2$
- Skewness: $sk = E(x - \mu)^3 = \int_{-\infty}^{\infty}(x - \mu)^3 f(x; \mu, \sigma^2) = 0$ Skewness coefficient: $\frac{E(x-\mu)^3}{\sigma^3} = 0$
- Kyrtosis: $E(x-\mu)^4 = \int_{-\infty}^{\infty}(x-\mu)^4 f(x; \mu, \sigma^2) = 3\sigma^4$ Kyrtosis coefficient: $\frac{E(x-\mu)^4}{\sigma^4} = \frac{3\sigma^4}{\sigma^4} = 3$

Standard Normal Distribution

Consider $z = \frac{x-\mu}{\sigma} \Rightarrow z \sim N(0,1)$ with $f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}$, which implies that $E(z) = 0$ and $Var(z) = 1$. The functional form of the standard normal distribution can be obtained from the normal distribution using the *change in variable* technique. This says that if

$$x \sim f_X(x) : x\text{'s pdf}$$

then $y = g(x)$ is distributed as

$$y\text{'s pdf: } f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dy}{dx}^{-1} \right|$$

To see this consider that

$$x \sim f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} : x \sim N(\mu, \sigma^2)$$

and that $z = \frac{x-\mu}{\sigma}$ (which implies that $x = \sigma z = \mu$). Then by applying the change in variables technique we get:

$$f_Z(z) = f_X(g^{-1}(y)) \left| \frac{dy}{dx}^{-1} \right| = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\sigma z+\mu-\mu)^2}{2\sigma^2}} \left| \frac{1}{\sigma}^{-1} \right| = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2}} |\sigma| = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

2.3 Other Distributions

- Chi-squared χ^2
  One Degree of freedom: $\chi^2(1) = z^2$, where $z \sim N(0,1)$
  n-degrees of freedom: $\chi^2(n) = \sum_{i=1}^{n} z_i$, where $z_1 \ldots z_n$ are independent $N(0,1)$ distributions.

- F-Distribution: F_{n_1,n_2}

  $$F_{n_1,n_2} = \frac{\chi^2(n_1)/n_1}{\chi^2(n_2)/n_2}$$

  where $\chi^2(n_1)$ and $\chi^2(n_2)$ are independent χ^2 distributions with n_1 and n_2 degrees of freedom respectively.

- t-distribution
$$t(n) = \frac{z}{\sqrt{\chi^2(n)/n}}$$
where $z \sim N(0,1)$ and $\chi^2(n)$ is χ^2 with n degrees of freedom.

- Log-normal distribution $LN(\mu, \sigma^2)$
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{1}{2\sigma^2}(lnx-\mu)^2}$$
$$E(x) = E^{\mu+\frac{\sigma^2}{2}}, \quad Var(x) = e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$$
Properties of the log-normal:
(i) Find that if $x \sim LN(\mu, \sigma^2)$ then $y = lnx \sim N(\mu, \sigma^2)$, x is positive.
Proof: By using the change in variables technique we have:
$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dy}{dx}^{-1} \right| =$$
where $y = g(x) = lnx$ or $x = e^y$.
$$= \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{1}{2\sigma^2}(lne^y-\mu)^2} \left| \frac{1}{x}^{-1} \right| = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} |x| =$$
$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} \Rightarrow y \sim N(\mu, \sigma^2)$$
(ii) If $x \sim LN(\mu, \sigma^2)$, then $lnx^v \sim N(v\mu, \sigma^2 v^2)$
(iii) If Y_1, Y_2 are independent lognormal variables with $Y_1 \sim LN(\mu_1, \sigma_1^2)$ and $Y_2 \sim LN(\mu_2, \sigma_2^2)$ then $Y_1 Y_2 \sim LN(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

2.4 Joint Probabilities

The *joint probability distributions* give the probabilities that two or more events happen simultaneously. e.g. in the case of two variables: X: years of being unemployed; Y: Criminal Records

$f(x,y)$	X			$f_Y(y)$
Y	0	1	2	
0	0.05	0.1	0.03	0.18
1	0.21	0.11	0.19	0.51
2	0.08	0.15	0.08	0.21
$f_X(x)$	0.34	0.36	0.30	1.00

$$pdf: \quad f(x,y) = Prob(a \leq x \leq b, c \leq y \leq d) = \int_a^b \int_c^d f(x,y)dydx$$

$$cdf: \quad F(x,y) = Prob(X \leq x, Y \leq y) = \int_{-\infty}^b \int_{-\infty}^d f(x,y)dydx$$

The *marginal distribution* is defined with respect to an individual variable (x or y)

$$f_X(x) = \int_y f(x,s)ds$$

$$f_Y(y) = \int_x f(s,y)ds$$

Expectations in a joint distribution

$$E(x) = \int_x x f_X(x)dx =$$

$$\int_x x \left[\int_y f(x,s)ds\right] dx =$$

$$\int_x \int_y x f(x,s)dsdx =$$

$$\int_x \int_y x f(x,y)dydx \quad \text{for notational convenience } y \equiv s$$

In general

$$E\left[g(x,y)\right] = \int_x \int_y g(x,y)f(x,y)dydx$$

Exercise 9 *Show that if two random variables are independent, i.e.*

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

then their covariance σ_{XY} is zero, i.e. $\sigma_{XY} = 0$

2.4.1 Conditional Distributions

In the bivariate case, there is a conditional distribution of Y given a value of $X = x$, i.e.

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

In the case that Y and X are independent $f(y|x) = f_Y(y)$.

2.4.2 Conditional Mean

$$E[Y|X] = \int_y yf(y|x)dy$$

This conditional mean of Y on X, $E(Y|X)$, is called the regression of Y on X. This happens since any random variable can be written as

$$Y = E[Y|X] - E[Y|X] + Y =$$

$$E[Y|X] + [Y - E[Y|X]] =$$

$$E[Y|X] + \epsilon$$

where ϵ is an error term.

2.4.3 Conditional Variance

$$Var[Y|X] = E[(Y - E(Y|X]))^2|X] =$$

$$\int_y (y - E[Y|X])^2 f(y|x)dy =$$

$$\int_y (y^2 + (E[Y|X])^2 - 2yE(Y|X))f(y|x)dy =$$

$$\int_y y^2 f(y|x)dy + \int_y (E[Y|X])^2 f(y|x)dy - \int_y 2yE[Y|X]f(y|x)dy =$$

$$E[Y^2|X] + (E[YX])^2 \int_y f(y|x)dy - \int_y 2yE[Y|X]f(y|x)dy =$$

$$E[Y^2|X] + (E[Y|X])^2 - 2E[Y|X]E[Y|X] =$$

$$E[Y^2|X] - (E[Y|X])^2$$

2.4.4 Law of Iterated Expectations

Denote by $E_X[.]$ the expectation over the values of X, i.e. $E_X[.] = \int_X [.] f_X(x)dx$. Then we can show that

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

Proof

$$E_X[E(Y|X)] = \int_x E(Y|X) f_X(x)dx =$$

$$\int_x \left(\int_y y f(y|x)dy \right) f_X(x)dx =$$

$$\int_x \int_y y f(y|x) f_X(x)dydx = E(Y)$$

Example 6 (The bivariate normal distribution)

$$f(x,y) = \frac{1}{2\pi\sigma_Y\sigma_X\sqrt{1-\rho^2}} e^{\frac{-1}{2(1-\rho^2)}[E_X^2 + E_Y^2 - 2\rho E_X E_Y]}$$

where

$$E_X = \frac{x-\mu_X}{\sigma_X}, \quad E_X = \frac{y-\mu_Y}{\sigma_Y}$$

and $\rho = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$ *is the correlation coefficient between* X *and* Y*. The marginal distributions are normal*

$$f_X(x) = \int_y f(x,y)dy = \frac{1}{\sqrt{2\pi}\sigma_X} e^{\frac{-1}{2\sigma_X^2}(x-\mu_X)^2}$$

$$f_Y(y) = \int_x f(x,y)dx = \frac{1}{\sqrt{2\pi}\sigma_Y} e^{\frac{-1}{2\sigma_Y^2}(y-\mu_Y)^2}$$

Proof*:*

$$f_X(x) = \int_y f(x,y)dy = \int_y \frac{1}{2\pi\sigma_Y\sigma_X\sqrt{1-\rho^2}} e^{\frac{-1}{2(1-\rho^2)}[E_X^2 + E_Y^2 - 2E_X E_Y]} dy =$$

$$\frac{1}{\sqrt{2\pi}\sigma_X} e^{\frac{-1}{2\sigma_X^2}(x-\mu_X)^2} \int_y \frac{1}{2\pi\sigma_Y\sigma_X\sqrt{1-\rho^2}} e^{\frac{-1}{2(1-\rho^2)}[E_X^2 + E_Y^2 - 2\rho E_X E_Y]} e^{1/2E_X^2} dy =$$

$$\frac{1}{\sqrt{2\pi}\sigma_X} e^{\frac{-1}{2\sigma_X^2}(x-\mu_X)^2} 1 \sim N(\mu_X, \sigma_X^2)$$

The conditional distributions are normal and have a linear mean in x

$$f(y|x) \equiv N(\alpha + \beta x, \sigma_Y^2(1-\rho^2))$$

Proof*:*

$$f(y|x) = \frac{f(y,x)}{f_X(x)} = \frac{\frac{1}{2\pi\sigma_Y\sigma_X\sqrt{1-\rho^2}} e^{\frac{-1}{2(1-\rho^2)}[E_X^2+E_Y^2-2\rho E_X E_Y]}}{\frac{1}{\sqrt{2\pi}\sigma_X} e^{\frac{-1}{2\sigma_X^2}(x-\mu_X)^2}} =$$

$$\frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}} e^{\frac{-1}{2(1-\rho^2)}[E_X^2+E_Y^2-2\rho E_X E_Y]} e^{\frac{1}{2\sigma_X^2}(x-\mu_X)^2} =$$

$$\frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}} e^{\frac{-1}{2(1-\rho^2)}[E_X^2+E_Y^2-2\rho E_X E_Y-(1-\rho^2)E_X^2]} =$$

$$\frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}} e^{\frac{-1}{2(1-\rho^2)}[E_Y-\rho E_X]^2} =$$

$$\frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}} e^{\frac{-1}{2(1-\rho^2)\sigma_Y^2}[y-\mu_Y-\frac{\sigma_{XY}}{\sigma_X}\frac{x-\mu_X}{\sigma_X}]^2} =$$

$$\frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}} e^{\frac{-1}{2(1-\rho^2)\sigma_Y^2}[y-(\mu_Y+\frac{\sigma_{XY}}{\sigma_X}\frac{x-\mu_X}{\sigma_X})]^2} =$$

$$\frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}} e^{\frac{-1}{2(1-\rho^2)\sigma_Y^2}[y-(\alpha+\beta x)]^2} =$$

where $\alpha = \mu_Y - \beta\mu_X$ *and* $\beta = \frac{\sigma_{XY}}{\sigma_X}$*. This is a normal distribution with mean* $\alpha + \beta x$ *and variance* $(1-\rho^2)\sigma_Y^2$*, i.e.* $f(y|x) \sim N(\alpha + \beta x, \sigma_Y^2(1-\rho^2))$*.*

2.4.5 The multivariate normal distribution

$$f(x_1, x_2, \ldots, x_n) = f(\mathbf{x}) = (2\pi)^{-n/2}(\sigma_1\sigma_2\ldots\sigma_n)^{-1}|R|^{-1/2}e^{-1/2\boldsymbol{\epsilon}'R^{-1}\boldsymbol{\epsilon}}$$

where $\boldsymbol{\epsilon}' = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$ with elements $\epsilon_i = \frac{x_i-\mu_i}{\sigma_i}$ and

$$R = \begin{pmatrix} 1 & \rho_{12} & \ldots & \rho_{1n} \\ \rho_{21} & 1 & \ldots & \rho_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ \rho_{n1} & \rho_{n2} & \ldots & \rho_{nn} \end{pmatrix}$$

is the correlation matrix.

TWO SPECIAL CASES:
(i) $R = I$, i.e. $x_1, x_2, \ldots, x_n$ are uncorrelated. Then

$$f(\mathbf{x}) = (2\pi)^{-n/2}(\sigma_1\sigma_2 \ldots \sigma_n)^{-1}e^{-1/2\boldsymbol{\epsilon}'\boldsymbol{\epsilon}}$$

since $|R| = |I| = 1$ and $I\boldsymbol{\epsilon} = I$. Then

$$f(\mathbf{x}) = (2\pi)^{-n/2}(\sigma_1\sigma_2 \ldots \sigma_n)^{-1}e^{-1/2(\epsilon_1^2+\ldots\epsilon_n^2)} =$$

$$(2\pi)^{-n/2}(\sigma_1\sigma_2 \ldots \sigma_n)^{-1}e^{-1/2\epsilon_1^2}e^{-1/2\epsilon_2^2} \ldots e^{-1/2\epsilon_n^2} =$$

$$\frac{1}{\sigma_1\sqrt{2\pi}}e^{-1/2\epsilon_1^2}\frac{1}{\sigma_2\sqrt{2\pi}}e^{-1/2\epsilon_2^2} \ldots \frac{1}{\sigma_n\sqrt{2\pi}}e^{-1/2\epsilon_n^2}$$

$$f(x_1)f(x_2)\ldots f(x_n) = \prod_{i=1}^{n} f(x_i)$$

(ii) $R = I$, $\boldsymbol{\mu} = (0, 0 \ldots, 0)'$ and $\sigma_i = \sigma\ \forall i$ Then

$$f(\mathbf{x}) = (2/pi)^{-n/2}\sigma^{-n}e^{\mathbf{x}'\mathbf{x}/2\sigma^2}$$

Exercise 10 (Decomposition of Variance) *Show that*

$$Var(Y) = E_X[Var(Y|X)] + Var_X[E(Y|X)]$$

Chapter 3

Estimation and Statistical Inference

3.1 Introduction

Denote the estimator of a parameter θ of the distribution (probability model) of the population as $\hat{\theta}$, then in econometrics we often look for the best linear unbiased estimator (BLUE). The properties of this estimator are:

- *Unbiased estimator*

  $$E(\hat{\theta}) = \theta$$

  E.g. consider the sample mean estimator of a random sample

  $$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

  drawn from $x \sim N(\mu, \sigma^2)$, then

  $$E(\bar{x}) = E\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{\sum_{i=1}^{n} E(x_i)}{n} = \frac{\sum_{i=1}^{n} \mu}{n} = \mu$$

- *Efficient Unbiased Estimator*

  $$Var(\hat{\theta}) < Var(\hat{\theta}^*)$$

  i.e. the variance of the estimator is smaller than the variance of any other unbiased estimator.

Theorem 4 *The variance of an unbiased estimator of a parameter θ will always be at least as large as*

$$(I(\theta))^{-1} = \left(-E\left[\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}\right]\right)^{-1} = \left(E\left[\frac{\partial \ln L(\theta)}{\partial \theta}\right]^2\right)^{-1}$$

i.e.

$$Var(\theta) \geq (I(\theta))^{-1}$$

where $L(\theta)$ is the likelihood function of θ gien thet data at $\mathbf{x} = (x_1, \dots x_n)'$, i.e. $L(\theta) = f(x_1, \dots, x_n; \theta)$.

Proof 4 *For a proof see Appendix 1.*

In *MATRIX NOTATION* we have

- Random sample (DATA): $(x_1, \dots, x_n)' = \mathbf{x}_{n\times 1}$
- vector of parameters of the underlying probability model: $(\theta_1, \dots, \theta_k) = \boldsymbol{\theta}_{k\times 1}$, e.g. $\boldsymbol{\theta} = (\mu, \sigma^2)'$ for the normal probability model.
- Variance of $\hat{\boldsymbol{\theta}}$:

$$Var(\hat{\boldsymbol{\theta}}) = E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'] =$$

$$E\begin{pmatrix} (\hat{\theta}_1 - \theta_1)^2 & (\hat{\theta}_1 - \theta_1)(\hat{\theta}_2 - \theta_2) & \dots & (\hat{\theta}_1 - \theta_1)(\hat{\theta}_k - \theta_k) \\ (\hat{\theta}_2 - \theta_2)(\hat{\theta}_1 - \theta_1) & (\hat{\theta}_2 - \theta_2)^2 & \dots & (\hat{\theta}_2 - \theta_2)(\hat{\theta}_k - \theta_k) \\ \dots & \dots & \dots & \dots \\ (\hat{\theta}_k - \theta_k)(\hat{\theta}_1 - \theta_1) & \dots & \dots & (\hat{\theta}_k - \theta_k)^2 \end{pmatrix} =$$

$$\begin{pmatrix} Var(\hat{\theta}_1) & Cov(\hat{\theta}_1, \hat{\theta}_2) & \dots & Cov(\hat{\theta}_1, \hat{\theta}_k) \\ Cov(\hat{\theta}_2, \hat{\theta}_1) & Var(\hat{\theta}_2) & \dots & Cov(\hat{\theta}_2, \hat{\theta}_k) \\ \dots & \dots & \dots & \dots \\ Cov(\hat{\theta}_k, \hat{\theta}_1) & \dots & \dots & Var(\hat{\theta}_k) \end{pmatrix}$$

3.2 Efficient Estimation: Maximum Likelihood

Consider a random sample of $(x_1, \dots, x_n)'$ observations. These observations are drawn independently from a probability density function (pdf) eith vector of parameters $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$. This means that

$$L(\boldsymbol{\theta}) = f(x_1, \dots, x_n; \boldsymbol{\theta}) = f(x_1, ; \boldsymbol{\theta}) \dots f(x_n; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta})$$

The *maximum likelihood estimator* requires to find the value $\hat{\boldsymbol{\theta}}_{ML}$ (ML estimator) which

$$\max_{\boldsymbol{\theta}} \ln L(\boldsymbol{\theta}) = \ln \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(x_i;\boldsymbol{\theta})$$

$$F.O.C. \quad \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \theta_1} \\ \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \theta_2} \\ \dots \\ \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \theta_k} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \end{pmatrix} = \mathbf{0}_{k\times 1}$$

$$S.O.C. \quad \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \begin{pmatrix} \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_1^2} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} & \dots & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_k} \\ \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_2^2} & \dots & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_k} \\ \dots & \dots & \dots & \dots \\ \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_1} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_2} & \dots & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \theta_k^2} \\ \end{pmatrix} = \boldsymbol{H}$$

$\boldsymbol{H}$ is the Hessian matrix which should be negative definite at $\hat{\boldsymbol{\theta}}$. Consider the random sample of $x_1, x_2, \dots x_n$ observations drawn independently from the normal distribution: $N(\mu, \sigma^2)$. Then ML estimation proceeds as follows:

$$L(\boldsymbol{\theta}) = f(x_1;\boldsymbol{\theta}) \dots f(x_n;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_1-\mu)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_2-\mu)^2}{2\sigma^2}} \dots \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_n-\mu)^2}{2\sigma^2}}$$

Then

$$\ln L(\boldsymbol{\theta}) = \frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - 1/2 \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^2}$$

$$F.O.C. \quad \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \mu} \\ \frac{\partial \ln L(\boldsymbol{\theta})}{\partial \sigma^2} \end{pmatrix} = \mathbf{0}$$

$$\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \mu} = \sum_{i=1}^{n} \frac{(x_i-\mu)}{\sigma^2} = 0 \Rightarrow \hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i-\mu)^2 = 0 \Rightarrow \hat{\sigma}^2 = 1/n \sum_{i=1}^{n} (x_i-\hat{\mu})^2$$

$$S.O.C. \quad \boldsymbol{H} = \begin{pmatrix} \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \mu^2} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \sigma^4} \end{pmatrix} = \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

$$\boldsymbol{H} = \begin{pmatrix} \frac{-n}{\sigma^2} & -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu) \\ -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i - \mu)^2 \end{pmatrix}$$

The information matrix

$$I(\boldsymbol{\theta}) = -E\left[\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right] = -E[\boldsymbol{H}] = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$$

Since

$$E(-n/\sigma^2) = -n/\sigma^2$$

$$E\left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \mu \partial \sigma^2}\right) = -1/\sigma^4 E(\sum_{i=1}^{n} x_i - \mu) = -1/\sigma^4 \sum_{i=1}^{n} E(x_i) - \mu = 0$$

$$E\left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \sigma^4}\right) = n/2\sigma^4 - 1/\sigma^6 \sum_{i=1}^{n} E(x_i - \mu)^2 = n/2\sigma^4 - n/\sigma^4 = -n/2\sigma^4$$

The ML estimator has the property that the variance of $\hat{\boldsymbol{\theta}}_{ML}$ equals the Cramer-Rao lower bound. Thus

$$Var(\hat{\boldsymbol{\theta}}_{ML}) = [I(\boldsymbol{\theta})]^{-1} = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}$$

For this reason we can treat ML estimation as efficient estimation. To confirm that the ML is the most efficient estimator, consider the alternative sample moment estimators of the mean and variance of a population.

$$\bar{x} = 1/n \sum_{i=1}^{n} x_i$$

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

Note that $\bar{x}$ is the same as the ML estimator. but s^2 is not. $s^2 = \frac{n}{n-1}\hat{\sigma}^2_{ML}$. We will show that the variance of $\boldsymbol{\theta}^* = (\bar{x}, s^2)'$ differs from $Var(\hat{\boldsymbol{\theta}}_{ML}) = [I(\boldsymbol{\theta})]^{-1}$ by a positive definite matrix. To show this we need to find the variance covariance matrix of $\boldsymbol{\theta}^* = (\bar{x}, s^2)'$, denoted by $Var(\boldsymbol{\theta}^*)$. We use the results

$$Var(\bar{x}) = 1/n^2 Var(\sum_{i=1}^{n} x_i) = 1/n^2 \sum_{i=1}^{n} Var(x_i) = 1/n^2 \sum_{i=1}^{n} \sigma^2 = \sigma^2/n$$

$$Var(s^2) = 2\sigma^4/n - 1$$

$$Cov(\bar{x}, s^2) = 0$$

Then ,

$$Var(\boldsymbol{\theta}^*) - Var(\hat{\boldsymbol{\theta}}_{ML}) = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n-1} \end{pmatrix} - \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & \frac{2\sigma^4}{n(n-1)} \end{pmatrix}$$

which is a positive semidefinite matrix. We need to prove the results

$$Var(s^2) = 2\sigma^4/n - 1$$

$$Cov(\bar{x}, s^2) = 0$$

To prove them we need to establish a number of useful results for quadratic forms.

Theorem 5 *If $\boldsymbol{X}_{k\times 1} \sim N(0, \sigma^2 I)$ then $C'\boldsymbol{X}_{k\times 1} \sim N(0, \sigma^2 I)$ where $C'C = I$*

Proof 5 *Define $\boldsymbol{Y} = C\boldsymbol{X}$. Since $\boldsymbol{Y}$ is a linear combination of $\boldsymbol{X}$, $\boldsymbol{Y}$ is normally distributed too. Then,*

$$E(\boldsymbol{Y}) = C'E(\boldsymbol{X}) = 0$$

The variance of $\boldsymbol{Y}$ is given by

$$Var(\boldsymbol{Y}) = E(\boldsymbol{Y} - E(\boldsymbol{Y}))(\boldsymbol{Y} - E(\boldsymbol{Y}))' = E(\boldsymbol{Y}\boldsymbol{Y}') = E(C'\boldsymbol{X}\boldsymbol{X}'C) =$$

$$C'E(\boldsymbol{X}\boldsymbol{X}')C = C'Var(\boldsymbol{X})C = C'\sigma^2 IC = \sigma^2 C'C = \sigma^2 I$$

Q.E.D.

Theorem 6 *If* $\boldsymbol{X} \sim N(0, I)$ *then the quadratic form* $q = \boldsymbol{X}'M\boldsymbol{X} \sim \chi^2(n-1)$

Proof 6 *Using the spectral decomposition of* M *we have*

$$q = \boldsymbol{X}'M\boldsymbol{X} = \boldsymbol{X}'C\Lambda C'\boldsymbol{X}' = \boldsymbol{Y}'\Lambda\boldsymbol{Y} = \sum_{i=1}^{n} \lambda_i y_i^2 = \sum_{i=1}^{J} y_i^2$$

where $\boldsymbol{Y} = C'\boldsymbol{X}$ *Note that* $\sum_{i=1}^{n} \lambda_i y_i^2 = \sum_{i=1}^{J} y_i^2$ *because* M *is idempotent with* J *unit and* $n - J$ *zero eigenvalues. The above result implies that* q *is the sum of squared random variables which are independent and normally distrubted. with mean zero and variance unity since* $\boldsymbol{X} \sim NM(0, I)$*. Thus* $q \sim \chi^2(J)$*.* J *is equal to the number of non-zero eigenvalues of* M*, which is equal to the rank of* M*. Since* M *is idempotent*

$$rank(M) = tr(M) = tr \begin{pmatrix} 1-1/n & -1/n & \dots & -1/n \\ -1/n & 1-1/n & \dots & -1/n \\ \dots & \dots & \dots & \dots \\ -1/n & -1/n & \dots & 1-1/n \end{pmatrix} = n(1-1/n) = n-1$$

Q.E.D.

Theorem 7 *If* $\boldsymbol{X} \sim N(0, I)$*, then* $\boldsymbol{i}'\boldsymbol{X}$ *and* $q = \boldsymbol{X}'M\boldsymbol{X}$ *are independent*

Proof 7 *Write*

$$q = \boldsymbol{X}'M\boldsymbol{X}$$

as

$$\boldsymbol{X}MM\boldsymbol{X} = \boldsymbol{X}_1'\boldsymbol{X}_1$$

where $\boldsymbol{X}_1 = M\boldsymbol{X}$ *Denote the function* $\boldsymbol{i}'\boldsymbol{X}$ *as*

$$C = \boldsymbol{i}'\boldsymbol{X} = X'\boldsymbol{i}$$

Note that

$$Cov(\boldsymbol{X}_1, C) = E[\boldsymbol{i}'\boldsymbol{X}(M\boldsymbol{X})'] = E[\boldsymbol{i}'\boldsymbol{X}\boldsymbol{X}'M] =$$
$$\boldsymbol{i}'E(\boldsymbol{X}\boldsymbol{X}')M = \boldsymbol{i}'IM = \boldsymbol{i}'M = (M\boldsymbol{i})' = 0$$

Thus q *is made of normal vectors which are uncorrelated to the linear function* $\boldsymbol{i}'\boldsymbol{X}$*. This means that* q *and the linear function* $\boldsymbol{i}'\boldsymbol{X}$ *are independent.*
Q.E.D.

Having established the results of the above Theorems we can now easily prove

$$Var(s^2) = 2\sigma^4/n - 1$$
$$Cov(\bar{x}, s^2) = 0$$

For $Var(s^2) = 2\sigma^4/n - 1$ we have for $\boldsymbol{X} \sim N(0, \sigma^2 I)$,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\boldsymbol{X}'M\boldsymbol{X} = \frac{1}{n-1}\sigma^2(\boldsymbol{X}'/\sigma)M(\boldsymbol{X}/\sigma)$$

$$\frac{1}{n-1}\sigma^2\boldsymbol{Z}'M\boldsymbol{Z} = \frac{1}{n-1}\sigma^2 q \tag{3.1}$$

where $\boldsymbol{Z} \sim N(0, I)$, $\boldsymbol{Z} = \boldsymbol{X}/\sigma$ and $q \sim \chi^2(n-1)$. Thus

$$E(s^2) = \frac{1}{n-1}\sigma^2 E(q) = \frac{1}{n-1}\sigma^2(n-1) = \sigma^2$$

$$Var(s^2) = \frac{\sigma^4}{(n-1)^2}Var(q) = \frac{\sigma^4}{(n-1)^2}2(n-1) = \frac{2\sigma^4}{n-1}$$

For

$$Cov(\bar{x}, s^2) = 0$$

we have:

$Cov(\bar{x}, s^2) = Cov(\boldsymbol{i}'\boldsymbol{X}, \sigma^2/n{-}1q) = \sigma^2/n{-}1Cov(\boldsymbol{i}'\boldsymbol{X}, q) = \sigma^2/n{-}1Cov(\boldsymbol{i}'\boldsymbol{X}, \boldsymbol{X}'M\boldsymbol{X}) = 0$

Exercise 11 *Show that the $\hat{\sigma}^2_{ML}$ is a biased estimator of σ^2*

Exercise 12 *Show that the variance of $\hat{\sigma}^2_{ML}$ is less than the variance of s^2.*

Exercise 13 *Show that the MSE of $\hat{\sigma}^2_{ML}$ is less than the MSE of s^2.*

3.3 Efficient Estimation of a Multivariate Normal

Consider n independent m-dimensional random variables. Denote by $\boldsymbol{X}_i$ the i-th observation across the m variables, i.e.

$$\boldsymbol{X}_i = (X_{i1}, \ldots, X_{im})'$$

For each observation the pdf can be written as

$$f(\boldsymbol{X}_i) = (2\pi)^{-m/2}|\Sigma|^{-1/2}e^{-1/2(\boldsymbol{X}_i-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X}_i-\boldsymbol{\mu})}$$

Assume that $X_{i1}, \ldots, X_{im}$ are independent with common variance σ^2. Then $|\Sigma|^{-1/2} = \sigma^2|I|$ and thus

$$f(\boldsymbol{X}_i) = (2\pi)^{-m/2}\sigma^2|I|^{-1/2}e^{-1/2\sigma^2(\boldsymbol{X}_i-\boldsymbol{\mu})'(\boldsymbol{X}_i-\boldsymbol{\mu})}$$

Take the joint pdf across $\boldsymbol{X}_i$, i.e.

$$f(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = f(\boldsymbol{X}_1)\ldots f(\boldsymbol{X}_n) = \prod_{i=1}^{n}(2\pi)^{-m/2}\sigma^2|I|^{-1/2}e^{-1/2\sigma^2(\boldsymbol{X}_i-\boldsymbol{\mu})'(\boldsymbol{X}_i-\boldsymbol{\mu})}$$

Then the log-likelihood is

$$L(\boldsymbol{\theta}) = ln\left[\prod_{i=1}^{n}(2\pi)^{-m/2}\sigma^2|I|^{-1/2}e^{-1/2\sigma^2(\boldsymbol{X}_i-\boldsymbol{\mu})'(\boldsymbol{X}_i-\boldsymbol{\mu})}\right] =$$

$$\frac{-nm}{2}ln(2\pi) - \frac{nm}{2}ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\boldsymbol{X}_i-\boldsymbol{\mu})'(\boldsymbol{X}_i-\boldsymbol{\mu})$$

To obtain the ML estimator we require

$$F.O.C.\quad \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\mu}} = \frac{\partial \frac{-1}{2\sigma^2}\sum_{i=1}^{n}(\boldsymbol{X}_i-\boldsymbol{\mu})'(\boldsymbol{X}_i-\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = \frac{-1}{2\sigma^2}\sum_{i=1}^{n}\frac{\partial(\boldsymbol{X}_i-\boldsymbol{\mu})'(\boldsymbol{X}_i-\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} =$$

$$\frac{-1}{2\sigma^2}\sum_{i=1}^{n}-2(\boldsymbol{X}_i-\boldsymbol{\mu}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(\boldsymbol{X}_i-\boldsymbol{\mu}) = 0 \tag{3.2}$$

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \sigma^2} = \frac{-mn}{2\sigma^2} + \frac{1}{\sigma^4}\sum_{i=1}^{n}(\boldsymbol{X}_i-\boldsymbol{\mu})'(\boldsymbol{X}_i-\boldsymbol{\mu}) = 0 \tag{3.3}$$

These conditions imply:

$$\sum_{i=1}^{n}\boldsymbol{X}_i - n\boldsymbol{\mu}) = 0 \Rightarrow \hat{\boldsymbol{\mu}}_{ML} = 1/n\sum_{i=1}^{n}\boldsymbol{X}_i$$

i.e.

$$\hat{\mu}_{1,ML} = 1/n\sum_{i=1}^{n}\boldsymbol{X}_{i1}$$

$$\hat{\mu}_{2,ML} = 1/n \sum_{i=1}^{n} \boldsymbol{X}_{i2}$$

$$\vdots$$

$$\hat{\mu}_{m,ML} = 1/n \sum_{i=1}^{n} \boldsymbol{X}_{im}$$

(3.3) implies:

$$-mn\sigma^2 + \sum_{i=1}^{n} (\boldsymbol{X}_i - \boldsymbol{\mu})'(\boldsymbol{X}_i - \boldsymbol{\mu}) = 0 \Rightarrow \sigma^2 = 1/nm \sum_{i=1}^{n} (\boldsymbol{X}_i - \boldsymbol{\mu})'(\boldsymbol{X}_i - \boldsymbol{\mu})$$

$$1/nm \sum_{i=1}^{n} \left[(X_{i1} - \mu_1)^2 + (X_{i2} - \mu_2)^2 + \ldots + (X_{im} - \mu_m)^2 \right] = 1/nm \sum_{i=1}^{n} \sum_{j=1}^{m} (X_{ij} - \mu_j)^2 =$$

$$1/m \sum_{j=1}^{m} 1/n \sum_{i=1}^{n} (X_{ij} - \mu_j)^2 = 1/m \sum_{j=1}^{m} \hat{\sigma}_j^2$$

The information matrix is given by

$$I(\boldsymbol{\theta}) = -E(\boldsymbol{H}) = E\left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right) = \begin{pmatrix} \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\mu} \partial \boldsymbol{\mu}'} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\mu} \partial \sigma^2} \\ \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \sigma^2 \partial \boldsymbol{\mu}'} & \frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \sigma^4} \end{pmatrix}$$

Given

$$\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\mu} \partial \boldsymbol{\mu}'} = \frac{1}{\sigma^2} \left[\frac{\partial \left(\sum_{i=1}^{n} (\boldsymbol{x}_- \boldsymbol{\mu}) \right)}{\partial \boldsymbol{\mu}'} \right] = \frac{1}{\sigma^2} (-nI)$$

$$\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \sigma^4} = \frac{nm}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^{n} (\boldsymbol{x}_- \boldsymbol{\mu})'(\boldsymbol{x}_- \boldsymbol{\mu})$$

$$\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\mu} \partial \sigma^2} = \frac{1}{\sigma^4} \sum_{i=1}^{n} (\boldsymbol{x}_- \boldsymbol{\mu})$$

$I(\boldsymbol{\theta})$ becomes

$$I(\boldsymbol{\theta}) = -E(\boldsymbol{H}) = \begin{pmatrix} n/\sigma^2 I & 0 \\ 0 & nm/2\sigma^4 \end{pmatrix}$$

Since

$$E\left(\frac{1}{\sigma^4}\sum_{i=1}^{n}(\boldsymbol{x}_{-}\boldsymbol{\mu})\right) = 0$$

and

$$E\left(\frac{1}{\sigma^6}\sum_{i=1}^{n}(\boldsymbol{x}_{-}\boldsymbol{\mu})'(\boldsymbol{x}_{-}\boldsymbol{\mu})\right) = nm\sigma^2$$

Finally, we have

$$Var(\hat{\boldsymbol{\theta}}_{ML}) = \{I(\boldsymbol{\theta})\}^{-1} = \begin{pmatrix} \sigma^2/nI & 0 \\ 0 & 2\sigma^4/nm \end{pmatrix}$$

Chapter 4

Large Sample Distribution Theory and Estimation

If we do not know the underlying probability model of an estimator (or statistic) or we cannot derive its distribution we can derive approximate results as the sample becomes large, ($n \to \infty$). To see how, we need first to pressent some results on convergence of random variables.

4.1 Modes of convergence

4.1.1 Convergence in probability

x_n converges in probability to c, denoted $x_n \xrightarrow{p} c$ or $plim_{n\to\infty} x_n = c$, if for all $\epsilon > 0$,

$$\lim_{n\to\infty} Prob(|x_n - c| > \epsilon) = 0$$

4.1.2 Convergence in mean square

x_n converges in mean square to c, denoted $x_n \xrightarrow{m.s.} c$, if

$$\lim_{n\to\infty} E(x_n - c)^2 = 0$$

4.1.3 Almost Sure Convergence

x_n converges almost surely to c, denoted $x_n \xrightarrow{as} c$ if, for all $\epsilon > 0$,

$$Prob(\lim_{n\to\infty} x_n = c) \leq \epsilon$$

The difference between convergence in probability and convergence almost surely is that convergence in probability deals with the probabilities of events involving individual random variables whereas convergence almost surely deals with probabilities of events involving infinite sequences of random variables Both almost sure convergence and convergence in mean square imply convergence in probability. An estimator is consistent if

$$plim_{n\to\infty}\hat{\theta} = \theta$$

Example 7 *Show that the mean of a random variable from any population is a consistent estimator of the population mean*
We need to show that $plim_{n\to\infty}\bar{x}_n = \mu$. *Convergence in mean square implies convergence in probability and so we show convergence in mean square. This requires*

$$E(\bar{x}_n) \to \mu$$

and

$$Var(\bar{x}_n) \to 0$$

But

$$E(\bar{x}_n) = 1/nE(\sum_{i=1}^{n} x_i) = \mu$$

$$Var(\bar{x}_n) = \sigma^2/n \to 0$$

This is an example of a Law of Large Numbers (LLN)

4.1.4 Rules of probability limits

Slutsky's Theorem

$$plim_{n\to\infty}g(x_n) = g(plim_{n\to\infty}x_n)$$

$g(.)$ is a continuous function which is not a function of n.

- If $plim_{n\to\infty}x_n = c$ and $plim_{n\to\infty}y_n = d$ then

$$plim_{n\to\infty}(x_n + y_n) = c + d$$

$$plim_{n\to\infty}x_n y_n = cd$$

$$plim_{n\to\infty}x_n/y_n = c/d \quad d \neq 0$$

$$plim_{n\to\infty}W_n = \Omega \Rightarrow plim_{n\to\infty}W_n^{-1} = \Omega^{-1}$$

$$plim_{n\to\infty}X_{n\ (k\times k)}Y_{n\ (k\times k)} = AB$$

4.1.5 Convergence in Distribution

Let x_n be a sequence of random variables indexed by sample size and assume that x_n has cumulative density function (cdf) $F_n(x)$. Then if

$$\lim_{n\to\infty} |F_n(x) - F(x)| = 0$$

where $F(x)$ is the cdf of a random variable x we say that x_n converges in distribution to a random variale with distribution $F(x)$. We denote this by

$$x_n \xrightarrow{d} x$$

Rules of limiting distributions

- Let $x_n \xrightarrow{d} x$ and $plim_{n\to\infty} y_n = c$ then

$$x_n y_n \xrightarrow{d} cx$$

$$x_n + y_n \xrightarrow{d} x + c$$

$$x_n/y_n \xrightarrow{d} x/c, \quad c \neq= 0$$

- (Continuous mapping theorem) If $x_n \xrightarrow{d} x$ then

$$g(x_n) \xrightarrow{d} g(x)$$

  where $g(.)$ is continuous.

- If $x_n \xrightarrow{d} x$ and $plim_{n\to\infty} y_n = x$ then $y_n \xrightarrow{d} x$

4.2 Univariate Central Limit Theorem

If x_i has mean μ and finite variance σ^2 then

$$1/\sqrt{n} \sum_{i=1}^{n} x_i - \mu \xrightarrow{d} N(0, \sigma^2)$$

This implies that

$$\sqrt{n}/n \sum_{i=1}^{n} x_i - \mu \xrightarrow{d} N(0, \sigma^2)$$

or

$$\sqrt{n}\left[1/n\sum_{i=1}^{n}x_i-\mu\right]\xrightarrow{d}N(0,\sigma^2)$$

or

$$\sqrt{n}(\bar{x}_n-\mu)\xrightarrow{d}N(0,\sigma^2)$$

This justifies the use of the normal distribution (as a limiting distribution) to test hypotheses about the sample mean. This can be regardless of the form of the population distribution of x.

4.3 Multivariate Central Limit Theorem

Let k random variables $x_1,\dots x_n$ from a multivariate distribution with finite mean vector $\boldsymbol{\mu}=(\mu_1,\dots,\mu_k)'$ and finite covariance matrix Σ then

$$\sqrt{n}(\bar{\boldsymbol{x}}_n-\boldsymbol{\mu})\Rightarrow N(0,\Sigma)$$

where $\bar{\boldsymbol{x}}_n=(\bar{x}_1,\dots\bar{x}_n)'$.

4.4 Limiting normal distribution of a function

4.4.1 Linear case

Theorem 8 *If*

$$\sqrt{n}(x_n-\mu)\Rightarrow N(0,\sigma^2)$$

and $y=cx$ is a continuous function of x, then

$$\sqrt{n}(y_n-c\mu)\Rightarrow N(0,c^2\sigma^2)$$

Proof 8 *First notice that by the continuous mapping theorem*

$$y_n=g(x_n)\xrightarrow{d}y=g(x)$$

Since x is $N(0,\sigma^2)$ by applying the change in variables technique we can show

$$y=g(x)\sim N(0,c^2\sigma^2)$$

4.4.2 Nonlinear case

Theorem 9 *If*

$$\sqrt{n}(x_n - \mu) \Rightarrow N(0, \sigma^2)$$

and $y = g(x)$ is a continuous nonlinear function of x, then

$$\sqrt{n}(y_n - g(\mu)) \Rightarrow N(0, g'(\mu)^2\sigma^2)$$

Proof 9 *Approximate $g(x)$ around the mean value of x as*

$$g(x) \approx g(\mu) + g(\mu)'(x - \mu) = g(\mu) - g'(\mu)\mu + g'(\mu)x$$

So

$$\sqrt{n}(g(x) - g(\mu)) \approx g'(\mu)\sqrt{n}(x - \mu) \sim N(0, g'(\mu)^2\sigma^2)$$

4.5 Limiting distribution of a a set of functions

If $\boldsymbol{x}_n$ satisfies

$$\sqrt{n}(\boldsymbol{x}_n - \boldsymbol{\mu}) \xrightarrow{d} N(0, \Sigma)$$

and if $\boldsymbol{c}(\boldsymbol{x}_n)$ is a set of J continuous functions of $\boldsymbol{x}_n$ then

$$\sqrt{n}(\boldsymbol{c}(\boldsymbol{x}_n) - \boldsymbol{c}(\boldsymbol{\mu})) \xrightarrow{d} N(0, Q(\boldsymbol{\mu})\Sigma Q(\boldsymbol{\mu})')$$

where $Q(\boldsymbol{\mu})$ is the Jacobian matrix of c.

4.6 Asymptotic Distribution

The asymptotic distribution is constructed from the known limiting distribution of a random variable and it is the distribution that is used to approximate the finite sample distribution.

Example 8 *If*

$$\sqrt{n}(\bar{x}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

then $\bar{x}_n$ is asymptotically distributed as

$$\bar{x}_n \xrightarrow{d} N(\mu, \sigma^2/n)$$

This result means that whatever the finite sample disrtibution of $\bar{x}_n$ the normal distribution provides an approximation to it in finite samples. The proof of this is easy to establish

$$\sqrt{n}(\bar{x}_n - \mu) \xrightarrow{d} N(0, \sigma^2) \Rightarrow$$

$$\sqrt{n}\bar{x}_n - \sqrt{n}\mu \xrightarrow{d} N(0, \sigma^2) \Rightarrow$$

$$\sqrt{n}\bar{x}_n \xrightarrow{d} N(\sqrt{n}\mu, \sigma^2) \Rightarrow$$

$$\bar{x}_n \xrightarrow{d} N(\mu, \sigma^2/n)$$

4.7 Application of Large Sample Distribution Theory

4.7.1 The Asymptotic Distribution of the ML Estimator

Consider a pdf, $f(x; \boldsymbol{\theta})$ The log-likelihood is

$$\ln L = \sum_{i=1}^{n} \ln f(x_i; \boldsymbol{\theta})$$

The ML estimator is defined as the estimator for which

$$g(\boldsymbol{\theta}) = \frac{\partial \ln L}{\partial \boldsymbol{\theta}} = \frac{\partial \left(\sum_{i=1}^{n} \ln f(x_i; \boldsymbol{\theta})\right)}{\partial \boldsymbol{\theta}} =$$

$$\sum_{i=1}^{n} \frac{\partial \left(\ln f(x_i; \boldsymbol{\theta})\right)}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} g_i = 0$$

where $g_i = \frac{\partial(\ln f(x_i;\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}$ is the gradient (or score) vector for each observation. Also

$$H = \frac{\partial^2 \ln L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \frac{\partial}{\partial \boldsymbol{\theta}} \left[\sum_{i=1}^{n} \frac{\partial \left(\ln f(x_i; \boldsymbol{\theta})\right)}{\partial \boldsymbol{\theta}} \right] =$$

$$\sum_{i=1}^{n} \frac{\partial^2 \left(\ln f(x_i; \boldsymbol{\theta})\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \sum_{i=1}^{n} H_i$$

should be non-positive. To derive the limiting distribution of the ML estimator ($\hat{\boldsymbol{\theta}}_{ML}$) take a first order Taylor expansion of the gradient around the vector of the true parameter $\boldsymbol{\theta}$, i.e.

$$g(\hat{\boldsymbol{\theta}}_{ML}) \approx g(\boldsymbol{\theta}) + \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}) = g(\boldsymbol{\theta}) + H(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta})$$

Solving out the above equation with respect to $(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta})$ gives

$$(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}) = H(\boldsymbol{\theta})^{-1}\left(g(\hat{\boldsymbol{\theta}}_{ML}) - g(\boldsymbol{\theta})\right) =$$

$$-H(\boldsymbol{\theta})^{-1}g(\boldsymbol{\theta})$$

since $g(\hat{\boldsymbol{\theta}}_{ML}) = 0$. Scaling by $\sqrt{n}$ gives

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}) = -H(\boldsymbol{\theta})^{-1}\sqrt{n}g(\boldsymbol{\theta}) = -H(\boldsymbol{\theta})^{-1}\sqrt{n}\sum_{i=1}^{n} g_i(\boldsymbol{\theta}) =$$

$$-\left[\frac{H(\boldsymbol{\theta})}{n}\right]^{-1} 1/\sqrt{n}\sum_{i=1}^{n} g_i(\boldsymbol{\theta})$$

Notice that $g_i(\boldsymbol{\theta})$ is an i.i.d. random variable with mean 0 and variance which can be obtained as

$$Var[g_i(\boldsymbol{\theta})] = 1/nVar[g(\boldsymbol{\theta})] = 1/nE(g(\boldsymbol{\theta})g(\boldsymbol{\theta})') = 1/nE(\frac{\partial \ln L}{\partial \boldsymbol{\theta}}\frac{\partial \ln L'}{\partial \boldsymbol{\theta}}) =$$

$$-1/nE(\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}) = -1/nE(H(\boldsymbol{\theta}))$$

Then by a straightforward application of the central limit theorem we get

$$1/\sqrt{n}\sum_{i=1}^{n} g_i(\boldsymbol{\theta}) \xrightarrow{d} N(0, -1/nE(H(\boldsymbol{\theta})))$$

Using this result and noticing that

$$plim\,[1/nH(\boldsymbol{\theta})] = 1/nE(H(\boldsymbol{\theta}))$$

which is a constant matrix the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta})$ can be derived by the limiting distribution of

$$[1/nE(H(\boldsymbol{\theta}))]^{-1} 1/\sqrt{n}\sum_{i=1}^{n} g_i(\boldsymbol{\theta})$$

which is

$$N\left(0,\left[1/nE(H(\boldsymbol{\theta}))\right]^{-1}\left[-1/nE(H(\boldsymbol{\theta}))\right]\left[1/nE(H(\boldsymbol{\theta}))\right]^{-1}\right)\sim N\left(0,-n\left[E(H(\boldsymbol{\theta}))\right]^{-1}\right)$$

or

$$N\left(0,n\left[I(\boldsymbol{\theta})\right]^{-1}\right)$$

That is

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{ML}-\boldsymbol{\theta})\xrightarrow{d}N\left(0,n\left[I(\boldsymbol{\theta})\right]^{-1}\right)$$

which implies that the asymptotic distribution of $\hat{\boldsymbol{\theta}}_{ML}$ is

$$\sqrt{n}\hat{\boldsymbol{\theta}}_{ML}-\sqrt{n}\boldsymbol{\theta}\xrightarrow{d}N\left(0,n\left[I(\boldsymbol{\theta})\right]^{-1}\right)$$

or

$$\sqrt{n}\hat{\boldsymbol{\theta}}_{ML}\xrightarrow{d}N\left(\sqrt{n}\boldsymbol{\theta},n\left[I(\boldsymbol{\theta})\right]^{-1}\right)$$

or

$$\hat{\boldsymbol{\theta}}_{ML}\xrightarrow{d}N\left(\boldsymbol{\theta},\left[I(\boldsymbol{\theta})\right]^{-1}\right)$$

which implies that $\hat{\boldsymbol{\theta}}_{ML}$ satisifes the Cramer-Rao lower variance bound.

4.7.2 Asymptotically Equivalent Test Procedures

The Likelihood ratio, Wald and Lagrange multiplier test

Likelihood Ratio test

This test checks whether the likelihood is reduced siginificantly by imposing the restreiction $C(\boldsymbol{\theta})=0$ The unrestricted and restricted likelihoods are denoted by L_U and L_R respectively. Define

$$\lambda=\frac{L_R}{L_U}$$

Then

$$-2ln\lambda=-2(\ln L_R-\ln L_U)\sim\chi_q^2$$

where q is the number of restrictions.

Wald test

The Wald tests checks whether $C(\hat{\boldsymbol{\theta}}_{ML})$ is significantly different from zero. The test statistic is

$$W = C(\hat{\boldsymbol{\theta}}_{ML}) \left[Var(C(\hat{\boldsymbol{\theta}}_{ML})) \right]^{-1} C(\hat{\boldsymbol{\theta}}_{ML})' \sim x_q^2$$

Example 9 *Consider a linear set of restrictions*

$$R_{q \times n} \boldsymbol{\theta} - \boldsymbol{r} = 0$$

Then

$$W = (R\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{r})' \left[Var(R\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{r}) \right]^{-1} (R\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{r}) =$$

$$(R\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{r})' \left[RVar(\hat{\boldsymbol{\theta}}_{ML}) R' \right]^{-1} (R\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{r}) =$$

$$(R\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{r})' \left[RI(\hat{\boldsymbol{\theta}}_{ML})^{-1} R' \right]^{-1} (R\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{r})$$

Exercise 14 *Derive the Wald statistic for the hypothesis that the mean of two variables is distributed with the same variance σ^2 and independently are the same*

Lagrange Multiplier (LM) Test

The LM test tests whether the restricted estimator is close to the estimator maximising the log-likelihood. This translates into the slope of the log-likelihood at the restricted estimator being close to zero. So we test whether

$$\frac{\partial \ln L\hat{\boldsymbol{\theta}}_R}{\partial \boldsymbol{\theta}} = \hat{g}_R = 0$$

The test statistic is

$$LM = \frac{\partial \ln L\hat{\boldsymbol{\theta}}_R}{\partial \boldsymbol{\theta}}' \left[I(\hat{\boldsymbol{\theta}}_R)]^{-1} \right] \frac{\partial \ln L\hat{\boldsymbol{\theta}}_R}{\partial \boldsymbol{\theta}} =$$

$$\hat{g'}_R \left[I(\hat{\boldsymbol{\theta}}_R)]^{-1} \right] \hat{g}_R \sim \chi_q^2$$

Exercise 15 *Find the variance of the LS estimator for the model*

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{u}$$

where $E(u_i^2) = \sigma^2$, $i = 1, \ldots, n$.

Exercise 16 *Find the distribution of the LS estimator for the model*

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{u}$$

where $E(u_i^2) = \sigma^2$, $i = 1, \ldots, n$.

Chapter 5

Solutions to Exercises

Solution 1

$$(A \otimes B)(A^{-1} \otimes B^{-1}) = AA^{-1} \otimes BB^{-1} = I_{n\times n} \otimes I_{m\times m} = I_{mn\times mn}$$

Solution 2

$$A'A = \begin{pmatrix} a_{11}^2 + a_{12}^2 + \ldots + a_{1n}^2 & \ldots & \ldots \\ \ldots & a_{21}^2 + a_{22}^2 + \ldots + a_{2n}^2 & \ldots \\ \ldots & \ldots & \ldots \\ \ldots & \ldots & a_{n1}^2 + a_{n2}^2 + \ldots + a_{nn}^2 \end{pmatrix}$$

So

$$tr(A'A) = a_{11}^2+a_{12}^2+\ldots+a_{1n}^2+a_{21}^2+a_{22}^2+\ldots+a_{2n}^2+\ldots+a_{n1}^2+a_{n2}^2+\ldots+a_{nn}^2 = \sum_{i=1}^{n}\sum_{j=1} a_{ij}^2$$

Solution 3

$$X'MX = X'(I - 1/n\boldsymbol{ii}')X = X'X - n(1/nX'\boldsymbol{i})(1/n\boldsymbol{i}'X)$$

But

$$1/nX'\boldsymbol{i} = \bar{\boldsymbol{x}}$$

Solution 4

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \ldots & a_{1k}B \\ a_{21}B & a_{22}B & \ldots & a_{2k}B \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1}B & a_{n2}B & \ldots & a_{nk}B \end{pmatrix}$$

So

$$tr(A\otimes B) = a_{11}b_{11}+a_{11}b_{22}+\ldots+a_{11}b_{mm}+a_2b_{11}+a_{22}b_{22}+\ldots+a_{22}b_{mm}+\ldots+$$

$$a_{nn}b_{11} + a_{nn}b_{22} + \ldots + a_{nn}b_{mm} =$$

$$(a_{11} + a_{22} + \ldots + a_{nn})(b_{11} + b_{22} + \ldots + b_{nn}) = tr(A)tr(B)$$

Solution 5

$$X'X = \begin{pmatrix} 4 & 0 \\ 0 & 54 \end{pmatrix}$$

$$(X'X)^{-1} = \begin{pmatrix} 1/4 & 0 \\ 0 & 1/54 \end{pmatrix}$$

$$X(X'X^{-1}X = 1/108 \begin{pmatrix} 59 & 11 & 51 & -13 \\ 11 & 35 & 15 & 47 \\ 51 & 15 & 45 & -3 \\ -13 & 47 & -3 & 77 \end{pmatrix} = P$$

For X

$$P = X(X'X)^{-1}X', \quad M = (I - P)$$

For XQ

$$P = XQ(Q'X'XQ)^{-1}Q'X' = XQ[Q^{-1}(X'X)^{-1}Q'^{-1}]Q'X' = X'(X'X)^{-1}X'$$

Since P and M are idempotent ($PP = X'(X'X)^{-1}X'X'(X'X)^{-1}X' = X'(X'X)^{-1}X'$, $MM = (I - P)(I - P) = I - P - P + P = I - P$), their characteristic roots are either 0 or 1. So the trace is equal to the number of unit eigenvalues. Both M and P have 2 unit and 2 zero eigenvalues.

Solution 6 *The solution is*

$$2[\boldsymbol{x}'A\boldsymbol{x}/\boldsymbol{x}'B\boldsymbol{x}][\boldsymbol{x}'A\boldsymbol{g}/\boldsymbol{x}'A\boldsymbol{x} - \boldsymbol{x}'B\boldsymbol{g}/\boldsymbol{x}'B\boldsymbol{x}]$$

Solution 7 *For a 2×2 matrix*

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

So

$$A^{-1} = \frac{1}{n\sum_{i=1} x_i^2 - (\sum_{i=1}^n x_i)^2} \begin{pmatrix} \sum_{i=1} x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}$$

So

$$F_2 = \frac{\boldsymbol{i}'\boldsymbol{i}}{\boldsymbol{i}'\boldsymbol{i}\boldsymbol{x}'\boldsymbol{x} - \boldsymbol{x}'\boldsymbol{i}\boldsymbol{i}'\boldsymbol{x}} = [\boldsymbol{x}'\boldsymbol{x} - \boldsymbol{x}'\boldsymbol{i}\boldsymbol{i}'\boldsymbol{x}/\boldsymbol{i}'\boldsymbol{i}]^{-1}$$
$$= [\boldsymbol{x}'(I - \boldsymbol{i}\boldsymbol{i}'/n)\boldsymbol{x}]^{-1} = [\boldsymbol{x}'M\boldsymbol{x}]^{-1}$$

Now note that $\boldsymbol{x}'M\boldsymbol{x} = \boldsymbol{x}'MM\boldsymbol{x}$ *But* $\boldsymbol{x}'M = (x_1 - \bar{x}, \ldots, x_n - \bar{x})$ *So*

$$[\boldsymbol{x}'M\boldsymbol{x}]^{-1} = \frac{1}{\sum_{i=1}(x_i - \bar{x})^2}$$

Solution 8

$$\boldsymbol{u}'\boldsymbol{u} = (\boldsymbol{y} - \boldsymbol{b}'X)'(\boldsymbol{y} - \boldsymbol{b}'X)$$

Taking the derivative w.r.t $\boldsymbol{b}$ *gives the first order condition*

$$2(X'X)\boldsymbol{b} - 2X'\boldsymbol{y} = 0 \Rightarrow \boldsymbol{b} = (X'X)^{-1}X\boldsymbol{y}$$

Solution 9

$$\sigma_{XY} = Cov[X, Y] = E[(X - \mu_X)(Y - \mu_Y)] =$$

$$E[XY] - \mu_X E(Y) - \mu_Y E(Y) + \mu_X\mu_Y =$$
$$E[XY] - \mu_X\mu_Y - \mu_X\mu_Y + \mu_X\mu_Y =$$
$$E[XY] - \mu_X\mu_Y = \sigma_{XY}$$

To show that $\sigma_{XY} = 0$*, we need to to prove that*

$$E[XY] = E(X)E(Y) = \mu_X\mu_Y$$

Then

$$\sigma_{XY} = \mu_X\mu_Y - \mu_X\mu_Y = 0$$
$$E[XY] = \int_x\int_y xyf(x, y)dydx =$$
$$\int_x\int_y xyf_X(x)f_Y(y)dydx =$$

$$\int_x x \int_y y f_Y(y) dy f_X(x) dx =$$

$$\int_x x E(Y) f_X(x) dx =$$

$$E(Y) \int_x x f_X(x) dx =$$

$$E(Y)E(X) = \mu_X \mu_Y$$

The above result can be generalised *for any function of* X *and* Y, $g_1(X)$, $g_2(Y)$ *i.e.*

$$E[g_1(X)g_2(Y)] = E[g_1(X)]E[g_2(Y)]$$

if X *and* Y *are independent random variables.*

Solution 10 *From the definition of conditional variance* $Var(Y|X)$ *we have:*

$$Var(Y|X) = E(Y^2|X) - (E(Y|X))^2$$

Take the $E_X(.)$ *of the above expression*

$$E_X(Var(Y|X)) = E_X(E(Y^2|X)) - E_X((E(Y|X))^2) =$$

$$\int_x E(Y^2|X)) f_X(x) dx - E_X((E(Y|X))^2) =$$

$$\int_x \left(\int_Y y^2 f(y|x) dy \right) f_X(x) dx - E_X((E(Y|X))^2) =$$

$$\int_x \int_Y y^2 f(y|x) f_X(x) dy dx - E_X((E(Y|X))^2) =$$

$$\int_x \int_Y y^2 f(y,x) dy dx - E_X((E(Y|X))^2) =$$

$$E(Y^2) - E_X((E(Y|X))^2) =$$

$$E(Y^2) - (E(Y))^2 + (E(Y))^2 - E_X((E(Y|X))^2) =$$

$$Var(Y) - \left(E_X((E(Y|X))^2) - (E(Y))^2 \right) =$$

$$Var(Y) - \left(E_X((E(Y|X))^2) - (E_X(E(Y|X)))^2 \right) =$$

$$Var(Y) - Var_X(E(Y|X))$$

Solution 11 *From (3.1)*

$$s^2 \sim \frac{1}{n-1}\sigma^2 q$$

where $q \sim \chi^2(n-1)$. *But* $E(q) = n-1$ *from the properties of a* χ^2 *variable and therefore*

$$E(s^2) = \sigma^2$$

But

$$\hat{\sigma}^2_{ML} = \frac{n-1}{n}s^2$$

So

$$E(\hat{\sigma}^2_{ML}) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

Solution 12

$$\hat{\sigma}^2_{ML} = \frac{n-1}{n}s^2$$

Therefore

$$Var(\hat{\sigma}^2_{ML}) = \left(\frac{n-1}{n}\right)^2 Var(s^2) < Var(s^2)$$

Solution 13

$$MSE(\hat{\sigma}^2_{ML}) = (Bias(\hat{\sigma}^2_{ML}))^2 + Var(\hat{\sigma}^2_{ML}) = \sigma^4/n^2 + \left(\frac{n-1}{n}\right)^2 \frac{2\sigma^4}{n-1} = \frac{(2n-1)\sigma^4}{n^2}$$

$$MSE(\hat{\sigma}^2_{ML}) - MSE(s^2) = \frac{(2n-1)\sigma^4}{n^2} - \frac{2\sigma^4}{n-1} = \frac{(2n-1)(n-1)\sigma^4}{n^2(n-1)} - \frac{2n^2\sigma^4}{n^2(n-1)} =$$

$$-\frac{(3n-1)\sigma^4}{n^2(n-1)} < 0$$

Solution 14 *The restrictions that we would like to test are*

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix} - \mu \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$Var(\hat{\boldsymbol{\mu}}) = I(\hat{\boldsymbol{\mu}})^{-1} = \left[\sigma^2/n \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right]$$

Thus

$$W = [(\hat{\mu}_1, \hat{\mu}_2) - \mu(1,1)] \left[\sigma^2/n \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right]^{-1} \left[\begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix} - \mu \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right] =$$

$$n/\sigma^2(\hat{\mu}_1-\mu,\hat{\mu}_2-\mu)\begin{pmatrix}\hat{\mu}_1-\mu\\ \hat{\mu}_2-\mu\end{pmatrix}=n/\sigma^2\left[(\hat{\mu}_1-\mu)^2,(\hat{\mu}_2-\mu)^2\right]\sim\chi^2_1$$

Solution 15

$$\hat{\boldsymbol{\beta}}=(X'X)^{-1}X'\boldsymbol{y}\Rightarrow\hat{\boldsymbol{\beta}}=(X'X)^{-1}X'(X\hat{\boldsymbol{\beta}}+\boldsymbol{u})=$$

$$(X'X)^{-1}X'X\boldsymbol{\beta}+(X'X)^{-1}X'\boldsymbol{u}=\boldsymbol{\beta}+(X'X)^{-1}X'\boldsymbol{u}$$

The variance of $\hat{\boldsymbol{\beta}}$ *is*

$$V(\hat{\boldsymbol{\beta}})=E\left[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right]=$$

$$E\left[(X'X)^{-1}X'\boldsymbol{u}((X'X)^{-1}X'\boldsymbol{u})'\right]=$$
$$E\left[(X'X)^{-1}X'\boldsymbol{u}\boldsymbol{u}'X(X'X)^{-1}\right]=$$
$$(X'X)^{-1}X'E(\boldsymbol{u}\boldsymbol{u}')X(X'X)^{-1}=$$
$$(X'X)^{-1}X'\sigma^2IX(X'X)^{-1}=$$
$$\sigma^2(X'X)^{-1}X'X(X'X)^{-1}=$$
$$\sigma^2(X'X)^{-1}$$

This estimator obtains the lower Cramer-Rao bound

Solution 16

$$\hat{\boldsymbol{\beta}}=(X'X)^{-1}X'\boldsymbol{y}\Rightarrow\hat{\boldsymbol{\beta}}=(X'X)^{-1}X'(X\hat{\boldsymbol{\beta}}+\boldsymbol{u})=$$

$$(X'X)^{-1}X'X\boldsymbol{\beta}+(X'X)^{-1}X'\boldsymbol{u}=\boldsymbol{\beta}+(X'X)^{-1}X'\boldsymbol{u}\Rightarrow$$
$$\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}=(X'X)^{-1}X'\boldsymbol{u}=C\boldsymbol{u}$$

Then if

$$\boldsymbol{u}\sim N(0,\sigma^2I)$$
$$C\boldsymbol{u}\sim N(0,C\sigma^2IC')$$

or

$$C\boldsymbol{u}\sim N(0,(X'X)^{-1}X'\sigma^2I((X'X)^{-1}X)')$$

or

$$C\boldsymbol{u}\sim N(0,\sigma^2(X'X)^{-1})$$

Then

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

Appendix

Proof of Theorem 4
For simplicity denote the likelihood by L. Then

$$\int \dots \int L dx_1 \; dx_2 \dots dx_n = 1$$

Differentiate both sides w.r.t. θ and interchange differentiation and integration to give

$$\int \dots \int \frac{\partial L}{\partial \theta} \; dx_1 \; dx_2 \dots dx_n = 0$$

or

$$E(\frac{\partial \ln L}{\partial \theta}) = \int \dots \int \left(\frac{1}{L} \frac{\partial L}{\partial \theta} \right) L \; dx_1 \; dx_2 \dots dx_n = 0$$

Differentiate again to get

$$\int \dots \int \left[\left(\frac{1}{L} \frac{\partial L}{\partial \theta} \right) \frac{\partial L}{\partial \theta} + L \frac{\partial}{\partial \theta} \left(\frac{1}{L} \frac{\partial L}{\partial \theta} \right) \right] \; dx_1 \; dx_2 \dots dx_n = 0$$

which is

$$\int \dots \int \left[\left(\frac{1}{L} \frac{\partial L}{\partial \theta} \right)^2 + \frac{\partial^2 \ln L}{\partial \theta^2} \right] L \; dx_1 \; dx_2 \dots dx_n = 0$$

Rearranging gives

$$E \left(\frac{\partial \ln L}{\partial \theta} \right)^2 = -E \left(\frac{\partial^2 \ln L}{\partial \theta^2} \right) = I(\theta)$$

If $\hat{\theta}$ is unbiased then

$$\int \dots \int \hat{\theta} L \; dx_1 \; \dots dx_n = \theta$$

or

$$\int \dots \int \hat{\theta} \frac{\partial \ln L}{\partial \theta} L \; dx_1 \; \dots dx_n = 1$$

But

$$E\left(\frac{\partial \ln L}{\partial \theta}\right) = 0$$

So

$$E\left(\theta \frac{\partial \ln L}{\partial \theta}\right) = \theta E\left(\frac{\partial \ln L}{\partial \theta}\right) = 0$$

$$\int \dots \int (\hat{\theta} - \theta) \frac{\partial \ln L}{\partial \theta} L \; dx_1 \; \dots dx_n = 1$$

By the Cauchy-Schwartz inequality we get that

$$1 \leq \left[\int \dots \int (\hat{\theta} - \theta)^2 L \; dx_1 \; \dots dx_n\right] \left[\int \dots \int \left[\frac{\partial \ln L}{\partial \theta}\right]^2 L \; dx_1 \; \dots dx_n\right]$$

But the first term

$$\int \dots \int (\hat{\theta} - \theta)^2 L \; dx_1 \; \dots dx_n = E\left((\hat{\theta} - \theta)^2\right)$$

The second is the information matrix.
So

$$\left[\int \dots \int (\hat{\theta} - \theta)^2 L \; dx_1 \; \dots dx_n\right] > \left[\int \dots \int \left[\frac{\partial \ln L}{\partial \theta}\right]^2 L \; dx_1 \; \dots dx_n\right]^{-1}$$