Chapter 1

What Is Econometrics

1.1 Data

1.1.1 Accounting

Definition 1.1 *Accounting data* is routinely recorded data (that is, records of transactions) as part of market activities. Most of the data used in econmetrics is accounting data. □

Acounting data has many shortcomings, since it is not collected specifically for the purposes of the econometrician. An econometrician would prefer data collected in experiments or gathered for her.

1.1.2 Nonexperimental

The econometrician does not have any control over nonexperimental data. This causes various problems. Econometrics is used to deal with or solve these problems.

1.1.3 Data Types

Definition 1.2 *Time-series data* are data that represent repeated observations of some variable in subseqent time periods. A time-series variable is often subscripted with the letter t. □

Definition 1.3 *Cross-sectional data* are data that represent a set of observations of some variable at one specific instant over several agents. A cross-sectional variable is often subscripted with the letter i. □

Definition 1.4 *Time-series cross-sectional data* are data that are both time-series and cross-sectional. □

An special case of time-series cross-sectional data is *panel data*. Panel data are observations of the same set of agents over time.

1.1.4 Empirical Regularities

We need to look at the data to detect regularities. Often, we use stylized facts, but this can lead to over-simplifications.

1.2 Models

Models are simplifications of the real world. The data we use in our model is what motivates theory.

1.2.1 Economic Theory

By choosing assumptions, the

1.2.2 Postulated Relationships

Models can also be developed by postulating relationships among the variables.

1.2.3 Equations

Economic models are usually stated in terms of one or more equations.

1.2.4 Error Component

Because none of our models are exactly correct, we include an error component into our equations, usually denoted u_i. In econometrics, we usually assume that the error component is stochastic (that is, random).

It is important to note that the error component cannot be modeled by economic theory. We impose assumptions on u_i, and as econometricians, focus on u_i.

1.3 Statistics

1.3.1 Stochastic Component

Some stuff here.

1.3.2 Statistical Analysis

Some stuff here.

1.3.3 Tasks

There are three main tasks of econometrics:

1. estimating parameters;
2. hypothesis testing;
3. forecasting.

Forecasting is perhaps the main reason for econometrics.

1.3.4 Econometrics

Since our data is, in general, nonexperimental, econometrics makes use of economic theory to adjust for the lack of proper data.

1.4 Interaction

1.4.1 Scientific Method

Econometrics uses the scientific method, but the data are nonexperimental. In some sense, this is similar to astronomers, who gather data, but cannot conduct experiments (for example, astronomers predict the existence of black holes, but have never made one in a lab).

1.4.2 Role Of Econometrics

The three components of econometrics are:

1. theory;
2. statistics;
3. data.

These components are interdependent, and each helps shape the others.

1.4.3 Ocam's Razor

Often in econometrics, we are faced with the problem of choosing one model over an alternative. The simplest model that fits the data is often the best choice.

Chapter 2

Some Useful Distributions

2.1 Introduction

2.1.1 Inferences

Statistical Statements

As statisticians, we are often called upon to answer questions or make statements concerning certain random variables. For example: is a coin fair (i.e. is the probability of heads = 0.5) or what is the expected value of GNP for the quarter.

Population Distribution

Typically, answering such questions requires knowledge of the distribution of the random variable. Unfortunately, we usually do not know this distribution (although we may have a strong idea, as in the case of the coin).

Experimentation

In order to gain knowledge of the distribution, we draw several realizations of the random variable. The notion is that the observations in this *sample* contain information concerning the population distribution.

Inference

Definition 2.1 The process by which we make statements concerning the population distribution based on the sample observations is called *inference*. □

Example 2.1 We decide whether a coin is fair by tossing it several times and observing whether it seems to be heads about half the time. □

2.1.2 Random Samples

Suppose we draw n observations of a random variable, denoted $\{x_1, x_2, ..., x_n\}$ and each x_i is independent and has the same (marginal) distribution, then $\{x_1, x_2, ..., x_n\}$ constitute a *simple random sample*.

Example 2.2 We toss a coin three times. Supposedly, the outcomes are independent. If x_i counts the number of heads for toss i, then we have a simple random sample. □

Note that not all samples are simple random.

Example 2.3 We are interested in the income level for the population in general. The n observations available in this class are not indentical since the higher income individuals will tend to be more variable. □

Example 2.4 Consider the aggregate consumption level. The n observations available in this set are not independent since a high consumption level in one period is usually followed by a high level in the next. □

2.1.3 Sample Statistics

Definition 2.2 Any function of the observations in the sample which is the basis for inference is called a *sample statistic*. □

Example 2.5 In the coin tossing experiment, let S count the total number of heads and $P = \frac{S}{3}$ count the sample proportion of heads. Both S and P are sample statistics. □

2.1.4 Sample Distributions

A sample statistic is a random variable — its value will vary from one experiment to another. As a random variable, it is subject to a distribution.

Definition 2.3 The distribution of the sample statistic is the *sample distribution* of the statistic. □

Example 2.6 The statistic S introduced above has a multinomial sample distribution. Specifically Pr(S = 0) = 1/8, Pr(S = 1) = 3/8, Pr(S = 2) = 3/8, and Pr(S = 3) = 1/8. □

2.2 Normality And The Sample Mean

2.2.1 Sample Sum

Consider the simple random sample $\{x_1, x_2, ..., x_n\}$, where x measures the height of an adult female. We will assume that $\mathrm{E}\, x_i = \mu$ and $\mathrm{Var}(x_i) = \sigma^2$, for all $i = 1, 2, ..., n$

Let $S = x_1 + x_2 + \cdots + x_n$ denote the sample sum. Now,

$$\mathrm{E}\, S = \mathrm{E}(\, x_1 + x_2 + \cdots + x_n \,) = \mathrm{E}\, x_1 + \mathrm{E}\, x_2 + \cdots + \mathrm{E}\, x_n = n\mu \tag{2.1}$$

Also,

$$\begin{aligned}
\mathrm{Var}(S) &= \mathrm{E}(\, S - \mathrm{E}\, S\,)^2 \\
&= \mathrm{E}(\, S - n\mu\,)^2 \\
&= \mathrm{E}(\, x_1 + x_2 + \cdots + x_n - n\mu\,)^2 \\
&= \mathrm{E}\left[\sum_{i=1}^{n}(\, x_i - \mu\,)\right]^2 \\
&= \mathrm{E}[(\, x_1 - \mu\,)^2 + (\, x_1 - \mu\,)(\, x_2 - \mu\,) + \cdots + \\
&\quad (\, x_2 - \mu\,)(\, x_1 - \mu\,) + (\, x_2 - \mu\,)^2 + \cdots + (\, x_n - \mu\,)^2] \\
&= n\sigma^2. \qquad (2.2)
\end{aligned}$$

Note that $\mathrm{E}[(\, x_i - \mu\,)(\, x_j - \mu\,)] = 0$ by independence.

2.2.2 Sample Mean

Let $\overline{x} = \frac{S}{n}$ denote the *sample mean* or average.

2.2.3 Moments Of The Sample Mean

The mean of the sample mean is

$$\mathrm{E}\, \overline{x} = \mathrm{E}\, \frac{S}{n} = \frac{1}{n}\, \mathrm{E}\, S = \frac{1}{n} n\mu = \mu. \tag{2.3}$$

The variance of the sample mean is

$$\begin{aligned}
\mathrm{Var}(\overline{x}) &= \mathrm{E}(\overline{x} - \mu)^2 \\
&= \mathrm{E}(\, \frac{S}{n} - \mu\,)^2 \\
&= \mathrm{E}[\, \frac{1}{n}\, (\, S - n\mu\,)\,]^2 \\
&= \frac{1}{n^2}\, \mathrm{E}(\, S - n\mu\,)^2
\end{aligned}$$

$$\begin{aligned} &= \frac{1}{n^2} n\sigma^2 \\ &= \frac{\sigma^2}{n}. \end{aligned} \tag{2.4}$$

2.2.4 Sampling Distribution

We have been able to establish the mean and variance of the sample mean. However, in order to know its complete distribution precisely, we must know the probability density function (pdf) of the random variable x.

2.3 The Normal Distribution

2.3.1 Density Function

Definition 2.4 A continuous random variable x_i with the density function

$$f\,(\,x_i\,) = \frac{1}{\sqrt{2\pi\sigma^2}}\,e^{-\frac{1}{2\sigma^2}\,(\,x_i - \mu\,)^2} \tag{2.5}$$

follows the normal distribution, where μ and σ^2 are the mean and variance of x_i, respectively.□

Since the distribution is characterized by the two parameters μ and σ^2, we denote a normal random variable by $x_i \sim \mathrm{N}\,(\,\mu, \sigma^2\,)$.

The normal density function is the familiar "bell-shaped" curve, as is shown in Figure 2.1 for $\mu = 0$ and $\sigma^2 = 1$. It is symmetric about the mean μ. Approximately $\frac{2}{3}$ of the probability mass lies within $\pm\sigma$ of μ and about .95 lies within $\pm 2\sigma$. There are numerous examples of random variables that have this shape. Many economic variables are assumed to be normally distributed.

2.3.2 Linear Transformation

Consider the transformed random variable

$$Y_i = a + bx_i$$

We know that

$$\mu_Y = \mathrm{E}\,Y_i = a + b\mu_x$$

and

$$\sigma_Y^2 = \mathrm{E}(Y_i - \mu_Y)^2 = b^2\sigma_x^2$$

If x_i is normally distributed, then Y_i is normally distributed as well. That is,

$$Y_i \sim \mathrm{N}\,(\,\mu_Y, \sigma_Y^2\,)$$

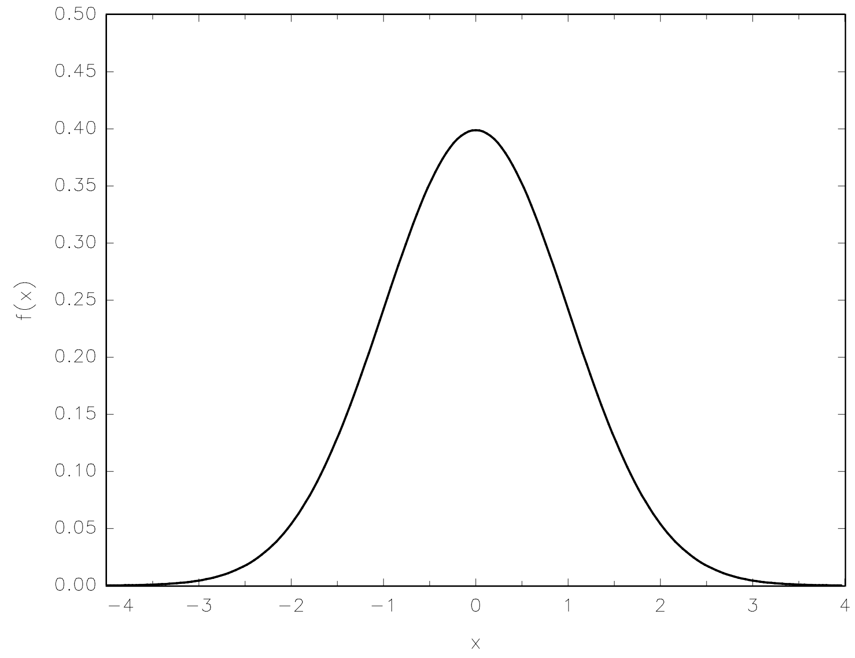

Figure 2.1: The Standard Normal Distribution

Moreover, if $x_i \sim \mathrm{N}\,(\,\mu_x, \sigma_x^2\,)$ and $z_i \sim \mathrm{N}\,(\,\mu_z, \sigma_z^2\,)$ are independent, then

$$Y_i = a + bx_i + cz_i \sim \mathrm{N}\,(\,a + b\mu_x + c\mu_z, b^2\sigma_x^2 + c^2\sigma_z^2\,)$$

These results will be formally demonstrated in a more general setting in the next chapter.

2.3.3 Distribution Of The Sample Mean

If, for each $i = 1, 2, \ldots, n$, the x_i's are independent, identically distributed (iid) normal random variables, then

$$\overline{x}_i \sim \mathrm{N}\,(\,\mu_x, \frac{\sigma_x^2}{n}\,) \tag{2.6}$$

2.3.4 The Standard Normal

The distribution of $\overline{x}$ will vary with different values of μ_x and σ_x^2, which is inconvenient. Rather than dealing with a unique distribution for each case, we perform the following transformation:

$$\begin{aligned} Z &= \frac{\overline{x} - \mu_x}{\sqrt{\frac{\sigma_x^2}{n}}} \\ &= \frac{\overline{x}}{\sqrt{\frac{\sigma_x^2}{n}}} - \frac{\mu_x}{\sqrt{\frac{\sigma_x^2}{n}}} \end{aligned} \tag{2.7}$$

Now,

$$\begin{aligned} \mathrm{E}\, Z &= \frac{\mathrm{E}\,\overline{x}}{\sqrt{\frac{\sigma_x^2}{n}}} - \frac{\mu_x}{\sqrt{\frac{\sigma_x^2}{n}}} \\ &= \frac{\mu_x}{\sqrt{\frac{\sigma_x^2}{n}}} - \frac{\mu_x}{\sqrt{\frac{\sigma_x^2}{n}}} \\ &= 0. \end{aligned}$$

Also,

$$\begin{aligned} \mathrm{Var}(\, Z \,) &= \mathrm{E}\left(\frac{\overline{x}}{\sqrt{\frac{\sigma_x^2}{n}}} - \frac{\mu_x}{\sqrt{\frac{\sigma_x^2}{n}}} \right)^2 \\ &= \mathrm{E}\left[\frac{n}{\sigma_x^2} \left(\overline{x} - \mu_x \right)^2 \right] \\ &= \frac{n}{\sigma_x^2} \frac{\sigma_x^2}{n} \\ &= 1. \end{aligned}$$

Thus $Z \sim \mathrm{N}(0,1)$. The $\mathrm{N}(\,0,1\,)$ distribution is the standard normal and is well-tabulated. The probability density function for the standard normal distribution is

$$f\,(\,z_i\,) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}\,(\,z_i\,)^2} = \varphi(z_i) \tag{2.8}$$

2.4 The Central Limit Theorem

2.4.1 Normal Theory

The normal density has a prominent position in statistics. This is not only because many random variables appear to be normal, but also because most any sample mean appears normal as the sample size increases.

Specifically, suppose $x_1, x_2, \ldots, x_n$ is a simple random sample and $\mathrm{E}\, x_i = \mu_x$ and $\mathrm{Var}(\, x_i\,) = \sigma_x^2$, then as $n \to \infty$, the distribution of $\overline{x}$ becomes normal. That is,

$$\lim_{n\to\infty} f\left(\frac{\overline{x}-\mu_x}{\sqrt{\frac{\sigma_x^2}{n}}}\right)=\varphi\left(\frac{\overline{x}-\mu_x}{\sqrt{\frac{\sigma_x^2}{n}}}\right) \tag{2.9}$$

2.5 Distributions Associated With The Normal Distribution

2.5.1 The Chi-Squared Distribution

Definition 2.5 Suppose that $Z_1, Z_2, \ldots, Z_n$ is a simple random sample, and $Z_i \sim N(0,1)$. Then

$$\sum_{i=1}^{n} Z_i^2 \sim \mathcal{X}_n^2, \tag{2.10}$$

where n are the degrees of freedom of the Chi-squared distribution. □

The probability density function for the $\mathcal{X}_n^2$ is

$$f_{\chi^2}(x)=\frac{1}{2^{n/2}\Gamma(n/2)}x^{n/2-1}e^{-x/2}, x>0 \tag{2.11}$$

where $\Gamma(x)$ is the gamma function. See Figure 2.2. If $x_1, x_2, \ldots, x_n$ is a simple random sample, and $x_i \sim \mathrm{N}(\mu_x, \sigma_x^2)$, then

$$\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2 \sim \mathcal{X}_n^2. \tag{2.12}$$

The chi-squared distribution will prove useful in testing hypotheses on both the variance of a single variable and the (conditional) means of several. This multivariate usage will be explored in the next chapter.

Example 2.7 Consider the estimate of σ^2

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

Then

$$(n-1)\frac{s^2}{\sigma^2} \sim \mathcal{X}_{n-1}^2. \tag{2.13}$$

□

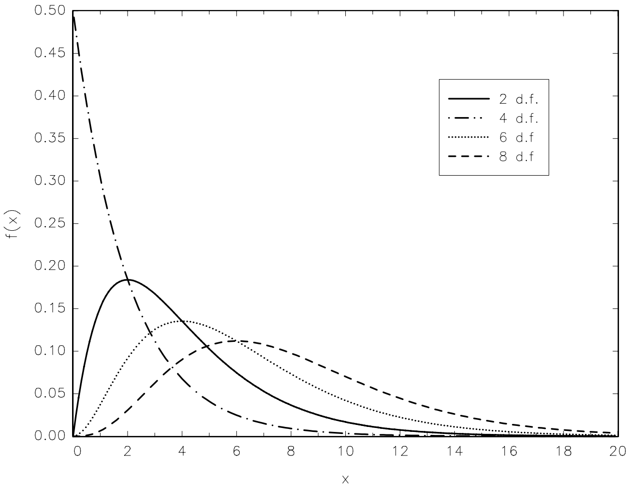

Figure 2.2: Some Chi-Squared Distributions

2.5.2 The t Distribution

Definition 2.6 Suppose that $Z \sim \mathrm{N}(0, 1)$, $Y \sim \mathcal{X}_k^2$, and that Z and Y are independent. Then

$$\frac{Z}{\sqrt{\frac{Y}{k}}} \sim t_k, \tag{2.14}$$

where k are the degrees of freedom of the t distribution. □

The probability density function for a t random variable with n degrees of freedom is

$$f_t(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)\left(1+\frac{x^2}{n}\right)^{(n+1)/2}}, \tag{2.15}$$

for $-\infty < x < \infty$. See Figure 2.3

The t (also known as Student's t) distribution, is named after W.S. Gosset, who published under the pseudonym "Student." It is useful in testing hypotheses concerning the (conditional) mean when the variance is estimated.

Example 2.8 Consider the sample mean from a simple random sample of normals. We know that $\overline{x} \sim \mathrm{N}(\mu, \sigma^2/n)$ and

$$Z = \frac{\overline{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim \mathrm{N}(0, 1).$$

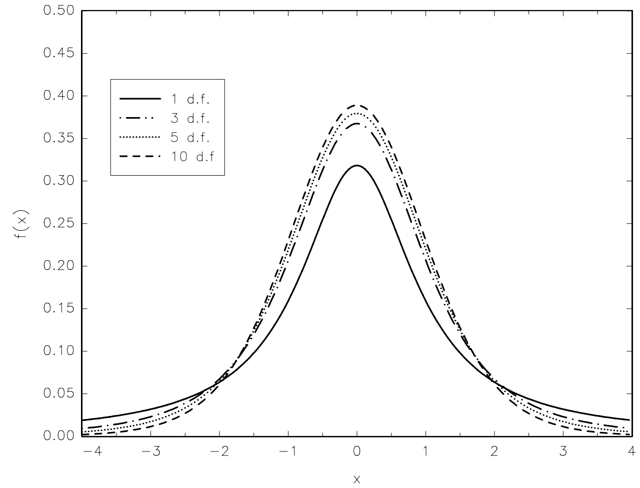

Figure 2.3: Some t Distributions

Also, we know that

$$Y = (n-1)\frac{s^2}{\sigma^2} \sim \mathcal{X}^2_{n-1},$$

where s^2 is the unbiased estimator of σ^2. Thus, if Z and Y are independent (which, in fact, is the case), then

$$\begin{aligned}\frac{Z}{\sqrt{\frac{Y}{(n-1)}}} &= \frac{\frac{\overline{x}-\mu}{\sqrt{\frac{\sigma^2}{n}}}}{\sqrt{\frac{(n-1)\frac{s^2}{\sigma^2}}{(n-1)}}} \\ &= (\overline{x}-\mu)\sqrt{\frac{\frac{n}{\sigma^2}}{\frac{s^2}{\sigma^2}}} \\ &= \frac{\overline{x}-\mu}{\sqrt{\frac{s^2}{n}}} \sim t_{n-1} \qquad (2.16)\end{aligned}$$

□

2.5.3 The F Distribution

Definition 2.7 Suppose that $Y \sim \mathcal{X}_m^2$, $W \sim \mathcal{X}_n^2$, and that Y and W are independent. Then

$$\frac{\frac{Y}{m}}{\frac{W}{n}} \sim F_{m,n}, \tag{2.17}$$

where m,n are the degrees of freedom of the F distribution. □

The probability density function for a F random variable with m and n degrees of freedom is

$$f_{\mathrm{F}}(\,x\,) = \frac{\Gamma\left(\frac{m+n}{2}\right)(m/n)^{m/2}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \frac{x^{(m/2)-1}}{(1+mx/n)^{(m+n)/2}} \tag{2.18}$$

The F distribution is named after the great statistician Sir Ronald A. Fisher, and is used in many applications, most notably in the analysis of variance. This situation will arise when we seek to test multiple (conditional) mean parameters with estimated variance. Note that when $x \sim t_n$ then $x^2 \sim F_{1,n}$. Some examples of the F distribution can be seen in Figure 2.4.

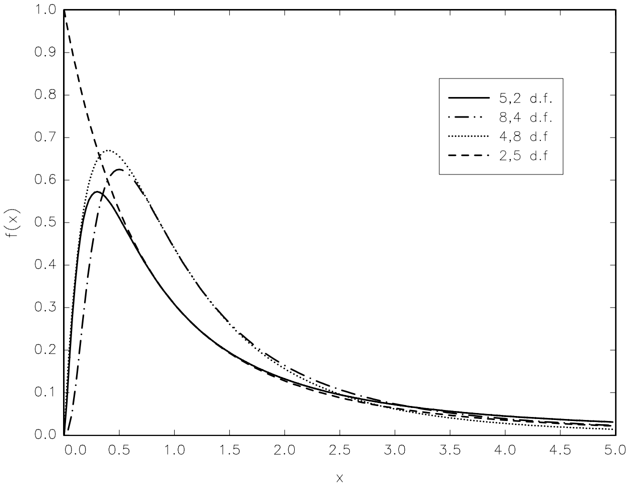

Figure 2.4: Some F Distributions

Chapter 3

Multivariate Distributions

3.1 Matrix Algebra Of Expectations

3.1.1 Moments of Random Vectors

Let

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} = \mathbf{x}$$

be an $m \times 1$ vector-valued random variable. Each element of the vector is a scalar random variable of the type discussed in the previous chapter.

The expectation of a random vector is

$$\mathrm{E}[\mathbf{x}] = \begin{bmatrix} \mathrm{E}[x_1] \\ \mathrm{E}[x_2] \\ \vdots \\ \mathrm{E}[x_m] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{bmatrix} = \boldsymbol{\mu}. \tag{3.1}$$

Note that $\boldsymbol{\mu}$ is also an $m \times 1$ column vector. We see that the mean of the vector is the vector of the means.

Next, we evaluate the following:

$$\mathrm{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})']$$

$$= \mathrm{E} \begin{bmatrix} (x_1 - \mu_1)^2 & (x_1 - \mu_1)(x_2 - \mu_2) & \cdots & (x_1 - \mu_1)(x_m - \mu_m) \\ (x_2 - \mu_2)(x_1 - \mu_1) & (x_2 - \mu_2)^2 & \cdots & (x_2 - \mu_2)(x_m - \mu_m) \\ \vdots & \vdots & \ddots & \vdots \\ (x_m - \mu_m)(x_1 - \mu_1) & (x_m - \mu_m)(x_2 - \mu_2) & \cdots & (x_m - \mu_m)^2 \end{bmatrix}$$

$$
\begin{aligned}
&= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_{mm} \end{bmatrix} \\
&= \boldsymbol{\Sigma}. \qquad (3.2)
\end{aligned}
$$

$\boldsymbol{\Sigma}$, the covariance matrix, is an $m \times m$ matrix of variance and covariance terms. The variance $\sigma_i^2 = \sigma_{ii}$ of x_i is along the diagonal, while the cross-product terms represent the covariance between x_i and x_j.

3.1.2 Properties Of The Covariance Matrix

Symmetric

The variance-covariance matrix $\boldsymbol{\Sigma}$ is a symmetric matrix. This can be shown by noting that

$$\sigma_{ij} = \mathrm{E}(\, x_i - \mu_i\,)(\, x_j - \mu_j\,) = \mathrm{E}(\, x_j - \mu_j\,)(\, x_i - \mu_i\,) = \sigma_{ji}.$$

Due to this symmetry $\boldsymbol{\Sigma}$ will only have $m(m+1)/2$ unique elements.

Positive Semidefinite

$\boldsymbol{\Sigma}$ is a positive semidefinite matrix. Recall that any $m \times m$ matrix is positive semidefinite if and only if it meets any of the following three equivalent conditions:

1. All the principle minors are nonnegative;
2. $\boldsymbol{\lambda}'\boldsymbol{\Sigma}\boldsymbol{\lambda} \geq 0$, for all $\underbrace{\boldsymbol{\lambda}}_{m\times 1} \neq 0$;
3. $\boldsymbol{\Sigma} = \mathbf{PP}'$, for some $\underbrace{\mathbf{P}}_{m\times m}$.

The first condition (actually we use negative definiteness) is useful in the study of utility maximization while the latter two are useful in econometric analysis.

The second condition is the easiest to demonstrate in the current context. Let $\boldsymbol{\lambda} \neq \mathbf{0}$. Then, we have

$$
\begin{aligned}
\boldsymbol{\lambda}'\boldsymbol{\Sigma}\boldsymbol{\lambda} &= \boldsymbol{\lambda}'\,\mathrm{E}[(\,\mathbf{x}-\boldsymbol{\mu}\,)(\,\mathbf{x}-\boldsymbol{\mu}\,)']\boldsymbol{\lambda} \\
&= \mathrm{E}[\boldsymbol{\lambda}'(\,\mathbf{x}-\boldsymbol{\mu}\,)(\,\mathbf{x}-\boldsymbol{\mu}\,)'\boldsymbol{\lambda}\,] \\
&= \mathrm{E}\{\,[\,\boldsymbol{\lambda}'(\,\mathbf{x}-\boldsymbol{\mu}\,)]^2\,\} \geq 0,
\end{aligned}
$$

since the term inside the expectation is a quadratic. Hence, $\mathbf{\Sigma}$ is a positive semidefinite matrix.

Note that $\mathbf{P}$ satisfying the third relationship is not unique. Let $\mathbf{D}$ be any $m \times m$ orthonormal martix, then $\mathbf{DD}' = \mathbf{I}_m$ and $\mathbf{P}^* = \mathbf{PD}$ yields $\mathbf{P}^*\mathbf{P}^{*\prime} = \mathbf{PDD}'\mathbf{P}' = \mathbf{PI}_m\mathbf{P}' = \mathbf{\Sigma}$. Usually, we will choose $\mathbf{P}$ to be an upper or lower triangular matrix with $m(m+1)/2$ nonzero elements.

Positive Definite

Since $\mathbf{\Sigma}$ is a positive semidefinite matrix, it will be a positive definite matrix if and only if $\det(\mathbf{\Sigma}) \neq 0$. Now, we know that $\mathbf{\Sigma} = \mathbf{PP}'$ for some $m \times m$ matrix $\mathbf{P}$. This implies that $\det(\mathbf{P}) \neq 0$.

3.1.3 Linear Transformations

Let $\underbrace{\mathbf{y}}_{m\times 1} = \underbrace{\mathbf{b}}_{m\times 1} + \underbrace{\mathbf{B}}_{m\times m}\overbrace{\mathbf{x}}^{m\times 1}$. Then

$$\begin{aligned} \mathrm{E}[\mathbf{y}] &= \mathbf{b} + \mathbf{B}\,\mathrm{E}[\mathbf{x}] \\ &= \mathbf{b} + \mathbf{B}\boldsymbol{\mu} \\ &= \boldsymbol{\mu}_y \qquad (3.3) \end{aligned}$$

Thus, the mean of a linear transformation is the linear transformation of the mean.

Next, we have

$$\begin{aligned} \mathrm{E}[(\mathbf{y}-\boldsymbol{\mu}_y)(\mathbf{y}-\boldsymbol{\mu}_y)'] &= \mathrm{E}\{[\mathbf{B}(\mathbf{x}-\boldsymbol{\mu})][(\mathbf{B}(\mathbf{x}-\boldsymbol{\mu}))']\} \\ &= \mathbf{B}\,\mathrm{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})']\mathbf{B}' \\ &= \mathbf{B\Sigma B}' \\ &= \mathbf{B\Sigma B}' \qquad (3.4) \\ &= \mathbf{\Sigma}_y \qquad (3.5) \end{aligned}$$

where we use the result $(ABC)' = C'B'A'$, if conformability holds.

3.2 Change Of Variables

3.2.1 Univariate

Let x be a random variable and $f_x(\cdot)$ be the probability density function of x. Now, define $y = h(x)$, where

$$h'(x) = \frac{d\,h(x)}{d\,x} > 0.$$

That is, $h(\,x\,)$ is a strictly monotonically increasing function and so y is a one-to-one transformation of x. Now, we would like to know the probability density function of y, $f_y(y)$. To find it, we note that

$$\Pr(\,y \leq h(\,a\,)\,) = \Pr(\,x \leq a\,), \tag{3.6}$$

$$\Pr(\,x \leq a\,) = \int_{-\infty}^{a} f_x(\,x\,)\;dx = F_x(\,a\,), \tag{3.7}$$

and,

$$\Pr(\,y \leq h(\,a\,)\,) = \int_{-\infty}^{h(\,a\,)} f_y(\,y\,)\;dy = F_y(\,h(\,a\,)), \tag{3.8}$$

for all a.

Assuming that the cumulative density function is differentiable, we use (3.6) to combine (3.7) and (3.8), and take the total differential, which gives us

$$\begin{aligned} dF_x(\,a\,) &= dF_y(\,h(\,a\,)) \\ f_x(\,a\,)da &= f_y(\,h(\,a\,))h'(\,a\,)da \end{aligned}$$

for all a. Thus, for a small perturbation,

$$f_x(\,a\,) = f_y(\,h(\,a\,))h'(\,a\,) \tag{3.9}$$

for all a. Also, since y is a one-to-one transformation of x, we know that $h(\cdot)$ can be inverted. That is, $x = h^{-1}(y)$. Thus, $a = h^{-1}(y)$, and we can rewrite (3.9) as

$$f_x(\,h^{-1}(\,y\,)) = f_y(\,y\,)h'(\,h^{-1}(\,y\,)).$$

Therefore, the probability density function of y is

$$f_y(\,y\,) = f_x(\,h^{-1}(\,y\,))\frac{1}{h'(\,h^{-1}(\,y\,))}. \tag{3.10}$$

Note that $f_y(\,y\,)$ has the properties of being nonnegative, since $h'(\cdot) > 0$. If $h'(\cdot) < 0$, (3.10) can be corrected by taking the absolute value of $h'(\cdot)$, which will assure that we have only positive values for our probability density function.

3.2.2 Geometric Interpretation

Consider the graph of the relationship shown in Figure 3.1. We know that

$$\Pr[\,h(\,b\,) > y > h(\,a\,)] = \Pr(\,b > x > a\,).$$

Also, we know that

$$\Pr[\,h(\,b\,) > y > h(\,a\,)] \simeq f_y[\,h(\,b\,)][\,h(\,b\,) - h(\,a\,)],$$

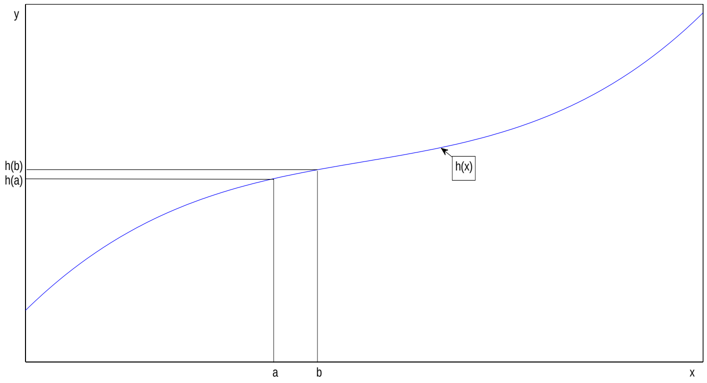

Figure 3.1: Change of Variables

and

$$\Pr(\, b > x > a\,) \simeq f_x(\, b\,)(\, b - a\,).$$

So,

$$\begin{aligned} f_y[\, h(\, b\,)][\, h(\, b\,) - h(\, a\,)] &\simeq f_x(\, b\,)(\, b - a\,) \\ f_y[\, h(\, b\,)] &\simeq f_x(\, b\,)\frac{1}{[\, h(\, b\,) - h(\, a\,)]/(\, b - a\,)]} \qquad (3.11) \end{aligned}$$

Now, as we let $a \rightarrow b$, the denominator of (3.11) approaches $h'(\cdot)$. This is then the same formula as (3.10).

3.2.3 Multivariate

Let

$$\underbrace{\mathbf{x}}_{m\times 1} \sim f_x(\, \mathbf{x}\,).$$

Define a one-to-one transformation

$$\underbrace{\mathbf{y}}_{m\times 1} = \overbrace{\mathbf{h}}^{m\times 1}(\, \mathbf{x}\,).$$

Since $\mathbf{h}(\cdot)$ is a one-to-one transformation, it has an inverse:

$$\mathbf{x} = \mathbf{h}^{-1}(\, \mathbf{y}\,).$$

We also assume that $\frac{\partial \mathbf{h}(\mathbf{x})}{\partial \mathbf{x}'}$ exists. This is the $m \times m$ Jacobian matrix, where

$$\begin{aligned}
\frac{\partial \mathbf{h}(\mathbf{x})}{\partial \mathbf{x}'} &= \frac{\partial \begin{bmatrix} h_1(\mathbf{x}) \\ h_2(\mathbf{x}) \\ \vdots \\ h_m(\mathbf{x}) \end{bmatrix}}{\partial(x_1 x_2 \cdots x_m)} \\
&= \begin{bmatrix} \frac{\partial h_1(\mathbf{x})}{\partial x_1} & \frac{\partial h_2(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial h_m(\mathbf{x})}{\partial x_1} \\ \frac{\partial h_1(\mathbf{x})}{\partial x_2} & \frac{\partial h_2(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial h_m(\mathbf{x})}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_1(\mathbf{x})}{\partial x_m} & \frac{\partial h_2(\mathbf{x})}{\partial x_m} & \cdots & \frac{\partial h_m(\mathbf{x})}{\partial x_m} \end{bmatrix} \\
&= \mathrm{J}_x(\mathbf{x}) \qquad (3.12)
\end{aligned}$$

Given this notation, the multivariate analog to (3.11) can be shown to be

$$f_y(\mathbf{y}) = f_x[\mathbf{h}^{-1}(\mathbf{y})]\frac{1}{|\det(J_x[\mathbf{h}^{-1}(\mathbf{y})])|} \qquad (3.13)$$

Since $h(\cdot)$ is differentiable and one-to-one then $\det(J_x[\mathbf{h}^{-1}(\mathbf{y})]) \neq 0$.

Example 3.1 Let $y = b_0 + b_1 x$, where x, b_0, and b_1 are scalars. Then

$$x = \frac{y - b_0}{b_1}$$

and

$$\frac{dy}{dx} = b_1.$$

Therefore,

$$f_y(y) = f_x\left(\frac{y - b_0}{b_1}\right)\frac{1}{|b_1|}. \ \square$$

Example 3.2 Let $\mathbf{y} = \mathbf{b} + \mathbf{B}\mathbf{x}$, where $\mathbf{y}$ is an $m \times 1$ vector and $\det(\mathbf{B}) \neq 0$. Then

$$\mathbf{x} = \mathbf{B}^{-1}(\mathbf{y} - \mathbf{b})$$

and

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}'} = \mathbf{B} = J_x(\mathbf{x}).$$

Thus,

$$f_y(\mathbf{y}) = f_x\left(\mathbf{B}^{-1}(\mathbf{y} - \mathbf{b})\right)\frac{1}{|\det(\mathbf{B})|}. \ \square$$

3.3 Multivariate Normal Distribution

3.3.1 Spherical Normal Distribution

Definition 3.1 An $m \times 1$ random vector $\mathbf{z}$ is said to be spherically normally distributed if

$$f(\mathbf{z}) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\mathbf{z}'\mathbf{z}}. \ \square$$

Such a random vector can be seen to be a vector of independent standard normals. Let $z_1, z_2, \ldots, z_m$, be i.i.d. random variables such that $z_i \sim \mathrm{N}(0,1)$. That is, z_i has pdf given in (2.8), for $i = 1, ..., m$. Then, by independence, the joint distribution of the z_i's is given by

$$\begin{aligned} f(z_1, z_2, \ldots, z_m) &= f(z_1) f(z_2) \cdots f(z_m) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} z_i^2} \\ &= \frac{1}{(2\pi)^{m/2}} e^{-\frac{1}{2}\sum_{i=1}^{n} z_i^2} \\ &= \frac{1}{(2\pi)^{m/2}} e^{-\frac{1}{2}\mathbf{z}'\mathbf{z}}, \end{aligned} \tag{3.14}$$

where $\mathbf{z}' = (z_1 \ z_2 \ \ldots \ z_m)$.

3.3.2 Multivariate Normal

Definition 3.2 The $m \times 1$ random vector $\mathbf{x}$ with density

$$f_x(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}[\det(\mathbf{\Sigma})]^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \tag{3.15}$$

is said to be distributed multivariate normal with mean vector $\boldsymbol{\mu}$ and positive definite covariance matrix $\mathbf{\Sigma}$. $\square$

Such a distribution for x is denoted by $\mathbf{x} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$. The spherical normal distribution is seen to be a special case where $\boldsymbol{\mu} = \mathbf{0}$ and $\mathbf{\Sigma} = \mathbf{I}_m$.

There is a one-to-one relationship between the multivariate normal random vector and a spherical normal random vector. Let $\mathbf{z}$ be an $m \times 1$ spherical normal random vector and

$$\underbrace{\mathbf{x}}_{m\times 1} = \boldsymbol{\mu} + \mathbf{A}\mathbf{z},$$

where $\mathbf{z}$ is defined above, and $\det(\mathbf{A}) \neq 0$. Then,

$$\mathrm{E}\,\mathbf{x} = \boldsymbol{\mu} + \mathbf{A}\,\mathrm{E}\,\mathbf{z} = \boldsymbol{\mu}\ , \tag{3.16}$$

since $\mathrm{E}[\mathbf{z}] = \mathbf{0}$.

Also, we know that

$$\mathrm{E}(\,\mathbf{z}\mathbf{z}'\,) = \mathrm{E}\begin{bmatrix} z_1^2 & z_1 z_2 & \cdots & z_1 z_m \\ z_2 z_1 & z_2^2 & \cdots & z_1 z_m \\ \vdots & \vdots & \ddots & \vdots \\ z_m z_1 & z_m z_2 & \cdots & z_m^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \mathbf{I}_m\,, \quad (3.17)$$

since $\mathrm{E}[z_i z_j] = 0$, for all $i \neq j$, and $\mathrm{E}[z_i^2] = 1$, for all i. Therefore,

$$\begin{aligned} \mathrm{E}[(\,\mathbf{x}-\boldsymbol{\mu}\,)(\,\mathbf{x}-\boldsymbol{\mu}\,)'] &= \mathrm{E}(\,\mathbf{A}\mathbf{z}\mathbf{z}'\mathbf{A}'\,) \\ &= \mathbf{A}\,\mathrm{E}(\,\mathbf{z}\mathbf{z}'\,)\mathbf{A}' \\ &= \mathbf{A}\mathbf{I}_m\mathbf{A}' = \boldsymbol{\Sigma}, \qquad (3.18) \end{aligned}$$

where $\boldsymbol{\Sigma}$ is a positive definite matrix (since $\det(\mathbf{A}) \neq 0$).

Next, we need to find the probability density function $f_x(\,\mathbf{x}\,)$ of $\mathbf{x}$. We know that

$$\mathbf{z} = \mathbf{A}^{-1}(\,\mathbf{x}-\boldsymbol{\mu}\,)\ ,$$

$$\mathbf{z}' = (\,\mathbf{x}-\boldsymbol{\mu}\,)'\mathbf{A}^{-1\prime}\ ,$$

and

$$J_z(\,\mathbf{z}\,) = \mathbf{A},$$

so we use (3.13) to get

$$\begin{aligned} f_x(\,\mathbf{x}\,) &= f_z[\,\mathbf{A}^{-1}(\,\mathbf{x}-\boldsymbol{\mu}\,)]\frac{1}{|\det(\,\mathbf{A}\,)|} \\ &= \frac{1}{(\,2\pi\,)^{m/2}}e^{-\frac{1}{2}(\,\mathbf{x}-\boldsymbol{\mu}\,)'\mathbf{A}^{-1\prime}\mathbf{A}^{-1}(\,\mathbf{x}-\boldsymbol{\mu}\,)}\frac{1}{|\det(\,\mathbf{A}\,)|} \\ &= \frac{1}{(\,2\pi\,)^{m/2}}e^{-\frac{1}{2}(\,\mathbf{x}-\boldsymbol{\mu}\,)'(\,\mathbf{A}\mathbf{A}'\,)^{-1}(\,\mathbf{x}-\boldsymbol{\mu}\,)}\frac{1}{|\det(\,\mathbf{A}\,)|} \qquad (3.19) \end{aligned}$$

where we use the results $(\mathbf{ABC}) = \mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1}$ and $\mathbf{A}'^{-1} = \mathbf{A}^{-1\prime}$. However, $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}'$, so $\det(\boldsymbol{\Sigma}) = \det(\mathbf{A})\cdot\det(\mathbf{A})$, and $|\det(\mathbf{A})| = [\,\det(\boldsymbol{\Sigma})]^{1/2}$. Thus we can rewrite (3.19) as

$$f_x(\,\mathbf{x}\,) = \frac{1}{(\,2\pi\,)^{m/2}[\,\det(\,\boldsymbol{\Sigma}\,)]^{1/2}}e^{-\frac{1}{2}(\,\mathbf{x}-\boldsymbol{\mu}\,)'\boldsymbol{\Sigma}^{-1}(\,\mathbf{x}-\boldsymbol{\mu}\,)} \qquad (3.20)$$

and we see that $\mathbf{x} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

Since this process is completely reversable the relationship is one-to-one.

3.3.3 Linear Transformations

Theorem 3.1 *Suppose* $\mathbf{x} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *with* $\det(\boldsymbol{\Sigma}) \neq 0$ *and* $\mathbf{y} = \mathbf{b} + \mathbf{B}\mathbf{x}$ *with* $\mathbf{B}$ *square and* $\det(\mathbf{B}) \neq 0$. *Then* $\mathbf{y} \sim \mathrm{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$. □

Proof: From (3.3) and (3.4), we have $\mathrm{E}\,\mathbf{y} = \mathbf{b} + \mathbf{B}\boldsymbol{\mu} = \boldsymbol{\mu}_y$ and $\mathrm{E}[(\,\mathbf{y} - \boldsymbol{\mu}_y\,)(\,\mathbf{y} - \boldsymbol{\mu}_y\,)'] = \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}' = \boldsymbol{\Sigma}_y$. To find the probability density function $f_y(\,\mathbf{y}\,)$ of $\mathbf{y}$, we again use (3.13), which gives us

$$\begin{aligned}
f_y(\,\mathbf{y}\,) &= f_x[\,\mathbf{B}^{-1}(\,\mathbf{y} - \mathbf{b}\,)]\frac{1}{|\det(\,\mathbf{B}\,)|} \\
&= \frac{1}{(\,2\pi\,)^{m/2}[\,\det(\,\boldsymbol{\Sigma}\,)]^{1/2}} e^{-\frac{1}{2}[\,\mathbf{B}^{-1}(\,\mathbf{y}-\mathbf{b}\,)-\boldsymbol{\mu}]'\boldsymbol{\Sigma}^{-1}[\,\mathbf{B}^{-1}(\,\mathbf{y}-\mathbf{b}\,)-\boldsymbol{\mu}\,]} \frac{1}{|\det(\,\mathbf{B}\,)|} \\
&= \frac{1}{(\,2\pi\,)^{m/2}[\,\det(\,\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}')]^{1/2}} e^{-\frac{1}{2}(\,\mathbf{y}-\mathbf{b}-\mathbf{B}\boldsymbol{\mu}\,)'(\,\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}'\,)^{-1}(\,\mathbf{y}-\mathbf{b}-\mathbf{B}\boldsymbol{\mu}\,)} \\
&= \frac{1}{(\,2\pi\,)^{m/2}[\,\det(\,\boldsymbol{\Sigma}_y\,)]^{1/2}} e^{-\frac{1}{2}(\,\mathbf{y}-\boldsymbol{\mu}_y\,)'\,\boldsymbol{\Sigma}_y^{-1}(\,\mathbf{y}-\boldsymbol{\mu}_y\,)} \qquad (3.21)
\end{aligned}$$

So, $\mathbf{y} \sim \mathbf{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$. □

Thus we see, as asserted in the previous chapter, that linear transformations of multivariate normal random variables are also multivariate normal random variables. And any linear combination of independent normals will also be normal.

3.3.4 Quadratic Forms

Theorem 3.2 *Let* $\mathbf{x} \sim \mathrm{N}(\,\boldsymbol{\mu}, \boldsymbol{\Sigma}\,)$, *where* $\det(\boldsymbol{\Sigma}) \neq \mathbf{0}$, *then* $(\,\mathbf{x} - \boldsymbol{\mu}\,)'\boldsymbol{\Sigma}^{-1}(\,\mathbf{x} - \boldsymbol{\mu}\,) \sim \mathcal{X}^2_m$. □

Proof Let $\boldsymbol{\Sigma} = \mathbf{P}\mathbf{P}'$. Then

$$(\,\mathbf{x} - \boldsymbol{\mu}\,) \sim \mathrm{N}(\,0, \boldsymbol{\Sigma}\,),$$

and

$$\mathbf{z} = \mathbf{P}^{-1}(\,\mathbf{x} - \boldsymbol{\mu}\,) \sim \mathrm{N}(\,0, \mathbf{I}_m\,).$$

Therefore,

$$\begin{aligned}
\mathbf{z}'\mathbf{z} &= \sum_{i=1}^{n} z_i^2 \sim \mathcal{X}^2_m \\
&= \mathbf{P}^{-1}(\,\mathbf{x} - \boldsymbol{\mu}\,)'\mathbf{P}^{-1}(\,\mathbf{x} - \boldsymbol{\mu}\,) \\
&= (\,\mathbf{x} - \boldsymbol{\mu}\,)'\mathbf{P}^{-1\prime}\mathbf{P}^{-1}(\,\mathbf{x} - \boldsymbol{\mu}\,) \\
&= (\,\mathbf{x} - \boldsymbol{\mu}\,)'\boldsymbol{\Sigma}^{-1}(\,\mathbf{x} - \boldsymbol{\mu}\,) \sim \mathcal{X}^2_m. \; \square \qquad (3.22)
\end{aligned}$$

$\boldsymbol{\Sigma}^{-1}$ is called the *weight matrix*. With this result, we can use the $\mathcal{X}^2_m$ to make inferences about the mean $\boldsymbol{\mu}$ of $\mathbf{x}$.

3.4 Normality and the Sample Mean

3.4.1 Moments of the Sample Mean

Consider the $m \times 1$ vector $\mathbf{x}_i \sim i.i.d.$ jointly, with $m \times 1$ vector mean $\mathrm{E}[\mathbf{x}_i] = \boldsymbol{\mu}$ and $m \times m$ covariance matrix $\mathrm{E}[(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})'] = \boldsymbol{\Sigma}$. Define $\overline{\mathbf{x}}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$ as the *vector sample mean* which is also the vector of scalar sample means. The mean of the vector sample mean follows directly:

$$\mathrm{E}[\overline{\mathbf{x}}_n] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[\mathbf{x}_i] = \boldsymbol{\mu}.$$

Alternatively, this result can be obtained by applying the scalar results element by element. The second moment matrix of the vector sample mean is given by

$$\begin{aligned}
\mathrm{E}[(\overline{\mathbf{x}}_n - \boldsymbol{\mu})(\overline{\mathbf{x}}_n - \boldsymbol{\mu})'] &= \frac{1}{n^2}\,\mathrm{E}\left[\left(\textstyle\sum_{i=1}^n \mathbf{x}_i - n\boldsymbol{\mu}\right)\left(\textstyle\sum_{i=1}^n \mathbf{x}_i - n\boldsymbol{\mu}\right)'\right] \\
&= \frac{1}{n^2}\,\mathrm{E}[\{(\mathbf{x}_1 - \boldsymbol{\mu}) + (\mathbf{x}_2 - \boldsymbol{\mu}) + ... + (\mathbf{x}_n - \boldsymbol{\mu})\} \\
&\qquad\qquad \{(\mathbf{x}_1 - \boldsymbol{\mu}) + (\mathbf{x}_2 - \boldsymbol{\mu}) + ... + (\mathbf{x}_n - \boldsymbol{\mu})\}'] \\
&= \frac{1}{n^2} n\boldsymbol{\Sigma} = \frac{1}{n}\boldsymbol{\Sigma}
\end{aligned}$$

since the covariances between different observations are zero.

3.4.2 Distribution of the Sample Mean

Suppose $\mathbf{x}_i \sim i.i.d.\, \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ jointly. Then it follows from joint multivariate normality that $\overline{\mathbf{x}}_n$ must also be multivariate normal since it is a linear transformation. Specifically, we have

$$\overline{\mathbf{x}}_n \sim \mathrm{N}(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$$

or equivalently

$$\begin{aligned}
\overline{\mathbf{x}}_n - \boldsymbol{\mu} &\sim \mathrm{N}(0, \frac{1}{n}\boldsymbol{\Sigma}) \\
\sqrt{n}(\overline{\mathbf{x}}_n - \boldsymbol{\mu}) &\sim \mathrm{N}(0, \boldsymbol{\Sigma}) \\
\sqrt{n}\boldsymbol{\Sigma}^{-1/2}(\overline{\mathbf{x}}_n - \boldsymbol{\mu}) &\sim \mathrm{N}(0, \mathbf{I}_m)
\end{aligned}$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2\prime}$ and $\boldsymbol{\Sigma}^{-1/2} = (\boldsymbol{\Sigma}^{1/2})^{-1}$.

3.4.3 Multivariate Central Limit Theorem

Theorem 3.3 *Suppose that (i)* $\mathbf{x}_i \sim i.i.d$ *jointly, (ii)* $\mathrm{E}[\mathbf{x}_i] = \boldsymbol{\mu}$*, and (iii)* $\mathrm{E}[(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})'] = \boldsymbol{\Sigma}$*, then*

$$\sqrt{n}(\overline{\mathbf{x}}_n - \boldsymbol{\mu}) \rightarrow_d \mathrm{N}(0, \boldsymbol{\Sigma})$$

or equivalently

$$\mathbf{z} = \sqrt{n}\boldsymbol{\Sigma}^{-1/2}(\overline{\mathbf{x}}_n - \boldsymbol{\mu}) \rightarrow_d \; \mathrm{N}(0, \mathbf{I}_m) \; \square \; .$$

These results apply even if the original underlying distribution is not normal and follow directly from the scalar results applied to any linear combination of $\overline{\mathbf{x}}_n$.

3.4.4 Limiting Behavior of Quadratic Forms

Consider the following quadratic form

$$\begin{aligned} n \cdot (\overline{\mathbf{x}}_n - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}}_n - \boldsymbol{\mu}) &= n \cdot (\overline{\mathbf{x}}_n - \boldsymbol{\mu})'(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2\prime})^{-1}(\overline{\mathbf{x}}_n - \boldsymbol{\mu}) \\ &= [n \cdot (\overline{\mathbf{x}}_n - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1/2\prime}\boldsymbol{\Sigma}^{-1/2}(\mathbf{x}_n - \boldsymbol{\mu}) \\ &= [\sqrt{n}\boldsymbol{\Sigma}^{-1/2}(\mathbf{x}_n - \boldsymbol{\mu})]'[\sqrt{n}\boldsymbol{\Sigma}^{-1/2}(\mathbf{x}_n - \boldsymbol{\mu})] \\ &= \mathbf{z}'\mathbf{z} \rightarrow_d^2 \chi^2_m. \end{aligned}$$

This form is convenient for asymptotic joint test concerning more than one mean at a time.

3.5 Noncentral Distributions

3.5.1 Noncentral Scalar Normal

Definition 3.3 Let $x \sim \mathrm{N}(\mu, \sigma^2)$. Then,

$$z^* = \frac{x}{\sigma} \sim \mathrm{N}(\,\mu/\sigma, 1\,) \tag{3.23}$$

has a noncentral normal distribution. □

Example 3.3 When we do a hypothesis test of mean, with known variance, we have, under the null hypothesis $H_0 : \mu = \mu_0$,

$$\frac{x - \mu_0}{\sigma} \sim \mathrm{N}(\,0, 1\,) \tag{3.24}$$

and, under the alternative $H_1 : \mu = \mu_1 \neq \mu_0$,

$$\begin{aligned} \frac{x - \mu_0}{\sigma} &= \frac{x - \mu_1}{\sigma} + \frac{\mu_1 - \mu_0}{\sigma} \\ &= \mathrm{N}(\,0, 1\,) + \frac{\mu_1 - \mu_0}{\sigma} \sim \mathrm{N}\left(\frac{\mu_1 - \mu_0}{\sigma}, 1\right). \end{aligned} \tag{3.25}$$

Thus, the behavior of $\frac{x-\mu_0}{\sigma}$ under the alternative hypothesis follows a noncentral normal distribution. □

As this example makes clear, the noncentral normal distribution is especially useful when carefully exploring the behavior of the alternative hypothesis.

3.5.2 Noncentral t

Definition 3.4 Let z*$\sim$N$(\mu/\sigma, 1)$, w $\sim \chi^2_{\mathrm{k}}$, and let z^* and w be independent. Then

$$\frac{z^*}{\sqrt{w/k}} \sim t_k(\mu) \tag{3.26}$$

has a noncentral t distribution. □

The noncentral t distribution is used in tests of the mean, when the variance is unknown.

3.5.3 Noncentral Chi-Squared

Definition 3.5 Let $\mathbf{z}^*\sim$N(μ, I_m). Then

$$\mathbf{z}^{*\prime}\mathbf{z}^* \sim \mathcal{X}^2_m(\delta), \tag{3.27}$$

has a noncentral chi-aquared distribution, where $\delta = \boldsymbol{\mu}'\boldsymbol{\mu}$ is the noncentrality parameter. □

In the noncentral chi-squared distribution, the probability mass is shifted to the right as compared to a regular chi-squared distribution.

Example 3.4 When we do a test of $\boldsymbol{\mu}$, with known $\boldsymbol{\Sigma}$, we have

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$$

$$(\mathbf{x} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_0) \sim \mathcal{X}^2_m \tag{3.28}$$

$$H_1 : \boldsymbol{\mu} = \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_0$$

Let $\mathbf{z}^*$=$\mathbf{P}^{-1}$($\mathbf{x}$−$\boldsymbol{\mu}_0$). Then, we have

$$\begin{aligned}(\mathbf{x} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_0) &= (\mathbf{x} - \boldsymbol{\mu}_0)'\mathbf{P}^{-1\prime}\mathbf{P}^{-1}(\mathbf{x} - \boldsymbol{\mu}_0) \\ &= \mathbf{z}^{*\prime}\mathbf{z}^* \\ &\sim \mathcal{X}^2_m[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)] \end{aligned} \tag{3.29}$$

□

3.5.4 Noncentral F

Definition 3.6 Let Y $\sim \chi^2_m(\delta)$, W $\sim\chi^2_n$, and let Y and W be independent random variables. Then

$$\frac{Y/m}{W/n} \sim F_{m,n}(\delta), \tag{3.30}$$

has a noncentral F distribution, where δ is the noncentrality parameter. □

The noncentral F distribution is used in tests of mean vectors, where the variance-covariance matrix is unknown and must be estimated.

Chapter 4

Asymptotic Theory

4.1 Convergence Of Random Variables

4.1.1 Limits And Orders Of Sequences

Definition 4.1 A sequence of real numbers $a_1, a_2, \dots, a_n, \dots$, is said to have a limit of α if for every $\delta > 0$, there exists a positive real number N such that for all $n > N$, $|\, a_n - \alpha \,| < \delta$. This is denoted as

$$\lim_{n\to\infty} a_n = \alpha\,. \qquad \Box$$

Definition 4.2 A sequence of real numbers $\{\, a_n \,\}$ is said to be of at most order n^k, and we write $\{\, a_n \,\}$ is $\mathrm{O}(\, n^k \,)$, if

$$\lim_{n\to\infty} \frac{1}{n^k} a_n = c,$$

where c is any real constant. □

Example 4.1 Let $\{\, a_n \,\} = 3 + 1/n$, and $\{\, b_n \,\} = 4 - n^2$. Then, $\{\, a_n \,\}$ is $\mathrm{O}(\,1\,)$ $= \mathrm{O}(\, n^0 \,)$*, since*

$$\lim_{n\to\infty} \frac{1}{n} a_n = 3,$$

and $\{\, b_n \,\}$ is $\mathrm{O}(\, n^2 \,)$, since

$$\lim_{\mathrm{n}\to\infty} \frac{1}{n^2} \mathrm{b}_n\text{=-1.} \qquad \Box$$

Definition 4.3 A sequence of real numbers $\{\, a_n \,\}$ is said to be *of order smaller than* n^k, and we write $\{\, a_n \,\}$ is $\mathrm{o}(\, n^k \,)$, if

$$\lim_{n\to\infty} \frac{1}{n^k} a_n = 0. \qquad \Box$$

Example 4.2 Let $\{ a_n \} = 1/n$. Then, $\{ a_n \}$ is o(1), since

$$\lim_{n\to\infty} \frac{1}{n^0} \mathrm{a}_n{=}0. \qquad \square$$

4.1.2 Convergence In Probability

Definition 4.4 A sequence of random variables $y_1, y_2, \ldots, y_n, \ldots$, with distribution functions $F_1(\cdot), F_2(\cdot), \ldots, F_n(\cdot), \ldots$, is said to *converge weakly in probability* to some constant c if

$$\lim_{n\to\infty} \Pr[\, |\, y_n - c\, | > \epsilon\,] = 0. \qquad \square \tag{4.1}$$

for every real number $\epsilon > 0$.

Weak convergence in probability is denoted by

$$\operatorname*{plim}_{n\to\infty} y_n = c, \tag{4.2}$$

or sometimes,

$$y_n \xrightarrow{p} c, \tag{4.3}$$

or

$$y_n \rightarrow_p c.$$

This definition is equivalent to saying that we have a sequence of tail probabilities (of being greater than $c+\epsilon$ or less than $c-\epsilon$), and that the tail probabilities approach 0 as $n \to \infty$, regardless of how small ϵ is chosen. Equivalently, the probability mass of the distribution of y_n is collapsing about the point c.

Definition 4.5 A sequence of random variables $y_1, y_2, \ldots, y_n, \ldots$, is said to *converge strongly in probability to some constant* c if

$$\lim_{N\to\infty} \Pr[\, \sup_{n>N} \, |\, y_n - c\, | > \epsilon\,] = 0, \tag{4.4}$$

for any real $\epsilon > 0$. □

Strong convergence is also called *almost sure convergence* and is denoted

$$y_n \xrightarrow{a.s.} c, \tag{4.5}$$

or

$$y_n \rightarrow_{a.s.} c.$$

Notice that if a sequence of random variables converges strongly in probability, it converges weakly in probability. The difference between the two is that almost sure convergence involves an element of uniformity that weak convergence does not. A sequence that is weakly convergent can have $\Pr[|\, y_n - c\,| > \epsilon]$ wiggle-waggle above and below the constant δ used in the limit and then settle down to subsequently be smaller and meet the condition. For strong convergence, once the probability falls below δ for a particular N in the sequence it will subsequently be smaller for all larger N.

Definition 4.6 A sequence of random variables $y_1, y_2, \ldots, y_n, \ldots,$ is said to *converge in quadratic mean* if

$$\lim_{n\to\infty} \mathrm{E}[\, y_n\,] = c$$

and

$$\lim_{n\to\infty} \mathrm{Var}[\, y_n\,] = 0. \qquad \Box$$

By Chebyshev's inequality, convergence in quadratic mean implies weak convergence in probability. For a random variable x with mean μ and variance σ^2, Chebyshev's inequality states $\Pr(|x-\mu| \geq k\sigma) \leq \frac{1}{k^2}$. Let σ_n^2 denote the variance on y_n, then we can write the condition for the present case as $\Pr(|y_n - \mathrm{E}[y_n]| \geq k\sigma_n) \leq \frac{1}{k^2}$. Since $\sigma_n^2 \to 0$ and $\mathrm{E}[y_n] \to c$ the probability will be less than $\frac{1}{k^2}$ for sufficiently large n for any choice of k. But this is just weak convergence in probability to c.

4.1.3 Orders In Probability

Definition 4.7 Let $y_1, y_2, \ldots, y_n, \ldots$ be a sequence of random variables. This sequence is said to be *bounded in probability* if for any $1 > \delta > 0$, there exist a $\Delta < \infty$ and some N sufficiently large such that

$$\Pr(|\, y_n\,| > \Delta\,) < \delta,$$

for all $n > N$. $\Box$

These conditions require that the tail behavior of the distributions of the sequence not be pathological. Specifically, the tail mass of the distributions cannot be drifting away from zero as we move out in the sequence.

Definition 4.8 The sequence of random variables $\{\, y_n\,\}$ is said to be *at most of order in probability* n^λ, and is denoted $\mathrm{O}_p(\, n^\lambda\,)$, if $n^{-\lambda} y_n$ is bounded in probability. $\Box$

Example 4.3 Suppose $z \sim \mathrm{N}(0,1)$ and $y_n = 3 + n \cdot z$, then $n^{-1} y_n = 3/n + z$ is a bounded random variable since the first term is asymptotically negligible and we see that $y_n = \mathrm{O}_p(\, n\,)$.

Definition 4.9 The sequence of random variables $\{ y_n \}$ is said to be *of order in probability smaller than* n^λ, and is denoted $\mathrm{o}_p(n^\lambda)$, if $n^{-\lambda} y_n \xrightarrow{p} 0$. □

Example 4.4 Convergence in probability can be represented in terms of order in probability. Suppose that $y_n \xrightarrow{p} c$ or equivalently $y_n - c \xrightarrow{p} 0$, then $n^0 (y_n - c) \xrightarrow{p} 0$ and $y_n - c = \mathrm{o}_p(1)$.

4.1.4 Convergence In Distribution

Definition 4.10 A sequence of random variables $y_1, y_2, \ldots , y_n, \ldots$, with cumulative distribution functions $F_1(\cdot), F_2(\cdot), \ldots , F_n(\cdot), \ldots$, is said to *converge in distribution* to a random variable y with a cumulative distribution function $F(y)$, if

$$\lim_{n \to \infty} F_n(\cdot) = F(\cdot), \tag{4.6}$$

for every point of continuity of $F(\cdot)$. The distribution $F(\cdot)$ is said to be the *limiting distribution* of this sequence of random variables. □

For notational convenience, we often write $y_n \xrightarrow{d} F(\cdot)$ or $y_n \to_d F(\cdot)$ if a sequence of random variables converges in distribution to $F(\cdot)$. Note that the moments of elements of the sequence do not necessarily converge to the moments of the limiting distribution.

4.1.5 Some Useful Propositions

In the following propositions, let $\mathbf{x}_n$ and $\mathbf{y}_n$ be sequences of random vectors.

Proposition 4.1 *If $\mathbf{x}_n - \mathbf{y}_n$ converges in probability to zero, and $\mathbf{y}_n$ has a limiting distribution, then $\mathbf{x}_n$ has a limiting distribution, which is the same.* □

Proposition 4.2 *If $\mathbf{y}_n$ has a limiting distribution and $\operatorname*{plim}_{n \to \infty} \mathbf{x}_n = 0$, then for $\mathbf{z_n} = {\mathbf{y}_n}' \mathbf{x}_n$,*

$$\operatorname*{plim}_{n \to \infty} \mathbf{z}_n = 0. \qquad \square$$

Proposition 4.3 *Suppose that $\mathbf{y}_n$ converges in distribution to a random variable $\mathbf{y}$, and $\operatorname*{plim}_{n \to \infty} \mathbf{x}_n = \mathbf{c}$. Then ${\mathbf{x}_n}' \mathbf{y}_n$ converges in distribution to $\mathbf{c}' y$.* □

Proposition 4.4 *If $g(\cdot)$ is a continuous function, and if $\mathbf{x}_n - \mathbf{y}_n$ converges in probability to zero, then $g(\mathbf{x}_n) - g(\mathbf{y}_n)$ converges in probability to zero.* □

Proposition 4.5 *If $g(\cdot)$ is a continuous function, and if $\mathbf{x}_n$ converges in probability to a constant $\mathbf{c}$, then $\mathbf{z}_n = g(\mathbf{x}_n)$ converges in distribution to the constant $g(\mathbf{c})$.* □

Proposition 4.6 *If $g(\cdot)$ is a continuous function, and if $\mathbf{x}_n$ converges in distribution to a random variable $\mathbf{x}$, then $\mathbf{z}_n = g(\mathbf{x}_n)$ converges in distribution to a random variable $g(\mathbf{x})$.* □

4.2 Estimation Theory

4.2.1 Properties Of Estimators

Definition 4.11 An *estimator* $\widehat{\boldsymbol{\theta}}_n$ of the $p \times 1$ parameter vector $\boldsymbol{\theta}$ is a function of the sample observations $x_1, x_2, ..., x_n$. □

It follows that $\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2, \ldots, \widehat{\boldsymbol{\theta}}_n$ form a sequence of random variables.

Definition 4.12 The estimator $\widehat{\boldsymbol{\theta}}_n$ is said to be *unbiased* if $\mathrm{E}\widehat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}$, for all n. □

Definition 4.13 The estimator $\widehat{\boldsymbol{\theta}}_n$ is said to be *asympotically unbiased* if $\lim\limits_{n\to\infty} \mathrm{E}\widehat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}$. □

Note that an estimator can be biased in finite samples, but asymptotically unbiased.

Definition 4.14 The estimator $\widehat{\boldsymbol{\theta}}_n$ is said to be *consistent* if $\operatorname*{plim}\limits_{n\to\infty} \widehat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}$. □

Consistency neither implies nor is implied by asymptotic unbiasedness, as demonstrated by the following examples.

Example 4.5 Let

$$\widetilde{\theta}_n = \begin{cases} \theta, \ \textit{with probability}\ 1 - 1/n \\ \theta + nc, \ \textit{with probability}\ 1/n \end{cases}$$

We have $\mathrm{E}\,\widetilde{\theta}_n = \theta + c$, so $\widetilde{\theta}_n$ is a biased estimator, and $\lim_{n\to\infty} \mathrm{E}\,\widetilde{\theta}_n = \theta + c$, so $\widetilde{\theta}_n$ is asymptotically biased as well. However, $\lim_{n\to\infty} \Pr(|\,\widetilde{\theta}_n - \theta\,| > \epsilon\,) = 0$, so $\widetilde{\theta}_n$ is a consistent estimator. □

Example 4.6 Suppose $x_i \sim i.i.d.\, \mathrm{N}(\mu, \sigma^2)$ for $i = 1, 2, ..., n, ...$ and let $\widetilde{x}_n = x_n$ be an estimator of μ. Now $\mathrm{E}[\widetilde{x}_n] = \mu$ *so t*he estimator is unbiased but

$$\Pr(|\widetilde{x}_n - \mu| > 1.96\sigma) = .05$$

so the probability mass is not collapsing about the target point μ so the estimator is not consistent.

4.2.2 Laws Of Large Numbers And Central Limit Theorems

Most of the large sample properties of he estimators considered in the sequel derive from the fact the estimators involve sample averages and the asymptotic behavior of averages is well studies. In addition to the central limit theorems presented in the previous two chapters we have the following two laws of large numbers:

Theorem 4.1 *If $x_1, x_2, \ldots, x_n$ is a simple random sample, that is, the x_i's are i.i.d., and $\mathrm{E}\, x_i = \mu$ and $\mathrm{Var}(\, x_i\,) = \sigma^2$, then by Chebyshev's Inequality,*

$$\operatorname*{plim}_{n\to\infty} \overline{x}_n = \mu \qquad \square \tag{4.7}$$

Theorem 4.2 (Khitchine) *Suppose that $x_1, x_2, \ldots, x_n$ are i.i.d. random variables, such that for all $i = 1, \ldots, n$, $\mathrm{E}\, x_i = \mu$, then,*

$$\operatorname*{plim}_{n\to\infty} \overline{x}_n = \mu \qquad \square \tag{4.8}$$

Both of these results apply element-by-element to vectors of estimators. For sake of completeness we repeat the following scalar central linit theorem.

Theorem 4.3 (Linberg-Levy) *Suppose that $x_1, x_2, \ldots, x_n$ are i.i.d. random variables, such that for all $i = 1, \ldots, n$, $\mathrm{E}\, x_i = \mu$ and $\mathrm{Var}(\, x_i\,) = \sigma^2$, then $\sqrt{n}(\,\overline{x}_n - \mu\,) \xrightarrow{d} \mathrm{N}(\,0, \sigma^2\,)$, or*

$$\begin{aligned}\lim_{n\to\infty} f(\,\overline{x}_n - \mu\,) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{1}{2\sigma^2}(\,\overline{x}_n - \mu\,)^2} \\ &= \mathrm{N}(\,0, \sigma^2\,) \qquad \square \end{aligned} \tag{4.9}$$

This result is easily generalized to obtain the multivariate version given in Theorem 3.3.

Theorem 4.4 (Multivariate CLT) *Suppose that $m \times 1$ random vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are (i) jointly i.i.d., (ii) $\mathrm{E}\, \mathbf{x}_i = \mu$, and (iii) $\mathrm{Cov}(\, \mathbf{x}_i\,) = \Sigma$, then*

$$\sqrt{n}(\,\overline{\mathbf{x}}_n - \boldsymbol{\mu}\,) \xrightarrow{d} \mathrm{N}(\,\mathbf{0}, \Sigma\,) \tag{4.10}$$

4.2.3 CUAN And Efficiency

Definition 4.15 An estimator is said to be *consistently uniformly asymptotically normal (CUAN)* if it is consistent, and if $\sqrt{n}(\,\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\,)$ converges in distribution to $N(\,\mathbf{0}, \Psi\,)$, and if the convergence is uniform over some compact subset of the parameter space. □

Suppose that $\sqrt{n}(\,\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\,)$ converges in distribution to $\mathrm{N}(\,\mathbf{0}, \Psi\,)$. Let $\widetilde{\boldsymbol{\theta}}_n$ be an alternative estimator such that $\sqrt{n}(\,\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\,)$ converges in distribution to $\mathrm{N}(\,\mathbf{0}, \Omega\,)$.

Definition 4.16 If $\widehat{\boldsymbol{\theta}}_n$ is CUAN with asymptotic covariance Ψ and $\widetilde{\boldsymbol{\theta}}_n$ is CUAN with asymptotic covariance Ω, then $\widehat{\boldsymbol{\theta}}_n$ is *asymptotically efficient relative to* $\widetilde{\boldsymbol{\theta}}_{\mathrm{n}}$ if $\Psi{-}\Omega$ is a positive semidefinite matrix. □

Among other properties asymptotic relative efficiency implies that the diagonal elements of Ψ are no larger than those of Ω, so the asymptotic variances of $\widehat{\theta}_{n,i}$ are no larger than those of $\widetilde{\theta}_{n,i}$. And a similar result applies for the asymptotic variance of any linear combination.

Definition 4.17 A CUAN estimator $\widehat{\boldsymbol{\theta}}_n$ is said to be *asymptotically efficient* if it is asymptotically efficient relative to any other CUAN estimator. □

4.3 Asymptotic Inference

4.3.1 Normal Ratios

Now, under the conditions of the central limit theorem, $\sqrt{n}(\,\overline{x}_n - \mu\,) \xrightarrow{d} \mathrm{N}(\,0, \sigma^2\,)$, so

$$\frac{(\,\overline{x}_n - \mu\,)}{\sqrt{\sigma^2/n}} \xrightarrow{d} \mathrm{N}(\,0,1\,) \tag{4.11}$$

Suppose that $\widehat{\sigma^2}$ is a consistent estimator of σ^2. Then we also have

$$\begin{aligned} \frac{(\,\overline{x}_n - \mu\,)}{\sqrt{\widehat{\sigma}^2/n}} &= \frac{\sqrt{\sigma^2/n}}{\sqrt{\widehat{\sigma}^2/n}}\frac{(\,\overline{x}_n - \mu\,)}{\sqrt{\sigma^2/n}} \\ &= \sqrt{\frac{\sigma^2}{\widehat{\sigma}^2}}\frac{(\,\overline{x}_n - \mu\,)}{\sqrt{\sigma^2/n}} \xrightarrow{d} \mathrm{N}(\,0,1\,) \end{aligned} \tag{4.12}$$

since the term under the square root converges in probability to one and the remainder converges in distribution to $\mathrm{N}(\,0,1)$.

Most typically, such ratios will be used for inference in testing a hypothesis. Now, for $\mathrm{H}_0 : \mu = \mu_0$, we have

$$\frac{(\,\overline{x}_n - \mu_0\,)}{\sqrt{\widehat{\sigma}^2/n}} \xrightarrow{d} \mathrm{N}(\,0,1\,), \tag{4.13}$$

while under $\mathrm{H}_1 : \mu = \mu_1 \neq \mu_0$, we find that

$$\begin{aligned} \frac{(\,\overline{x}_n - \mu_0\,)}{\sqrt{\widehat{\sigma}^2/n}} &= \frac{\sqrt{n}(\,\overline{x}_n - \mu_1\,)}{\sqrt{\widehat{\sigma}^2}} + \frac{\sqrt{n}(\,\mu_1 - \mu_0\,)}{\sqrt{\widehat{\sigma}^2}} \\ &= \mathrm{N}(\,0,1\,) + \mathrm{O}_p(\,\sqrt{n}\,) \end{aligned} \tag{4.14}$$

Thus, extreme values of the ratio can be taken as rare events under the null or typical events under the alternative.

Such ratios are of interest in estimation and inference with regard to more general parameters. Suppose that $\boldsymbol{\theta}$ is the parameter vector

$$\sqrt{n}(\underset{p\times 1}{\widehat{\boldsymbol{\theta}}} - \boldsymbol{\theta}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \underset{p\times p}{\Psi}).$$

Then, if θ_i is the parameter of particular interest we consider

$$\frac{(\widehat{\theta}_i - \theta_i)}{\sqrt{\psi_{ii}/n}} \xrightarrow{d} \mathrm{N}(0,1), \tag{4.15}$$

and duplicating the arguments for the sample mean

$$\frac{(\widehat{\theta}_i - \theta_i)}{\sqrt{\widehat{\psi}_{ii}/n}} \xrightarrow{d} \mathrm{N}(0,1), \tag{4.16}$$

where $\widehat{\Psi}$ is a consistent estimator of Ψ. This ratio will have a similar behavior under a null and alternative hypotesis with regard to θ_i.

4.3.2 Asymptotic Chi-Square

Suppose that

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \Psi),$$

where $\widehat{\Psi}$ is a consistent estimator of the nonsingular $p\times p$ matrix Ψ. Then, from the previous chapter we have

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\Psi^{-1}\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = n\cdot(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\Psi^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \chi^2_p \tag{4.17}$$

and

$$n\cdot(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\widehat{\Psi}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \chi^2_p \tag{4.18}$$

for $\widehat{\Psi}$ a consistent estimator of Ψ.

This result can be used to conduct infence by testing the entire parmater vector. If $\mathrm{H}_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^0$, then

$$n\cdot(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)'\widehat{\Psi}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) \xrightarrow{d} \chi^2_p, \tag{4.19}$$

and large positive values are rare events. while for $\mathrm{H}_1 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^1 \neq \boldsymbol{\theta}_1^0$, we can show (later)

$$\begin{aligned} n\cdot(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)'\widehat{\Psi}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) &= n\cdot((\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^1) + (\boldsymbol{\theta}^1 - \boldsymbol{\theta}^0))'\Psi^{-1}((\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^1) + (\boldsymbol{\theta}^1 - \boldsymbol{\theta}^0)) \\ &= n\cdot(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^1)'\widehat{\Psi}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^1) + 2n\cdot(\boldsymbol{\theta}^1 - \boldsymbol{\theta}^0)'\Psi^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^1) \\ &\qquad + n\cdot(\boldsymbol{\theta}^1 - \boldsymbol{\theta}^0)'\widehat{\Psi}^{-1}(\boldsymbol{\theta}^1 - \boldsymbol{\theta}^0) \\ &= \chi^2_p + \mathrm{O}_p(\sqrt{n}) + \mathrm{O}_p(n) \end{aligned} \tag{4.20}$$

Thus, if we obtain a large value of the statistic, we may take it as evidence that the null hypothesis is incorrect.

This result can also be applied to any subvector of $\boldsymbol{\theta}$. Let

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{pmatrix},$$

where $\boldsymbol{\theta}_1$ is a $p_1 \times 1$ vector. Then

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) \xrightarrow{d} \mathrm{N}(0, \Psi_{11}), \tag{4.21}$$

where Ψ_{11} is the upper left-hand $q \times q$ submatrix of Ψ and,

$$n \cdot (\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1)' \widehat{\Psi}_{11}^{-1} (\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) \xrightarrow{d} \chi^2_{p_1} \tag{4.22}$$

4.3.3 Tests Of General Restrictions

We can use similar results to test general nonlinear restrictions. Suppose that $\mathbf{r}(\cdot)$ is a $q \times 1$ continuously differentiable function, and

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \Psi).$$

By the intermediate value theorem we can obtain the exact Taylor's series representation

$$\mathbf{r}(\widehat{\boldsymbol{\theta}}) = \mathbf{r}(\boldsymbol{\theta}) + \frac{\partial \mathbf{r}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}'}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

or equivalently

$$\begin{aligned} \sqrt{n}(\mathbf{r}(\widehat{\boldsymbol{\theta}}) - \mathbf{r}(\boldsymbol{\theta})) &= \frac{\partial \mathbf{r}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}'} \sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \\ &= \mathbf{R}(\boldsymbol{\theta}^*) \sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \end{aligned}$$

where $\mathbf{R}(\boldsymbol{\theta}^*) = \frac{\partial \mathbf{r}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}'}$ and $\boldsymbol{\theta}^*$ lies between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$. Now $\widehat{\boldsymbol{\theta}} \to_p \boldsymbol{\theta}$ so $\boldsymbol{\theta}^* \to_p \boldsymbol{\theta}$ and $\mathbf{R}(\boldsymbol{\theta}^*) \to \mathbf{R}(\boldsymbol{\theta})$. Thus, we have

$$\sqrt{n}[\mathbf{r}(\widehat{\boldsymbol{\theta}}) - \mathbf{r}(\boldsymbol{\theta})] \xrightarrow{d} \mathrm{N}(0, \mathbf{R}(\boldsymbol{\theta}) \Psi \mathbf{R}'(\boldsymbol{\theta})). \tag{4.23}$$

Thus, under $\mathrm{H}_0 : \mathbf{r}(\boldsymbol{\theta}) = 0$, assuming $\mathbf{R}(\boldsymbol{\theta}) \Psi \mathbf{R}'(\boldsymbol{\theta})$ is nonsingular, we have

$$n \cdot \mathbf{r}(\widehat{\boldsymbol{\theta}})' [\mathbf{R}(\boldsymbol{\theta}) \Psi \mathbf{R}'(\boldsymbol{\theta})]^{-1} \mathbf{r}(\widehat{\boldsymbol{\theta}}) \xrightarrow{d} \chi^2_q, \tag{4.24}$$

where q is the length of $\mathbf{r}(\cdot)$. In practice, we substitute the consistent estimates $\mathbf{R}(\widehat{\boldsymbol{\theta}})$ for $\mathbf{R}(\boldsymbol{\theta})$ and $\widehat{\Psi}$ for Ψ to obtain, following the arguments given above

$$n \cdot \mathbf{r}(\widehat{\boldsymbol{\theta}})' [\mathbf{R}(\widehat{\boldsymbol{\theta}}) \widehat{\Psi} \mathbf{R}'(\widehat{\boldsymbol{\theta}})]^{-1} \mathbf{r}(\widehat{\boldsymbol{\theta}}) \xrightarrow{d} \chi^2_q, \tag{4.25}$$

The behavior under the alternative hypothesis will be $\mathrm{O}_p(n)$ as above.

Chapter 5

Maximum Likelihood Methods

5.1 Maximum Likelihood Estimation (MLE)

5.1.1 Motivation

Suppose we have a model for the random variable y_i, for $i = 1, 2, \ldots, n$, with unknown $(p \times 1)$ parameter vector $\boldsymbol{\theta}$. In many cases, the model will imply a distribution $f(\,y_i, \boldsymbol{\theta}\,)$ for each realization of the variable y_i.

A basic premise of statistical inference is to avoid unlikely or rare models, for example, in hypothesis testing. If we have a realization of a statistic that exceeds the critical value then it is a rare event under the null hypothesis. Under the alternative hypothesis, however, such a realization is much more likely to occur and we reject the null in favor of the alternative. Thus in choosing between the null and alternative, we select the model that makes the realization of the statistic more likely to have occured.

Carrying this idea over to estimation, we select values of $\boldsymbol{\theta}$ such that the corresponding values of $f(\,y_i, \boldsymbol{\theta}\,)$ are not unlikely. After all, we do not want a model that disagrees strongly with the data. Maximum likelihood estimation is merely a formalization of this notion that the model chosen should not be unlikely. Specifically, we choose the values of the parameters that make the realized data most likely to have occured. This approach does, however, require that the model be specified in enough detail to imply a distribution for the variable of interest.

5.1.2 The Likelihood Function

Suppose that the random variables $y_1, y_2, \ldots, y_n$ are *i.i.d.* Then, the joint density function for n realizations is

$$\begin{aligned} f(\, y_1, y_2, \ldots, y_n|\boldsymbol{\theta}\,) &= f(\, y_1|\boldsymbol{\theta}\,) \cdot f(\, y_2, |\boldsymbol{\theta}\,) \cdot \ldots \cdot f(\, y_n|\boldsymbol{\theta}\,) \\ &= \prod_{i=1}^{n} f(\, y_i|\boldsymbol{\theta}\,) \end{aligned} \tag{5.1}$$

Given values of the parameter vector $\boldsymbol{\theta}$, this function allows us to assign local probability measures for various choices of the random variables $y_1, y_2, \ldots, y_n$. This is the function which must be integrated to make probability statements concerning the joint outcomes of $y_1, y_2, \ldots, y_n$.

Given a set of realized values of the random variables, we use this same function to establish the probability measure associated with various choices of the parameter vector $\boldsymbol{\theta}$.

Definition 5.1 The *likelihood function* of the parameters, for a particular sample of $y_1, y_2, \ldots, y_n$, is the joint density function considered as a function of $\boldsymbol{\theta}$ given the y_i's. That is,

$$L(\, \boldsymbol{\theta}|y_1, y_2, \ldots, y_n\,) = \prod_{i=1}^{n} f(\, y_i|\boldsymbol{\theta}\,) \qquad \square \tag{5.2}$$

5.1.3 Maximum Likelihood Estimation

For a particular choice of the parameter vector $\boldsymbol{\theta}$, the likelihood function gives a probability measure for the realizations that occured. Consistent with the approach used in hypothesis testing, and using this function as the metric, we choose $\boldsymbol{\theta}$ that make the realizations most likely to have occured.

Definition 5.2 The *maximum likelihood estimator* of $\boldsymbol{\theta}$ is the estimator obtained by maximizing $L(\, \boldsymbol{\theta}|y_1, y_2, \ldots, y_n\,)$ with respect to $\boldsymbol{\theta}$. That is,

$$\max_{\boldsymbol{\theta}} L(\, \boldsymbol{\theta}|y_1, y_2, \ldots, y_n\,) = \widehat{\boldsymbol{\theta}}, \tag{5.3}$$

where $\widehat{\boldsymbol{\theta}}$ is called the MLE of $\boldsymbol{\theta}$. □

Equivalently, since $\log(\cdot)$ is a strictly monotonic transformation, we may find the MLE of $\boldsymbol{\theta}$ by solving

$$\max_{\boldsymbol{\theta}} \mathcal{L}(\, \boldsymbol{\theta}|y_1, y_2, \ldots, y_n\,), \tag{5.4}$$

where

$$\mathcal{L}(\, \boldsymbol{\theta}|y_1, y_2, \ldots, y_n\,) = \log \mathrm{L}(\, \boldsymbol{\theta}|y_1, y_2, \ldots, y_n\,)$$

is denoted the log-likelihood function. In practice, we obtain $\widehat{\boldsymbol{\theta}}$ by solving the first-order conditions (FOC)

$$\frac{\partial \mathcal{L}(\, \boldsymbol{\theta}; y_1, y_2, \ldots, y_n \,)}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{\partial \log f(\, y_i; \widehat{\boldsymbol{\theta}} \,)}{\partial \boldsymbol{\theta}} = 0.$$

The motivation for using the log-likelihood function is apparent since the summation form will result, after division by n, in estimators that are approximately averages, about which we know a lot. This advantage is particularly clear in the folowing example.

Example 5.1 Suppose that y_i ∼i.i.d.$N(\, \mu, \sigma^2 \,)$, for $i = 1, 2, \ldots, n$. Then,

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

for $i = 1, 2, \ldots, n$. Using the likelihood function (5.2), we have

$$L(\, \mu, \sigma^2 | y_1, y_2, \ldots, y_n \,) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}\,(\, y_i - \mu \,)^2}. \tag{5.5}$$

Next, we take the logarithm of (5.5), which gives us

$$\begin{aligned} \log L(\, \mu, \sigma^2 | y_1, y_2, \ldots, y_n \,) &= \sum_{i=1}^{n} \left[-\frac{1}{2} \log(\, 2\pi\sigma^2 \,) - \frac{1}{2\sigma^2} (\, y_i - \mu \,)^2 \right] \\ &= -\frac{n}{2} \log(\, 2\pi\sigma^2 \,) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\, y_i - \mu \,)^2 \qquad (5.6) \end{aligned}$$

We then maximize (5.6) with respect to both μ and σ^2. That is, we solve the following first order conditions:

(A) $\frac{\partial \log L(\cdot)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (\, y_i - \mu \,) = 0;$

(B) $\frac{\partial \log L(\cdot)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (\, y_i - \mu \,)^2 = 0.$

By solving (A), we find that $\mu = \frac{1}{n} \sum_{i=1}^{n} y_i = \overline{y}_n$. Solving (B) gives us $\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (\, y_i - \overline{y}_n \,)$. Therefore, $\widehat{\mu} = \overline{y}_n$, and $\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (\, y_i - \widehat{\mu} \,)$. □

Note that $\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (\, y_i - \widehat{\mu} \,) \neq s^2$, where $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\, y_i - \widehat{\mu} \,)$. s^2 is the familiar unbiased estimator for σ^2, and $\widehat{\sigma}^2$ is a biased estimator.

5.2 Asymptotic Behavior of MLEs

5.2.1 Assumptions

For the results we will derive in the following sections, we need to make five assumptions:

1. The y_i's are iid random variables with density function $f(\, y_i,\boldsymbol{\theta}\,)$ for $i = 1, 2, \ldots , n$;

2. $\log f(\, y_i,\boldsymbol{\theta}\,)$ and hence $f(\, y_i,\boldsymbol{\theta}\,)$ possess derivatives with respect to $\boldsymbol{\theta}$ up to the third order for $\boldsymbol{\theta} \in \Theta$;

3. The range of y_i is independent of $\boldsymbol{\theta}$ hence differentiation under the integral is possible;

4. The parameter vector $\boldsymbol{\theta}$ is globally identified by the density function.

5. $\partial^3 \log f(\, y_i, \boldsymbol{\theta}\,)/\partial\theta_i\partial\theta_j\partial\theta_k$ is bounded in absolute value by some function $\mathrm{H}_{ijk}(\, y\,)$ for all y and $\boldsymbol{\theta} \in\theta$, which, in turn, has a finite expectation for all $\boldsymbol{\theta} \in \Theta$.

The first assumption is fundamental and the basis of the estimator. If it is not satisfied then we are misspecifying the model and there is little hope for obtaining correct inferences, at least in finite samples. The second assumption is a regularity condition that is usually satisfied and easily verified. The third assumption is also easily verified and guarateed to be satisfied in models where the dependent variable has smooth and infinite support. The fourth assumption must be verified, which is easier in some cases than others. The last assumption is crucial and bears a cost and really should be verified before MLE is undertaken but is usually ignored.

5.2.2 Some Preliminaries

Now, we know that

$$\int \mathrm{L}(\, \boldsymbol{\theta}^0|\mathbf{y}\,)d\mathbf{y} = \int f(\, \mathbf{y}|\boldsymbol{\theta}^0\,)d\mathbf{y} = 1 \tag{5.7}$$

for any value of the true parameter vector $\boldsymbol{\theta}^0$. Therefore,

$$\begin{aligned} 0 &= \frac{\partial \int \mathrm{L}(\boldsymbol{\theta}^0|\mathbf{y})d\mathbf{y}}{\partial\boldsymbol{\theta}} \\ &= \int \frac{\partial \mathrm{L}(\boldsymbol{\theta}^0|\mathbf{y})}{\partial\boldsymbol{\theta}} d\mathbf{y} \\ &= \int \frac{\partial f(\mathbf{y}|\boldsymbol{\theta}^0)}{\partial\boldsymbol{\theta}} d\mathbf{y} \\ &= \int \frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta}^0)}{\partial\boldsymbol{\theta}} f(\mathbf{y}|\boldsymbol{\theta}^0) d\mathbf{y}, \end{aligned} \tag{5.8}$$

and

$$\begin{aligned} 0 &= \mathrm{E}\left[\frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta}^0)}{\partial\boldsymbol{\theta}}\right] \\ &= \mathrm{E}\left[\frac{\partial \log \mathrm{L}(\boldsymbol{\theta}^0|\mathbf{y})}{\partial\boldsymbol{\theta}}\right] \end{aligned} \tag{5.9}$$

for any value of the true parameter vector $\boldsymbol{\theta}^0$.

Differentiating (5.8) again yields

$$0 = \int \left[\frac{\partial^2 \log f(\mathbf{y}|\boldsymbol{\theta}^0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} f(\mathbf{y}|\boldsymbol{\theta}^0) + \frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta}^0)}{\partial\boldsymbol{\theta}} \frac{\partial f(\mathbf{y}|\boldsymbol{\theta}^0)}{\partial\boldsymbol{\theta}'}\right] d\mathbf{y}. \tag{5.10}$$

Since

$$\frac{\partial f(\mathbf{y}|\boldsymbol{\theta}^0)}{\partial\boldsymbol{\theta}'} = \frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta}^0)}{\partial\boldsymbol{\theta}'} f(\mathbf{y}|\boldsymbol{\theta}^0), \tag{5.11}$$

then we can rewrite (5.10) as

$$0 = \mathrm{E}\left[\frac{\partial^2 \log f(\mathbf{y}|\boldsymbol{\theta}^0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] + \mathrm{E}\left[\frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta}^0)}{\partial\boldsymbol{\theta}} \frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta}^0)}{\partial\boldsymbol{\theta}'}\right]. \tag{5.12}$$

or, in terms of the likelihood function,

$$\vartheta(\boldsymbol{\theta}^0) = \mathrm{E}\left[\frac{\partial \log \mathrm{L}(\boldsymbol{\theta}^0|\mathbf{y})}{\partial\boldsymbol{\theta}} \frac{\partial \log \mathrm{L}(\boldsymbol{\theta}^0|\mathbf{y})}{\partial\boldsymbol{\theta}'}\right] = -\mathrm{E}\left[\frac{\partial \log \mathrm{L}(\boldsymbol{\theta}^0|\mathbf{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]. \tag{5.13}$$

The matrix $\vartheta(\boldsymbol{\theta}^0)$ is called the *information matrix* and the relationship given in (5.13) the *information matrix equality.*

Finally, we note that

$$\begin{aligned} \mathrm{E}\left[\frac{\partial \log \mathrm{L}(\boldsymbol{\theta}^0|\mathbf{y})}{\partial\boldsymbol{\theta}} \frac{\partial \log \mathrm{L}(\boldsymbol{\theta}^0|\mathbf{y})}{\partial\boldsymbol{\theta}'}\right] &= \mathrm{E}\left[\sum_{i=1}^{n} \frac{\partial \log f_i(\mathbf{y}|\boldsymbol{\theta}^0)}{\partial\boldsymbol{\theta}} \sum_{i=1}^{n} \frac{\partial \log f_i(\mathbf{y}|\boldsymbol{\theta}^0)}{\partial\boldsymbol{\theta}'}\right] \\ &= \sum_{i=1}^{n} \mathrm{E}\left[\frac{\partial \log f_i(\mathbf{y}|\boldsymbol{\theta}^0)}{\partial\boldsymbol{\theta}} \frac{\partial \log f_i(\mathbf{y}|\boldsymbol{\theta}^0)}{\partial\boldsymbol{\theta}'}\right]. \end{aligned} \tag{5.14}$$

since the covariances between different observations is zero.

5.2.3 Asymptotic Properties

Consistent Root Exists

Consider the case where $p = 1$. Then, expanding in a Taylor's series and using the intermediate value theorem on the quadratic term yields

$$\begin{aligned}\frac{1}{n}\frac{\partial \log \mathrm{L}(\,\theta|\mathbf{y}\,)}{\partial\theta} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log f(\,y_i|\theta^0\,)}{\partial\theta} \\ &\quad +\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(\,y_i|\theta^0\,)}{\partial\theta^2}(\,\widehat{\theta}-\theta^0\,) \\ &\quad +\frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^3 \log f(\,y_i|\theta^*\,)}{\partial\theta^3}(\,\widehat{\theta}-\theta^0\,)^2 \qquad (5.15)\end{aligned}$$

where θ^* lies between $\widehat{\theta}$ and θ^0. Now, by assumption 5, we have

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^3 \log f(\,y_i|\theta^*\,)}{\partial\theta^3} = k\sum_{i=1}^{n}H(\,y_i\,), \qquad (5.16)$$

for some $|k| < 1$. So,

$$\frac{1}{n}\frac{\partial \log \mathrm{L}(\,\theta|\mathbf{y}\,)}{\partial\theta} = a\delta^2 + b\delta + c, \qquad (5.17)$$

where

$$\begin{aligned}\delta &= \widehat{\theta}-\theta^0, \\ a &= \frac{k}{2}\frac{1}{n}\sum_{i=1}^{n}H(\,y_i\,), \\ b &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(\,y_i|\theta^0\,)}{\partial\theta^2}, \text{and} \\ c &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log f(\,y_i|\theta^0\,)}{\partial\theta}.\end{aligned}$$

Note that $|a| \leq \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}H(\,y_i\,) = \frac{1}{2}\,\mathrm{E}[H(\,y_i\,)] + o_p(1) = O_p(1)$, $\operatorname*{plim}_{n\to\infty} c = 0$, and $\operatorname*{plim}_{n\to\infty} b = -\vartheta(\,\theta^0\,)$.

Now, since $\partial \log \mathrm{L}(\,\widehat{\theta}|\mathbf{y}\,)/\partial\theta = 0$, we have $a\delta^2 + b\delta + c = 0$. There are two possibilities. If $a \neq 0$ with probability 1, which will occur when the FOC are nonlinear in a, then

$$\delta = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \tag{5.18}$$

Since $ac = o_p(1)$, then $\delta \xrightarrow{p} 0$ for the plus root while $\delta \xrightarrow{p} \vartheta(\,\theta^0\,)/\alpha$ for the negative root if $\operatorname*{plim}_{n\to\infty} a = \alpha \neq 0$ exists. If $a = 0$, then the FOC are linear in δ whereupon $\delta = -\frac{c}{b}$ and again $\delta \xrightarrow{p} 0$. If the *F.O.C.* are nonlinear but asymptotically linear then $a \xrightarrow{p} 0$ and ac in the numerator of (5.18) will still go to zero faster than a in the denominator and $\delta \xrightarrow{p} 0$. Thus there exits at least one consistent solution $\widehat{\theta}$ which satisfies

$$\operatorname*{plim}_{n\to\infty}(\,\widehat{\theta} - \theta^0\,) = 0. \tag{5.19}$$

and in the asymptotically nonlinear case there is also a possibly inconsistent solution.

For the case of $\boldsymbol{\theta}$ a vector, we can apply a similar style proof to show there exists a solution $\widehat{\boldsymbol{\theta}}$ to the FOC that satisfies $\operatorname*{plim}_{n\to\infty}(\,\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\,) = 0$. And in the event of asymptotically nonlinear FOC there is at least on other possibly inconsistent root.

Global Maximum Is Consistent

In the event of multiple roots, we are left with the problem of selecting between them. By assumption 4, the parameter $\boldsymbol{\theta}$ is globally identified by the density function. Formally, this means that

$$f(\,y, \boldsymbol{\theta}\,) = f(\,y, \boldsymbol{\theta}^0\,), \tag{5.20}$$

for all y implies that $\boldsymbol{\theta} = \boldsymbol{\theta}^0$. Now,

$$\mathrm{E}\left[\frac{f(\,y, \boldsymbol{\theta}\,)}{f(\,y, \boldsymbol{\theta}^0\,)}\right] = \int \frac{f(\,y, \boldsymbol{\theta}\,)}{f(\,y, \boldsymbol{\theta}^0\,)} f(\,y, \boldsymbol{\theta}^0\,) = 1. \tag{5.21}$$

Thus, by Jensen's Inequality, we have

$$\mathrm{E}\left[\log \frac{f(\,y, \boldsymbol{\theta}\,)}{f(\,y, \boldsymbol{\theta}^0\,)}\right] < \log \mathrm{E}\left[\frac{f(\,y, \boldsymbol{\theta}\,)}{f(\,y, \boldsymbol{\theta}^0\,)}\right] = 0, \tag{5.22}$$

unless $f(\,y, \boldsymbol{\theta}\,) = f(\,y, \boldsymbol{\theta}^0\,)$ for all y, or $\boldsymbol{\theta} = \boldsymbol{\theta}^0$. Therefore, $\mathrm{E}\left[\log f(\,y, \boldsymbol{\theta}\,)\right]$ achieves a maximum if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}^0$.

However, we are solving

$$\max \frac{1}{n} \sum_{i=1}^{n} \log f(\,y_i, \boldsymbol{\theta}\,) \xrightarrow{p} \mathrm{E}\left[\log f(\,y_i, \boldsymbol{\theta}\,)\right], \tag{5.23}$$

and if we choose an inconsistent root, we will not obtain a global maximum. Thus, asymptotically, the global maximum is a consistent root. This choice of the global root has added appeal since it is in fact the MLE among the possible alternatives and hence the choice that makes the realized data most likely to have occured.

There are complications in finite samples since the value of the likelihood function for alternative roots may cross over as the sample size increases. That is the global maximum in small samples may not be the global maximum in larger samples. An added problem is to identify all the alternative roots so we can choose the global maximum. Sometimes a solution is available in a simple consistent estimator which may be used to start the nonlinear MLE optimization.

Asymptotic Normality

For $p = 1$, we have $a\delta^2 + b\delta + c = 0$, so

$$\delta = \widehat{\theta} - \theta^0 = \frac{-c}{a\delta + b} \tag{5.24}$$

and

$$\begin{aligned} \sqrt{n}(\,\widehat{\theta} - \theta^0\,) &= \frac{-\sqrt{n}c}{a(\,\widehat{\theta} - \theta^0\,) + b} \\ &= \frac{-1}{a(\,\widehat{\theta} - \theta^0\,) + b}\sqrt{n}c. \end{aligned}$$

Now since $a = O_p(1)$ and $\widehat{\theta} - \theta^0 = o_p(1)$, then $a(\,\widehat{\theta} - \theta^0\,) = o_p(1)$ and

$$a(\,\widehat{\theta} - \theta^0\,) + b \xrightarrow{p} -\vartheta(\,\theta^0\,). \tag{5.25}$$

And by the CLT we have

$$\sqrt{n}c = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f(\,y_i|\theta^0\,)}{\partial\theta} \xrightarrow{d} \mathrm{N}(\,0, \vartheta(\,\theta^0\,)). \tag{5.26}$$

Substituing these two results in (5.25), we find

$$\sqrt{n}(\,\widehat{\theta} - \theta^0\,) \xrightarrow{d} \mathrm{N}(\,0, \vartheta(\,\theta^0\,)^{-1}).$$

In general, for $p > 1$, we can apply the same scalar proof to show $\sqrt{n}(\,\boldsymbol{\lambda}'\widehat{\boldsymbol{\theta}} - \boldsymbol{\lambda}'\boldsymbol{\theta}^0\,) \xrightarrow{d} \mathrm{N}(\,0, \boldsymbol{\lambda}'\boldsymbol{\vartheta}(\,\boldsymbol{\theta}^0\,)^{-1}\boldsymbol{\lambda})$ for any vector $\boldsymbol{\lambda}$, which means

$$\sqrt{n}(\,\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\,) \xrightarrow{d} \mathrm{N}(\,\mathbf{0}, \boldsymbol{\vartheta}^{-1}(\,\boldsymbol{\theta}^0\,)), \tag{5.27}$$

if $\widehat{\boldsymbol{\theta}}$ is the global maximum.

Cramér-Rao Lower Bound

In addition to being the covariance matrix of the MLE, $\boldsymbol{\vartheta}^{-1}(\,\boldsymbol{\theta}^0\,)$ defines a lower bound for covariance matrices with certain desirable properties. Let $\widetilde{\boldsymbol{\theta}}(\,\mathbf{y}\,)$ be any unbiased estimator, then

$$\mathrm{E}\,\widetilde{\boldsymbol{\theta}}(\,\mathbf{y}\,) = \int \widetilde{\boldsymbol{\theta}}(\,\mathbf{y}\,) f(\,\mathbf{y};\boldsymbol{\theta}\,) d\mathbf{y} = \boldsymbol{\theta}, \tag{5.28}$$

for any underlying $\boldsymbol{\theta} = \boldsymbol{\theta}^0$. Differentiating both sides of this relationship with respect to $\boldsymbol{\theta}$ yields

$$\begin{aligned} \mathbf{I}_p &= \frac{\partial\,\mathrm{E}\,\widetilde{\boldsymbol{\theta}}(,\mathbf{y}\,)}{\partial\boldsymbol{\theta}'} \\ &= \int \widetilde{\boldsymbol{\theta}}(\,\mathbf{y}\,)\frac{\partial f(\,\mathbf{y};\boldsymbol{\theta}^0\,)}{\partial\boldsymbol{\theta}'} d\mathbf{y} \\ &= \int \widetilde{\boldsymbol{\theta}}(\,\mathbf{y}\,)\frac{\partial \log f(\,\mathbf{y};\boldsymbol{\theta}^0\,)}{\partial\boldsymbol{\theta}'} f(\,\mathbf{y};\boldsymbol{\theta}^0\,) d\mathbf{y} \\ &= \mathrm{E}\left[\widetilde{\boldsymbol{\theta}}(\,\mathbf{y}\,)\frac{\partial \log f(\,\mathbf{y};\boldsymbol{\theta}^0\,)}{\partial\boldsymbol{\theta}'}\right] \\ &= \mathrm{E}\left[(\,\widetilde{\boldsymbol{\theta}}(\,\mathbf{y}\,) - \boldsymbol{\theta}^0\,)\frac{\partial \log f(\,\mathbf{y};\boldsymbol{\theta}^0\,)}{\partial\boldsymbol{\theta}'}\right]. \end{aligned} \tag{5.29}$$

Next, we let

$$\mathrm{C}(\,\boldsymbol{\theta}^0\,) = \mathrm{E}\left[(\,\widetilde{\boldsymbol{\theta}}(\,\mathbf{y}\,) - \boldsymbol{\theta}^0\,)(\,\widetilde{\boldsymbol{\theta}}(\,\mathbf{y}\,) - \boldsymbol{\theta}^0\,)'\right]. \tag{5.30}$$

be the covariance matrix of $\widetilde{\boldsymbol{\theta}}(\,\mathbf{y}\,)$, then,

$$\mathrm{Cov}\begin{pmatrix} \widetilde{\boldsymbol{\theta}}(\,\mathbf{y}\,) \\ \frac{\partial \log L}{\partial\boldsymbol{\theta}} \end{pmatrix} = \begin{pmatrix} \mathrm{C}(\,\boldsymbol{\theta}^0\,) & \mathbf{I}_p \\ \mathbf{I}_p & \boldsymbol{\vartheta}(\,\boldsymbol{\theta}^0\,) \end{pmatrix}, \tag{5.31}$$

where $\mathbf{I}_p$ is a $p \times p$ identity matrix, and (5.31) as a covariance matrix is positive semidefinite.

Now, for any $(p \times 1)$ vector $\mathbf{a}$, we have

$$\begin{pmatrix} \mathbf{a}' & \mathbf{a}'\vartheta(\,\boldsymbol{\theta}^0\,)^{-1} \end{pmatrix} \begin{pmatrix} \mathrm{C}(\,\boldsymbol{\theta}^0\,) & \mathbf{I}_p \\ \mathbf{I}_p & \boldsymbol{\vartheta}(\,\boldsymbol{\theta}^0\,), \end{pmatrix} \begin{pmatrix} \mathbf{a}' \\ \mathbf{a}'\boldsymbol{\vartheta}(\,\boldsymbol{\theta}^0\,)^{-1} \end{pmatrix} = \mathbf{a}'[\,\mathrm{C}(\,\boldsymbol{\theta}^0\,) - \boldsymbol{\vartheta}(\,\boldsymbol{\theta}^0\,)^{-1}\,]\mathbf{a} \geq 0. \tag{5.32}$$

Thus, any unbiased estimator $\widetilde{\boldsymbol{\theta}}(\,\mathbf{y}\,)$ has a covariance matrix that exceeds $\boldsymbol{\vartheta}(\,\boldsymbol{\theta}^0\,)^{-1}$ by a positive semidefinite matrix. And if the MLE estimator is unbiased, it is efficient within the class of unbiased estimators. Likewise, any CUAN estimator will have a covariance exceeding $\boldsymbol{\vartheta}(\,\boldsymbol{\theta}^0\,)^{-1}$. Since the asymptotic covariance of MLE is, in fact, $\boldsymbol{\vartheta}(\,\boldsymbol{\theta}^0\,)^{-1}$, it is efficient (asymptotically). $\boldsymbol{\vartheta}(\,\boldsymbol{\theta}^0\,)^{-1}$ is called the Camér-Rao lower bound.

5.3 Maximum Likelihood Inference

5.3.1 Likelihood Ratio Test

Suppose we wish to test $\mathrm{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}^0$ against $\mathrm{H}_0 : \boldsymbol{\theta} \neq \boldsymbol{\theta}^0$. Then, we define

$$L_u = \max_{\boldsymbol{\theta}} L(\,\boldsymbol{\theta}|\mathbf{y}\,) = L(\,\widehat{\boldsymbol{\theta}}|\mathbf{y}\,) \tag{5.33}$$

and

$$L_r = L(\,\boldsymbol{\theta}^0|\mathbf{y}\,), \tag{5.34}$$

where L_u is the unrestricted likelihood and L_r is the restricted likelihood. We then form the likelihood ratio

$$\lambda = \frac{L_r}{L_u}. \tag{5.35}$$

Note that the restricted likelihood can be no larger than the unrestricted which maximizes the function.

As with estimation, it is more convenient to work with the logs of the likelihood functions. It will be shown below that, under H_0,

$$\begin{aligned} \mathrm{LR} &= -2\log\lambda \\ &= -2\left[\log\frac{L_r}{L_u}\right] \\ &= 2[\,\mathcal{L}(\,\widehat{\boldsymbol{\theta}}|\mathbf{y}\,) - \mathcal{L}(\,\boldsymbol{\theta}^0|\mathbf{y}\,)] \xrightarrow{d} \chi^2_p, \end{aligned} \tag{5.36}$$

where $\widehat{\boldsymbol{\theta}}$ is the unrestricted MLE, and $\boldsymbol{\theta}^0$ is the restricted MLE. If H_1 applies, then $\mathrm{LR} = O_p(\,n\,)$. Large values of this statistic indicate that the restrictions make the observed values much less likely than the unrestricted and we prefer the unrestricted and reject the restictions.

In general, for $\mathrm{H}_0 : \mathbf{r}(\,\boldsymbol{\theta}\,) = 0$, and $\mathrm{H}_1 : \mathbf{r}(\,\boldsymbol{\theta}\,) \neq 0$, we have

$$L_u = \max_{\boldsymbol{\theta}} L(\,\boldsymbol{\theta}\ \mathbf{y}\,) = L(\,\widehat{\boldsymbol{\theta}}\,|\,\mathbf{y}\,), \tag{5.37}$$

and

$$\begin{aligned} L_r &= \max_{\boldsymbol{\theta}} L(\,\boldsymbol{\theta}\ \mathbf{y}\,) \text{ s.t. } \mathbf{r}(\,\boldsymbol{\theta}\,) = 0 \\ &= L(\,\widetilde{\boldsymbol{\theta}}\,|\,\mathbf{y}\,). \end{aligned} \tag{5.38}$$

Under H_0,

$$\mathrm{LR} = 2[\,\mathcal{L}(\,\widehat{\boldsymbol{\theta}}|\mathbf{y}\,) - \mathcal{L}(\,\widetilde{\boldsymbol{\theta}}|\mathbf{y}\,)] \xrightarrow{d} \chi^2_q, \tag{5.39}$$

where q is the length of $\mathbf{r}(\cdot)$.

Note that in the general case, the likelihood ratio test requires calculation of both the restricted and the unrestricted MLE.

5.3.2 Wald Test

The asymptotic normality of MLE may be used to obtain a test based only on the unrestricted estimates.

Now, under $\mathrm{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}^0$, we have

$$\sqrt{n}(\,\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\,) \xrightarrow{d} \mathrm{N}(\,0, \vartheta^{-1}(\,\boldsymbol{\theta}^0\,)). \tag{5.40}$$

Thus, using the results on the asymptotic behavior of quadratic forms from the previous chapter, we have

$$\mathrm{W} = n(\,\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\,)'\vartheta(\,\boldsymbol{\theta}^0\,)(\,\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\,) \xrightarrow{d} \chi^2_p, \tag{5.41}$$

which is the Wald test. As we discussed for quadratic tests, in general, under $\mathrm{H}_0 : \boldsymbol{\theta} \neq \boldsymbol{\theta}^0$, we would have $\mathrm{W} = O_p(\,n\,)$.

In practice, since

$$\frac{1}{n}\frac{\partial^2 \mathcal{L}(\,\widehat{\boldsymbol{\theta}}|\mathbf{y}\,)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \sum_{i=1}^{n}\frac{1}{n}\frac{\partial^2 \log f(\,\widehat{\boldsymbol{\theta}}|\mathbf{y}\,)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \xrightarrow{p} -\vartheta(\,\boldsymbol{\theta}^0\,), \tag{5.42}$$

we use

$$\mathrm{W}^* = -(\,\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\,)'\frac{\partial^2 \mathcal{L}(\,\widehat{\boldsymbol{\theta}}|\mathbf{y}\,)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}(\,\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\,) \xrightarrow{d} \chi^2_p. \tag{5.43}$$

Aside from having the same asymptotic distribution, the Likelihood Ratio and Wald tests are asymptotically equivalent in the sense that

$$\operatorname*{plim}_{n\rightarrow\infty}(\,\mathrm{LR} - \mathrm{W}^*\,) = 0. \tag{5.44}$$

This is shown by expanding $\mathcal{L}(\,\boldsymbol{\theta}^0|\mathbf{y}\,)$ in a Taylor's series about $\widehat{\boldsymbol{\theta}}$. That is,

$$\begin{aligned}
\mathcal{L}(\,\boldsymbol{\theta}^0\,) &= \mathcal{L}(\,\widehat{\boldsymbol{\theta}}\,) + \frac{\partial\mathcal{L}(\,\widehat{\boldsymbol{\theta}}\,)}{\partial\boldsymbol{\theta}}(\,\boldsymbol{\theta}^0 - \widehat{\boldsymbol{\theta}}\,) \\
&+ \frac{1}{2}(\,\boldsymbol{\theta}^0 - \widehat{\boldsymbol{\theta}}\,)'\frac{\partial^2\mathcal{L}(\,\widehat{\boldsymbol{\theta}}\,)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}(\,\boldsymbol{\theta}^0 - \widehat{\boldsymbol{\theta}}\,) \\
&\quad + \frac{1}{6}\sum_i\sum_j\sum_k\frac{\partial^3\mathcal{L}(\,\boldsymbol{\theta}^*\,)}{\partial\theta_i\partial\theta_j\partial\theta_k}(\theta^0_i - \widehat{\theta}_i)(\theta^0_j - \widehat{\theta}_j)(\theta^0_k - \widehat{\theta}_k).
\end{aligned} \tag{5.45}$$

where the third line applies the intermediate value theorem for $\boldsymbol{\theta}^*$ between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^0$. Now $\frac{\partial\mathcal{L}(\,\widehat{\boldsymbol{\theta}}\,)}{\partial\boldsymbol{\theta}} = 0$, and the third line can be shown to be $O_p(1/\sqrt{n})$ under assumption 5, whereupon we have

$$\mathcal{L}(\,\widehat{\boldsymbol{\theta}}\,) - \mathcal{L}(\,\boldsymbol{\theta}^0\,) = -\frac{1}{2}(\,\boldsymbol{\theta}^0 - \widehat{\boldsymbol{\theta}}\,)'\frac{\partial^2\mathcal{L}(\,\widehat{\boldsymbol{\theta}}\,)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}(\,\boldsymbol{\theta}^0 - \widehat{\boldsymbol{\theta}}\,) + O_p(\,1/\sqrt{n}\,) \tag{5.46}$$

and

$$\mathrm{LR} = \mathrm{W}^* + O_p(1/\sqrt{n}). \tag{5.47}$$

In general, we may test $\mathrm{H}_0 : \mathbf{r}(\boldsymbol{\theta}) = 0$ with

$$\mathrm{W}^* = -\mathbf{r}(\widehat{\boldsymbol{\theta}})' \left[R(\widehat{\boldsymbol{\theta}}) \left(\frac{\partial^2 \mathcal{L}(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right)^{-1} R'(\widehat{\boldsymbol{\theta}}) \right]^{-1} \mathbf{r}(\widehat{\boldsymbol{\theta}}) \xrightarrow{d} \chi^2_p. \tag{5.48}$$

5.4 Lagrange Multiplier

Alternatively, but in the same fashion, we can expand $\mathcal{L}(\widehat{\boldsymbol{\theta}})$ about $\boldsymbol{\theta}^0$ to obtain

$$\begin{aligned} \mathcal{L}(\widehat{\boldsymbol{\theta}}) &= \mathcal{L}(\boldsymbol{\theta}^0) + \frac{\partial \mathcal{L}(\boldsymbol{\theta}^0)}{\partial \boldsymbol{\theta}} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) \\ &+ \frac{1}{2} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)' \frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}^0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) + O_p(1/\sqrt{n}). \end{aligned} \tag{5.49}$$

Likewise, we can also expand $\frac{1}{n} \frac{\partial \mathcal{L}(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}$ about $\boldsymbol{\theta}^0$, which yields

$$0 = \frac{1}{n} \frac{\partial \mathcal{L}(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} = \frac{1}{n} \frac{\partial \mathcal{L}(\boldsymbol{\theta}^0)}{\partial \boldsymbol{\theta}} + \frac{1}{n} \frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}^0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) + O_p(1/n), \tag{5.50}$$

or

$$(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) = -\left(\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}^0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right)^{-1} \frac{\partial \mathcal{L}(\boldsymbol{\theta}^0)}{\partial \boldsymbol{\theta}} + O_p(1/n). \tag{5.51}$$

Substituting (5.51) into (5.49) gives us

$$\mathcal{L}(\widehat{\boldsymbol{\theta}}) - \mathcal{L}(\boldsymbol{\theta}^0) = -\frac{1}{2} \frac{\partial \mathcal{L}(\boldsymbol{\theta}^0)}{\partial \boldsymbol{\theta}'} \left(\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}^0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right)^{-1} \frac{\partial \mathcal{L}(\boldsymbol{\theta}^0)}{\partial \boldsymbol{\theta}} + O_p(1/\sqrt{n}), \tag{5.52}$$

and $\mathrm{LR} = \mathrm{LM} + O_p(1/\sqrt{n})$, where

$$\mathrm{LM} = -\frac{\partial \mathcal{L}(\boldsymbol{\theta}^0)}{\partial \boldsymbol{\theta}'} \left(\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}^0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right)^{-1} \frac{\partial \mathcal{L}(\boldsymbol{\theta}^0)}{\partial \boldsymbol{\theta}}, \tag{5.53}$$

is the Lagrange Multiplier test.

Thus, under $\mathrm{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}^0$,

$$\operatorname*{plim}_{n \to \infty} (\mathrm{LR} - \mathrm{LM}) = 0. \tag{5.54}$$

and

$$\mathrm{LM} \xrightarrow{d} \chi^2_p. \tag{5.55}$$

Note that the Lagrange Multiplier test only requires the restricted values of the parameters.

In general, we may test $\mathrm{H}_0 : \mathbf{r}(\boldsymbol{\theta}) = 0$ with

$$\mathrm{LM} = -\frac{\partial \mathcal{L}(\widetilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}'} \left(\frac{\partial^2 \mathcal{L}(\widetilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right)^{-1} \frac{\partial \mathcal{L}(\widetilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \xrightarrow{d} \chi^2_q, \tag{5.56}$$

where $\mathcal{L}(\cdot)$ is the unrestricted log-likelihood function, and $\widetilde{\boldsymbol{\theta}}$ is the restricted MLE.

5.5 Choosing Between Tests

The above analysis demonstrates that the three tests: likelihood ratio, Wald, and Lagrange multiplier are asymptotically equivalent. In large samples, not only do they have the same limiting distribution, but they will accept and reject together. This in not the case in finite samples where one can reject when the other does not. This might lead a cynical analyst to use one rather than the other by choosing the one that yields the results (s)he wants to obtain. Making an informed choice based on their finite sample behavior is beyond the scope of this course.

In many cases, however, one of the tests is a much more natural choice than the others. Recall that Wald test only requires the unrestricted estimates while the Lagrange multiplier test only requires the restricted estimates. In some cases the unrestricted estimates are much easier to obtain than the restricted and in other cases the reverse is true. In the first case we might be inclined to use the Wald test while in the latter we would prefer to use the Lagrange multiplier.

Another issue is the possible sensitivity of the test results to how the restrictions are written. For example, $\theta_1 + \theta_2 = 0$ can also be written $-\theta_2/\theta_1 = 1$. The Wald test, in particular is sensitive to how the restriction is written. This is yet another situation where a cynical analyst might be tempted to choose the "normalization" of the restriction to force the desired result. The Lagrange multiplier test, as presented above, is also sensitive to the normalization of the restriction but can be modified to avoid this difficulty. The likelihood ratio test however will be unimpacted by the choice of how to write the restriction.